Quality Fairness-oriented Cross-layer Resource Allocation for Scalable Video Delivery over OFDMA Wireless Networks

QUALITY FAIRNESS-ORIENTED CROSS-LAYER RESOURCE ALLOCATION FOR SCALABLE VIDEO DELIVERY OVER OFDMA WIRELESS NETWORKS

BY

KUAN LIN, B.Eng.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2015) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Quality Fairness-oriented Cross-layer Resource Allocation for Scalable Video Delivery over OFDMA Wireless Networks
AUTHOR:	Kuan Lin B.Eng., (Electronic Information Engineering) Univ. of Elect. Sci. and Tech. of China, Chengdu, China
SUPERVISOR:	Dr. Sorina Dumitrescu
NUMBER OF PAGES:	xv, 122

This thesis is dedicated to my parents Kexing and Pingli, and my sister Ying, for their support and love during my graduate studies.

Abstract

This thesis addresses the problem of scalable video delivery to multiple users over OFDMA wireless networks, where quality fairness and system efficiency are jointly considered. Fairness deals with the provision of a similar video quality to all users, while system efficiency concerns the maximization of the overall received video quality.

This problem was recently tackled by Cicalò and Tralli who proposed a cross-layer optimization framework with the aim of maximizing the sum of the ergodic rates while minimizing the distortion difference among multiple videos. The optimization problem was "vertically" decomposed into two subproblems: a source adaptation problem at the application (APP) layer and a resource allocation problem at the medium access control (MAC) layer. An iterative local approximation (ILA) algorithm was proposed to solve the two subproblems iteratively until the optimal solution is obtained. One drawback of the above work is that the APP layer algorithm to solve the source adaptation problem is unnecessarily complex. Moreover, the optimal solution may not be accurate since the adopted semi-analytical rate-distortion (R-D) model used for source adaptation is only an approximation of the empirical R-D data.

Our first main contribution is to overcome the aforementioned two drawbacks. To this end, we propose a quality fairness-oriented cross-layer optimization framework that solves a joint resource allocation and source adaptation (JRASA) problem

where the objective is to maximize the sum of the PSNRs while minimizing the PSNR difference among the received videos. The JRASA problem is equivalent to the aforementioned sum-rate maximization problem and capable of being solved by the ILA approach. On the other hand, it has a different formulation which naturally leads to the development of a considerably faster APP layer algorithm based on the bisection search method. Furthermore, we show that the above optimization framework can be extended to solve efficiently the JRASA problem based on accurate empirical R-D models, as well. The solution to the JRASA problem using the empirical R-D models can be used as a benchmark to assess the performance of solutions based on approximate R-D models, such as the semi-analytical R-D model.

Our second main contribution is an adjustable quality-fair cross-layer optimization framework, which is able to achieve trade-offs between quality fairness and system efficiency, aspect which was not considered by Cicalò and Tralli. Our procedure consists of two steps. First the aforementioned JRASA problem is solved. The second step seeks to maximize the sum of the PSNRs while limiting the absolute value of the relative difference between the PSNR of each video and the common PSNR value obtained in the first step. We show that the second problem is a general utility-based resource allocation problem, for which efficient algorithms are available to obtain an almost surely optimal solution.

Numerical results show that the proposed quality-fair optimization framework provides significantly better performance in terms of quality fairness and the provision of better quality to high-complexity videos with respect to an equal-rate adaptation scheme. Moreover, various trade-offs between fairness and system efficiency can be achieved using the adjustable quality-fair cross-layer optimization framework.

Acknowledgements

There are so many people I want to thank for making my experience as a graduate student at McMaster University one of the most exhilarating and challenging periods in my life. First of all, I would especially like to thank my supervisor Prof. Sorina Dumitrescu for her support and guidance, over the past two years, which make the work presented in this thesis possible. She is the model of integrity, work ethic and kindness. She always shows great faith in letting me to work independently on the challenging research project and shows great patience in helping me to overcome all the challenges. The experience of being her student will be always remembered.

I would also like to thank my other supervisory committee members, Prof. Dongmei Zhao and Prof. Jun Chen, for reading my thesis, participating in my thesis defence and assessing my research work. Their comments and advices are valuable input for my work.

I am very grateful to Dr. Sergio Cicalò from University of Ferrara, Italy. Each time I sent Dr. Sergio Cicalò an email to ask questions about video coding and research-related simulation setups, he always showed great patience and replied to me an email with detailed answers to my questions. His feedback and suggestions really helped me in the initial simulations of my research.

I also want to express my gratitude to all my friends in Hamilton who made my

life enjoyable and meaningful.

Finally, I would like to thank my parents and sister for their love, support and encouragement during my graduate studies at McMaster.

Notation and abbreviations

Notation	Definition	Notation	Definition
$\mathbf{0}$	All-zero vector	$\mathbf{1}$	All-one vector
$\boldsymbol{x}^T$	Transpose of column vector $\boldsymbol{x}$	$\|\boldsymbol{x}\|_p$	p-norm of column vector $\boldsymbol{x}$
$\vee$	OR	$\wedge$	AND

Abbr.	Full name	Abbr.	Full name
APP	Application	ARA	Adaptive Resource Allocator
AU	Access Unit	BL	Base Layer
CP	Cyclic Prefix	CSI	Channel State Information
EL	Enhancement Layer	FST	Frame Significance Throughput
FDD	Frequency Division Duplex	FEC	Forward Error Correction
GOP	Group of Pictures	IDR	Instantaneous Decoding Refresh
ILA	Iterative Local Approximation	LTE	Long Term Evolution
MCP	Motion-compensated Prediction	MSE	Mean Square Error
MAC	Medium Access Control	MS	Multimedia Server
OSI	Open System Interconnection	PHY	Physical
QoS	Quality of Service	QoE	Quality of Experience
QL	Quality Layer	UXP	Unequal Erasure Protection
R-D	Rate-Distortion	RI	Relative Importance
SVC	Scalable Video Coding	TB	Transmission Block
TDD	Time Division Duplex	UXP	Unequal Erasure Protection

Notation	Definition
$[x]^+$	The maximum value between x and zero
$[x]^+_\epsilon$	The maximum value between x and an arbitrary samll value ϵ
$\prec$	Strictly componentwise inequality
$\preceq$	Componentwise inequality
$\precneqq$	Componentwise inequality where at least one pair of related components are unequal
$\mathbb{E}_{\boldsymbol{y}}[\cdot]$	Expectation with respect to the random process $\boldsymbol{y}$

Abbr.	Full name
AWGN	Additive White Gaussian Noise
CGS	Coarse-grain Quality Scalable Coding
FDMA	Frequency Division Multiple Access
FGS	Fine-grain Quality Scalable Coding
JRASA	Joint Resource Allocation and Source Adaptation
MANE	Media-aware Network Element
MGS	Medium-grain Quality Scalable Coding
NALU	Network Abstraction Layer Unit
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PSNR	Peak Signal-to-Noise Ratio
RTP	Real-time Transport Protocol
TDMA	Time Division Multiple Access
WLAN	Wireless Local Access Network

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and Literature Review

1.1.1 Background

With the advancement of video compression technology and the rapid development and deployment of network infrastructure, recent years have witnessed an unprecedented growth in demand for video services. Today's video services range from Internet video streaming, video conferencing, real-time sport broadcasting, HDTV to DVD and Blue-ray Discs. According to recent forecasts (Cisco, 2015), video will represent 72% of the total mobile data traffic by 2019, compared to 55% in 2014. In order to support the rapidly increasing demands for various video services, a number of different video transmission systems (e.g., real-time streaming, video-on-demand, peer-to-peer streaming systems) may be employed. Whichever video transmission system is employed, when the broadcast operators deliver different compressed video programs to multiple users sharing a resource-limited wireless network, the design

and optimization of the video communications should consider two essential service objectives, namely, the fairness and efficiency. To achieve fairness, the system should provide fair services typically in terms of video quality to all users subscribing to video services with the same quality level. The second objective, efficiency, is to attain the highest overall video quality with constraints on the system resources. To attain the highest efficiency while providing fairness whenever needed, cross-layer optimization is one of the approaches that can be exploited (van Der Schaar and Sai, 2005; Cicalò and Tralli, 2014). In this thesis, we will limit our discussion of cross-layer approaches to the three layers of the open system interconnection (OSI) stack, namely, the physical (PHY), medium access control (MAC) and the application (APP) layers.

Orthogonal frequency division multiple access (OFDMA) is one of the key physical layer techniques for the current wireless standards such as IEEE 802.16e (Committee *et al.*, 2006) and 3GPP-Long Term Evolution (LTE) (3GP, 2006). It has become the workhorse for wireless broadband applications due to its ability to provide high-rate wireless connectivity. To fully exploit the temporal, frequency and multiuser diversities, and thus improve the system performances of an OFDMA system, a highly adaptive allocation scheme should be adopted to jointly allocate the system resources, e.g., subcarriers and transmission power, according to the varying channel conditions and user requirements. In practice, the channel statistics of different users are not identical at any given time. By exploiting such characteristics, the channel-aware or opportunistic resource allocation schemes (Munaretto and Zorzi, 2012) assign the system resources in favor of the users with better channel conditions, e.g., users close to the base station, and thus can maximize the throughput of the network. Even though high network throughput can be achieved, such opportunistic allocation

schemes typically sacrifice the transmissions of the cell-edge users experiencing poor channel quality, and thus result in an unfair rate assignment. Consequently, the promised quality of service (QoS), e.g., the start-up delay of video playback and video quality, of the sacrificing users cannot be met. Then, their quality of experience (QoE) can be significantly degraded if they pay the same price as the favored users and are expecting the same level of quality. The provision of acceptable QoE can be achieved based on a MAC-centric cross-layer optimization framework (van Der Schaar and Sai, 2005). That is, the QoE requirements, which will be forwarded to the MAC layer, are specified according to specific utilities and constraints defined in the APP layer. An adaptive resource allocator (ARA) at the MAC layer will decide how the system resources are distributed among users so that the QoE requirements are satisfied. The MAC layer is also responsible for selecting the optimal PHY layer parameters, e.g., modulation and coding schemes, based on the available channel information.

Modern video transmission systems typically consist of users with different needs (difference in video formats, continuity of playback, timing requirements, etc) and devices with diverse capabilities (diversity in computational and display capabilities, battery capacity, etc). In addition to the heterogeneous users, the time-varying network channel conditions determining the available throughput for each user at a given time necessitate the use of a source rate adaptation entity at the APP layer. With the information about the users and channel conditions, the videos are delivered to users adaptively to improve transmission stability, to provide desired video formats, to avoid buffer overflow, etc. Scalable Video Coding (SVC) (Schwarz *et al.*, 2007) is a highly attractive tool to achieve source rate adaptation. Within SVC, each video sequence is encoded into a single multi-layer stream consisting of one base layer (BL)

and one or more enhancement layers (ELs). The enhancement layers can be dropped and the remaining substream forms another valid bit stream which could be decoded by a target decoder. The resulting substream represents a source content with a degraded reconstruction quality (in terms of frame rate, frame size and fidelity, etc) compared to that of the original stream, but is high by taking into account the low data quantity of the substream. Another benefit of SVC is that it can be used in conjunction with unequal erasure protection (UXP) to combat packet losses in the case of video transmissions over error-prone channels with unpredictably varying throughput. With SVC, the encoded bit stream intrinsically consists of data segments with different importance in terms of the reconstructed video quality. Therefore, the bit stream can be partitioned into segments of diminishing importance and protected with progressively weaker error-correcting code, e.g., forward error correction (FEC) schemes based on Reed Solomon (RS) codes. In this way, the chance of losing more important data portions is decreased, and thus the overall robustness of video communications is enhanced. The work in (Schierl *et al.*, 2005) and the following (Mansour *et al.*, 2008; Ha and Yim, 2008; Maani and Katsaggelos, 2010; Cicalò *et al.*, 2012) have shown the effectiveness of using SVC in conjunction with UXP schemes for video transmission in error-prone environments. In practical systems, media-aware network elements (MANEs) defined in (Wenger and Stockhammer, 2005), which receive feedback about the user devices capabilities and channel information from the wireless network, are usually deployed to remove the needless data portions from the original streams before forwarding the remaining substreams to the users.

Due to the different spatio-temporal complexities of frames in a video sequence, the relationship between the rate and distortion can be considerably different among

different videos. The rate-distortion (R-D) model of a video sequence is a function that estimates the relationship between the required rate and the reconstructed distortion of the received video. In the scenario of error-free transmission, the R-D model enables us to predict the minimum encoding rate to achieve an objective distortion incurred by the lossy encoding process. On other hand, when videos are transmitted in error-prone channels, the rate in the R-D model takes into account the overhead introduced by the protection scheme, e.g., FEC, and the distortion includes both the encoding distortion and the additional distortion resulting from packet losses during transmission. In general, R-D models can be organized into three categories, namely, analytical, semi-analytical and empirical models(Hsu and Hefeeda, 2008). Analytical models exploit the distribution of the discrete consine transform coefficients of the input video to derive the R-D equation, without needing to go through the whole encoding process, based on a number of simplifying assumptions. As a result, the accuracy of analytical models is, in general, lower than that of empirical models, which directly measure the achieved distortions by decoding the video at a set of rates, and thus are disadvantageous from the complexity perspective. The trade-off between the low accuracy of analytical models and the high complexity of empirical models is achieved by the semi-analytical models. They consist of parameterized functions where the parameters are predicted, using curve-fitting methods, from several empirical R-D points. The use of R-D models that can be efficiently constructed considerably facilitates the process of source rate adaptation at the APP layer. Moreover, the selection of a suitable and accurate R-D model for a targeted video streaming system has a significant impact on the system performance. Recently, many R-D models have been proposed for real time and non-real time streaming systems (Dai *et al.*,

2006; Kwon *et al.*, 2007; Seferoglu *et al.*, 2007; Haseeb *et al.*, 2012).

