Machine Learning in Radio Resource Scheduling

Chapter · January 2019

DOI: 10.4018/978-1-5225-7458-3.ch002

6 authors, including:

Ioan Sorin Comsa
Brunel University London
29 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Sijing Zhang
University of Bedfordshire
55 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Mehmet Emin Aydin
University of the West of England, Bristol
96 PUBLICATIONS **948** CITATIONS
SEE PROFILE

Pierre Kuonen
The University of Applied Sciences Western Switzerland, Fribourg
107 PUBLICATIONS **946** CITATIONS
SEE PROFILE

Machine Learning in Radio Resource Scheduling

Ioan-Sorin Comșa[1], Sijing Zhang[2], Mehmet Emin Aydin[3], Pierre Kuonen[4], Ramona Trestian[5] and Gheorghiță Ghinea[1]

[1]*Brunel University London, United Kingdom*

[2]*University of Bedfordshire, United Kingdom*

[3]*University of West of England, United Kingdom*

[4]*University of Applied Sciences of Western Switzerland, Switzerland*

[5]*Middlesex University London, United Kingdom*

ABSTRACT

In access networks, the radio resource management is designed to deal with the system capacity maximization while the Quality of Service (QoS) requirements need be satisfied for different types of applications. In particular, the radio resource scheduling aims to allocate users' data packets in frequency domain at each predefined Transmission Time Intervals (TTIs), time windows used to trigger the user requests and to respond them accordingly. At each TTI, the scheduling procedure is conducted based on a scheduling rule that aims to focus only on particular scheduling objective such as: fairness, delay, packet loss or throughput requirements. The purpose of this chapter is to formulate and solve an aggregate optimization problem that selects at each TTI the most convenient scheduling rule in order to maximize the satisfaction of all scheduling objectives concomitantly TTI-by-TTI. The use of reinforcement learning is proposed to solve such complex multi-objective optimization problem and to ease the decision making on which scheduling rule should be applied at each TTI.

Keywords: Scheduling, Scheduler State Space, Utility Functions, Optimization, Reinforcement Learning.

1. INTRODUCTION

The continuous growth of mobile data usage and the increased interest for immersive video applications are pushing network operators to find suitable solutions to accommodate these services with very stringent Quality of Service (QoS) demands (Cisco, 2017). According to Trestian, Comsa, and Tuysuz (2018), the Quality of Experience provisioning will become the main differentiator between network operators, in which, the satisfaction of heterogeneous QoS requirements is playing a crucial role. In this context, the 5th Mobile Generation (5G) technology comes up with the promise of very low end-to-end

latencies and much higher system capacity while implementing some important features such as (G. Andrews et al., 2014): new waveforms, densification of access networks, higher frequency bands, mass scale antennas and millimeter-wave communications. An issue to be addressed when delivering these bandwidth-hungry applications in multi-user scenario refers to the management of radio resources that can strongly affects the overall performance of QoS provisioning (Li et al., 2017). The responsible entity is the Radio Resource Management (RRM) that aims to ensure an efficient allocation of the disposable system bandwidth in order to maximize the QoS satisfaction while implementing advanced technologies (as provided by Olwal, Djouani, and Kurien, 2017) able to: save the energy, control the mobility and power allocation, mitigate interference, schedule users' packets in frequency domain at each TTI.

According to the performance of the scheduling process, the operator is struggled to provide the requested services while using the disposable radio infrastructure, regardless of the spatial/time positions of mobile terminals, user preferences, devices' types and application requirements (Comşa, 2014a). A major concern is to increase the system capacity or data rates of all active users while satisfying the application requirements. Users located in the proximity of base stations experience better channel quality and consequently can get higher data rates than those users with poorer channel quality located farer away from any available base station. By providing the disposable spectrum to those users with better channel conditions, other users are starved in receiving the requested data for longer time. Then, the fairness measure between different users with the same QoS profiles is impaired. Certain tradeoff measures between system throughput and user fairness can be adopted (Jain, Chiu, and Hawe, 1984; Comşa, 2014a). Together with these aspects, the requested services should be provided under some predefined QoS requirements (as imposed by 3GPP, 2012) in terms of Guaranteed Bit Rate (GBR), Head of Line (HoL) packet delay and Packet Loss Rate (PLR). The QoS requirements become more restrictive with the evolution of cellular standards, system architectures and applications. By encompassing the above discussed aspects, the packet scheduler is dealing with an optimization problem aiming to maximize the system throughput and being constrained by fairness and QoS requirements. If we further consider that the constraints' satisfaction refer to fairness and QoS objectives, then the Multi-Objective Optimization (MOO) is addressed (as stated by Comşa, 2014a).

At each TTI, the scheduling procedure is conducted based on the scheduling rule that aims to prioritize active users in frequency domain (Toufik and Knopp, 2011). In literature, these rules are various, and each of them is targeting particular scheduling objective (as surveyed by Capozzi, Piro, Grieco, Boggia, and Camarda, 2013). Actually, the scheduling rule quantifies the benefit of allocating each user in the frequency domain from the perspective of the addressed objective. For this reason, almost all state-of-the-art scheduling rules impact differently when considering the multi-objective satisfaction criterion (Comşa, 2014a). For example, some scheduling rules are oriented more on the tradeoff between system throughput maximization and user fairness satisfaction while degrading the performance of QoS objectives (Proebster, Mueller, and Bakker, 2010). Other rules aim to focus more on particular QoS objective (i.e. delay) while harming the performance of other QoS objectives (i.e. GBR, PLR), system throughput and user fairness (Liu, Tian, and Xu, 2013). Therefore, the multi-objective performance strongly depends on the adopted scheduling rule.

Alongside of the exploited scheduling rule, the multi-objective satisfaction depends also on the momentary scheduler states, observed at each TTI and being composed by: channel conditions, number of users, QoS parameters, queue loads, application types, etc. Actually, when the scheduler conditions permit (reduced number of users, very good channel conditions, low traffic load), the multi-objective target can be attained by almost all existing scheduling rules. However, under more generalized scheduler state space (variable traffic types, channel conditions, number of user, etc.) particular scheduling rules are unable to reach an acceptable level for the multi-objective satisfaction. Hence, it can be envisioned that a mixture of scheduling rules can be used instead of a single one adopted across the entire process, in

which, the scheduling rule that gives the highest multi-objective outcome in each scheduler state must be found and applied in order to maximize the multi-objective satisfaction measure over time.

This chapter deals with the multi-objective optimization problem where the joint assignation of radio resources and scheduling rules is considered in downlink scheduling. The general idea is to find in each instantaneous scheduler state the most convenient scheduling rule to be applied in order to maximize the overall system throughput while keeping the QoS constraints satisfied as long as possible. Due to the increased complexity of the proposed aggregate optimization problem, we adopt the use of machine learning tools in order to find suitable solutions in each scheduler state. In this sense, the Reinforcement Learning (RL) framework is proposed to learn over time the most convenient scheduler rule for each momentary state (as initially proposed by Comşa, Aydin, Zhang, Kuonen, and Wagen, 2011, 2012). The implementation of RL framework for downlink Orthogonal Frequency Division Multiple Access (OFDMA) scheduling is also addressed. The choice of OFDMA is due to its simplicity, efficiency and its wide deployment, being one of the multiple access schemes to be considered in 5G networks (FANTASTIC-5G, 2016).

2. BACKGROUND

In general, one scheduling rule is focused first on the main objective; once the main objective is satisfied, the static scheduling rule can optimize other objectives with the amendment that the first objective is always satisfied. In this case, the multi-objective optimization problem becomes Sequential MOO (SMOO). As some studies show (Lundevall et al., 2004; Ning, Ying, and Ping, 2006; Zhang, Yuan, and Zhang, 2011), the SMOO problems are considering the GBR satisfaction as a primary objective, whereas the user fairness satisfaction or user throughput maximization is considered once the GBR objective is satisfied for all active users TTI-by-TTI. The same principles of SMOO are targeted by the proposed scheduling rules in (Rhee, Holtzman, and Kim; 2003; Sadiq, Madan, and Sampath, 2009; Bae, Choi, and Chung, 2011) where the delay is the first objective. Another scheduler (as proposed by Khan, Martini, Bharucha, and Auer, 2012) aims to focus first on the PLR satisfaction followed by other objectives such as fairness and throughput. The biggest disadvantage of these rules is the lack of consideration of all QoS objectives. Since these rules are static over the entire scheduling process, the optimization problem is entitled Static Scheduling Rule based SMOO (SSR-SMOO).

When a different scheduling rule is applied at each TTI, then the Dynamic Scheduling Rule (DSR) principle is addressed. Schwarz, Mehlfuhrer, and Rupp (2011) address the DSR-SMOO problem in the sense that the Generalized Proportional Fair (GPF) is parameterized in order to optimize the throughput-fairness trade-off in terms of Jain fairness index (first introduced by Jain et al., 1984). The RL framework is proposed by Comşa, Zhang, Aydin, Kuonen, and Wagen (2012) to achieve different tradeoff levels between system throughput and user fairness. But this quantitative fairness objective used in these approaches are static and do not depends on channel or network conditions. Instead, the qualitative fairness measures based on channel statistics can be used, such as Next Generation Mobile Networks (NGMN) fairness (Proebster et al., 2010). According to this metric, the GPF scheduling rule should be adapted at certain TTIs, such that, the Cumulative Distribution Function (CDF) of normalized user throughputs is adjusted to respect the requirement from the CDF domain, as studied by Proebster et al. (2010). However, when using actor-critic RL algorithm (as proposed by Comşa et al., 2014b) to parameterize the GPF scheduling rule based on dynamic scheduler conditions in order to increase the time when the NGMN fairness requirement is satisfied, the obtained gain is higher than 10% when compared with the method proposed by Schwarz et al. (2011), and higher than 20% when compared to the adaptation technique proposed by Proebster et al. (2010). But all these approaches use a simple parameterization of GPF scheduling rule (considering only α parameter adaptation) when matching against the NGMN fairness requirement. In the study conducted by Comşa et al. (2014c), the same actor-

critic RL algorithm is used to perform the double parameterization of the GPF scheduling rule in terms of both α and β parameters. The results indicate a gain larger than 5% when compared to actor-critic RL framework that uses the simple parameterization. The advantages of these RL-based frameworks reveal a higher capacity to adapt to a more generalized scheduler state space and an increased level of satisfaction for the NGMN fairness requirement. The QoS objectives are not considered in these optimization models.

In other circumstances, one static scheduling can optimize multiple objectives at the same time when applied TTI-by-TTI. The optimization procedure is called Static Scheduling Rule based Concurrent MOO (SSR-CMOO). Another scheduling proposed by Khan et al. (2012) considers the joint optimization of PLR and delay objectives, while the fairness objective is considered once the primary multi-objective satisfaction is achieved. Other schedulers aim to split the scheduling process in two stages (as indicated by Monghal, Laselva, Michaelsen, and Wigard, 2010; Chung, Chang, and Wang, 2012; Wang, Li, Ji, and Zhang, 2013; Avocanh, Abdennebi, and Ben-Othman, 2014): a) time domain where the group of users with more stringent QoS requirements is prioritized to be scheduled in b) frequency domain, where the radio resources are allocated for the preselected users. In these cases, the QoS objectives are targeted in time domain, whereas the frequency domain deals more with the trade-off between system throughput maximization and user fairness provisioning. Although these rules aims to satisfy the entire set of objectives, there is not a clear evidence how these schedulers will perform under generalized RRM state space (channel types, variable number of users and traffic types, arrival rates in data queues, etc).

By combining the concurrent multi-objective optimization and the dynamic scheduling rule selection (DSR-CMOO), a mixture of different scheduling rules is used, in which one rule is applied at each TTI in order to improve the satisfaction measure when multiple scheduling objectives are addressed. Different RL algorithms are implemented by Comşa et al. (2018) for Constant Bit Rate and Variable Bit Rate traffic types in order to maximize the number of TTIs when the PLR and delay constraints are satisfied. The proposed framework makes use of four scheduling rules that are oriented only on the delay objective. Gains higher than 10% are obtained for the proposed RL framework when compared to other conventional static scheduling rules. In other studies (Comşa, De-Domenico, and Ktenas, 2017; Comşa, Trestian, and Ghinea, 2018), the actor-critic RL framework is used to optimize the scheduling rule selection in each state when PLR, GBR and delay objectives are considered. These proposals aim to minimize the disadvantages of each particular scheduling rule while maximizing their advantages by applying on each scheduler state the best scheduling rule such that the multi-objective satisfaction is maximized. The throughput maximization and user fairness objectives are not taken into account by these proposals.