1.1.2 Literature Review

The cross-layer optimization for multi-video delivery over wireless networks has been a very active research area because it can provide better multimedia performance in comparison with the "layered" optimization schemes. The authors of (Huang *et al.*, 2008) proposed a framework for joint resource allocation, source adaptation and deadline-oriented scheduling for multi-user video streaming over a CDMA wireless network. With the users utility functions, the resource allocation at the PHY layer is performed in a distributed fashion based on the Lagrangian dual decomposition. Then, source adaptation at the APP layer is achieved based on video content analysis and summarization, and transcoding, while scheduling of the transmission of packets is performed in a centralized fashion to meet individual deadline requirements. It was shown that such a joint optimization scheme achieved much better overall delivered video quality and resource utilization efficiency, with low complexity, small communication overhead and satisfied deadline constraints, compared with the heuristic schemes considering no multi-user content diversity and interaction between different network layers. However, the resource allocation in a CDMA-based network with a time-division multiplexing (TDM) fashion cannot exploit the frequency and multi-user diversities provided by system such as OFDMA.

To take the advantages of OFDMA, (Ha *et al.*, 2008) presented a cross-layer multiuser resource allocation algorithm for video transmission in downlink OFDM networks. The algorithm comprises two separate steps: subcarrier assignment and power allocation. Both steps exploit an R-D function taking into account the temporal error

propagation effect from packet loss, and a set of relative importance (RI) imposing constraints on the required individual user rate and quality of service. Under the assumption of identical power level for each subcarrier, the subcarrier assignment step assigns individual subcarrier to the user achieving the largest distortion reduction on that subcarrier. Based on the results of the subcarrier assignment step, the power allocation step distributes the power among the subcarriers to maximize the sum of distortion reduction without changing the RI for each user. As already mentioned, the R-D relationship can be significantly different among different videos. In addition, the RI imposes constraints on the required rate rather than the required video quality. Consequently, even a non-differentiated service, i.e., with the same RI value for all users, can lead to large quality variation among users. Moreover, formulating a problem where the objective is to maximize the sum of distortion reductions, under a set of resource constraints, without addressing fairness usually leads to the provisions of higher quality to the low-complexity videos and considerably lower quality to the high-complexity videos (Guan *et al.*, 2009).

To ensure fairness in terms of video quality, the authors in (Li *et al.*, 2009) proposed a content-aware distortion-fair video delivery scheme for multihop video communications. Instead of providing bandwidth fairness, it assures max-min distortion-fair sharing among users. Without modeling the R-D relationship of videos, the cross-layer resource allocation is guided by exploiting the temporal prediction structure of the video sequences and a frame drop distortion metric based on frame importance. The main drawback of such a scheme is that the source rate adaptation is based on a coarse distinction of the data importance at frame level and could lead to a waste of bandwidth if the thresholds of dropping frames are not carefully selected. In (Khan

et al., 2012), a scheduling strategy, relying on the concept of Nash equilibrium, for scalable video transmission to multiple users over OFDMA systems, is devised. It is based on a novel metric named frame significance throughput (FST) considering the spatio-temporal dependencies among frames in a video sequence. The FST is incorporated into a payoff metric. Then, a scheduler at the MAC layer exploits the payoff metric to guide the quality-driven resource allocation and scheduling procedure where the aim is to maximize the Nash product of the received video quality values of each user and achieved fairness in terms of objective video quality. In (Khan *et al.*, 2013), the strategy is extended to exploit also a multi-user time-averaged diversity which represents the statistically independent variations among the video traffic rate of different users. That is, the resource allocation and scheduling decisions are made by exploiting both the payoff metric and the achieved time-averaged bit throughput. More recent works have been proposed for cross-layer video transmission optimization whose goal is to maximize the minimum video quality across users and provide max-min quality fairness (Chen *et al.*, 2010; Khalek *et al.*, 2015) or to maximize (minimize respectively) the overall received video quality (distortion) without addressing fairness (Maani *et al.*, 2008; He and Liu, 2014). As already mentioned, the cross-layer design approach shows its advantages for better resource allocation and QoE provision. However, it requires, without appropriate designs, extensive exchange of information across different network layers and introduces high communication overhead. All the works in (Huang *et al.*, 2008; Ha *et al.*, 2008; Li *et al.*, 2009; Khan *et al.*, 2012, 2013; Chen *et al.*, 2010; Khalek *et al.*, 2015; Maani *et al.*, 2008; He and Liu, 2014) require that the MAC and APP layers interact with each other directly, e.g., the MAC layer must directly manipulate or utilize the utility functions defined

for the applications. Such direct and frequent interactions among layers prevent the application of abovementioned works to layering transmission systems where only a limited amount of information can be exchanged across layers.

In reality, the utilities and constraints defined for the applications should be functions of the rate averaged over a time window, not the instantaneous one (Song and Li, 2005). The authors in (Cicalò and Tralli, 2014) proposed a novel cross-layer optimization framework for scalable video delivery to multiple users over OFDMA wireless networks. The optimization seeks to maximize the sum of the ergodic (average) rate assigned to users while minimizing the distortion difference among the received videos. The optimization problem is "vertically" decomposed into a resource allocation problem at the MAC layer and a source adaptation problem at the APP layer. Then, an iterative local approximation (ILA) algorithm, with provable optimality and convergence, is proposed to obtain the optimal solution. The authors also presented the algorithms to solve the resource allocation and source adaptation problems, respectively. Since the iterative procedure to solve the resource allocation and source adaptation problems requires only limited scalar information exchange between the MAC and APP layers, it is applicable for layering transmission systems. The global optimal solution under the distortion-fairness constraint aims to attain zero distortion difference between any two users' received videos if (i) the R-D relationship of each video is continuous and (ii) there are no constraints on the maximum or minimum individual video distortion values. However, paradoxically, such a totally fair solution can be unfair for users experiencing good channel quality or requesting low-complexity videos. This is because, under such a totally fair scheme, a majority

of the available resources should be allocated to the users having poor channel conditions or requiring high-complexity videos in order to make sure that they can achieve the same quality level as other users who are likely to achieve much higher quality improvement if assigned the same amount of resources. In other words, achieving pure fairness among users usually comes at the cost of sacrificing the video qualities of a set of users and decreasing the system efficiency in terms of overall received video quality. Therefore, there is an inherent conflict between fairness and efficiency.

In (Su *et al.*, 2006), the authors proposed a cross-layer framework for sending multiple scalable videos over OFDM networks where trade-offs between quality fairness and system efficiency can be achieved. Within the framework, the video streams are transmitted across J transmission intervals. The optimization problem is broke down into J sequential problems, each of which is solved during a transmission interval to either ensure fairness or improve efficiency. To ensure fairness, the problem is formulated to minimize the maximal end-to-end distortion received among all users. To improve efficiency, the problem is formulated to minimize the overall end-to-end distortion among all users. Due to the NP-hard nature of the fairness and efficiency problems, the authors proposed two suboptimal algorithms to the above two problems, respectively. Then, the authors proposed to apply the fairness algorithm in the first x transmission intervals to ensure the baseline fairness, and then the efficiency algorithm in the rest $J - x$ transmission intervals to improve the overall efficiency. In this way, a desired trade-off between fairness and efficiency can be achieved by varying the value of x. However, such a transmission interval-based optimization, without considering the ergodic rate, does not allow to fully exploit the temporal diversity. In addition, the framework supports only $J + 1$ trade-off points, and thus

is disadvantageous when finer trade-offs are required.

1.2 Contribution and Organization of the Thesis

In this thesis, we address the problem of quality fairness-oriented cross-layer resource allocation for scalable video delivery over OFDMA wireless networks. The optimization seeks to maximize the system efficiency in terms of overall received video quality and to provide quality fairness by limiting the video quality deviation among users. To this end, we first formulate a joint resource allocation and source adaptation (JRASA) problem where the objective is to maximize the sum of the user PSNRs while minimizing the PSNR difference among the received videos. Motivated by the optimization framework presented in (Cicalò and Tralli, 2014), the sum-PSNR maximization is "vertically" decomposed into two coupled subproblems, namely, the resource allocation at the MAC layer and the source adaptation at the APP layer, and the ILA algorithm is used to obtain the optimal solution. Differently, the formulation of the JRASA problem allows us to develop a low-complexity APP layer algorithm to solve the source adaptation problem, which is considerably faster than the algorithm presented in (Cicalò and Tralli, 2014).

When solving the JRASA problem, a continuous semi-analytical R-D model is used in the source adaptation procedure as in (Cicalò and Tralli, 2014). However, the semi-analytical model is only an approximation of the empirical R-D points. Consequently, the performance of the whole optimization framework is directly influenced by the accuracy of the semi-analytical R-D model. Given the high accuracy of empirical R-D models, we extend the cross-layer optimization framework and the ILA algorithm to cover the case of source adaptation using a discrete empirical R-D model.

The optimal solution to the JRASA problem produced by the empirical approach can then be used as the benchmark for assessing the performance of solutions based on approximate R-D models, such as the semi-analytical R-D model.

Solving the JRASA problem results in a totally fair solution, which minimizes the PSNR difference among the videos. In order to achieve the trade-off between fairness and system efficiency, we propose an adjustable quality-fair cross-layer optimization framework that aims to maximize the sum of the PSNRs while limiting the PSNR difference among the received videos within an acceptable and adjustable range. The optimization consists of two steps. First the aforementioned JRASA problem is solved to obtain a common target PSNR value for all videos. The second step seeks to maximize the sum of the PSNRs under the constraint that the absolute value of the relative difference between the target PSNR and the achieved PSNR of each video is bounded from above by a nonnegative scalar. In this way, the larger the scalar is, the looser the fairness constraints will be. Consequently, the scalar can be considered as a parameter, which controls the trade-off between fairness and efficiency. Such a framework allows to achieve an infinite number of trade-off points. Finally, we show that the adjustable quality-fair sum-PSNR maximization is a general utility-based resource allocation problem where a low-complexity algorithm has been proposed to obtain an almost surely optimal solution (Wang and Giannakis, 2011).

Finally, extensive simulation results and the related discussions are presented to show the advantages of the quality fairness-oriented cross-layer resource allocation framework over a simple equal-rate adaptation scheme, and the flexibility of the adjustable quality-fair cross-layer optimization framework in terms of fairness and efficiency trade-off.

The remainder of the thesis is organized as follows. Chapter 2 provides a brief overview of SVC. In Chapter 3, we discuss briefly the concepts related to ODFMA within the context of LTE and present some recent related works in resource allocation for OFDMA wireless networks. In Chapter 4, we present the architecture of the video transmission system and two R-D models that are suitable for both error-free and packet-erasure video transmissions. In Chapter 5, we review the cross-layer optimization framework for SVC video transmission presented in (Cicalò and Tralli, 2014), which is very closely related to our work. In Chapter 6, we formulate the JRASA problem with totally quality-fair constraints, and discuss the techniques and algorithms used to obtain the global solutions. In Chapter 7, we extend the JRASA problem to cover the case of source adaptation based on discrete empirical R-D models. In Chapter 8, we discuss the adjustable quality-fair optimization problem, and the techniques and algorithms used to obtain the global solutions. The performance of the proposed optimization frameworks and schemes is evaluated in Chapter 9. Chapter 10 concludes the thesis.

Chapter 2

Overview of Scalable Video Coding

This chapter briefly presents the concepts, terms and techniques that are related to SVC.

2.1 Concepts of SVC Extension of the H.264/AVC Standard

The H.264/AVC video standard (Wiegand *et al.*, 2003a) has achieved a significant improvement in rate-distortion efficiency relative to all previous standards (Wiegand *et al.*, 2003b). As the most recent video coding standard, it has been adopted by a variety of application standards and is expected to be continuously used by most video applications in the near future. Given the popularity of H.264/AVC, the Joint Video Team of the ITU-T VCEG and ISO/IEC MPEG published the standard of SVC extension of the H.264/AVC standard (Schwarz *et al.*, 2007). Unlike the prior standards, SVC provides a higher degree of scalability with a little loss in coding

efficiency and a little increase in decoding complexity in comparison to the nonscalable counterparts. As an extension of H.264/AVC, SVC reuses most of the key coding tools of H.264/AVC and introduces new tools only when they can efficiently support the required type of scalabilities. The core design of SVC that distinguishes it from H.264/AVC is the layered structure of the encoded bit stream.

Similar to the bit stream of H.264/AVC, the bit stream of SVC is organized into packets called Network Abstraction Layer Units (NALUs). Each NALU consists of an integer number of bytes, and contains header bytes signaling the type of the carrying data and payload bytes which are the actual encoded video data. An access unit (AU) is formed by grouping a consecutive set of NALUs with particular characteristics and can be decoded to obtain exactly one frame. A set of successive AUs with specific properties is considered to be a coded video sequence which represents an independent decodable part of the bit stream. A coded video sequence always starts with an intra-coded frame (I-frame), or similarly, an instantaneous decoding refresh (IDR) AU, which signals that the decoding of the current IDR AU as well as the following AUs is independent of the previous AUs in the bit stream. In practice, we will refer to the frame interval between any two consecutive I-frames as an IDR period. For more comprehensive descriptions of NALUs, AUs and other related concepts of H.264/AVC and SVC, we refer the interested readers to the overview papers (Schwarz *et al.*, 2007) and (Wiegand *et al.*, 2003a).

2.2 Types of Scalability

Within the SVC standard, the three common scalable modes are temporal, spatial and quality scalability. With temporal and spatial scalability, a substream can be

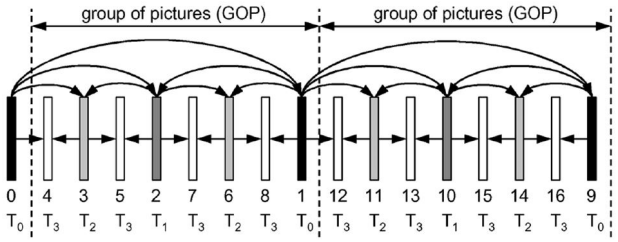
decoded to obtain a source content with reduced frame rate (temporal resolution) and frame size (spatial resolution) in comparison to that represented by the original bit stream. Quality scalability enables to reconstruct a video sequence with the same frame rate and size as the original video sequence, but a degraded fidelity which is typically measured in signal-to-noise ratio (SNR). Therefore, it is commonly to refer to quality scalability as fidelity or SNR scalability. In addition to the three commonly used modes, SVC also supports the so-called region-of-interest and object-based scalability through which the quality of particular regions or objects in the reconstructed frames can be selectively enhanced. The region-of-interest and object-based scalabilities are of crucial interest in the application scenarios where some regions or objects in the video frames are more important or interesting than the remaining area. The aforementioned basic modes of scalability can be combined to produce substreams that represent reconstructed video sequences with different temporal-spatial resolutions and fidelity (or bit rate).

2.2.1 Temporal Scalability

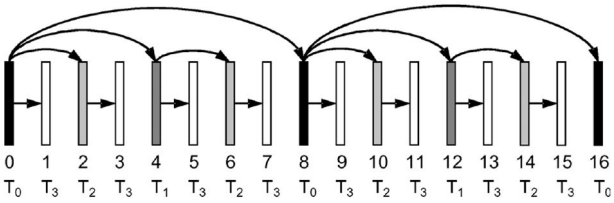
Temporal scalability refers to the capability of scaling a video sequence via its temporal resolution, or equivalently, frame rate. The alteration of frame rate can be achieved by dropping frames in the video sequence. However, randomly discarding frames should be avoided because other frames may depend on the discarded frames for motion-compensated prediction (MCP) and cannot be successfully decoded without the discarded frames being available at the decoder. Within SVC, temporal scalability is achieved based on the concept of hierarchical prediction structure. For

example, the hierarchical coding structures with B-frame or inter-coded frame (P-frame) (Schwarz *et al.*, 2005), (Schwarz *et al.*, 2006) are shown in Fig. 1(a) and (b), which also show the frame ordering within the coded bit stream and the dependencies in terms of MCP. Let T_i denote the temporal layer identifier associated with the ith temporal layer. The values of i and T_i start from 0 (e.g., T_0 is associated with the temporal base layer and considered to be the highest level in the hierarchy) and increase by one from one layer to the next layer. We refer to the frames between any two consecutive base layer frames plus the following base layer frame as a group of pictures (GOP). Every frame in the GOP is assigned a temporal identifier and the successful decoding of frame(s) associated with a temporal identifier T_i depends only on previous and forward frame(s) with temporal identifiers smaller than or equal to T_i. Therefore, for each integer number t, the bit stream obtained by removing from the original stream the set of AUs, representing the frames associated with all temporal layers with temporal layer identifiers larger than k, represents a reconstructed video sequence with a particular frame rate.

Fig. 2.1(a) shows the hierarchical B-pictures prediction structure. The different levels of hierarchy are marked in different shades of gray and denoted by different temporal layer identifiers T_i. The structure specifies that the enhancement layer frames (B-frames) are predicted from both the preceding and successive frames. Such a bi-direction prediction structure enables to achieve a better coding efficiency. Fig. 2.1(b) shows another hierarchical prediction structure having the same degree of temporal scalability as the hierarchical B-pictures structure. This structure restricts the MCP to reference frames that precede the frames to be predicted. Consequently, it achieves a zero delay, compared to 7 frames for the hierarchical B-pictures structure, in the



(a) Hierarchical B-pictures



(b) Zero-delay coding structure

Figure 2.1: Hierarchical coding structures for supporting temporal scalability. From (Schwarz *et al.*, 2007, Fig. 1). (a) Coding with hierarchical B-pictures. (b) Hierarchical prediction structure with a structural encoding/decoding delay of zero. The numbers directly underneath the frames denote their coding/decoding order and T_i specify the temporal layer identifier associated with the ith temporal layer.

decoding process. However, such a zero delay comes at the cost of a lower coding efficiency. It is worth noting that H.264/AVC has already provided a high degree of flexibility of temporal scalability through the use of hierarchical motion prediction, and its effective reference frame selection and controlling mechanisms. What distinguishes temporal scalability in the SVC design from that supported by H.264/AVC is that SVC additionally provides signaling information by which the enhancement layers can be easily identified and removed (if necessary) without too much effort.

2.2.2 Spatial Scalability

Spatial scalability enables the representations of a video sequence in different spatial resolutions or frame sizes within a single scalable bit stream. SVC supports spatial

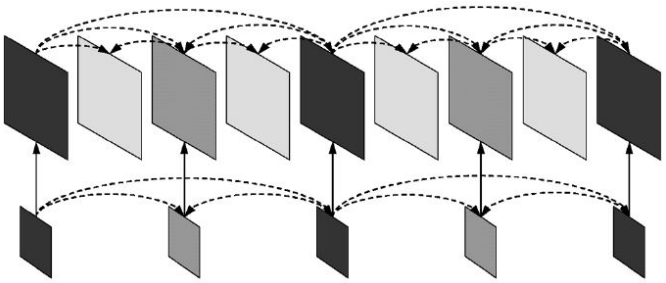


Figure 2.2: An example of spatial scalability/multilayer structure with inter-layer prediction. From (Schwarz *et al.*, 2007, Fig. 4).

scalability via multilayer coding, which has been used by several prior standards, e.g., H.262 | MPEG-2 Video, H.263, and MPEG-4 Visual. Similar to temporal scalability, a spatially scalable bit stream is partitioned into one base layer and several enhancement layers. Each spatial layer i is associated with a spatial or dependency layer identifier D_i. The values of i and D_i of the spatial base layer are 0 and increased by 1 from one spatial layer to the next. By decoding the base layer, the user can display the video sequence at a smaller frame size. Every new layer will enlarge the frame size until the full spatial resolution is achieved. Fig. 2.2 shows an example of a spatial scalable video sequence with one base layer and one enhancement layer. It could be found that in each single spatial layer, MCP and inter-prediction are employed as for single-layer coding. Moreover, from a layer to the next layer, not only the frame size, but also the number of frames increases. Therefore, it is possible to combine spatial and temporal scalability.

The core design that improves the coding efficiency of spatial scalability in SVC over simulcasting the video in each spatial resolution separately is the inter-layer prediction mechanisms, which are also shown in Fig. 2.2 where the vertical arrows indicate the inter-layer prediction between layers. The inter-layer prediction tools strive to reuse as much information as possible from one layer to the next. This

helps to reduce redundancies between layers, and thus improves the coding efficiency accordingly.

The most straightforward way to perform inter-layer prediction is based on the fully decoded frames from the lower layer directly underneath the layer to be predicted. This is adopted by the previous standards such as H.262 | MPEG-2 Video, H.263, and MPEG-4 Visual. However, this direct method will considerably increase the decoding complexity because of the need to completely decode the lower-layer frames. In order to restrict the decoding complexity, the SVC design introduces three inter-layer prediction mechanisms, namely, inter-layer motion prediction, inter-layer residual prediction and inter-layer intra-prediction. With these three prediction mechanisms, the layer to be predicted collects and reuses the information from the lower layer without all lower-layer frames being completely decoded. Before we give a brief overview of the three mechanisms, let us recall that in conventional motion prediction methods, two main components are encoded for every B- or P-frame macroblock: the motion vectors and the related residual information.

Inter-layer Motion Prediction

In general, it is unlikely that the motion vectors will change significantly from one layer to the next. Therefore, with the inter-layer motion prediction mechanism, each additional enhancement layer will collect and reuse the motion information (motion vectors, macroblock partition and reference frame indices) of the underlying layers. In some cases, the motion information can be directly used for motion compensation. If the motion information is not exactly reused, it is possible that it functions in the prediction process for the actual motion vectors, i.e., a motion vector predictor can

be formed based on the motion vectors of the co-located macroblocks in the reference layer.

Inter-layer Residual Prediction

The inter-layer residual prediction mechanism provides means for reusing the residual information obtained by coding the reference layer to predict the residual information of the enhancement layers. For example, the residual information of macroblocks in the enhancement layer can be predicted by upsampling the residual information of the corresponding macroblocks in the underlying layers.

Inter-layer Intra Prediction

Unlike the aforementioned two inter-layer prediction mechanisms, inter-layer intra prediction requires the decoding of the frames of the reference layers because these reconstructed frames will be used as prediction. For example, in spatial scalability, the reconstructed pixels in the reference layer will be upscaled to form the prediction for the enhancement layer. However, to prevent complete decoding of the lower layers, such direct predictions are only allowed for the enhancement layer macroblocks whose co-located macroblocks are intra-coded in the reference layer. These particular set of macroblocks in the reference layer are encoded without references to other frames, and thus can be decoded independently of other frames without running a separate motion compensation loop.

2.2.3 Quality Scalability

Quality scalability is defined as representing a video sequence with different fidelities and details. Recall that during video encoding, the texture information of a macroblock is transformed to the frequency domain and the corresponding transform coefficients are quantized before being finally encoded. In this way, the transform coefficients are represented by a set of quantization levels whose number is determined by the quantization step size. The smaller the quantization step size is, the more the quantization levels we have, and the more accurate representation and finer graduation of the transform coefficients will be. Therefore, a video can be reconstructed with a higher degree of details and fidelity with a smaller quantization step size during encoding. Basically, what quality scalability does is to carefully select a set of non-increasing quantization step sizes for different quality layers, and thus can be achieved by decreasing the quantization step size from one quality layer to the next. Similar to temporal and spatial scalability, we associate each quality layer by a quality layer identifier, i.e., Q_i is the quality layer identifier of the ith quality layer. Again, the values of i and D_i of the quality base layer are 0 and increased by 1 from one spatial layer to the next.

Different techniques for quality scalability have been well-studied for the prior standards and SVC. In this section, we give a brief overview of three available techniques, namely, coarse-grain quality scalable coding (CGS), medium-grain quality scalable coding (MGS) and fine-grain quality scalable coding (FGS).

Coarse-grain Quality Scalable Coding

Within CGS, the provision of a video with different qualities is enabled by dropping quality layers one by one until the target quality or bit rate is achieved. This can be considered as a special case of spatial scalability where the spatial resolution remains unchanged from one layer to the next. But different from spatial scalability, the inter-layer prediction in CGS is applied without the corresponding upsampling or scaling operations, e.g., scale the motion vectors and upsample the residual information from the reference layer. Instead, the inter-layer prediction mechanism enables the enhancement layer to collect the textual information in the lower layer and perform re-quantization of the textual information with a smaller quantization step size relative to that used for the lower layer.

However, such layer-based quality scalability coding has a drawback which is already shown by the name of CGS: the number of supported bit rates is limited to the number (up to eight) of CGS quality layers (De Cock *et al.*, 2009). Another drawback of CGS, inherited from spatial scalability, is the low flexibility of bit stream switching. That is to say, switch between any two CGS layers can be done only in particular points of the bit stream, e.g., at IDR AU.

Fine-grain Quality Scalable Coding

During the standardization of SVC, a complicated FGS design was investigated but not included in the SVC specification eventually due to its high design complexity and large syntax overhead (Wien *et al.*, 2007). FGS is a packet-based quality scalable coding technique where any quality enhancement layer NALU (no base layer NALU is allowed to be dropped) can be discarded from the scalable bit stream. Therefore,

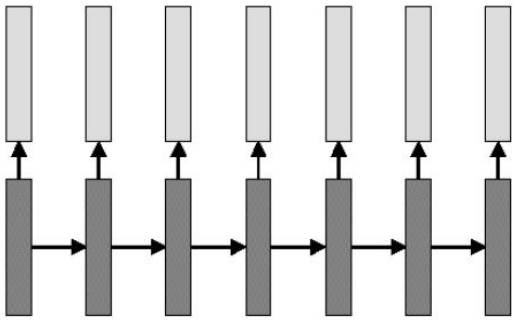


Figure 2.3: FGS in MPEG-4 Visual. From (Schwarz *et al.*, 2007, Fig. 8(a)).