This chapter considers the DSR-CMOO model that aims to combine the static scheduling rules with the main focus on the following objectives: throughput maximization, NGMN fairness requirement and QoS constraints satisfaction in terms of PLR, GBR and delay. In Section 3, the OFDMA scheduling elements are presented in terms of: scheduler state, resource allocation procedure, utility and objective functions. Section 4 presents the general form of SSR-SMOO problems when each scheduling objective is considered separately. Section 5 formulates the proposed DSR-CMOO aggregate problem based on each particular SSR-SMOO problems. Section 6 provides a sub-optimal solution to this complex problem based on the RL approach. Finally, this chapter concludes with Section 7.

3. ODFMA SCHEDULING PROCESS

The scheduler process is conducted based four main components: scheduler state space, scheduling procedure, scheduling rule, and MOO performance evaluation. The scheduler state space contains all parameters necessary to conduct the scheduling procedure at each TTI. The scheduling procedure in OFDMA systems includes: the user selection for each radio resource according to the exploited scheduling rule; the Modulation and Coding Scheme selection (MCS) for each scheduled user; and the

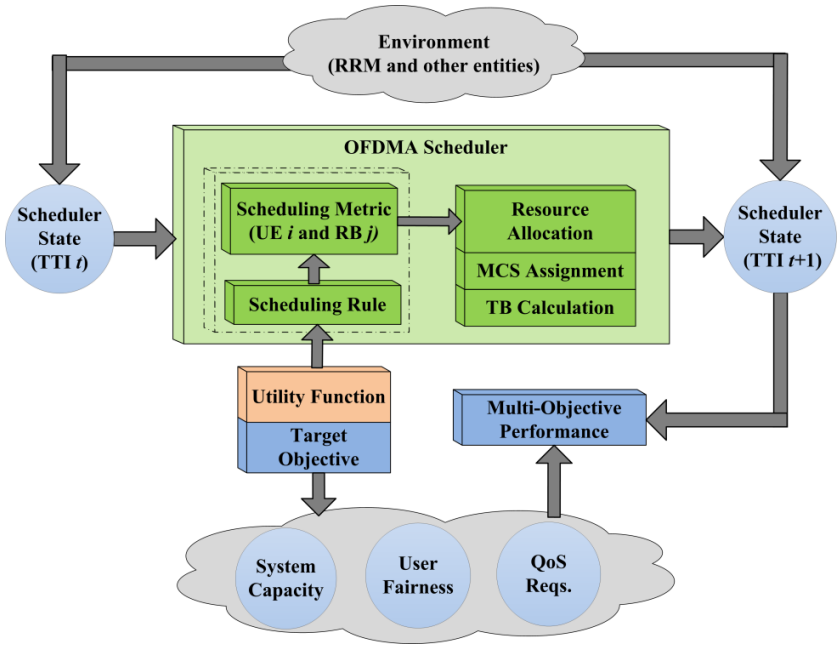

Figure 1. OFDMA Scheduling Process

Transport Block (TB) computation that actually gives the total number of bits to be transmitted to each scheduled user based on the amount of allocated radio resources and supportable MCS. A scheduling rule is associated with the resource allocation process in order to satisfy given objectives. The MOO performance evaluation gives a satisfaction measure for all scheduling objectives at each TTI according to the applied scheduling rule in the previous state. The OFDMA scheduling process is presented in Fig. 1.

In OFDMA scheduling, the available bandwidth is divided in B number of resource blocks every TTI. If we also consider a number of I_t number of users that can change at each TTI t, then the most pretentious task in OFDMA scheduling is to find the potential benefit of allocating each Resource Block (RB) $j \in \mathcal{B}=\{1,2,...,B\}$ to certain users $i \in \mathcal{I}_t = \{1,2,...,I_t\}$ when a given performance criterion is addressed based on the exploited scheduling rule. More precisely, being given a certain performance criterion, the scheduler should be aware about the exact price or cost value of allocating RB $j \in \mathcal{B}$ to UE $i \in \mathcal{I}_t$ for its target objective satisfaction. So, the scheduler is responsible for optimizing the obtained pricing structure problem every TTI.

The potential benefit quantification of using some limited resources is inherited from the utility theory in economics which has been applied with great success in wireless networks in order to guarantee the QoS requirements and to exploit the multi-user diversity principle in opportunistic scheduling (Liu, Chong, and Shroff, 2001). In OFDMA networks, the proposed scheduling procedure maps the performance criteria in some utility metrics for each user $i \in \mathcal{I}_t$ and for each RB $j \in \mathcal{B}$. Then, the instantaneous optimization problems resume to the sum maximization of each user utility TTI-by-TTI.

Adopting the performance criteria in order to evaluate the performance of user centric objectives for different type of services represents a crucial task. As mentioned earlier, by using classical scheduling procedures, it is difficult to satisfy multiple objectives simultaneously. Therefore, some priorities in satisfying particular objectives are given by adopting different performance criteria at once as denoted by SSR-SMOO problems. However, the performance criteria indicate basically the types of utility functions.

For instance, if the utility function addresses the HoL packet delay performance, the scheduler is designed in such a way that the packet delay budget should be satisfied in certain requirements. If the optimality of the first condition is satisfied, then other objectives can be considered depending on the particularity of the utility function.

In Figure 1, based on the selected performance criterion (or scheduling objective to be satisfied) and the type of utility function, a scheduling rule is selected to conduct the scheduling procedure each TTI. At TTI *t*, the scheduler momentary state is observed and according to the decided scheduling rule, a number of $B \times I_t$ metrics is calculated by considering the metrics of all users $i \in \mathcal{I}_t$ for each RB $j \in \mathcal{B}$. The resource allocation performs for each RB from $\mathcal{B}$ the metric ordering of all users metrics calculated on that particular RB. Only users with the highest metric for each RB are selected to be scheduled at TTI *t*. This means actually the urgency for those users to be scheduled in order to maximize the given performance criterion. Then, the MCS levels are assigned and the TB determines actually the amount of data to be broadcasted to those selected users. After the scheduling procedure is completed, the system moves to the next state at TTI *t*+1. Only at this stage, the real multi-objective measure is determined in order to evaluate the scheduler performance in previous state, at TTI *t*.

3.1 Scheduler State Space

An important role in OFDMA scheduling is represented by the scheduler state space since both optimization approaches, SMOO and CMOO, perform the scheduling procedure at each TTI based on the momentary scheduler conditions. The scheduler state space is exploited in different ways based on the optimization type:

- In the SSR-SMOO optimization, the scheduler state space provides the necessary parameters for the utility function computation in the scheduling procedure;
- For the DSR-SMOO and DSR-CMOO problems, alongside the utility parameters provision, the scheduler state space is responsible for choosing a proper scheduling rule in order to meet the proposed objective.

Inevitably, the selected scheduling rule affects part of the scheduler state space evolution. The scheduler state space is divided into two disjoint sub-spaces:

- ***Uncontrollable scheduler state space***: The CQI reports, HARQ retransmissions indicators, arrival bit rates, QoS requirements are included. We define the measurable uncontrollable state space $\mathcal{S}_U$ and $\mathrm{Z}[t] \in \mathcal{S}_U$ the momentary uncontrollable vector at TTI *t*.
- ***Controllable scheduler state space***: The parameters responsible for the objective performance evaluation, such as HoL delay, average user rate, normalized user rate, packet loss rate, and queue size are included. Similarly, we define the measurable controllable state space $\mathcal{S}_C$ and $\mathrm{C}[t] \in \mathcal{S}_C$ the momentary controllable vector at TTI *t*.

The overall measurable state space is defined as $\mathcal{S} = \mathcal{S}_U \cup \mathcal{S}_C$ and the momentary scheduler state at TTI *t* is $\mathrm{S}[t] = [\mathrm{Z}, \mathrm{C}] \in \mathcal{S}$. Each of these elements is detailed in the following sub-sections.

3.1.1 Uncontrollable Scheduler State

The uncontrollable scheduler state space represents the indices and parameters that reflect mainly the channel conditions, the service parameters from the upper layers, and the QoS requirements for each active data flow. Even if these parameters are modeled as random processes rather than the scheduling

procedure results, the obtained subspace plays a crucial role in achieving satisfaction of different objectives. The uncontrollable state space $\mathcal{S}_U$ encompasses the following elements:

- **Channel Quality Indicator (CQI) Reports**: It is assumed that at each TTI *t,* each user $i \in \mathcal{I}_t$ reports the CQI value for each RB $j \in \mathcal{B}$ through the PUCCH control channel. Let us define $CQI_{i,j}[t]$ as the CQI report value of user $i \in \mathcal{I}_t$ and resource block $j \in \mathcal{B}$ at TTI *t*. Then, $CQI[t]$ is the momentary vector of all CQI reports for all active users defined as follows:

$$CQI[t] = \left[CQI_{i,j}[t] \right]_{\substack{j=1,\ldots,B \\ i=1,\ldots,I_t}} \quad (1)$$

- **Achievable user rate**: Based on the $CQI_{i,j}[t]$ reports, the achievable user rate $r_{i,j}[t]$ is computed for the scheduling decision. For each $CQI_{i,j}[t]$, a MCS is determined in order to provide the number of bits that can be transmitted if RB $j \in \mathcal{B}$ would be allocated to user $i \in \mathcal{I}_t$. Similarly to CQI reports, the vector of instantaneous rates is obtained based on:

$$\mathrm{r}[t] = \left[r_{i,j}[t] \right]_{\substack{j=1,\ldots,B \\ i=1,\ldots,I_t}} \quad (2)$$

- **Instantaneous arrival rate**: We define the arrival rate in data queue for user $i \in \mathcal{I}_t$ at TTI *t* as $\lambda_i[t]$. Then, the instantaneous vector of arrival rates is defined by $\Lambda[t] = \left[\lambda_1, \lambda_2, \ldots, \lambda_{I_t} \right]$ with the amendment that for each user is considered only one data queue.

The momentary uncontrollable scheduler state is obtained by taking into account the indicators introduced above such as: $Z[t] = [CQI, \mathrm{r}, \Lambda]$.

3.1.2 Controllable Scheduler State

The controllable scheduler subspace denotes the set of indices which are used for the multi-objective performance evaluation. Basically, when $\mathcal{S}_C$ is optimal, the entire scheduler state is considered optimal. Under these circumstances, other subspace $\mathcal{S}_U$ provides the necessary information to maintain the scheduler in the optimal region by satisfying all scheduling objectives. The controllable subspace being considered here comprises the following elements:

- **Instantaneous user rate**: When performing the scheduling procedure, the instantaneous user rate $R_i[t]$ represents the total number of bits associated to the scheduled user $i \in \mathcal{I}_t$. The associated vector is: $\mathrm{R}[t] = [R_1, R_2, \ldots, R_I]$.
- **Instantaneous user throughput**: If the transmitted packets in the previous TTI were correctly decoded by each scheduled active user (HARQ report is zero), then the instantaneous user rate becomes the instantaneous user throughput $T_i[t]$. Consequently, the instantaneous throughput vector becomes $\mathrm{T}[t] = [T_1, T_2, \ldots, T_I]$.
- **Average user throughput**: It is used to improve the fairness among users. If $T_i[t]$ is used as the fairness satisfaction metric, then the scheduler should be fair at each TTI. This aspect is undesirable because it affects the spectral efficiency performance. Therefore, it is preferred to evaluate the fairness performance by using a time window or a predefined number of TTIs. So, the average user throughput $\overline{T}_i[t]$ is determined as follows:

$$\overline{T}_i[t] = (1-\beta) \cdot \overline{T}_i[t-1] + \beta \cdot T_i[t] \quad (3)$$

where β represents the forgetting factor which impacts in the scheduler performance. The lower values for parameter β implies in fact lower impacts of the current scheduling procedure of the optimization problem. The instantaneous vector is: $\overline{\mathrm{T}}[t]=\left[\overline{T}_1,\overline{T}_2,...,\overline{T}_I\right]$.