FGS provides a higher degree of granularity for quality scalability in comparison to CGS. However, when quality scalability is packet-based, the MCP process should be carefully designed to attain a good trade-off between enhancement layer coding efficiency and drift. In video coding, drift refers to a situation when the encoder and the decoder MCP loops are running out of synchronization. That is, the motion compensation loops are no longer working on the reference images with the same quality level because of the loss of quality refinement packets during transmission. With FGS in MPEG-4 Visual, drift can be completely eliminated since the motion compensation at both encoder and decoder always use the base layer reconstruction as reference as illustrated in Fig. 2.3. However, such drift elimination comes at the expense of a lower coding efficiency of enhancement layers relative to the single-layer coding because the more accurate information from the enhancement layers is not exploited in the process of MCP.

Medium-grain Quality Scalable Coding

MGS is introduced in the SVC design as a trade-off between CGS and FGS. It is a packet-based quality scalable coding technique using the concepts of dependency layers as in CGS. In comparison to CGS, MGS allows more extractable rate points

(up to 128) by dividing each dependency layer into up to 16 MGS layers, each of which can be dropped for the purpose of rate adaptation. Particularly, with the MGS, the transform coefficients of a given macroblock are split into groups, each of which is distributed to a selected MGS layer. Moreover, it is shown in (Schwarz and Wiegand, 2007) that MGS can provide a similar coding efficiency with respect to FGS, while keeping the design complexity and drift at an acceptable level. The improvement of MGS relative to FGS lies on the flexibility to perform the MCP using either the base layer reconstruction or the enhancement layer reconstruction of the reference frames. Fig. 2.4 shows MGS with the concept of key picture (Schwarz *et al.*, 2004) combined with hierarchical prediction structures where the hatched boxes mark the key pictures. In Fig. 2.4, the frames of the temporal base layer are considered as key pictures which use only the frames of the base layer for MPC. Therefore, drift can be completely avoided in the MCP loop of the temporal base layer. By contrast, the frames of the temporal enhancement layers typically use the highest available quality of the reference frames for MPC, and thus achieve a higher coding efficiency for them. With the key picture concept, MPC is conducted in the enhancement layers, but with a periodic update in the base layer. In this way, resynchronization of the motion compensation loops at the encoder and decoder is enabled periodically and the propagation effect of drift is kept within any two consecutive frames of the temporal base layers. Therefore, MGS enables a better trade-off between drift and enhancement layers coding efficiency.

To extract a MGS bit stream that meets a target bit rate, the JSVM reference software (Reichel *et al.*, 2007) provides two extraction methods. The straightforward method discards NALUs from the scalable bit stream randomly until the target bit

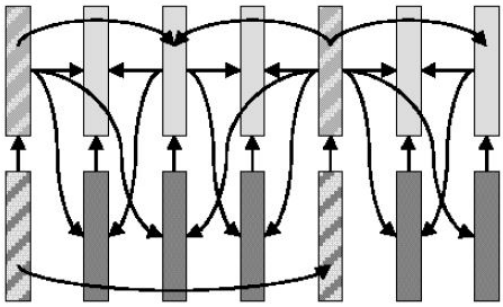


Figure 2.4: MGS with Key picture combined with hierarchical prediction structure. From (Schwarz *et al.*, 2007, Fig. 8(d)).

rate is reached. The second extraction method is based on the concept of Quality Layers (QLs). The idea of QLs is to determine the priority orders of various data units that forming the scalable bit stream. The priority orders can then be used to optimize the rate-distortion efficiency of a substream when extracting it. In SVC, a post-processing is carried out to perform rate-distortion analysis on a scalable bit stream to determine the priority level, in terms of contribution to the quality of the reconstructed video, of each NALU in the bit stream. A quality level assigner is responsible for the rate-distortion analysis and embedding the priority level information into the header of the corresponding NALUs. During the extraction process, the NALUs are discarded sequentially from the lowest priority level to the highest priority level until the target bit rate is achieved. The interested readers can refer to (Amonou *et al.*, 2007) for more details about optimized bit stream extraction.

2.2.4 Combined Scalability

Temporal, spatial and quality scalability can be combined to support the representation of a video sequence with various frame rates, frame sizes and bit rates within a single scalable bit stream. In Fig. 2.5, an example of an SVC encoder that supports

combined scalability is shown.

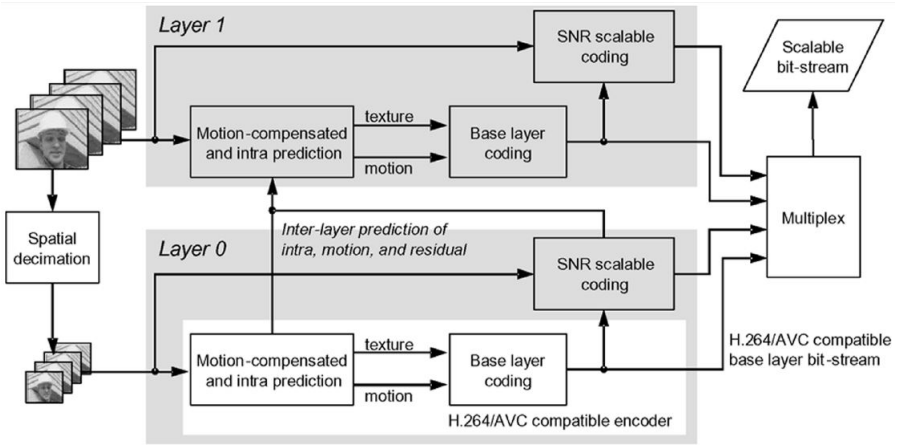


Figure 2.5: An example of a SVC encoder supporting combined scalability. From (Schwarz *et al.*, 2007, Fig. 12).

Chapter 3

Overview of Resource Allocation for OFDMA Wireless Systems

3.1 OFDMA and Its Application in 3GPP-LTE Downlink

OFDMA is a frequency-division multiplexing and multiple access scheme which has been adopted for both uplink and downlink transmissions to cope with frequency-selective fading and to support high data rate in current wireless standards, e.g., IEEE 802.16 and 3GPP-LTE. Even though OFDMA can be considered as a multi-user version of OFDM, OFDMA is distinct from OFDM because of its highly flexible resource scheduling and allocation mechanism. For a multi-user OFDM system with a time division multiple access (TDMA) scheme, the entire bandwidth is divided into a large number of orthogonal narrowband subcarriers which are exclusively allocated to an individual user within a period of time. Similarly, for OFDM with a frequency

division multiple access (FDMA) scheme, the subcarriers are arranged into groups, each of which is allocated to a user for transmission at any time instant. By contrast, in OFDMA, all the subcarriers can be shared by all users simultaneously at any given time and dynamically over time, according to the varying channel statistics of different users. In this way, OFDMA enables concurrent data transmissions for a number of users and provides an extra diversity named multi-user diversity, which can be exploited in conjunction with the effectiveness of OFDMA in dealing with fading and inter-symbol interference caused by multipath to achieve higher spectrum efficiency compared to other modulation and multiple access schemes.

In this thesis, we will focus on the resource allocation for OFDMA wireless systems within the context of LTE. Before we delve into the resource allocation problem, we first discuss the structures of LTE frames and OFDMA frames. The key concepts for the discussion are summarized as follows (Zyren and McCoy, 2007):

- **Slot**. A time period that is 0.5 ms in duration and consists of either 6 or 7 OFDM symbols.
- **Resource element**. The smallest modulation structure in LTE that is one 15 kHz subcarrier by one OFDM symbol.
- **Subframe**. A time period that is 1 ms in duration and comprises 2 slots.
- **Frame**. A time period, as shown in Fig. 3.1, that is 10 ms in duration and composed of 10 subframes, or equivalently, 20 slots.
- **Resource block**. The smallest resource allocation structure in LTE that is formed by 12 consecutive subcarriers in frequency domain and 6 or 7 OFDM symbols in time domain.
- **Time-frequency OFDMA frame**. The signal that is transmitted in each downlink transmission interval and is represented by a time-frequency resource

grid, as illustrated in Fig. 3.2.

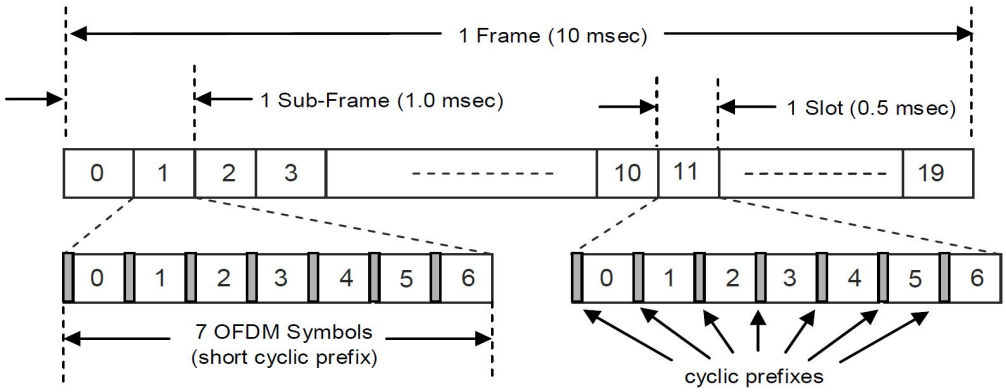


Figure 3.1: LTE frame structure. From (Zyren and McCoy, 2007, Fig. 2.3.2-1).

As shown in Fig. 3.1, a LTE frame consists of 10 subframes, each of which can be divided into 2 slots. Each slot is composed of either 6 or 7 OFDM symbols, depending on the type of cyclic prefix (CP) in use. If a normal CP is used, a slot contains 7 OFDM symbols. If an extended CP is used, a slot comprises 6 OFDM symbols. Note that the aforementioned frame structure is defined for the use in frequency division duplexing (FDD). An alternative frame structure used in time division duplexing (TDD) is not considered in this thesis.

In Fig. 3.2, an OFDMA frame represented by a time-frequency resource grid is shown. Each box in the resource grid represents a resource element and the resource grid is made up of M_{SC} subcarriers and N_S OFDM symbols, or equivalently, of $M \cdot N$ resource blocks, each of which is formed by m_c consecutive subcarriers within n_s successive OFDM symbols. Note that $M = M_{SC}/m_{sc}$ and $N = N_S/n_s$ are the number of subchannels and number of time slots in each frame, respectively. Within the context of LTE, the values of m_{sc} and n_s are 12 and 6, respectively, when an extended CP is employed. The value of N, determined by the duration of a downlink

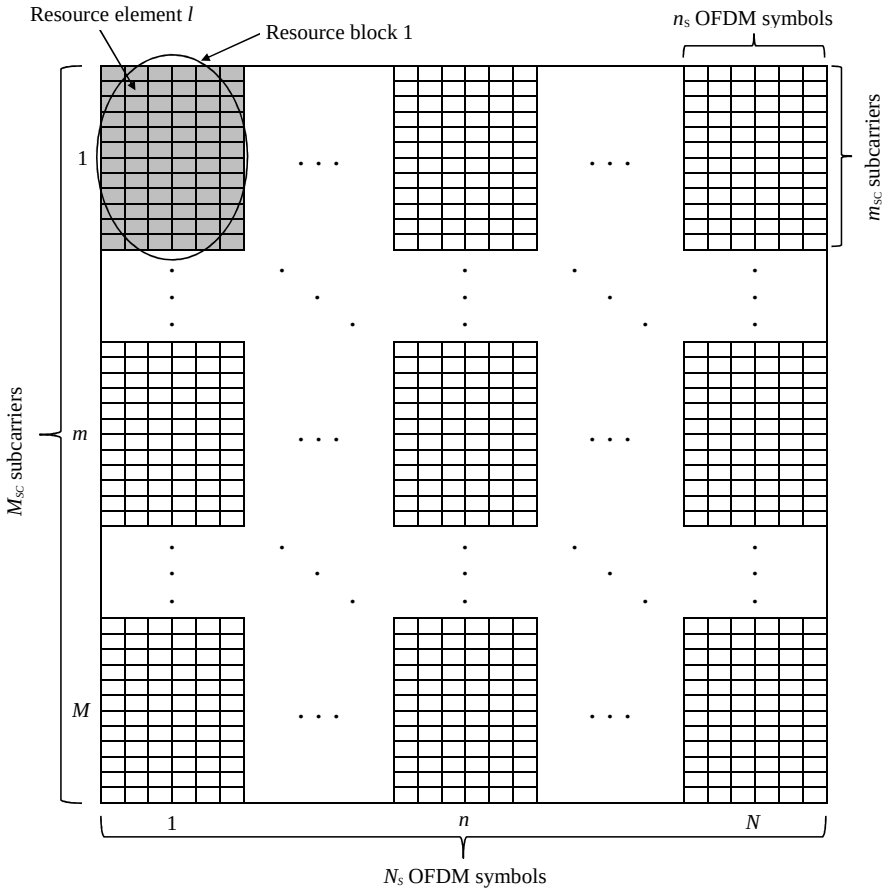


Figure 3.2: An example of an OFDMA transmission frame represented by a time-frequency resource grid.

transmission interval, is 20, as for instance, when the transmission interval is 10 ms. The value of M varies when the overall system channel bandwidth changes. The bandwidth configuration information about the LTE supported channel bandwidths, the maximum number of subchannels and the maximum occupied bandwidth are summarized in Table 3.1. Note that the maximum occupied bandwidth is the product of the maximum number of subchannels and the bandwidth of a single subchannel, which is 180 kHz.

Table 3.1: LTE Bandwidth Configuration Information

Channel Bandwidth (MHz)	1.4	3.0	5.0	10.0	15.0	20.0
Maximum Number of Subchannels	6	15	25	50	75	100
Maximum Occupied Bandwidth (MHz)	1.08	2.7	4.5	9.0	13.5	18.0

3.2 Resource Allocation for OFDMA Wireless Networks

3.2.1 Resource Allocation: Preliminaries

In this thesis, we consider a single-cell OFDMA WLAN, as illustrated in Fig. 3.3, where a base station (BS) broadcasts different video programs to different users through single-antenna links, i.e., the BS and all user devices are equipped with one antenna. Before the BS transmits the video data and packets, a resource allocator at the BS allocates the available resource blocks and power according to the channel state information (CSI) of the wireless channels between the BS and each user. The wireless channels between the BS and each user can be modeled as a set of M_{SC} parallel Rayleigh fading channels, each of which is characterized by a different channel gain.

To facilitate the analysis of the resource allocation problem, several assumptions are made as follows:

1) The channel gain of each wireless channel is fixed within one time slot and varies from one time slot to the next according to a random process.

2) The channel gains of all subcarriers belonging to a single subchannel are approximately the same within one time slot.

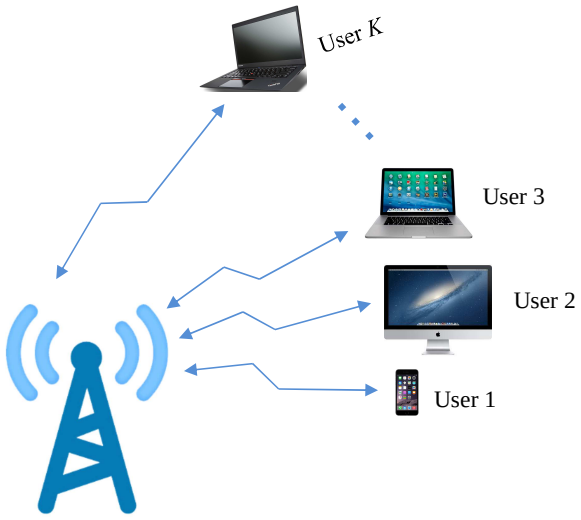


Figure 3.3: The architecture of the OFDMA WLAN

3) Each user knows exactly the CSI of the wireless channels between the BS and itself, and the BS has a perfect knowledge of the CSI of all wireless channels.

While the slot-based variation of channel gain is commonly assumed in the literature, i.e., by considering slowly changing channels, the second assumption is acceptable when the wireless channels between the BS and each user has a coherence bandwidth larger than the bandwidth of a single subchannel. A perfect knowledge of CSI at both the BS and user sides can be achieved by channel estimation at the user side, with a training phase with adequately long pilot sequences, and by proper mechanisms to feedback the estimated CSI to the BS.

Consider now that there are K users sharing a bandwidth of B in the OFDMA WLAN. The total bandwidth B is divided into M subchannels, each of which has a bandwidth $\triangle B = B/M$, or equivalently, into M_{SC} orthogonal subcarriers. Let us denote the sets of users and subchannels as $\mathcal{K} = \{1, 2, \cdots, K\}$ and $\mathcal{M} = \{1, 2, \cdots, M\}$, respectively. Moreover, the channel gain between the BS and user k on a subcarrier

belonging to subchannel m, denote by $h_{k,m}$, is modeled as a stationary and ergodic complex Gaussian random process (Rayleigh fading). Then, the normalized SNR, i.e., the SNR corresponding to unit transmission power, of user k on a subcarrier pertaining to subchannel m is given by:

$$\gamma_{k,m} = \frac{|h_{k,m}|^2}{\sigma^2} \tag{3.1}$$

where σ^2 is the power, or equivalently, the variance of the zero-mean AWGN at the user side. We further denote the set of KM realizations of the normalized SNR random process by $\boldsymbol{\gamma} = \{\gamma_{k,m}, k \in \mathcal{K}, m \in \mathcal{M}\}$. During each time slot, the resource allocator, depending on the channel realization $\boldsymbol{\gamma}$, strives to optimally allocate the resource blocks and power for the transmissions to users. Since we consider a slot-based resource allocation scheme, the allocation of resource blocks and the allocation of subchannels will be used interchangeably in the following analysis.

Let us first assume that a subcarrier can be shared by multiple users over nonoverlapping time fractions of the total time slot duration t_{slot}. Let $\tau_{k,m} \geq 0$ and $p_{k,m} \geq 0$ denote the nonnegative time fraction and the average power, respectively, allocated to user k for data transmission on a subcarrier belonging to subchannel m. Since the transmission to user k is only activated for a fraction of the time slot, the transmission power allocated to user k, during the active time fraction, is $p_{k,m}/\tau_{k,m}$. Taking into consideration the modulation and coding scheme adopted by the PHY layer, the maximum achievable rate of the transmission to user k on a subcarrier belonging to

subchannel m is given as:

$$r_{k,m}(\tau_{k,m}, p_{k,m}) = \begin{cases} \frac{\triangle B}{m_{sc}} \tau_{k,m} R\left(\frac{\gamma_{k,m} p_{k,m}}{\tau_{k,m}}\right) & \tau_{k,m} > 0 \\ 0 & \tau_{k,m} = 0 \end{cases} \tag{3.2}$$

where $R(x) = a_1 \log_2(1 + x/a_2)$, and a_1 and a_2 are two parameters named *rate adjustment* and *SNR gap* that are introduced to account for the particular modulation and coding scheme in use (Mazzotti *et al.*, 2012). Because the rate in (3.2) is a function of $\tau_{k,m}$ and $p_{k,m}$, the objective of the resource allocator is to optimize the allocation by determining the set of allocation polices $\boldsymbol{\tau}(\boldsymbol{\gamma}) = \{\tau_{k,m}(\boldsymbol{\gamma}), \forall\, k, m\}$ and $\boldsymbol{p}(\boldsymbol{\gamma}) = \{p_{k,m}(\boldsymbol{\gamma}), \forall\, k, m\}$ per channel realization $\boldsymbol{\gamma}$. If the optimal $\boldsymbol{\tau}^*$ and $\boldsymbol{p}^*$ are found, the corresponding optimal rates, following from (3.2), will be $\boldsymbol{r}^*(\boldsymbol{\gamma}) = \left\{r_{k,m}\left(\tau^*_{k,m}(\boldsymbol{\gamma}), p^*_{k,m}(\boldsymbol{\gamma})\right), \forall\, k, m\right\}$.

As mentioned in Section 1.1.2, in practice, users are more concerned about the received rate averaged over a period of time related to the specific applications. Therefore, in the following analysis, we will focus on the rate averaged over a certain period of time that is associated with the video coding structure, i.e., the period is the length of one or more GOPs. Instead of being computed directly, the time-averaged rate will be approximated, through ergodicity, by the corresponding ensemble-averaged rate in regard to the the random process $\boldsymbol{\gamma}$. In particular, we consider a sufficiently long *application period* t_{ap}, over which the approximation of the time averages by the ensemble averages becomes reasonable. Then, the maximum achievable ergodic rate

per user k is given by:

$$R_k(\boldsymbol{\tau}, \boldsymbol{p}) = \frac{1}{N_{slot}} \sum_{t=1}^{N_{slot}} \left[\sum_{m=1}^{M} m_{sc} r_{k,m}(\tau_{k,m}[t], p_{k,m}[t]) \right]$$
$$\simeq \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{m=1}^{M} m_{sc} r_{k,m} \left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma}) \right) \right] \tag{3.3}$$

where $N_{slot} = \left\lfloor \frac{t_{ap}}{t_{slot}} \right\rfloor \gg 1$ indicates the number of time slots within an *application period*, and t is the index of the time slot and is used to emphasize that the values of the allocation variables and the related instantaneous rate are evaluated per time slot.

To formulate the resource allocation problem, let us first define the set of all possible allocation polices $\boldsymbol{\tau}(\boldsymbol{\gamma})$ and $\boldsymbol{p}(\boldsymbol{\gamma})$ which is specified by a set of basic constraints, i.e., the set is

$$\mathcal{S} = \left\{ (\boldsymbol{\tau}, \boldsymbol{p}) \mid \tau_{k,m}(\boldsymbol{\gamma}) \geq 0, p_{k,m}(\boldsymbol{\gamma}) \geq 0, \forall\, k, m, \sum_{k=1}^{K} \tau_{k,m}(\boldsymbol{\gamma}) \leq 1, \forall\, m \right\}. \tag{3.4}$$

Then, consider a practical scenario where the transmitter at the BS has an average power constraint, i.e., the total average transmission power is $\bar{P}$, the set of the achievable allocation polices considering the power constraint can be defined as

$$\mathcal{A} = \left\{ (\boldsymbol{\tau}, \boldsymbol{p}) \in \mathcal{S} \mid \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{k=1}^{K} \sum_{m=1}^{M} p_{k,m}(\boldsymbol{\gamma}) \right] \leq \bar{P} \right\}. \tag{3.5}$$

Denote by $\boldsymbol{R}(\boldsymbol{\tau}, \boldsymbol{p}) = [R_1(\boldsymbol{\tau}, \boldsymbol{p}), R_2(\boldsymbol{\tau}, \boldsymbol{p}), \cdots, R_K(\boldsymbol{\tau}, \boldsymbol{p})]^T$ the maximum achievable ergodic rate vector and by $\boldsymbol{R} = [R_1, R_2, \cdots, R_K]^T$ an ergodic rate vector. Then, the

ergodic rate region of the OFDMA downlink channel can be defined as

$$\mathcal{R} = \bigcup_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{A}} \{\boldsymbol{R} \mid \mathbf{0} \preceq \boldsymbol{R} \preceq \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})\}\,. \tag{3.6}$$

Given a concave function $f(x)$, its perspective defined as $g(x,t) = tf(x/t)$ is also concave since the perspective operation preserves concavity (Boyd and Vandenberghe, 2004). Accordingly, the rate $r_{k,m}(\tau_{k,m}, p_{k,m})$ in (3.2) is a jointly concave function of $\tau_{k,m}$ and $p_{k,m}$. Consequently, the ergodic rate region $\mathcal{R}$ in (3.6) is a convex set of the rate vectors (Wang and Giannakis, 2011).

3.2.2 Resource Allocation: Optimal Subcarrier and Power Allocation

Resource allocation for OFDMA wireless networks has been an area of active research. In general, the objective of the resource allocation problem is to maximize (alternatively minimize) the weighted sum of average rates (powers) allocated to users under a set of power (rate) constraints. In recent years, a great deal of research efforts have been dedicated to the weighted sum of average rates (WSAR) maximization problem (Wong and Evans, 2008b), (Wong and Evans, 2008a), (Wang and Giannakis, 2011) and reference therein. We retrace here in this section the formulation of the WSAR problem and the main results of (Wang and Giannakis, 2011).

The WSAR problem is formulated as follow:

$$\begin{aligned} &\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \quad \boldsymbol{w}^T\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \\ &s.t. \quad \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{R} \end{aligned} \tag{3.7}$$

where $\boldsymbol{w} = [w_1, w_2, \cdots, w_K] \succeq \mathbf{0}$ is the weight vector that can be used to prioritize different users or to enforce proportional fairness in terms of rate. Since the ergodic rate region $\mathcal{R}$ defined in (3.6) is a convex set of the rate vectors, solving the WSAR problem in (3.7) results in a optimal rate vector residing on the boundary of $\mathcal{R}$ (Boyd and Vandenberghe, 2004). Theoretically, varying the weight vector $\boldsymbol{w}$ allows us to achieve all the boundary points and thus trace out the ergodic rate region (Li and Goldsmith, 2001). However, it is worth noting that there is in general no an analytical formula for the set of the boundary points of the rate region for the OFDMA downlink scenario.

After substituting (3.3) and (3.6) into (3.7), the optimization problem can be rewritten as:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \quad \sum_{k=1}^{K} w_k \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{m=1}^{M} m_{sc} r_{k,m} \left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma}) \right) \right] \tag{3.8a}$$

$$s.t. \quad \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{k=1}^{K} \sum_{m=1}^{M} p_{k,m}(\boldsymbol{\gamma}) \right] \leq \bar{P}. \tag{3.8b}$$

The optimization problem in (3.8) is convex because (i) the objective function in (3.8a) is concave for the reason that $r_{k,m}(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma}))$ is a concave function of $\tau_{k,m}(\boldsymbol{\gamma})$ and $p_{k,m}(\boldsymbol{\gamma})$, and (ii) the average power constraint in (3.8b) is linear. Therefore, it could be solved efficiently using a Lagrangian dual approach (Boyd and Vandenberghe, 2004). Let λ be the Lagrangian multiplier associated with the average power

constraint, the Lagrangian associated with (3.8) is defined as:

$$\begin{aligned} L(\boldsymbol{\tau}, \boldsymbol{p}, \lambda) &= \sum_{k=1}^{K} w_k \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{m=1}^{M} m_{sc} r_{k,m} \left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma}) \right) \right] \\ &\quad + \lambda \left\{ \bar{P} - \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{k=1}^{K} \sum_{m=1}^{M} p_{k,m}(\boldsymbol{\gamma}) \right] \right\} \\ &= \lambda \bar{P} + m_{sc} \mathbb{E}_{\boldsymbol{\gamma}} \left[\sum_{k=1}^{K} \sum_{m=1}^{M} w_k r_{k,m}(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})) - \lambda p_{k,m}(\boldsymbol{\gamma}) \right]. \end{aligned} \tag{3.9}$$

Then, the Lagrangian dual problem of (3.8) is:

$$\min_{\lambda \geq 0} g(\lambda) \tag{3.10}$$

where $g(\lambda)$ is the Lagrangian dual function that is defined as the maximum value of the Lagrangian in (3.9) over $\mathcal{S}$ for a given λ, i.e., $g(\lambda) = \max_{(\boldsymbol{\tau}, \boldsymbol{p}) \in S} L(\boldsymbol{\tau}, \boldsymbol{p}, \lambda)$.