- **HoL packet delay**: We define by $d_i[t]$ the maximum waiting time for a given data packet in the MAC level data queue for user $i \in \mathcal{I}_t$. The momentary delay vector is $\mathrm{d}[t]=[d_1,d_2,\ldots,d_I]$.
- **Packet Loss Rate**: We denote by $L_i[t]$ the packet loss rate at TTI *t* that indicates the number of lost packets over the total number of sent packets in a given time window of T_w. We define the momentary vector of PLRs as $\mathrm{L}[t]=[L_1,L_2,\ldots,L_I]$.
- **Transmission queues size**: We define $q_i[t]$ the transmission queue size for user $i \in \mathcal{I}_t$ at TTI *t*. Consequently, the instantaneous queue vector is $\mathrm{q}[t]=[q_1,q_2,...,q_I]$.

Then, the momentary controllable scheduler state becomes $\mathrm{c}[t]=\left[\mathrm{R},\mathrm{T},\overline{\mathrm{T}},\mathrm{L},\mathrm{d},\mathrm{q}\right]\in\mathcal{S}_C$.

3.2 Radio Resource Allocation

The main focus for the OFDMA scheduler is to assign the available set of RBs to different active users in order to satisfy given scheduling objectives. The idea is to quantify the benefit (utility) of allocating each RB $j \in \mathcal{B}$ to user $i \in \mathcal{I}_t$ at each TTI *t*. In this sense, the utility function has to be defined.

For our scheduling purposes, the utility functions cannot be measured directly. The solution is to perform the instantaneous rate allocation based on the utility representation at each TTI *t* and to measure or to evaluate the allocation performance at each TTI *t*+1 by using the objective functions. We define by $\mathrm{r}=\left[\mathrm{r}_1,\mathrm{r}_2,...,\mathrm{r}_I\right]^T$ the instantaneous achievable rate matrix being obtained based on the CQI reports, where $\mathrm{r}_i=\left[r_{i,1},r_{i,2},...,r_{i,B}\right]$ is the vector of instantaneous rates of each user $i \in \mathcal{I}_t$.

For each user $i \in \mathcal{I}_t$, let us consider $U_i(\mathrm{r}_i)$ the utility function which is a benefit representation of allocating the vector of rates r_i to user $i \in \mathcal{I}_t$. In long term, the packet scheduler aims to maximize the aggregate user, such as:

$$\max_{t\to\infty,i\in\mathcal{I}_t} \sum_t \sum_i U_i\left(\mathrm{r}_i[t]\right) \tag{4}$$

In multi-user scenario, the disposable set of RBs is allocated only to some users at each TTI according to some allocations variables $b[t]=\left\{b_{i,j}[t]\right\}, i=1,...,I_t, j=1,...,B$ that take the binary values as follows:

$$b_{i,j}=\begin{cases}1, \ if \ RB \ j\in\mathcal{B} \ is \ allocated \ to \ UE \ i\in\mathcal{I}_t\\ 0, \ otherwise\end{cases} \tag{5}$$

The instantaneous data rates $R_i[t]$ for each user are obtained after performing the scheduling decision under the allocation variables *b* at each TTI *t*. Let us define the instantaneous rate region constrained by policy *b* such as $\mathcal{R}_b$. Therefore, the definition domain for the utility function is $U_i:\mathcal{R}_b\to\mathbb{R}$ and the long-term optimization problem becomes (Song and Li Geoffrey, 2005):

$$\begin{aligned} \max_{b} \ & \sum_i U_i\left(\sum_j b_{i,j}\cdot r_{i,j}\right) \\ s.t. \ & \sum_i b_{i,j}=1, \ \ j=1,...,B \\ & b_{i,j}\in\{0,1\}, \ \ \forall i\in\mathcal{I},\forall j\in\mathcal{B} \end{aligned} \tag{6}$$

According to Kelly (1997) the local maximum is also a global maximum in (6) if and only if the region $\mathcal{R}_b$ is a *convex set* and $U_i(R_i)$ is a *concave* function. However, the convexity problem of $\mathcal{R}_b$ in OFDM systems has been discussed intensively by Song (2005) and the authors came with the conclusion that the short-term optimization problem at each TTI *t* can be obtained by using the first order approximation of Taylor's expansion as expressed by (7) (as stated by Song, 2005)

$$\sum_i U_i\left(R_i[t]\right)-\sum_i U_i\left(R_i[t-1]\right)\approx\sum_i U_i'\left(R_i[t-1]\right)\cdot\left(R_i[t]-R_i[t-1]\right) \tag{7}$$

where $U_i'\left(R_i[t-1]\right)=\partial U_i\left(R_i[t-1]\right)/\partial R_i[t-1]$ is the marginal utility for user $i\in\mathcal{I}_t$. The instantaneous rate $R_i[t-1]$ at TTI *t*-1 is obtained after performing the scheduling procedure at TTI *t*-1 and this value is used in the optimization problem at TTI *t*. Therefore, the short-term optimization problem can be written under the following form:

$$\begin{aligned} \max_{b[t]} \ & \sum_i\sum_j b_{i,j}[t]\cdot U_i'\left(R_i[t]\right)\cdot r_{i,j}[t] \\ s.t. \ & \sum_i b_{i,j}[t]=1, j=1,...,B \\ & b_{i,j}[t]\in\{0,1\},\forall i\in\mathcal{I}_t,\forall j\in\mathcal{B} \end{aligned} \tag{8}$$

The optimization model being exposed in (8) represents a *linear programming model* where the unknown variables are the resource assignment variables $b_{i,j}[t]\in\{0,1\}$ that need to be determined at each TTI *t*. Due to the reduced number of resources that has to be allocated at each TTI, the assignment is performed by using the following equation:

$$m_j[t]=\underset{i\in\mathcal{I}_t}{arg\,max}\left\{U_i'\left(R_i[t]\right)\cdot r_{i,j}[t]\right\} \tag{9}$$

where, $m_j[t]$ indicates that RB $j\in\mathcal{B}$ is assigned or allocated to user $m\in\mathcal{I}_t$ $\forall m\neq i$ at TTI *t*. Consequently, $b_{m,j}[t]=1$ and $b_{i,j}[t]=0,\ \forall i\in\mathcal{I}_t$ and $\forall i\neq m$. This way, the user assignment is performed for each RB for a given bandwidth. Once the resource allocation is finished, the transport block size is determined for each selected user.

As mentioned earlier, the instantaneous rate for each user $i\in\mathcal{I}_t$ and for each RB $j\in\mathcal{B}$ is determined based on the CQI reports. The marginal function is positive because the utility function $U_i(R_i)$ must be concave (the second derivative is negative) in order to assure the linearity of the considered optimization problem. When the utility function $U_i(R_i)$ takes the polynomial form, the role of its marginal utility is to schedule those users with the highest instantaneous rates by increasing at the same time the total system capacity if the radio channels are errorless. In the case of re-transmissions, the marginal utility as a function of instantaneous user throughput $U_i'(T_i)$ should be used in order to provide more RBs to those users which require less retransmissions during the downlink scheduling session to avoid the waste of radio resources.

The linear programming model exposed in (8) is a typical SSR-SMOO problem being focused on the system throughput maximization without considering other objectives such as: user fairness, GBR, HoL delay, packet loss. The impact of the resource allocation problem in the scheduling objectives can be measured by using the objective functions. The objectives functions can be modeled by using the QoS

constraints. When different objective(s) is (are) analyzed, the performance of the optimization problem from (8) can be improved if the marginal utility function considers the performance parameter(s) of the addressed objective(s). More details about this aspect are presented in the following section.

3.3 Utility and Objective Functions in OFDMA Systems

Utility functions are designed to quantify the benefit of allocating a given and finite number of RBs to active mobile users. The type of utility function can influence the optimization problem in the direction of different scheduling objectives. The classification of utility functions can be achieved by considering three perspectives: the argument function, the utility weight and the manufacturing methodologies. Based on the manufacturing methods, there are two modes to obtaining utility functions (as provided by Song and Li Geoffrey, 2005; Song, 2005) exposed bellow together with the proposed methodology:

- **Application based utility functions**: One way is to develop utility functions that characterize a specific type of application which can be obtained by using sophisticated subjective surveys. These utilities can suffer from the imperfection of the measurements, and different parameters are fixed to some objective values denoting the inflexibility for those situations which are not covered by the considered surveys.
- **Traffic habits based utility functions**: statistics about the percentage of different traffic types that can co-exist at different moments of time in different urban scenarios. The utility functions are designed based on these statistics of heterogeneous traffic types.
- ***Scheduler state based aggregate utility function***: Based on the scheduler state $\mathrm{s} \in \mathcal{S}$, different utility functions (being already proposed in the literature) are applied in order to maximize the long term aggregate utility function and to solve the DSR-SMOO/CMOO combinatorial problems. More precisely, it maximizes the sum of some existing utility functions subject to some objective functions' requirements.

The short-term optimization for the resource allocation is obtained when performing the first order of Taylor's expansion between two time consecutive utility functions. This way, the marginal utility function or the first derivative utility function is obtained. The term of marginal can refer also to a small change that can appear in the optimization problem between two consecutive momentary scheduler states. In fact, the marginal utility indicates high gains for those users with poor objective performance, whereas other users with much better multi-objective performance are getting much lower gains. In this sense it is encouraged to allocate higher amount of resources to those users with higher gains in the marginal utility.

In the optimization problem exposed in (8), when selecting any gain in the marginal utility function leads to the system capacity maximization without any consideration about other objectives. By designing the marginal utility with proper weights, different objectives can be addressed. So, the role of the marginal utility in the optimization problem is to reduce the impact of the instantaneous achievable rate $r_{i,j}[t]$ (or to annihilate any variation of the radio channel) and to focus the entire optimization problem on scheduling different users based on their performance of satisfying different objectives addressed by different marginal utility weights. To conclude, the multi-objective performance depends on the type of marginal utility which is used in the optimization problem.

To generalize the utility function representation, we define $\mathrm{x}_i \in \mathcal{X}_i$ the argument of the utility function for user $i \in \mathcal{I}_t$ and $\mathrm{y}_i \in \mathcal{Y}_i$ the argument for the utility weight, where $\mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{S}_C$, $\mathcal{X}=\bigcup_i \mathcal{X}_i$ and $\mathcal{Y}=\bigcup_i \mathcal{Y}_i$. Therefore, the utility function for user $i \in \mathcal{I}_t$ can be decomposed as follows:

$$U_i(\mathrm{x}_i) = F_i(\mathrm{x}_i) \cdot W_i(\mathrm{y}_i) \tag{10}$$

where function $F_i : \mathcal{X}_i \to \mathbb{R}$ is concave and differentiable, and the utility weight $W_i : \mathcal{Y}_i \to \mathbb{R}$ of user $i \in \mathcal{I}_t$ is a constant, but it is represented as a function in order to highlight the objective indicator $\mathrm{y} = [\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_I] \in \mathcal{Y}$, where $\mathrm{y} = \{\mathrm{R}, \mathrm{T}, \overline{\mathrm{T}}, \mathrm{L}, \mathrm{d}\}$.

When one element for the controllable parameters set $\mathrm{y}_i = \{R_i, T_i, \overline{T}_i, L_i, d_i\}$ respects the corresponding QoS constraints, then user $i \in \mathcal{I}_t$ is satisfied from the viewpoint of the addressed objective. The first derivative for the utility function is determined by using $U_i'(\mathrm{x}_i) = F_i'(\mathrm{x}_i) \cdot W_i(\mathrm{y}_i)$, where $F_i'(\mathrm{x}_i) = \partial F_i(\mathrm{x}_i) / \partial \mathrm{x}_i$. If the marginal utility function is developed in such a way that the radio channel variations are compensated at each TTI t for each user $i \in \mathcal{I}_t$ $\left(r_{i,j}[t] \cdot F_i'(\mathrm{x}_i) \approx 1\right)$, then the optimization problem is focused more on the scheduling objective evaluated by the weight argument $\mathrm{y} \in \mathcal{Y}$.

4. SINGLE OBJECTIVE OPTIMIZATION PROBLEM

In this chapter, we study five objectives such as: throughput maximization, NGMN user fairness, guaranteed bit rate, packet delay satisfaction and packet loss minimization. We consider $\mathcal{O}$ the set of aforementioned objectives. For each objective $o \in \mathcal{O}$, let us define the pool of utilities $\mathcal{U}_o$, and then, the entire set of utilities for all objectives is defined as $\mathcal{U} = \bigcup_o \mathcal{U}_o$. As mentioned earlier, the type of marginal utility for objective $o \in \mathcal{O}$ is correlated with the utility weight. If $u_o \in \mathcal{U}_o$ is the weight index targeting the objective $o \in \mathcal{O}$, then the short-term optimization can be obtained for the general form of scheduling utilities by using the first order of Taylor's expansion and being similar to (7):

$$\sum_i U_{o,i}\left(\mathrm{x}_i[t]\right) - \sum_i U_{o,i}\left(\mathrm{x}_i[t-1]\right) \approx \sum_i W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) \cdot F_i'\left(\mathrm{x}_i[t]\right) \cdot \left(R_i[t] - R_i[t-1]\right) \tag{11}$$

where, $\mathrm{y}_{o,i}[t] \in \{R_i, T_i, \overline{T}_i, L_i, d_i\}$ is the controllable QoS parameter targeting objective $o \in \mathcal{O}$ and $W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right)$ is the weight function $u_o \in \mathcal{U}_o$ that aims to maximize the resource allocation from the perspective of the same objective $o \in \mathcal{O}$. Then, the optimization problem when both $o \in \mathcal{O}$ and $u_o \in \mathcal{U}_o$ are static becomes as follows:

$$\begin{aligned} &\max_{b[t]} \sum_i \sum_j b_{i,j}[t] \cdot W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) \cdot F_i'\left(\mathrm{x}_i[t]\right) \cdot r_{i,j}[t] \\ &s.t. \quad \sum_i b_{i,j}[t] = 1, \ j = 1, ..., B \\ &\qquad b_{i,j}[t] = \{0,1\}, \ \forall i \in \mathcal{I}_t, \forall j \in \mathcal{B} \end{aligned} \tag{12}$$

where the optimization problem is focusing on objective $o \in \mathcal{O}$ subject to the convex set of constraints. The optimization problem of (12) is a linear programming model. If $W_{o,u,i}[t] >> F_i'[t] \cdot r_{i,j}[t]$, then the addressed objective $o \in \mathcal{O}$ is evaluated based on the weight argument $\mathrm{y}_{o,i}[t]$ for each user $i \in \mathcal{I}_t$. At this extent, the weight matrix for objective $o \in \mathcal{O}$ is $\nabla U_o = \{U'_{o,u,i} = W_{o,u,i} \cdot F_i', u_o = 1, ..., U_o, i = 1, ..., I_t\}$. In SSR-SMOO problems, the weight matrix ∇U_o assigns the same type of marginal utility functions to each user $i \in \mathcal{I}_t$ for the entire scheduling session.

The optimal resource allocation when following the linear optimization problem from (12) for each RB $j \in \mathcal{B}$ and user $\forall i \in \mathcal{I}_t$ is given by the following metric calculation:

$$m_j[t] = \underset{i \in \mathcal{I}_t}{arg\,max} \left\{ W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) \cdot F_i'\left(\mathrm{x}_i[t]\right) \cdot r_{i,j}[t] \right\} \tag{13}$$

where $m_j[t]$ indicates that RB $j \in \mathcal{B}$ is allocated to user $m \in \mathcal{I}_t$, $\forall m \neq i$ at TTI *t*. Then, $b_{m,j}[t]=1$ and $b_{i,j}[t]=0, \forall m \in \mathcal{I}_t, \forall i \in \mathcal{I}_t$ and $\forall i \neq m$. Therefore, the scheduling rule function can be defined in the following manner:

$$D_{o,u,i}: \mathcal{Y} \rightarrow \mathbb{R} \quad D_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) = W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) \cdot F_i'\left(\mathrm{x}_i[t]\right) \tag{14}$$

If the objective $o \in \mathcal{O}$ and the utility weight $u_o \in \mathcal{U}_o$ stay fixed during the entire scheduling session, then the linear optimization problem from (12) is a SSR-SMOO/CMOO problem. The CMOO refers to the fact that the argument of $W_{o,u,i}\left(\mathrm{y}_i[t]\right)$ can be multi-dimensional and the optimization problem is focusing on two or more objectives, and the argument is a vector of $\mathrm{y}_i = \left[\mathrm{y}_{1,i}, \mathrm{y}_{2,i}, ..., \mathrm{y}_{O,i}\right]$.

The performance of the scheduling discipline $D_{o,u,i}$ can be evaluated by using the objective functions. Let us define the objective function $\phi_{o,i}\left(\mathrm{y}_{o,i}[t]\right)$ for objective $o \in \mathcal{O}$ and user $i \in \mathcal{I}_t$, where the definition domain is $\phi_{o,i}: \mathbb{R} \rightarrow \mathbb{R}$. The objective condition for a particular SSR-SMOO problem for each user $i \in \mathcal{I}_t$ at each TTI is given by:

$$\phi_{o,i}\left(\mathrm{y}_{o,i}[t]\right) \geq 0 \ , \forall o \in \mathcal{O}, \forall i \in \mathcal{I}_t \tag{15}$$

When the condition is satisfied for each user $i \in \mathcal{I}_t$, then the scheduler becomes optimal at TTI *t* from the viewpoint of objective $o \in \mathcal{O}$. Therefore, the aggregate function for all users and objective $\forall o \in \mathcal{O}$ becomes: $\phi_o\left(\mathrm{y}_o[t]\right) = (1/I_t) \cdot \sum_{i=1}^{I_t} \phi_{o,i}\left(\mathrm{y}_{o,i}[t]\right)$.

In DSR-CMOO problems, the impact of each scheduling rule in each objective is strongly required. In this sense, the aggregate multi-objective function $\Phi_{o,u_o}: \mathbb{R}^{I \times O} \rightarrow \mathbb{R}$ when applying the scheduling rule $D_{o,u,i}$, can be represented as indicated in (16):

$$\Phi_{o,u_o}\left(\mathrm{y}[t]\right) = \sum_{o_*}^{O} \delta_{o_*} \cdot \phi_{o_*,u}\left(\mathrm{y}_{o_*}[t]\right) \tag{16}$$

where $\mathrm{y} = [\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_O]$ and $\phi_{o_*,u_o}: \mathbb{R}^I \rightarrow \mathbb{R}$ is the aggregate function of objective $o_* \in \mathcal{O}$ when the scheduling rule $D_{o,u_o,i}$ is considered for each user $i \in \mathcal{I}_t$. Parameter δ_{o_*} is the weight for each particular objective function ϕ_{o_*,u_o}, $\forall o_* \in \mathcal{O}$. The necessary and not sufficient condition to be satisfied by the aggregate multi-objective function at each TTI *t* for the entire set of active users is:

$$\Phi_{o,w_o}\left(\mathrm{y}[t]\right) \geq 0 \tag{17}$$

Equation (17) should be satisfied only and only the condition (16) is met for each user $i \in \mathcal{I}_t$ and for each objective $o \in \mathcal{O}$. Precise details about each scheduling objective are provided in the following sub-sections.

4.1 Throughput Maximization

For throughput maximization ($o = 1$), the utility argument is $\mathrm{x}_i[t] = R_i$ and the weight function is $W_{1,u_1,i}\left(\mathrm{y}_i[t]\right) = 1$. The multi-objective evaluator grants the scheduling performance according to objective function $\phi_{1,i}: \mathbb{R} \rightarrow \mathbb{R}$, $\forall i \in \mathcal{I}_t$ where $\phi_{1,i}[t] = \sum_i^{I_t} T_i[t] - \sum_i^{I_t} T_i[t-1]$. The role of this particular optimization problem is to increase the total cell throughput at each TTI in such a way that $\phi_{1,i}[t] \geq 0$. If in (13) the product is $W_{o,u,i}\left(\mathrm{y}_{o,i}[t]\right) \cdot F_i'\left(\mathrm{x}_i[t]\right) = 1$, then we obtain the Maximum Rate scheduling rule, aiming to maximize at each TTI the total cell spectral efficiency.

4.2 User Fairness

The user fairness should be guaranteed at each TTI while affecting the system capacity maximization. Then, a new scheme is required in order to give more flexibility to the system throughput improvement. The time window (a given number of TTIs) used in user throughput computation constrains or relaxes the fairness performance depending on its length. By adopting $\mathrm{x}_i[t] = T_i$ as an argument for the marginal utility function, the resource allocation at TTI t depends on the allocation history in the previous TTIs. The averaging procedure can be achieved basically in two ways:

- By using the Exponential Moving Filter (EMF): The forgetting factor β is used to control the system throughput and user fairness tradeoff as indicated in (3), where $\beta = 1_{TTI} / T_{Ew}$, and T_{Ew} is the time window length. This means that a higher average throughput implies a lower priority for that user to be selected on the considered RB. The only condition is to set β larger than the channel correlation time in order to exploit the time diversity principle as revealed by Song (2005). When the time window T_{Ew} is too large, the cell spectral efficiency is affected whereas when the time window is too small, then the user fairness is not sensed anymore.
- By using the Median Moving Filter (MMF): The idea is to store the instantaneous user throughputs for a given time window T_{Mw} and to use the mean value of these observations at each TTI in order to balance the system throughput and user fairness tradeoff.

The NGMN fairness is one of many others criteria that can be used to compute the objective function According to this principle detailed by Proebster et al. (2010), the CDF calculated for a given set of average/instantaneous user throughputs should not exceed a given NGMN threshold. Based on some convergence studies presented by Song (2005), the local optimization considers $\mathrm{x}_{2,i}[t] = \overline{T}_i$ as an argument for the utility function and the weight function is $W_{2,u,i}\left(\mathrm{y}_{2,i}[t]\right) = 1$ since the QoS requirements are not included at this stage. A particular type of marginal utility function is $U'_{2,i}\left(\overline{T}_i[t]\right) = 1/\overline{T}_i[t]$ which implies the metric of $m_j[t] = arg\,max_{i\in\mathcal{I}}\left\{r_{i,j}[t]/\overline{T}_i[t]\right\}$, known in the literature as the Proportional Fair (PF) scheduling rule, proposed by Kelly (1997).

In particular, the NGMN fairness objective function can be determined according to the following formula: $\phi_{2,i}[t] = \psi_i^R\left(NT_i\right) - \psi_i\left(NT_i\right)$, where $NT_i = T_i / \sum T_m$ is the normalized user $i \in \mathcal{I}_t$ throughput, $\psi_i\left(NT_i\right)$ is the CDF function of user $i \in \mathcal{I}_t$ for a given distribution of input normalized observations and $\psi_i^R\left(NT_i\right)$ represents the NGMN fairness requirement.

4.3 Guaranteed Bit Rate

When the rate constraint satisfaction ($o = 3$) is considered in the optimization problem, the utility function weight is composed by: $\mathrm{x}_{3,i}[t] = \overline{T}_i$ and $\mathrm{y}_{3,i}[t] = \overline{\overline{T}}_i$, where $\overline{\overline{T}}_i[t]$ represents the average user throughput calculated by using the median moving filter. This way, the first objective is to satisfy all users' GBR requirements and then to focus more on user fairness by scheduling those users with lower throughputs. In this way, we measure the performance of such optimization problem by developing the following objective function: $\phi_{3,i}[t] = \overline{\overline{T}}_i[t] - T_i^R[t]$, where T_i^R is the GBR requirement of user $i \in \mathcal{I}_t$. According to (15), the mean user throughput should be greater than the GBR requirement at each TTI t in order to meet its objective.

4.4 HoL Packet Delay

If the utility weight depends on the instantaneous HoL packet delay ($o = 4$) such that $\mathrm{y}_{4,i}[t] = d_i$ and $\mathrm{x}_{4,i}[t] = \overline{T}_i$, then the optimization problem considers the HoL packet delay as the first priority in the satisfaction of the performance criterion. Then, the delay based objective function that has to be maximized for each active data queue at each TTI t becomes $\phi_{4,i}[t] = d_i^R[t] - d_i[t]$, where d_i^R is the HoL delay requirement.

4.5 Packet Loss Rate

The packet loss objective ($o = 5$) represents an important performance target in OFDMA packet scheduling. The utility weight depends on $\mathrm{y}_{5,i}[t] = L_i[t]$ and the utility argument keeps a similar form of $\mathrm{x}_{5,i}[t] = \overline{T}_i$. The objective function used to measure the performance of radio resource allocation from the viewpoint of PLR performance becomes: $\phi_{5,i}[t] = L_i^R[t] - L_i[t]$, where L_i^R is the PLR requirement. The PLR rate $L_i[t]$ is computed as a ratio between the number of lost packets and total number of transmitted packets. The same time moving window length used to compute the mean user throughput $\overline{\overline{T}}_i$ can be used also in this case to calculate the instantaneous packet loss at each TTI.