Since the problem in (3.8) is convex and Slater's condition is satisfied, the duality gap is zero (Boyd and Vandenberghe, 2004). Therefore, if we can find the optimal λ^* for the Lagrangian dual problem, the optimal allocation policies $(\boldsymbol{\tau}^*(\lambda^*), \boldsymbol{p}^*(\lambda^*))$ that maximize $L(\boldsymbol{\tau}, \boldsymbol{p}, \lambda^*)$ must also be the optimal solution for the primal problem in (3.8). As stated in *Lemma 1* of (Wang and Giannakis, 2011), for ergodic fading channels with continuous cumulative distribution function, the almost surely unique solution of $g(\lambda)$, given some λ, is attained when each subcarrier is exclusively assigned to one single user per time slot, i.e.,

$$\tau_{k,m}^*(\lambda, \boldsymbol{\gamma}) = \begin{cases} 1 & k = k_m^* \\ 0 & \forall k \neq k_m^* \end{cases}, \forall\, m \tag{3.11}$$

where $k_m^* = \arg\max_{k \in K} \varphi_{k,m}^*(\lambda, \boldsymbol{\gamma})$ and $\varphi_{k,m}^*(\lambda, \boldsymbol{\gamma})$ is defined as:

$$\varphi_{k,m}^*(\lambda, \boldsymbol{\gamma}) = \begin{cases} \frac{a_1 \triangle B}{m_{sc}} \frac{w_k}{\ln 2} \ln \left[\frac{a_1 \triangle B}{m_{sc} a_2} \frac{w_k \gamma_{k,m}}{\lambda \ln 2} \right] - \frac{a_1 \triangle B}{m_{sc}} \frac{w_k}{\ln 2} + \frac{a_2 \lambda}{\gamma_{k,m}} & \gamma_{k,m} > \frac{a_2 m_{sc}}{a_1 \triangle B} \frac{\lambda \ln 2}{w_k} \\ 0 & \gamma_{k,m} \leq \frac{a_2 m_{sc}}{a_1 \triangle B} \frac{\lambda \ln 2}{w_k}. \end{cases} \tag{3.12}$$

The corresponding optimal power allocation is:

$$p_{k,m}^*(\lambda, \boldsymbol{\gamma}) = \begin{cases} \left[\frac{a_1 \triangle B}{m_{sc}} \frac{w_k}{\lambda \ln 2} - \frac{a_2}{\gamma_{k,m}} \right]^+ & k = k_m^* \\ 0 & \forall k \neq k_m^* \end{cases}, \forall\, m. \tag{3.13}$$

Based on the optimal allocation policy in (3.11) and (3.13), the optimal solution for the primal problem in (3.8) can be obtained after we determine the optimal λ^*. From (3.13), it is clear that λ must be larger than zero, otherwise, the allocated power will be positive infinity which is impractical. According to the complementary slackness condition (Boyd and Vandenberghe, 2004), the optimal λ^*, the maximum average transmission power $\bar{P}$ and the sum of allocated average transmission power $\bar{P}_t(\lambda^*) = \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{k=1}^{K} \sum_{m=1}^{M} p_{k,m}^*(\lambda^*, \boldsymbol{\gamma}) \right]$ must satisfy: $\lambda^* \left(\bar{P} - \bar{P}_t(\lambda^*) \right) = 0$. Since $\lambda^* > 0$, we have $\bar{P} = \bar{P}_t(\lambda^*)$. According to *Lemma 2* of (Wang and Giannakis, 2011), the instantaneous power $p_{k^*,m}(\lambda, \boldsymbol{\gamma})$ and the related $\bar{P}_t$ are both nonincreasing function of λ. For this reason, the optimal λ^* can be obtained using numerical methods, e.g., bisection search.

Chapter 4

Quality Fairness-oriented Cross-layer Resource Allocation: Preliminaries

In this chapter, we present the preliminary results that are related to the quality fairness-oriented cross-layer resource allocation frameworks introduced in the following chapters. Specifically, we first show and discuss the architecture and functionality of a general multi-user video delivery system in Section 4.1. In Section 4.2, we present two R-D models, based only on the SVC encoding process, for MGS video streams supporting QL-based extraction. Finally, we present a quality-maximized unequal erasure protection (UXP) scheme and extend the two R-D models to cover the case of packet erasure with UXP in Section 4.3.

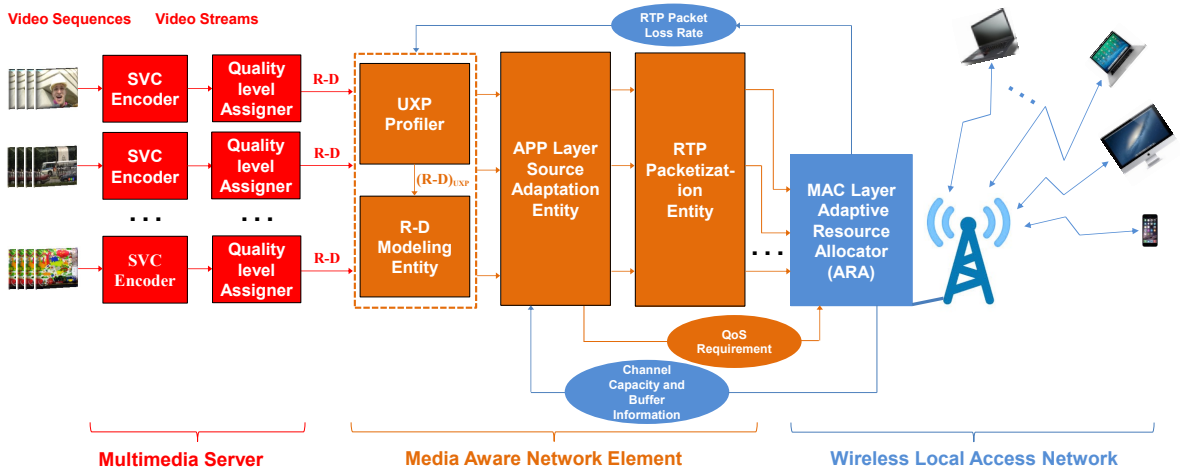


Figure 4.1: Architecture and components of the multi-user video delivery system.

4.1 Multi-user Video Delivery System: Architecture and Functionality

We consider a general multi-user video delivery system shown in Fig. 4.1. As illustrated in Fig. 4.1, the system components can be arranged into three groups, namely, the multimedia server (MS), the media-aware network element (MANE), and the wireless local access network (WLAN). The MS encodes a set of video sequences, each of which is requested by a user, to fully support MGS. Each encoded video stream is then organized into NALUs and post-processed by a quality level assigner. The quality level assigner evaluates the priority level of each NALU according to its contribution to the quality of the reconstructed video. Such priority level information is embedded into the header of the NALU and will be exploited by the source adaptation entity. It should be pointed out that the encoding and priority level assessment are carried out off-line. The pre-encoded video streams are stored in databases at the MS, whereas their R-D information will be forwarded to the MANE.

In the scenario of error-free transmission, the R-D modeling entity exploits the R-D information from the MS to construct R-D models that depict the relationship between the rates and the reconstructed qualities of the video streams. For video transmission over error-prone channels, the UXP profiler collects the R-D information from the MS and the estimated real-time transport protocol (RTP) packet loss rate from the BS periodically. Then, it computes, according to a predefined protection policy and the available information, the rates and the expected reconstructed video qualities, after error protection, of the videos. Such information is fed to the R-D modeling entity and is used for the R-D modeling for the UXP protected video streams. The source adaptation entity removes, according to the results of the source adaptation algorithm, the needless NALUs from each original video stream to form a valid substream intended for a user. It should be pointed out that the source adaptation algorithm requires as inputs the estimated channel capacity and buffer status information from the BS and the information about the R-D models at the MANE. Each outcoming substream is forwarded to the real-time transport protocol (RTP) packetization entity where the substream will be protected by a UXP scheme based on RS codes. The resulting RS codewords, containing both data and parity symbols, are arranged into a transmission block (TB) and interleaved over a number of RTP packets. Finally, the RTP packets would be sent to the lower layers, e.g., MAC/PHY, of the system.

The adaptive resource allocator (ARA) at the BS of the WLAN adaptively allocates the system resource among users with the aim to maximize overall average rates while satisfying the QoS requirement provided by the APP layer.

It is worth noting that whereas the processes of R-D modeling, UXP, source

adaptation and the RTP packetization at the MANE are executed per *application period* (in the order of seconds), the resource allocation process at the BS is carried out every time slot (in the order of milliseconds). The MANE and the BS exchange information about the channel capacity, buffer status, RTP packet-loss rate and QoS requirement in a cross-layer style at regular intervals, i.e., *application periods*.

4.2 Rate Distortion Models for MGS Video Streams

In this section, we present two R-D models for MGS video streams supporting QL-based extraction. The presented R-D models describe the relationship between the rate and quality of the reconstructed video at the video encoder. When a scalable video sequence is encoded and transmitted, it is a common practice to encode it with respect to a small number of frames and transmit it adaptively. In this thesis, we follow the aforementioned common practice and will focus on a IDR-based video transmission.

Let us consider that at the beginning of each *application period*, I_k successive frames of a video sequence, intended for user k, are encoded to generate an MGS video stream. Let C_k denote the cardinality of the set of valid substreams that can be extracted from the original stream. In general, the value of C_k is different for each stream and is determined by the available encoding schemes that support different temporal, spatial and quality scalability. We define $\mathcal{D}_k = \{d_{k,1}, d_{k,2}, \cdots, d_{k,C_k}\}$ as the sets of lossy encoding distortion values for the video stream where the distortion $d_{k,c}, \forall c = 1, 2, \cdots, C_k$, is given as the mean square error (MSE) between the original and reconstructed video frames averaged over all I_k frames. The reconstructed video quality is measured according to the peak signal-to-noise ratio (PSNR), which is a

commonly used objective quality measure in video coding, defined as:

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right). \tag{4.1}$$

The set of PSNR values for the video stream will be denoted by $\mathcal{Q}_k = \{q_{k,1}, q_{k,2}, \cdots, q_{k,C_k}\}$ where $q_{k,c}, \forall c = 1, 2, \cdots, C_k$ is calculated according to eq. (4.1).

In practice, the minimum rate $F_k(q_{k,c})$, in bit per second, required to transmit to user k the cth substream with a given PSNR $q_{k,c}$ is a strictly monotonically increasing discrete-value function. We define the set of the minimum required rates as $\mathcal{F}_k = \{F_k(q_{k,1}), F_k(q_{k,2}), \cdots, F_k(q_{k,C_k})\} = \{f_{k,1}, f_{k,2}, \cdots, f_{k,C_k}\}$. In the case of error-free transmission, $f_{k,c}, \forall c = 1, 2, \cdots, C_k$ depends only on the rate of the encoder. The first model we consider is the following empirical R-D model which maps the PSNR to the minimum required rate, i.e.,

$$\begin{aligned} F_k : \mathcal{Q}_k &\to \mathcal{F}_k \\ q_{k,c} &\mapsto f_{k,c} = F_k(q_{k,c}). \end{aligned} \tag{4.2}$$

In (Stuhlmüller *et al.*, 2000), the authors proposed a general continuos semi-analytical R-D model, which has been verified for SVC quality scalable videos in (Mansour *et al.*, 2008) and (Cicalò *et al.*, 2012), to estimate the relationship between the rate and distortion at the encoder side. The model is expressed as:

$$F_k(D) = \frac{\theta_k}{D + \alpha_k} + \beta_k,\ D \in [D_{k,min}, D_{k,max}] \tag{4.3}$$

where D is the distortion measured as MSE; $F_k(D)$ is the output rate of the video

encoder; $D_{k,min}$ and $D_{k,max}$ are the minimum and maximum distortion values of the video, respectively, after decoding the base layer (with rate $F_{k,max}$) and all layers (with rate $F_{k,min}$) of the video stream; the three parameters θ_k, α_k and β_k are video content and encoder dependent, and can be estimated using curve-fitting methods over a number of empirical R-D points. According to extensive simulations, a general curve-fitting algorithm needs at least six empirical R-D points and a certain number of iterations and function evaluations to guarantee high accuracy for most of the video sequences (Cicalò *et al.*, 2012). Combining eq. (4.3) with the relation between MSE and PSNR given in eq. (4.1), the relationship between the PSNR and rate can be described by a parametric function $F_k(Q)$ with a continuous variable Q. The second model we consider is a continuous semi-analytical R-D model defined by the following strictly monotonically increasing function:

$$F_k(Q) = \frac{\theta_k}{255^2 10^{-Q/10} + \alpha_k} + \beta_k,\ Q \in [Q_{k,min}, Q_{k,max}] \tag{4.4}$$

where $Q_{k,min}$ and $Q_{k,max}$ are the minimum and maximum PSNR values of the video, respectively, after decoding the base layer and all layers of the video stream.