5. MULTIPLE OBJECTIVE OPTIMIZATION PROBLEM

Optimization problems similar (12) address only particular SMOO problems and each problem is linear guaranteeing at the same time the global optimal solution when selecting the decision variables for the radio resource assignment. By adopting different utility functions, the scheduling procedure impacts differently in multi-objective problem. We formulate the aggregate function by considering the utility functions introduced in the previous sub-sections, such as:

$$U(\mathrm{x}) = 1/(U \cdot I_t) \cdot \sum_o^O \sum_{u_o}^{U_o} \sum_i^{I_t} U_{o,u_o,i}(x_{o,i}) \tag{18}$$

The proposed multi-objective problem in the long term purpose aims to maximize the sum of utilities for each scheduling objective as indicated by the following statement:

$$\max_{\mathcal{X} \subset \mathcal{S}_C} \left[\sum_t U(\mathrm{x}[t]) \right] \tag{19}$$

By adopting the first order approximation of Taylor's expansion for each of the utility functions, the instantaneous multi-objective optimization problem becomes:

$$\max_{r[t] \in \mathcal{R}_b} \left[\sum_o \sum_{u_o} \sum_i \sum_j W_{o,u_o,i}(y_{o,i}[t]) \cdot F'_{o,i}(x_{o,i}[t]) \cdot r_{i,j}[t] \right] \tag{20}$$

According to (20), we need a policy that selects different utility functions at each TTI from the pool of utilities $\mathcal{U}$. Similar to decision matrix $b[t]$, instead of users, the new decision matrix will consider the number of objectives and instead of resource blocks, the policy takes into account the existing scheduling rules for each objective. Also, at each TTI, the selection of only one objective is required while multiple users can be selected within one TTI when following the decision matrix $b[t]$. Then, we define by $c[t] = \{c_{o,u_o}[t]\}$ the scheduling rule decision matrix, where $o = 1,...,O,$ and $u_o = 1,...,U_o$. According to (20), for each active user $i \in \mathcal{I}_t$, the same marginal utility $u_o \in \mathcal{U}_o$ must be assigned at each TTI t. Also, we need an additional variable able to assign a marginal utility function to each active user. In this sense, the

$$
\begin{aligned}
&\max_{c,w,b} \sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\sum_{i=1}^{I_t}\sum_{j=1}^{B} c_{o,w_o}[t]\cdot w^{o}_{u_o,i}[t]\cdot b_{i,j}[t]\cdot W_{o,u_o,i}\left(y_{o,i}[t]\right)\cdot F'_{o,i}\left(x_{o,i}[t]\right)\cdot r_{i,j}[t] \\
&\quad \sum_o \sum_{u_o} c_{o,u_o}[t] = 1 && (a) \\
&\quad \sum_o \sum_{u_o} w^{o}_{u_o,i}[t] = 1,\ i = 1,...,I_t && (b) \\
&\quad \sum_i w^{o}_{u^*_o,i}[t] = U_t,\ u^*_o \in \mathcal{U}_o, o \in \mathcal{O} && (c) \\
&\quad \sum_i w^{o}_{u^{\otimes}_o,i}[t] = 0,\ u^{\otimes}_o = 1,...,U_o, o = 1,...,O, \forall u^{\otimes}_o \neq u^*_o && (d) \\
&s.t.\ \ \sum_i b_{i,j}[t] \le 1,\ j = 1,...,B && (e) \\
&\quad c_{o,u_o}[t]\cdot \Phi_{o,u_o}[t+1] \ge 0,\ o = 1,...,O, u_o = 1,...,U_o && (f) \\
&\quad c_{o,u_o}[t] \in \{0,1\},\ \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o \\
&\quad w^{o}_{u_o,i}[t] \in \{0,1\},\ \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o, \forall i \in \mathcal{I}_t \\
&\quad b_{i,j}[t] \in \{0,1\},\ \forall i \in \mathcal{I}_t, \forall j \in \mathcal{B}
\end{aligned}
\qquad (21)
$$

matrix $w[t] = \left\{ w_{u_o,i}[t], u_o = 1,...,U_o, i = 1,...,I_t \right\}$ assigns the scheduling rule for objective $\forall o \in \mathcal{O}$ to each user $i \in \mathcal{I}_t$ at TTI *t*. Also, this matrix differs from one TTI to another when addressing DSR-CMOO problems.

The multi-objective optimization problem in OFDMA scheduling is formulated in (21), where the first constraint denotes the necessary condition of selecting at each TTI *t* only one scheduling rule according to the addressed objective. Constraints (b) indicate that only one marginal utility function is selected for the entire set of active users at each TTI *t*. Constraints (c) and (d) indicate that the same marginal utility function is assigned to all users at each TTI. Constraints (e) are the well-known conditions of assigning resource blocks to different users. Finally, the set of constraints (f) considers the aggregate multi-objective condition from (16). This implies that for the selected rule ($c_{o,u_o}[t]=1$), the sum of aggregate functions for each objective at TTI *t*+1 should be greater than zero since the evaluation of the scheduling procedure is performed in the next TTI. If the objective conditions from (21.f) are satisfied $\forall o \in \mathcal{O}$ and $\forall u_o \in \mathcal{U}$, then the scheduler is optimal when a given DSR-CMOO problem is addressed.

The idea is to find at each TTI *t* the optimal set of decision variables in order to maximize the optimization problem and to respect the set of constraints. Due to the product $c_{o,u_o}[t]\cdot w^{o}_{u_o,i}[t]\cdot b_{i,j}[t]$, the optimization problem becomes non-linear and thus, the optimal solution in not guaranteed. In (21), the MOO optimization can be divided into three categories based on the dynamicity of the rule selection:

- **SSR-SMOO**: occurs when objective $\forall o \in \mathcal{O}$ and the marginal utility function $\forall u_o \in \mathcal{U}_o$ are static over the entire transmission session. In this case, the SSR-SMOO problem refers to the classical optimization problems from (12) and constraints (a-d) and (f) from (21) are not required;
- **DSR-SMOO**: if $\forall o \in \mathcal{O}$ is static over the entire downlink scheduling session and $u_o[t] \in \mathcal{U}_o$ is variable TTI-by-TTI. In this case, different marginal utility functions are used concurrently in order to achieve the same target or objective, and constraints (f) consider only the satisfaction of the particular aggregate objective function;
- **DSR-CMOO** when both $o[t] \in \mathcal{O}$ and $u_o[t] \in \mathcal{U}_o$ are variable over time. Different utility functions with different objective targets may be applied in order to achieve the aggregate objective concomitantly. Only in this particular case, the aggregate multi-objective conditions or constraints (f) are fully taken into account.

By selecting the decision variable $c_{o,u_o}[t]$ at each TTI, the scheduler evolves from state $\mathrm{S}[t] \in \mathcal{S}$ to the next one $\mathrm{S}[t+1] = \mathrm{S}' \in \mathcal{S}$. Due to the time dependence process, the optimization problem from (21) is *dynamic*. The newest state $\mathrm{S}' \in \mathcal{S}$ contains the momentary uncontrollable subspace $\mathrm{Z}[t+1] \in \mathcal{S}_U$ which is not depending on decided variables $\left(c_{o,u_o}[t], w^{o}_{u_o,i}[t], b_{i,j}[t]\right)$ at TTI *t*. For these reasons, the problem from (22) is considered ***dynamic and stochastic***. Solving these combinatorial problems is not trivial since to find the best decision variables requires consistent computational time and system complexity. Developing a policy of scheduling rules is one of the best ways to ease the decision making at each TTI. We define a policy of scheduling rules as $\pi = \left\{c_{o,u_o}[t]; o = 1,..,O, u_o = 1,..,U_o, t = 1,...,\infty\right\}$ represented by a generic set of scheduling rules or marginal utilities that are applied dynamically TTI-by-TTI based on the momentary scheduler states. An example of such policy is indicated in (22):

$$\pi = \left\{c_{1,3}[t], c_{3,2}[t+1], c_{2,5}[t+2], c_{4,3}[t+3]....\right\} \tag{22}$$

The optimality of such policies (π^*) relies the selection of the best decision variables at each TTI in such a way that the set of constraints from (21) is fully satisfied and the outcome of QoS satisfaction is maximized over time. The optimization and refinement of such sequences abovementioned are not trivial due to the stochastic nature of the process which requires an infinite state space for searching the optimal solution. Two main approaches can be proposed for the policy optimization:

1. **Evolutionary methods**: e.g., expression and evolutionary programming;
2. **Dynamic programming methodologies**: e.g., real-time dynamic programming and temporal difference based learning algorithms such as ***reinforcement learning*** techniques.

Under the assumptions of constant power allocation and the sub-optimal MCS allocation, the complexity of DSR-CMOO problems is $\mathrm{O}(O \times U \times I_t \times B)$. Each algorithm above-mentioned requires a reasonable number of scenarios in order to fine tune the final policy for the real-time downlink scheduling.

6. MULTI-OBJECTIVE SCHEDULING BASED ON REINFORCEMENT LEARNING

Once the scheduling rule variable $c_{o,u_o}[t]$ is fixed, the entire aggregate problem is reduced to a simple resource allocation procedure. To solve such non-linear optimization problems, three approaches can be adopted, as follows:

- **Sequential Problem Linearization**: converts the non-linear problem into its corresponding linear representation. Unfortunately, the computation complexity increases with the size of $\mathcal{U}$, and this approach becomes immediately unsuitable for real-time OFDMA scheduling.
- **Parallel Problem Linearization**: divides the non-linear multi-objective problem into *U* linear sub-problems. Basically, this approach aims to run different schedulers in parallel by performing different scheduling rules. After the assignment of RBs is performed for each parallel process, the scheduling rule which maximizes the optimization problem and respects the constraint set is selected. This approach becomes unsuitable when optimizing the fairness objective and infinite number of utilities is considered due to the continuous parameterization of the PF rule.
- **Sequential Problem Linearization in Two Stages**: divides the non-linear multi-objective problem in two different stages of linear optimization problems. The solution is sub-optimal, but with some optimization tools the optimal solution can be very well approximated.

Due to its reduced complexity, we adopt the last solution that actually is solving the multi-objective satisfaction maximization in the first instance, followed by the simple resource allocation problem in the second instance, being governed by the selected scheduling rule from the first optimization problem.

6.1 Sequential Linearization in Two Stages

The main task of the linear optimization problem in (21) is to determine the best decision variable $c_{o,u_o}[t]$ at each TTI t in order to maximize the problem and to respect the given set of constraints. But this procedure is not guaranteeing the satisfaction of constraints (f) which implicitly highlights the performance of the entire scheduling procedure when one scheduling rule has been applied. In order to tackle this problematic issue, these constraints must to be included in the optimization problem by using relaxation methods. In this sense, the Augmented Lagrangian function and the dual optimization problem are required (Nocedal and Wright, 2006). We define the augmented Lagrangian for our problem as:

$$\mathcal{L}_A : \mathbb{R}^{O\times U} \times \mathbb{R}^{U\times I} \times \mathbb{R}^{I\times B} \times \mathbb{R}^{O\times U} \to \mathbb{R}$$

$$\mathcal{L}_A\left(c,w,b,\Phi^{\mathsf{A}}\right) = \sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\sum_{i=1}^{I_t}\sum_{j=1}^{B} c_{o,u_o}[t]\cdot w_{u_o,i}^{o}[t]\cdot b_{i,j}[t]\cdot U'_{o,u_o,i}\left(x_{o,i}[t]\right)\cdot r_{i,j}[t] + \quad (1)$$

$$\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\Phi_{o,u_o}^{\mathsf{A}}[t]\cdot c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1] + \quad (2) \qquad (23)$$

$$\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\frac{\mu_{o,u_o}}{2}\cdot\left(c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1]\right)^2 \quad (3)$$

where the first term is the optimization function to be maximized from (21), the second term represents the Lagrange relaxation function and finally, the third one is the penalty function as given by Nocedal and Wright (2006). Basically, the augmented Lagrangian is considered to be a combination of Lagrange relaxation and penalty methods in solving complex constrained optimization problems. In (23), μ_{o,u_o} is the penalty factor and $\Phi_{o,u_o}^{\mathsf{A}}[t]$ is the accumulated Lagrange multiplier that has to be updated TTI-by-TTI. According to Nocedal and Wright (2006), the accumulated Lagrange multiplier is updated by using the following formula:

$$\Phi_{o,u_o}^{\mathsf{A}}[t+1] = \Phi_{o,u_o}^{\mathsf{A}}[t] + \mu_{o,u_o}\cdot c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1] \quad (24)$$

and $\Phi^{\mathsf{A}}[t] = \left\{\Phi_{o,u_o}^{\mathsf{A}}[t]\right\}$ is the matrix of Lagrange multipliers at TTI t and $\mu = \left\{\mu_{o,u_o}\right\}$ is the penalty matrix for each objective $o\in\mathcal{O}$ and for each marginal utility function $u_o \in \mathcal{U}_o$. Then, let us define the concave Lagrange dual function $\mathcal{G}_A\left(\Phi^{\mathsf{A}}\right)$ which is defined as shown in (25):

$$\mathcal{G}_A : \mathbb{R}^{O\times U} \to \mathbb{R},\quad \mathcal{G}_A\left(\Phi^{\mathsf{A}}\right) = \sup_{c,u,b}\mathcal{L}_A\left(c,w^{o},b,\Phi^{\mathsf{A}}\right) \quad (25)$$

The objective is to find the optimal Lagrange dual function $\mathcal{G}_A\left(\Phi^{\mathsf{A}*}\right)$ at each TTI t in such a way that:

$$\mathcal{G}_A\left(\Phi^{\mathsf{A}*}[t]\right) = \sup_{c,w,b}\left[\mathcal{L}_A\left(c[t],w^{o}[t],b[t],\Phi^{\mathsf{A}*}[t]\right)\right] \geq \mathcal{L}_A\left(c^{*}[t],w^{o*}[t],b^{*}[t],\Phi^{\mathsf{A}}[t]\right) \quad (26)$$

where $\left\{c^{*}[t],w^{o*}[t],b^{*}[t]\right\}$ are the optimal assignment matrices at TTI t and $\Phi^{\mathsf{A}*}[t]$ is the optimal matrix of Lagrange multipliers being calculated online at each TTI t. The role of the Lagrange dual function is to learn the optimal Lagrange multipliers and to take the assignment decisions based on their optimized values at each TTI. When the learned matrix of Lagrange multipliers is optimal, then the scheduling decision variables are optimal. Based on these aspects, the dual optimization problem is presented in (28).

The multi-objective problem exposed in (27) is a non-linear programming problem where term (1) and (2) in the optimization problem aim to select the best scheduling decision matrix in order to maximize the accumulated Lagrange multiplier and the aggregate multi-objective function at TTI t+1, whereas the third term is the typical resource allocation procedure performed based on the selected marginal utility function

$$\max_{c,w,b}\Bigg\{\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\Phi^{\mathsf{A}}_{o,u_o}[t]\cdot c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1]+ \quad (1)$$

$$\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\frac{\mu_{o,u_o}}{2}\cdot\left(c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1]\right)^2+ \quad (2)$$

$$\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\sum_{i=1}^{I_t}\sum_{j=1}^{B}c_{o,u_o}[t]\cdot w^{o}_{u_o,i}[t]\cdot b_{i,j}[t]\cdot U'_{o,u_o,i}\left(x_{o,i}[t]\right)\cdot r_{i,j}[t]\Bigg\} \quad (3)$$

$$s.t.\quad \begin{aligned}
&\textstyle\sum_o\sum_{u_o}c_{o,u_o}[t]=1 && (a)\\
&\textstyle\sum_o\sum_{u_o}w^{o}_{u_o,i}[t]=1,\ i=1,...,I_t && (b)\\
&\textstyle\sum_i w^{o}_{u^*_o,i}[t]=U_t,\ u^*_o\in\mathcal{U}_o,o\in\mathcal{O} && (c)\\
&\textstyle\sum_i w^{o}_{u^{\otimes}_o,i}[t]=0,\ u^{\otimes}_o=1,...,U_o,o=1,...,O,\forall u^{\otimes}_o\neq u^*_o && (d)\\
&\textstyle\sum_i b_{i,j}[t]\le 1,\ j=1,...,B && (e)\\
&c_{o,u_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall u_o\in\mathcal{U}_o\\
&w^{o}_{u_o,i}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall u_o\in\mathcal{U}_o,\forall i\in\mathcal{I}_t\\
&b_{i,j}[t]\in\{0,1\},\ \forall i\in\mathcal{I}_t,\forall j\in\mathcal{B} && (27)
\end{aligned}$$

By selecting the optimal matrix $c^*[t]$ in the first term, the second term is also maximized. Thus, the proposed sequential linearization method aims to split the non-linear optimization problem into two sub-optimal linear sub-problems as follows:

- **First stage**, the scheduling rule that maximizes the product between the accumulated Lagrange multiplier at TTI t and the aggregate multi-objective function at TTI t+1 must be selected as indicated in the following equation:

$$\begin{aligned}
&\max_{c}\sum_{o=1}^{O}\sum_{u_o=1}^{U_o}\Phi^{\mathsf{A}}_{o,u_o}[t]\cdot c_{o,u_o}[t]\cdot\Phi_{o,u_o}[t+1]\\
&s.t.\quad \sum_{o=1}^{O}\sum_{u_o=1}^{U_o}c_{o,u_o}[t]=1\\
&\qquad c_{o,u_o}[t]\in\{0,1\},\ \forall o\in\mathcal{O},\forall u_o\in\mathcal{U}_o
\end{aligned} \quad (28)$$

- **Second stage**, the allocation procedure of radio resources for the active users is performed based on the selected rule from the first stage and the optimization problem is similar to (12).

The linear optimization problem from (28) can be solved by selecting the decision variable $c_{o,u_o}[t]$ according to the accumulated Lagrange multiplier $\Phi^{\mathsf{A}}_{o,u_o}[t]$ such that, the instantaneous aggregate multi-objective function $\Phi_{o,u_o}[t+1]$ would be maximized. There are two main problems in selecting the optimal decision variable $c^*_{o,u_o}[t]$ at each TTI: a) the scheduling policy π has to be optimized and the accumulated Lagrange multiplier must be updated by many TTI-to-TTI iterations; b) the optimization process of the scheduling policy is practically impossible since the scheduler state space is not considered when the Lagrange multiplier $\Phi^{\mathsf{A}}_{o,u_o}$ and the aggregate multi-objective function Φ_{o,u_o} are computed.

6.2 Scheduler State Space in Multi-Objective Optimization

Let us define the set of neighbor states $\mathcal{N}(\mathrm{s}) \subset \mathcal{S}$ being composed by those possible states to which the current state $\mathrm{s}[t]$ could evolve based on different selections of $c_{o,u_o}[t]$, and then the next state is $\mathrm{s}' \in \mathcal{N}(\mathrm{s})$. By using the terminology from the machine learning domain as given by Sutton and Barto (2017), the action value $Q_{o,u_o}(\mathrm{s})$ and the reward $RW_{o,u_o}(\mathrm{s}')$ functions are obtained based on the accumulated Lagrange multiplier and aggregate multi-objective functions, respectively, as follows:

$$\begin{aligned} &\Phi_{o,u_o}[t+1] \mapsto RW_{o,u_o}(\mathrm{s}'), \qquad RW_{o,u_o} : \mathcal{O} \times \mathcal{U} \times \mathcal{N}(\mathrm{s}) \to \mathbb{R} \\ &\Phi^{\mathsf{A}}_{o,u_o}[t] \mapsto Q_{o,u_o}(\mathrm{s}), \qquad\quad Q_{o,u_o} : \mathcal{O} \times \mathcal{U} \times \mathcal{S} \to \mathbb{R} \end{aligned} \tag{29}$$

where the reward function $RW_{o,u_o}(\mathrm{s}')$ measures the performance of applying the scheduling rule corresponding to the decision variable $c_{o,u_o}[t]$ when the current state is state $\mathrm{s} \in \mathcal{S}$; $Q_{o,u_o}(\mathrm{s})$ is the accumulated reward for the decision variable $c_{o,u_o}[t]$ being applied only in the scheduler state $\mathrm{s} \in \mathcal{S}$ for an infinite number of visits. By using the above notations, the dual optimization problem is highlighted in the following equation:

$$\begin{aligned} &\max_{c,d} \sum_{o=1}^{O} \sum_{u_o=1}^{U_o} \sum_{s=1}^{O \times U} Q_{o,u_o}(\mathrm{s}) \cdot c_{o,w_o}[t] \cdot d^{o}_{u_o,s}[t] \cdot RW_{o,u_o,s}(\mathrm{s}') \\ &\quad \sum_{o=1}^{\mathcal{O}} \sum_{u_o=1}^{U_o} c_{o,w_o}[t] = 1 \\ &s.t. \ \sum_{s=1}^{O \times U} d^{o}_{u_o,s}[t] = 1, \ o = 1,...,O, u_o = 1,...,U_o \\ &\quad c_{o,u_o}[t] \in \{0,1\}, \ \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o \\ &\quad d^{o}_{u_o,s}[t] \in \{0,1\}, \ \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o, \forall s \in \mathcal{N}(\mathrm{s}) \end{aligned} \tag{30}$$

where $d^{o}_{u_o,s}[t]$ is the variable that decides the next state $\mathrm{s}' \in \mathcal{S}$ and the considered assignation matrix is $d^{o}[t] = \left\{ d^{o}_{u_o,s}[t], u_o = 1,...,U_o, s = 1,...,O \times U_o \right\}, \forall o \in \mathcal{O}$. The optimization problem in (30) is non-linear due to the product between the scheduling rule and the next state variables such as: $c_{o,u_o} \cdot d^{o}_{u_o,s}$. An additional variable is needed for linearization, such that $e_{o,u_o,s}[t] = c_{o,u_o}[t] \cdot d^{o}_{u_o,s}[t]$, where the matrix which indicates the scheduler state evolution at TTI *t*+1 when the decision variable $c_{o,w_o}[t]$ has been applied in the previous state is $e_o[t] = \left\{ e_{o,u_o,s}[t], u_o = 1,...,U_o, s = 1,...,O \times U_o \right\}$. Then, the obtained linearized optimization problem is exposed in (31), where the first set of constraints (a) acts as an AND gate, where the input variables $\left\{ c_{o,u_o}; d^{o}_{u_o,s} \right\} \in \{0,1\}$ count four possible binary combinations and the parameter $e_{o,u_o,s}[t] \in \{0,1\}$ is the output variable, where $l \in \mathbb{R}_+$ is a large positive number. If $e_{o,u_o,s}[t] = 1$, then $\sum_o \sum_{u_o} \sum_s e_{o,u_o,s}[t] = 1$ which means that the scheduler is evolving to an unique state at TTI *t*+1 based on the selected objective $\forall o \in \mathcal{O}$ and utility function $\forall u_o \in \mathcal{U}_o$. Constraints (b) indicate that only one scheduling rule focused on a particular objective is selected. Constraints (c) associate only one scheduler state $\mathrm{s}' \in \mathcal{N}(\mathrm{s})$ according to the selected scheduling rule in state $\mathrm{s} \in \mathcal{S}$. Constraints (d)-(f) make the entire problem combinatorial.

$$
\begin{aligned}
&\max_{c,d} \sum_{o=1}^{O} \sum_{u_o=1}^{U_o} \sum_{s=1}^{N(\mathrm{s})} Q_{o,u_o}(\mathrm{s}) \cdot e_{o,u_o,s}[t] \cdot RW_{o,u_o,s}(\mathrm{s}') \\
&\left.\begin{aligned}
&e_{o,u_o,s}[t] \le c_{o,u_o}[t] \\
&e_{o,u_o,s}[t] \le d^{o}_{u_o,s}[t] \cdot l \\
&e_{o,u_o,s}[t] \ge c_{o,u_o}[t] - \left(1 - d^{o}_{u_o,s}[t]\right) \cdot l \\
&\quad \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o, \forall s \in \mathcal{N}(\mathrm{s})
\end{aligned}\right\} (a) \\
&\text{s.t.} \quad \sum_{o=1}^{O} \sum_{w_o=1}^{U_o} c_{o,u_o}[t] = 1 \qquad (b) \\
&\quad \sum_{s=1}^{O \times U} d^{o}_{u_o,s}[t] = 1, \; o = 1,...,O, u_o = 1,...,U_o \qquad (c) \\
&\quad c_{o,u_o}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o \qquad (d) \\
&\quad d^{o}_{u_o,s}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o, \forall s \in \mathcal{N}(\mathrm{s}) \qquad (e) \\
&\quad f_{o,u_o,s}[t] \in \{0,1\}, \; \forall o \in \mathcal{O}, \forall u_o \in \mathcal{U}_o, \forall s \in \mathcal{N}(\mathrm{s}) \qquad (f)
\end{aligned}
\tag{31}
$$