4.3 Rate Distortion Models with Packet Erasure

In practice, the quality of a received video is heavily influenced by packet losses during transmission. UXP has been promoted for the protection of SVC video data to combat packet losses during transmission (Schierl *et al.*, 2005). With UXP and packet losses, it is necessary to develop R-D models that estimate the relationship between the rate including the UXP overhead and the expected quality that considers both encoding

losses and packet losses during transmission. To this end, we first discuss an expected PSNR-maximized UXP scheme for SVC video data. Then, by taking into account the effect of UXP and packet losses, we extend the R-D models in (4.2) and (4.4) to cover the case of error-prone transmission with UXP.

4.3.1 An Expected PSNR-maximized UXP Scheme

In this thesis, we consider an UXP scheme based on the use of RS codes. After each IDR-period of video sequences is encoded, the NALUs of the resulting scalable stream are sorted based on their priority levels. In general, given the available bandwidth, a substream is extracted from the original scalable stream and the NALUs, belonging to the substream, are sequentially inserted into a transmission block (TB) from upper left to lower right according to their priority levels as shown in Fig. 4.2.

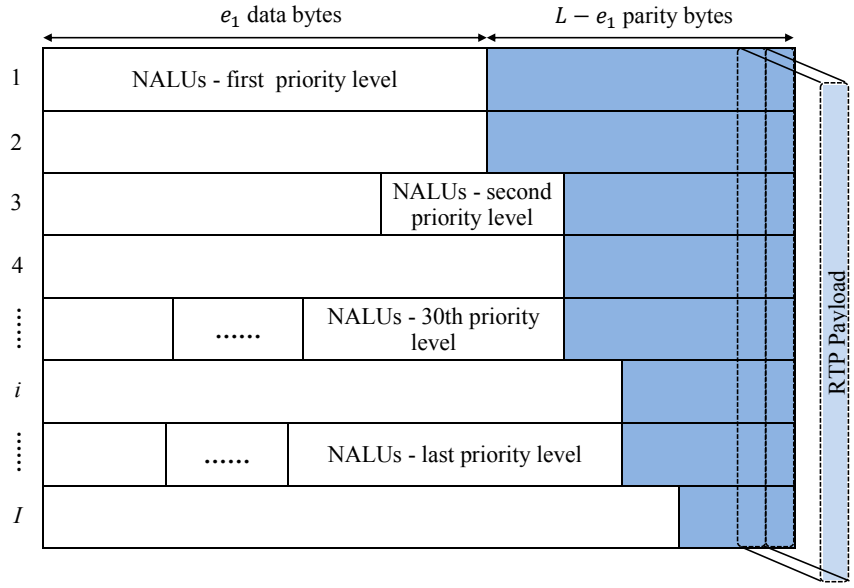


Figure 4.2: An example of transmission block structure where each row identifies an RS codeword and corresponds to a protection class, and each column represents an RTP packet.

Let L and I be the number of columns and rows of the TB, respectively. The

NALUs are partitioned into I consecutive data segments, each of which is protected by an (L, e_i) RS code where e_i is the number of data symbols (a symbol consists of a fixed number of bits, generally 8 bits) in the ith segment. The e_i data symbols followed by $f_i = L - e_i$ parity symbols form the ith row of the TB. An RTP packet is formed across the rows and thus is represented by a column of the TB. By applying the (L, e_i) RS code on data segment i, the e_i data symbols in segment i can be correctly recovered if at most f_i packets are lost during transmission provided that the orders of the lost packets are known. Due to the characteristics of the scalable stream, decoding the ith segment can further improve the reconstructed video quality only if the previous i-1 segments are available. Therefore, the number of parity symbols (corresponding to a particular protection class) allocated to each row should be monotonically non-increasing in the row number; in other words, the number of data symbols assigned to each row should be monotonically non-decreasing in the row number:

$$e_1 \leq e_2 \leq \cdots \leq e_I. \tag{4.5}$$

Let us first assume that the rate-quality function $q(r)$ of the scalable stream is a monotonically non-decreasing function in $r \in \{0, 1, 2, \cdots, R_{max}\}$ where r is the number of data symbols, and R_{max} is the total number of data symbols in the scalable stream to be transmitted, respectively. Let $p_L(l)$, $\forall l = 0, 1, \cdots, L$, denote the probability of losing l packets out of L packets. Furthermore, let $r_0, r_1, \cdots, r_I$ denote the data segment partitioning points, i.e., $r_0 = 0$ and $r_i = \sum_{j=1}^{i} e_j = iL - \sum_{j=1}^{i} f_j$. Then, the probability that the receiver can achieve the PSNR $q(r_i)$ is $\sum_{l=f_{i+1}+1}^{f_i} p_L(l)$ and

the expected PSNR of the received video is then given as follows (Mohr *et al.*, 2000):

$$\begin{aligned} Q(\boldsymbol{e}) &= P_L(L)q(r_0) + \sum_{i=1}^{I} P_L(f_i)(q(r_i) - q(r_{i-1})) \\ &= P_L(L)q(r_0) + \sum_{i=1}^{I} P_L(L - e_i)(q(r_i) - q(r_{i-1})) \end{aligned} \tag{4.6}$$

where $\boldsymbol{e} = [e_1, e_2, \cdots, e_I]^T$ is the vector whose elements are the number of data symbols allocated to the data segments and $P_L(l) = \sum_{k=1}^{l} p_L(k), \forall l = 0, 1, \cdots, L$. The objective of the expected PSNR-maximized UXP scheme is to find the assignment vector $\boldsymbol{e}$ that maximizes $Q(\boldsymbol{e})$, given I, L, $p_L(l)$ and $q(r)$. The corresponding optimization problem can be expressed as:

$$\max_{\boldsymbol{e}} \quad Q(\boldsymbol{e}) \tag{4.7a}$$

$$s.t. \quad e_1 \leq e_2 \leq \cdots \leq e_I. \tag{4.7b}$$

In (Dumitrescu *et al.*, 2007), the authors showed that solving problem (4.7) is equivalent to solving a maximum-weight path problem constrained on the number of edges, which can be solved using the Lagrangian method. Based on the assumptions that $q(r)$ is concave and the channel is an independent erasure channel with packet erasure rate no larger than $\frac{L}{2(L+1)}$, an algorithm is proposed to obtain the globally optimal solution $\boldsymbol{e}^*$, which can be computed in $O(\gamma LI)$ where γ is the number of iterations needed to find the optimal Lagrangian multiplier λ. Simulation results showed that on average γ increases at a rate close to $O(\log_2 I)$.

As shown in Section 4.2, SVC works only on a small discrete set of rate and PSNR points, i.e., the rate-quality function is $q(r)$ with $r \in \{r_1, r_2, \cdots, r_C\} \subset$

$\{0, 1, 2, \cdots, R_{max}\}$ where $r_C = R_{max}$. In addition, the rate-quality function is in general not concave. In order to have an exact concave rate-quality function $q(r)$ in $r \in \{0, 1, 2, \cdots, R_{max}\}$, we can approximate the real rate-quality curve with its upper convex hull. Based on the approximated rate-quality curve, the algorithm in (Dumitrescu *et al.*, 2007) is applied to obtain the optimal assignment vector $\boldsymbol{e}^*$. However, the expected PSNR $Q(\boldsymbol{e})$ is calculated, according to (4.6), using the optimal assignment vector $\boldsymbol{e}^*$ together with the following rate-quality function:

$$q'(r) = \begin{cases} 0 & r \in \{0, 1, r_1 - 1\} \quad (4.8a) \\ q(r_i) & r \in \{r_i, r_i + 1, \cdots, r_{i+1} - 1\}, i = 1, 2, \cdots, C - 1 \quad (4.8b) \\ q(r_C) & r = r_C. \quad (4.8c) \end{cases}$$

Given a certain packet-loss rate r_{rtp}, the expected PSNR is a non-decreasing function in the number of packets L, or equivalently, in the transmission budget LI, provided that the size of the packet is fixed. Therefore, varying the value of L, we can obtain a set of rate and expected PSNR points. Such expected R-D information can be exploited to extend the R-D models in (4.2) and (4.4) to cover the case of video transmission with packet losses and UXP. We summarize the changes as follows when the R-D models in (4.2) and (4.4) are extended using the expected R-D information:

1. The cardinality C_k of the set of R-D points is not the number of substreams, but rather, is determined by the number of available transmission budgets.

2. The PSNR value $q_{k,c}, \forall c = 1, 2, \cdots, C_k$ is the expected PSNR computed as in (4.6).

3. Each minimum required rate value $f_{k,c}, \forall c = 1, 2, \cdots, C_k$ corresponds to an available transmission budget including both the data and parity bits.

4. $F_{k,min}$ and $F_{k,max}$ are the minimum and maximum rates, respectively, corresponding to the minimum and maximum transmission budgets.

5. $Q_{k,min}$ is the expected PSNR value when the k-th video is transmitted at the minimum transmission budget. $Q_{k,max}$ is the expected PSNR value when the k-th video is transmitted at the maximum transmission budget.

6. For the R-D model (4.4), the three parameters θ_k, α_k and β_k depend on the video content, the encoder and r_{rtp}.

It is worth noting that the number of available transmission budgets is different for each video stream. Given a symbol size s, the maximum codeword length for an RS code is $L_{max} = 2^s - 1$. Therefore, the maximum number of RTP packets is L_{max}. In addition, in order to transmit at least the base layer, the minimum number of RTP packets L_{min} should be sufficiently large to map the base layer into the TB. Therefore, L_{min} is video content dependent. Since the number of available transmission budgets depends on the values of L_{max} and L_{min}, it is in general different for each stream.

Due to the variation of r_{rtp}, the main challenge of extending the R-D models to cover the case of packet-loss transmission is the periodical update of the expected R-D information. Given a certain packet-loss rate r_{rtp}, the UXP profiler needs to compute and store, in the worst case, L_{max} rate and expected PSNR values for each user. To overcome this challenge, we choose to compute and store only the needed expected R-D values on-the-fly. For example, for the semi-analytical model, we only compute the expected R-D values that are used in the curve-fitting process to estimate the three parameters of the model. For the empirical model, we compute the expected R-D values when they are needed during the source adaptation process.

Chapter 5

Distortion-fair Cross-layer Resource Allocation for SVC Video Delivery

Since our quality fairness-oriented cross-layer resource allocation framework is very closely related to the work in (Cicalò and Tralli, 2014), we retrace in this chapter the main results of (Cicalò and Tralli, 2014). In (Cicalò and Tralli, 2014), the authors presented a cross-layer optimization framework for SVC video transmission in OFDMA wireless networks. Within the framework, resource allocation and source adaptation are jointly addressed such that the sum of averaged (ergodic) rate assigned to users is maximized while the distortion difference among the received videos is minimized. The optimization problem is decomposed into two subproblems, namely, a resource allocation problem at the MAC layer and a rate adaptation problem at the APP layer. Then, an iterative local approximation (ILA) algorithm, with provable optimality and

convergence, is proposed to derive the optimal global solution. The iterative procedure requires only a limited scalar information exchange between the MAC and APP layers. The authors then present the algorithms to solve the source adaptation and resource allocation problems.

5.1 Problem Formulation

Based on model (4.3), the optimization problem is formulated to maximize the sum of the ergodic rates under a set of rate constraints:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \quad \|\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})\|_1 \tag{5.1a}$$

$$s.t. \quad \triangle(D_i, D_j) = 0, \forall i,j \in \mathcal{K} \wedge i \neq j \tag{5.1b}$$

$$V\boldsymbol{F}_{min} \preceq \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \preceq V\boldsymbol{F}_{max} \tag{5.1c}$$

$$\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{R} \tag{5.1d}$$

where $\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p})$ is the ergodic rate vector and $\mathcal{R}$ is the ergodic rate region given in (3.6). The distortion fairness constraints in (5.1b), to be explained shortly, impose constraints on individual rate through the relation $D_k = F_k^{-1}(R_k(\boldsymbol{\tau},\boldsymbol{p})/V), \forall k \in \mathcal{K}$. Moreover, $V \geq 1$ is a constant accounting for the overhead resulted from the communication among different network layers. Finally, $\boldsymbol{F}_{max} = [F_{1,max}, F_{2,max}, \cdots, F_{K,max}]^T$ and $\boldsymbol{F}_{min} = [F_{1,min}, F_{2,min}, \cdots, F_{K,min}]^T$, with $F_{k,max} = F_k(D_{k,min})$ and $F_{k,min} = F_k(D_{k,max})$, are vectors of the maximum and minimum rates, respectively, of the video streams.

With the fairness constraints in (5.1b), the optimization problem aims to equalize the distortion values of the received videos. However, this may not be possible because the distortion of each video is constrained to be within a range bounded by its minimum and maximum values. Consider that, if it is impossible to further reduce the distortion of the i-th video since its minimum value $D_{i,min}$ has already been achieved, then the available resources should be allocated for the transmission of the other videos to further decrease their distortion values even though those distortion values will be smaller than $D_{i,min}$. On the other hand, when the ith video has already reached its maximum distortion value $D_{i,max}$, but further reduction of rate is needed, it is mandatory to reduce the rate allocated to other videos, even though their distortion values would become larger than $D_{i,max}$. Motivated by the aforementioned considerations, the distortion difference $\triangle(D_i, D_j)$ in (5.1b) is defined as:

$$\triangle(D_i, D_j) = \begin{cases} 0 & (D_i, D_j) \in \mathbb{D} \vee (D_j, D_i) \in \mathbb{D} \quad (5.2a) \\ |D_i - D_j| & \text{otherwise} \quad (5.2b) \end{cases}$$

where $\mathbb{D} = \{(D_i, D_j) \mid (D_i = D_{i,max} \wedge D_j > D_i) \vee (D_i = D_{i,min} \wedge D_j < D_i)\}$.