6.3 Temporal Difference Learning for Multi-Objective Optimization Problem

The idea is to select the best decision variable $c_{o,u_o}[t]$ that maximizes the accumulated reward $Q_{o,u_o}(\mathrm{s})$ and to assign the state $\mathrm{s}' \in \mathcal{S}$ according to $d^{o}[t]$, such that, the RRM reward $RW_{o,u_o,s}(\mathrm{s}')$ at TTI *t*+1 is maximized. In real practice, when deciding the assignation variable $c_{o,u_o}[t]$, the future state $\mathrm{s}' \in \mathcal{S}$ is automatically determined at TTI *t*+1 as a result of the scheduling procedure. For this reason, the linear optimization problem exposed in (31) is valid only when the scheduling policy is optimal. In this case, by selecting the action (scheduling rule) with the maximum action value in state $\mathrm{s} \in \mathcal{S}$ then the reward maximization in the next state is guaranteed (Sutton and Barto, 2017). On this extent, in each momentary state, the best scheduling decision is determined such as $c^{*}_{o,u_o}[t] = arg\,max\, Q^{*}_{o,u_o}(\mathrm{s})$, where $Q^{*}_{o,u_o}(\mathrm{s})$ is the optimal accumulated reward value for objective $\forall o \in \mathcal{O}$ and scheduling rule $\forall u_o \in \mathcal{U}_o$ (or the optimal action value for action (o,u_o) in state $\mathrm{s} \in \mathcal{S}$). Then, the scheduler reward at TTI *t*+1 is maximized when the optimal decision is applied.

Similar to other control systems (as described by Sutton and Barto, 2017), the idea is to maximize the total expected return or the expected accumulated reward RW^{a}_{π} for a given policy of scheduling rules π starting from any initial state $\mathrm{s}[t] = \mathrm{s}$ until the optimal scheduler state $\mathrm{s}[t_o] = \mathrm{s}^{*}$ is reached, where t_o is the time needed to reach the optimal state $\mathrm{s}^{*} \in \mathcal{S}$ when given a certain order of applied scheduling rules following π from any random initial state $\mathrm{s}[t] \in \mathcal{S}$. The accumulated reward for the considered policy is discounted according to (32) (Sutton and Barto, 2017):

$$
\begin{aligned}
RW^{a}_{\pi}(\mathrm{s}[t]) &= RW_{\pi}(\mathrm{s}[t+1]) + ... + \gamma^{t_o} \cdot RW_{\pi}(\mathrm{s}[t+t_o+1]) = \sum_{t}^{t_o} \gamma^{t_o} \cdot RW_{\pi}(\mathrm{s}[t+t_o+1]) \\
&= RW_{\pi}(s[t+1]) + \gamma \cdot RW^{a}_{\pi}(\mathrm{s}[t+1])
\end{aligned}
\tag{32}
$$

where $\gamma \in [0,1]$ is the discount factor that sets the importance of future rewards. Equation (32) is the case of *temporal difference learning,* where the reward value can be deducted by following the reasoning:

$$RW_{\pi}\left(\mathrm{s}[t+1]\right) = RW_{\pi}^{a}\left(\mathrm{s}[t]\right) - \gamma \cdot RW_{\pi}^{a}\left(\mathrm{s}[t+1]\right) \tag{33}$$

We know that when the scheduling policy is optimal π^*, then the instantaneous scheduler reward $RW_{\pi}\left(\mathrm{s}[t+1]\right)$ is equivalent with the difference between the accumulated rewards from two consecutive states as indicated in (33). Actually, the accumulated reward RW_{π}^{a} in state $\mathrm{s} \in \mathcal{S}$ for a given policy π is similar to Q_{o,u_o} on that state. Under optimality conditions, the instantaneous reward from (34) can be re-written as follows:

$$RW_{\pi}\left(\mathrm{s}[t+1]\right) = Q_{o,u_o}^{*}\left(\mathrm{s}\right) - \gamma \cdot Q_{o',u_{o'}^{'}}^{*}\left(\mathrm{s}'\right) \tag{34}$$

where $\forall o' \in \mathcal{O}$ and $\forall u_{o'}^{'} \in \mathcal{U}_o$ are the selected objective and scheduling rule in state $\mathrm{s}' \in \mathcal{S}$.

The scheduling policy π needs to be improved by using many visits of state $\mathrm{s} \in \mathcal{S}$ in order to learn the optimal objective $o^* \in \mathcal{O}$ and the utility function $u_o^* \in \mathcal{U}_o$ that maximize the accumulated reward value $Q_{o,u_o}^{*}\left(\mathrm{s}\right)$. This stage is entitled the learning since all possible scheduling rules have to be tested for each given momentary scheduler states. Once this policy is refined and trained properly, the exploitation stage is performed. In this stage, the optimization problem exposed in (33) is satisfied since the scheduling rule that maximizes the accumulated reward $Q_{o,u_o}^{*}\left(\mathrm{s}\right)$ is selected and maximizes at the same time the reward value in next state $\mathrm{s}' \in \mathcal{S}$.

6.4 Reinforcement Learning in DSR-SMOO/CMOO Problems

The scheduler state space is continuous and multi-dimension and practically the size of the neighboring states $\left|\mathcal{N}\left(\mathrm{s}\right)\right| \to \infty$ is infinite due to the stochastic nature of the momentary states. Also, when optimizing the fairness criterion, an infinite number of scheduling rules are considered due to the continuous parameterization of PF scheduling rule. For these reasons the state-action pairs cannot be exhaustively enumerated and hence, the simplest look-up table is not suitable to store the $Q_{o,u_o}\left(\mathrm{s}\right)$ values for each momentary state and scheduling rule. Then, the proposed framework can only approximate the best decisions to be taken on each state. Function approximations such as neural networks, as proposed by Comşa et al. (2011) (2012).

The dimension of scheduler state is another important problem to be addressed. The main issue is the dependence on the number of active users and system bandwidth. For instance, the dimension of the controllable vector $\mathrm{c} = \left[\mathrm{c}_1, \mathrm{c}_2, ..., \mathrm{c}_{I_t}\right] \in \mathcal{S}_C$ at each TTI depends on the number of users I_t which is also variable given the fact that the amount of active users can change every TTI. Also, the CQI reports for each user depend on the number of resource blocks associated to a given bandwidth. Then, a procedure able to that compact the momentary scheduler state is absolutely necessary in order to avoid such dependencies and to enhance the learning procedure.

The machine learning framework must approximate the best scheduling selection at state-by-state. The Reinforcement Learning (RL) is a temporal-difference learning scheme able to provide very good solutions in various optimization problems (Sutton and Barto, 2017). The RL is considered a combination of Monte Carlo and Dynamic Programming methods. The Monte Carlo method provides the expected return (the sum of discounted rewards from state to state) only at the end of the learning stage and the

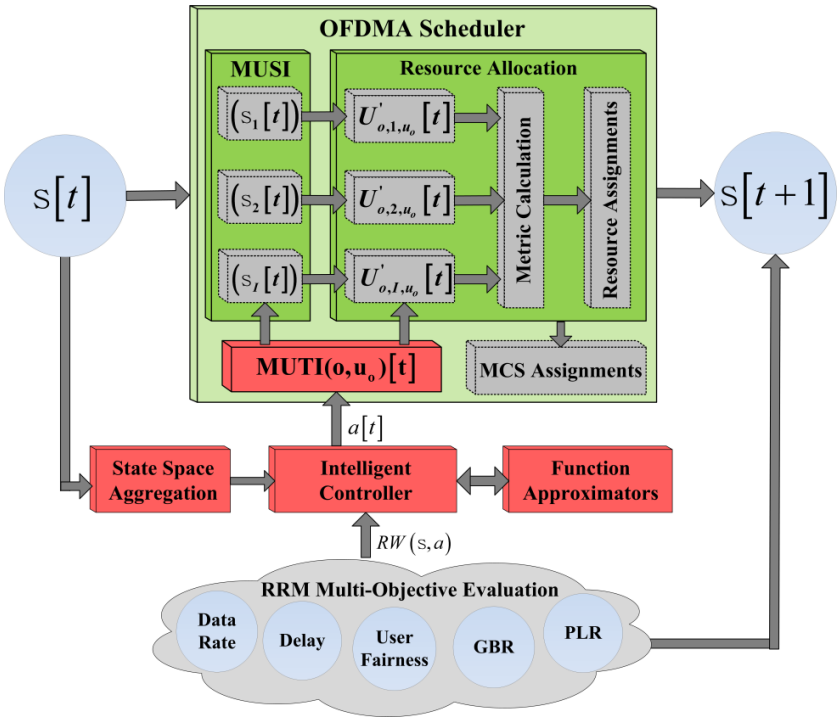

Figure 2. Reinforcement Learning based Solution for Multi-Objective Optimization Problem

tasks are non-episodic which means that the reward would eventually be maximized only at the end of the learning stage. On the other side, the dynamic programming method can approximate each action value function at each time and not at the end of the learning stage as is the case of Monte Carlo methods. Moreover, these action values can be updated according to the received reward in order to optimize a given policy. In DSR-SMOO/CMOO problems, the tasks are non-episodic for certain conditions (i.e. the satisfaction of all objectives may not be reached due to high number of users or/and poor channel conditions). However, the RL framework must approximate the best scheduling rule in each state such as the multi-objective satisfaction is maximized as much as possible.

Different RL algorithms can be used depending on the particularities of DSR-SMOO/CMOO problems, as stated in the research conducted by Comşa (2014a), Comşa et al. (2014b) (2014b), and Comşa et al. (2018). The role of the selected RL algorithm is to update function approximator for each selected action (o,u_o) by reinforcing each time the error $e=\overline{Q}_{o,u_o}(\mathrm{s})-Q_{o,u_o}(\mathrm{s})$, where the $\overline{Q}_{o,u_o}(\mathrm{s})$ is the approximated action value given by the approximator and $Q_{o,u_o}(\mathrm{s})$ is calculated based on (34) and reloaded here:

$$Q_{o,u_o}(\mathrm{s})=RW_{\pi}(\mathrm{s}')+\gamma\cdot\overline{Q}_{o',u_{o'}'}(\mathrm{s}') \tag{35}$$

It is important to notice that by updating the previous learned action value based on the temporal-difference error, $\overline{Q}_{o,u_o}(\mathrm{s})$ is decreased or increased by providing lower or higher probabilities for the decision variable $c_{o,u_o}[t]$ to be selected in the future when transiting nearly the same states from s to s'.

From the architectural point of view, the RL based DSR-CMOO approach defines two modules: Marginal Utility State Informer (MUSI) and Marginal Utility Type Informer (MUTI). MUTI converts the action which is provided by the intelligent controller in the corresponding scheduling rule such as $a[t]\rightarrow c_{o,u_o}[t]$

, where $a[t]$ is the controller action. Based on the MUTI decision, MUSI provides the necessary state parameters in order to compute the corresponding scheduling metrics for each user and for each resource block. Figure 2 presents the proposed RL–based DSR-CMOO architecture for the OFDMA downlink scheduling. The MCS assignment procedure for the transport block computation is executed in a separate stage once the radio resource allocation is performed. Studies conducted by Kwan, Leung, and Zhang (2009) indicate a degradation of the cell spectral efficiency of about 10% when the RB and MCS assignments are performed in separate stages, but at a much lower computational complexity when compared to the joint optimization approach.

As depicted in Fig. 2, in the first stage, the objective to be followed and the corresponding scheduling rule are selected in the first stage and then, the scheduling procedure is performed based on the data provided by MUSI and MUTI entities. In the learning stage, at TTI *t*, the momentary scheduler state $s[t] \in \mathcal{S}$ is observed. The state aggregation module aims to reduce the dimension of the initial state to a more compact representation in order to decrease the complexity of the proposed framework and to learn faster. Based on the aggregate momentary state, the function approximator provides the action values $\overline{Q}_{o,u_o}(s)$ for each objective $\forall o \in \mathcal{O}$ and utility $\forall u_o \in \mathcal{U}_o$. According to some probabilities, the controller may decide to apply the scheduling rule corresponding to $\left(o^*, u^*_{o^*}\right) = arg\,max_{o,u_o} Q_{o,u_o}(s)$ or to select a different scheduling rule in order to explore more action-state possibilities and to increase the quality of the learning stage. Once, the scheduling rule is decided, the resource allocation, MCS assignments and TB calculations are performed and the system evolves to the next state $s[t+1] = s' \in \mathcal{S}$. At TTI *t*+1, the RRM multi-objective evaluator determines the reward value, the error is computed according to (35) and reinforced in order to optimize the function approximator. The learning stage can continue for large number of state-to-state iterations until this error falls under a specified threshold. In the exploitation stage, the learnt function provides the scheduling rules to be applied in each state in order to maximize the multi-objective satisfaction.

7. CONCLUSIONS

This chapter proposes an aggregate multi-objective problem that aims to select in each scheduler state the most suitable scheduling rule in order to maximize the system throughput while increasing the satisfaction of scheduling objectives in terms of: user fairness, packet delay, user rate and packet loss rate. The proposed framework aims to minimize the drawback of each particular scheduling rule and maximize their benefits when applying each only on the best matching scheduler conditions. Due to the complexity of this aggregate optimization problem, the proposed solution splits the entire scheduling framework in two parts where, the first one deals with the selection of the scheduling rule in each momentary scheduler state and the second one performs the classical resource allocation problems. As part of machine learning domain, the reinforcement learning is proposed to learn over time the most appropriated scheduling rule to be applied in each state. Due to the complexity and the dimensionality problems of the scheduler state space, additional methods imported from artificial intelligence domain (i.e. data mining to compress the scheduler state, neural networks to approximate the best scheduling rule under each state) must be adopted in conjunction with the proposed reinforcement learning framework in order to make the proposed solution suitable for real-time downlink schedulers.

REFERENCES

Cisco. (2017). *Cisco Visual Networking Index: Global Mobile Data Traffic Fore-cast Update, 2016-2021 White Paper* Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ visual-networking-index-vni/mobile-white-paper-c11-520862.html.

Trestian, R., Comşa, I.-S., & Tuysuz, M. F. (2018). Seamless Multimedia Delivery within a Heterogeneous Wireless Networks Environment: Are We There Yet?. *IEEE Communications Surveys and Tutorials, 20*(2), 945 – 977.

G. Andrews, J., Buzzi, S., Choi, W., V. Hanly, S., Lozano, A., CK Soong, A., and Charlie Zhang, J. (2014). What will 5G be? *IEEE Journal on Selected Areas in Communications*, 1065-1082.

Li, Y., Pateromichelakis, E., Vucic, N., Luo, J., Xu, W., and Caire, G. (2017). Radio Resource Management Considerations for 5G Millimeter Wave Backhaul and Access Networks. *IEEE Communications Magazine*, 55(6), 86 – 92.

Olwal, T. O., Djouani, K. & Kurien, A. M. (2016). A Survey of Resource Management Toward 5G Radio Access Networks. *IEEE Communications Surveys and Tutorials*, 18(3), 1656 – 1686.

Comşa, I.-S. (2014a). *Sustainable Scheduling Policies for Radio Access Networks Based on LTE Technology*. University of Bedfordshire, Bedfordshire, U.K.

Jain, R., Chiu, D.M. & Hawe, W. A. (1984). Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System. In *Research Report*, DEC-TR-301, 1984.

3GPP. (2012). *Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture Release 12* . Retrieved from http://www.qtc.jp/3GPP/Specs/23203-a60.pdf.

Toufik, I. & Knopp, R (2011). Multi-User Scheduling and Interference Coordination. In S. Sesia et al. (Ed.), *LTE. The UMTS Long Term Evolution. From Theory To Practice* (pp. 279-292). John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom.

Capozzi, F., Piro, G., Grieco, L. A., Boggia, G. and Camarda, P. (2013). Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. *IEEE Communications Surveys Tutorials*, 15(2), pp. 678-700.

Proebster, M., Mueller, C.M. and Bakker, H. (2010). Adaptive Fairness Control for a Proportional Fair LTE Scheduler. *In Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)* (pp. 1504-1509). Istanbul, Turkey.

Liu, B. and Tian, H. & Xu, L. (2013). An Efficient Downlink Packet Scheduling Algorithm for Real Time Traffics in LTE Systems. *In Proceedings of IEEE Consumer Communications and Networking Conference (CCNC)* (vol. 1, pp. 364-369). Las Vegas, NV, USA.

Comşa, I. S., Aydin, M., Zhang, S., Kuonen, P., and Wagen, J.-F. (2011). Reinforcement learning based radio resource scheduling in LTE-advanced. In *Proceedings of 17th International Conference on Automation and Computing* (pp. 219 - 224). Huddersfield, UK.

Comşa, I. S., Aydin, M., Zhang, S., Kuonen, P., and Wagen, J.-F. (2012). Multi Objective Resource Scheduling in LTE Networks Using Reinforcement Learning. *International Journal of Distributed Systems and Technologies, 3*(2), 39-57.

FANTASTIC-5G. (2016). *D4.1: Technical Results for Service Specific Multi-Node/MultiAntenna Solutions*. Retrieved from http://fantastic5g.com/wp-content/uploads/2016/06/FANTASTIC-5G_D4.1_Final.pdf.

Lundevall, M., Olin, B., Olsson, J., Wiberg, N., Wanstedt, S., Eriksson, J., & Eng, F. (2004). Streaming Applications over HSDPA in Mixed Service Scenarios. In *Proceedings of IEEE Vehicular Technology Conference (VTC-Fall)* (vol.2, pp. 841 - 845). Los Angeles, CA, USA.

Ning, X., Ting, Z., Ying, W., & Ping, Z. (2006). A MC-GMR Scheduler for Shared Data Channel in 3GPP LTE System. In *Proceedings of IEEE Vehicular Technology Conference (VTC-Fall)* (pp. 1-5). Montreal, Quebec, Canada.

Zhang, J., Yuan, D., & Zhang, H. (2011). Joint Radio Resource Allocation and Scheduling in a Backhaul Constrained Multicell OFDMA Networks. *In Proceedings of* 13th *International Conference on Communication Technology (ICCT)* (pp. 47 – 51). Jinan, China.

Rhee, J.H., Holtzman, J.M., & Kim, D.-K. (2003). Scheduling of Real/Non-Real Time Services: Adaptive EXP/PF Algorithm. In *Proceedings of IEEE Semiannual Vehicular Technology Conference (VTC-Spring)* (vol.1, pp. 462 – 466). vol. 1 April 2003, pp. 462 – 466. Jeju, South Korea.

Sadiq, B., Madan, R., & Sampath, A. (2009). Downlink Scheduling for Multiclass Traffic in LTE. *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1-18.

Bae, S. J., Choi, B.-G., & Chung, M. (2011). Delay-Aware Packet Scheduling Algorithm for Multiple Traffic Classes in 3GPP LTE System. In *Proceedings of 17th Asia-Pacific Conference on Communications (APCC)* (pp. 33 – 37). Sabah, Malaysia.

Khan, N., Martini, M.G., Bharucha, Z., & Auer, G. (2012). Opportunistic Packet Loss Fair Scheduling for Delay-Sensitive Applications over LTE Systems. In *Proceedings of IEEE Wireless Communications and Networking Conference* (pp. 1456 – 1461). Shanghai, China.

Schwarz, S., Mehlfuhrer, C., & Rupp, M. (2011). Throughput Maximizing Multiuser Scheduling with Adjustable Fairness. In *Proceedings of IEEE International Conference on Communications (ICC)* (pp.1-5). Kyoto, Japan.

Comşa, I.-S., Zhang, S., Aydin, M., Kuonen, P., and Wagen, J. (2012). A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-advanced Technologies Using Neural Networks. In *Proceedings of IEEE Conference on Local Computer Networks (LCN)* (pp. 332 – 335). Clearwater, FL, USA.

Comşa, I.-S., Aydin, M., Zhang, S., Kuonen, P., Wagen, J.-F., & Lu, Y. (2014b). Scheduling Policies Based on Dynamic Throughput and Fairness Tradeoff Control in LTE-A Networks. In *Proceedings of IEEE Conference on Local Computer Networks (LCN)* (pp. 418-421). Edmonton, AB, Canada.

Comşa, I.-S., Zhang, S., Aydin, M., Chen, J., Kuonen, P., & Wagen, J.-F. (2014c). Adaptive Proportional Fair Parameterization Based LTE Scheduling Using Continuous Actor-Critic Reinforcement Learning. In *Proceedings of IEEE Global Communication Conference (GLOBECOM)* (pp. 4387 - 4393). Austin, TX, USA.

Monghal, G., Laselva, D., Michaelsen, P.-H., & Wigard, J. (2010). Dynamic Packet Scheduling for Traffic Mixes of Best Effort and VoIP Users in E-UTRAN Downlink. In *Proceedings of IEEE Vehicular Technology Conference (VTC-Spring)* (pp. 1-5). Taipei, Taiwan.

Chung, W.C., Chang, C. J., & Wang, L.C. (2012). An Intelligent Priority Resource Allocation Scheme for LTE-A Downlink Systems. *IEEE Wireless Communications Letters*, *1*(3), 241-244.

Wang, K., Li, X., Ji, H., & Zhang, X. (2013). Heterogeneous Traffic Scheduling in Downlink High Speed Railway LTE Systems. In *Proceedings of IEEE Global Communications Conference (GLOBECOM)* (pp. 1452-1457). Atlanta, GA, USA.

Avocanh, F.T.S., Abdennebi, M., & Ben-Othman, J. (2014). An Enhanced Two Level Scheduler to Increase Multimedia Services Performance in LTE Networks. In *Proceedings of IEEE International Conference on Communications (ICC)* (pp. 2351-2356). Sydney, NSW, Australia.

Comşa, I.-S., De Domenico, A., & Ktenas, D. (2017). QoS-Driven Scheduling in 5G Radio Access Networks - A Reinforcement Learning Approach. In *Proceedings of IEEE Global Communications Conference (GLOBECOM)* (pp. 1-7). Singapore, Singapore.

Comşa, I.-S., Trestian, R., & Ghinea, G. (2018). 360° Mulsemedia Experience over Next Generation Wireless Networks - A Reinforcement Learning Approach. In *Proceedings of Tenth International Conference on Quality of Multimedia Experience (QoMEX)* (pp. 1-6). Cagliari, Italy.

Comşa, I.-S., Zhang, S., Aydin, M. Kuonen, P., Lu, Y., Trestian, R., & Ghinea, G. (2018). Towards 5G: A Reinforcement Learning-based Scheduling Solution for Data Traffic Management. *IEEE Transactions on Network and Service Management (Early Access)*, 1-15.

Liu, X., Chong, E., and Shroff, N. (2001). Opportunistic Transmission Scheduling with Resource-Sharing Constraints in Wireless Networks. *IEEE Journal on Selected Areas in Communications, 19* (10), 2053-2064.

Song, G. and Li (Geoffrey), Y. (20015). Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks. *IEEE Communications Magazine*, *43*(12), 127-134.

Song, G. (2005). *Cross-Layer Resource Allocation and Scheduling in Wireless Multicarier Networks*. Georgia Institute of Technology, Atlanta, USA.

Kelly, F. (1997). Charging and Rate Control for Elastic Traffic. *European Transactions of Telecommunications*, *8*(1), 33-37.

Nocedal, J. and Wright, S. J. (2006). Penalty, Barrier, and Augmented Lagrangian Methods. In *Numerical Optimization* (pp. 488-524). Springer Series in Operations Research.

Sutton, R. S., and Barto, A. G. (2017). *Reinforcement Learning: An Introduction.* England/London: IMT Press Cambridge.

Kwan, R., Leung, C. and Zhang, J. (2009). Resource Allocation in an LTE Cellular Communication System. In *Proceedings of IEEE International Conference on Communications*, (pp. 1-5). Dresden, Germany.