The fairness constraints in (5.1b) restrict the feasible solutions to a set of rate vectors:

$$\mathcal{R}_f^c = \{\boldsymbol{R} \mid \triangle \left(F_i^{-1}(R_i/V), F_j^{-1}(R_j/V)\right) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j\}. \quad (5.3)$$

The set $\mathcal{R}_f^c$ describes a one-dimensional monotonically increasing manifold with boundary in the $\mathbb{R}^K$ space. Moreover, according to the constraints in (5.1c) and (5.1d), there exists achievable and nontrivial solutions to the optimization problem if and only if (i) transmitting all videos with the lowest quality is supported by the PHY layer,

i.e., $V\boldsymbol{F}_{min} \in \mathcal{R}$, and (ii) the PHY layer does not support the transmission of all videos with the highest quality, i.e., $V\boldsymbol{F}_{max} \notin \mathcal{R}$. This is because, on one hand, if we cannot guarantee that each user receives at least the video with the lowest quality, the optimization process must be terminated unsuccessfully and admission control should be considered, which is beyond the scope of (Cicalò and Tralli, 2014). On the other hand, if the channel capacity is sufficiently high to support the transmission of all original streams, there is no need for source adaptation. Therefore, the feasible and nontrivial solutions should lie in a rate region given by a nonempty set $\mathcal{R}_a = \{\boldsymbol{R} \in \mathcal{R} \mid V\boldsymbol{F}_{min} \preceq \boldsymbol{R} \preceq V\boldsymbol{F}_{max}\}$. The nonempty property of $\mathcal{R}_a$ can be guaranteed if $V\boldsymbol{F}_{min} \in \mathcal{R}$ and $V\boldsymbol{F}_{max} \notin \mathcal{R}$.

Since the objective function in (5.1) is concave and increasing, $\forall \boldsymbol{R} \in \mathcal{R}$ (Wang and Giannakis, 2011), the optimal solution $\boldsymbol{R}^*$ must be attained at the boundary of the rate region $\mathcal{R}$ defined by the Pareto efficient set:

$$\mathbf{bd}\,\mathcal{R} = \{\boldsymbol{R} \in \mathcal{R} \mid \nexists \boldsymbol{r} \in \mathcal{R} \text{ with } \boldsymbol{r} \succcurlyeq \boldsymbol{R}\} \tag{5.4}$$

where $\boldsymbol{r} = [r_1, r_2, \cdots, r_K]^T$ is a K-tuple, and given by the intersection of the boundary $\mathbf{bd}\,\mathcal{R}$ and the one-dimensional manifold described by $\mathcal{R}_f^c$. The optimal solution is unique due to the one-dimensionality and monotonicity of the manifold described by $\mathcal{R}_f^c$. In Fig. 5.1, we show an example of the optimization problem for a case of two-user video transmission.

Since the problem formulation in (5.1) involves optimization variables and constraints defined at both the MAC and APP layers, it requires in general the presence of a centralized controller to handle the variables and constraints jointly. In order to reduce the information exchange between the MAC and APP layers, an alternative

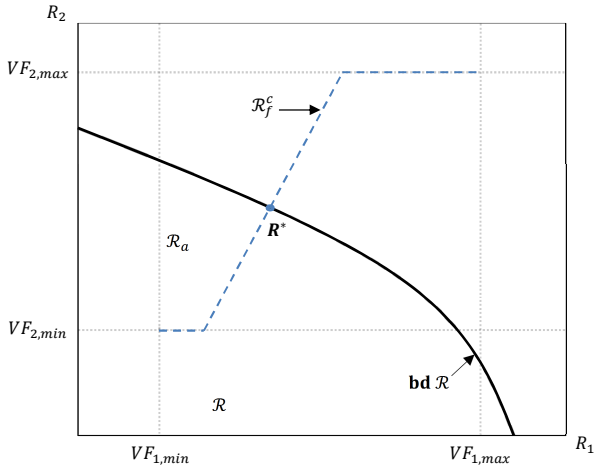


Figure 5.1: An example of the optimization problem for a system with two users. The optimal solution $\boldsymbol{R}^*$ is given by the intersection of the boundary of the rate region $\mathbf{bd}\,\mathcal{R}$ and the dash line described by the set $\mathcal{R}_f^c$.

way is to decompose the problem into two subproblems, each of which is addressed by an individual entity at the associated layer. In this way, the two subproblems can be solved repeatedly with only a limited information exchange between the MAC and APP layers until the optimal solution is attained.

5.2 Problem Decomposition

Supposing that a perfect knowledge of the boundary $\mathbf{bd}\,\mathcal{R}$ of the rate region $\mathcal{R}$ is available at the APP layer, the problem in (5.1) can be simplified into a constraint-satisfaction problem where the objective is to find $\boldsymbol{F}$ such that:

$$V\boldsymbol{F} \in \mathbf{bd}\,\mathcal{R} \cap \mathcal{R}_f^c. \tag{5.5}$$

The problem in (5.5) is a source adaptation problem that can be handled by the APP layer. Interestingly, there is no objective function in problem (5.5). However, the

objective of maximizing the sum of rates will be reached because the solution V$\boldsymbol{F}$ is on the boundary $\mathbf{bd}\,\mathcal{R}$.

On the other hand, assuming that the MAC layer has the information about the direction of the line passing through the optimal solution $\boldsymbol{R}^*$, problem (5.5) can be simplified into a resource allocation problem at the MAC layer. Particularly, the directional line departs from $\boldsymbol{R} = \mathbf{0}$ and has an intersection $\boldsymbol{R} = \boldsymbol{R}^*$ with the boundary $\mathbf{bd}\,\mathcal{R}$. Therefore, the line can be characterized by an equation $\boldsymbol{R}^* = \boldsymbol{\phi} r$ where $\boldsymbol{\phi} = [\phi_1, \phi_2, \cdots, \phi_K]^T \succeq \mathbf{0}$ defines the direction of the line and r is a positive real number. If we further assume that $\|\boldsymbol{\phi}\|_1 = 1$, we have $\|\boldsymbol{R}^*\|_1 = \|\boldsymbol{\phi}\|_1 r = r$. Then, the optimization problem in (5.1) can be simplified into a constrained sum-rate maximization that aims to find the optimal allocation policy $(\boldsymbol{\tau}^*, \boldsymbol{p}^*)$ under proportionality rate constraints implied by $\boldsymbol{\phi}$:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \quad r \tag{5.6a}$$

$$s.t. \quad \boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \succeq \boldsymbol{\phi} r \tag{5.6b}$$

$$\boldsymbol{R}(\boldsymbol{\tau},\boldsymbol{p}) \in \mathcal{R}. \tag{5.6c}$$

The problem in (5.6) is a well-investigated resource allocation problem (Wong and Evans, 2008a) that can be solved efficiently given the information on the directional vector $\boldsymbol{\phi}$. In fact, the vector $\boldsymbol{\phi}$ can be determined if the solution $\boldsymbol{F}^*$ to the problem (5.5) is known, i.e.,

$$\boldsymbol{\phi} = \frac{\boldsymbol{F}^*}{\|\boldsymbol{F}^*\|_1}. \tag{5.7}$$

Moreover, it is shown in (Wong and Evans, 2008a) that the Lagrangian dual problem

associated with problem (5.6) is similar to that related to the WSAR maximization problem in (3.7). Therefore, the optimal rate solution of problem (5.6) can be obtained through solving a WSAR problem and must be on the boundary $\mathbf{bd}\,\mathcal{R}$. However, differently from the WSAR problem, the weight vector $\boldsymbol{w}$ is not predefined, but rather, it is evaluated in the dual domain and constrained by $\boldsymbol{\phi}$.

5.3 Iterative Local Approximation Algorithm

The availability of a close-form formula for $\mathbf{bd}\,\mathcal{R}$ will enable us to easily derive the solution to problem (5.5). Even though such an explicit formula is in general not available for the OFDMA downlink scenario, each boundary point can be obtained by solving the WSAR problem in (3.7) with a given weight vector $\boldsymbol{w}$, as mentioned in Section 3.2.2. Let us assume that for each rate point $\boldsymbol{R} \in \mathbf{bd}\,\mathcal{R}$, there exists a tangent space $\mathcal{T}_{\mathcal{R}}(\boldsymbol{R})$ to the rate region $\mathcal{R}$ at $\boldsymbol{R}$. The key idea of the ILA algorithm is to exploit the tangent spaces as the local approximation of the boundary $\mathbf{bd}\,\mathcal{R}$ to establish an iterative procedure under which the two subproblems at the MAC and APP layers are solved repeatedly until the optimal rate point $\boldsymbol{R}^*$ is obtained. Let us denote $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{w}}$ as the optimal rate and the associated weight of the WSAR maximization problem, respectively. Then, the tangent space to the rate region $\mathcal{R}$ at the point $\tilde{\boldsymbol{R}}$ can be identified by the null space of $\tilde{\boldsymbol{w}}$, and thus defined by the set:

$$\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = \{\boldsymbol{R} \mid \tilde{\boldsymbol{w}}^T(\boldsymbol{R} - \tilde{\boldsymbol{R}}) = 0\}. \tag{5.8}$$

Then, the problem in (5.5) can be approximated by:

$$V\boldsymbol{F} \in \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \cap \mathcal{R}_f^c. \tag{5.9}$$

The ILA algorithm solves problem (5.6) followed by problem (5.9) iteratively. More specifically, starting with an initial direction vector $\tilde{\boldsymbol{\phi}}^{(0)}$, the MAC layer solves the problem in (5.6) to obtain the optimal rate $\tilde{\boldsymbol{R}}^{(0)}$ and the related weight $\tilde{\boldsymbol{w}}^{(0)}$, which are forwarded to the APP layer. The APP layer exploits the information of the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}^{(0)}, \tilde{\boldsymbol{w}}^{(0)})$ identified by $\tilde{\boldsymbol{R}}^{(0)}$ and $\tilde{\boldsymbol{w}}^{(0)}$ to derive the optimal solution $\tilde{\boldsymbol{F}}^{(1)}$, i.e., $\tilde{\boldsymbol{F}}^{(1)} = \boldsymbol{R}_f^c \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}^{(0)}, \tilde{\boldsymbol{w}}^{(0)})$. The resulting direction vector $\tilde{\boldsymbol{\phi}}^{(1)}$, computed from $\tilde{\boldsymbol{F}}^{(1)}/\|\tilde{\boldsymbol{F}}^{(1)}\|_1$, is then forwarded to the MAC layer, which projects the solution back to the boundary of $\mathcal{R}$ by solving the problem (5.6) to get $\tilde{\boldsymbol{R}}^{(1)}$ and the corresponding $\tilde{\boldsymbol{w}}^{(1)}$. Starting from an arbitrary $\tilde{\boldsymbol{R}}^{(0)} \in \mathbf{bd}\,\mathcal{R}$ and keeping moving toward $\boldsymbol{R}^*$, the above processes are repeated until convergence, according to a stopping criteria. Specifically, the repeated processes terminate when the error between APP and MAC layer solutions, i.e., $e^{(i)} = \|\tilde{\boldsymbol{R}}^{(i)} - V\tilde{\boldsymbol{F}}^{(i)}\|_1$, is sufficiently small. Interestingly, the error for the rates of user k at iteration i is in proportion to the value of $\phi_k^{(i)}$, i.e., $e_k^{(i)} = e^{(i)}\phi_k^{(i)}$ and the total error is given by $e^{(i)} = \|e^{(i)}\boldsymbol{\phi}\|_1$. A concise description of the ILA algorithm is given in **Algorithm 1** while an example of the first step of the ILA algorithm for a two-user case is shown in Fig. 5.2. The convergence and optimality of the ILA algorithm are stated in **Lemma 1**, which is proved in (Cicalò and Tralli, 2014).

Lemma 1. *Given that $V\boldsymbol{F}_{min} \in \mathcal{R}$ and $V\boldsymbol{F}_{max} \notin \mathcal{R}$, the ILA algorithm, starting from an initial $\boldsymbol{R} \succeq \mathbf{0}$, converges to the unique optimal rate solution $\boldsymbol{R}^* \in \boldsymbol{R}_f^c \cap \mathbf{bd}\,\mathcal{R}$,*

Algorithm 1 ILA algorithm

1: **Initialize:** $i = 0$; give an directional vector $\tilde{\boldsymbol{\phi}}^{(0)} \succcurlyeq \mathbf{0}$ and tolerance $\epsilon > 0$
2: Solve problem (5.6) to obtain $\tilde{\boldsymbol{R}}^{(0)}$ and $\tilde{\boldsymbol{w}}^{(0)}$
3: **while** $e^{(i)} > \epsilon$ **do**
4: $\quad i = i + 1$
5: $\quad$ Find $\tilde{\boldsymbol{F}}^{(i)}$ such that:
6: $\quad V\tilde{\boldsymbol{F}}^{(i)} \in \boldsymbol{R}_f^c \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}^{(i-1)}, \tilde{\boldsymbol{w}}^{(i-1)})$
7: $\quad \tilde{\boldsymbol{\phi}}^{(i)} = \tilde{\boldsymbol{F}}^{(i)} / \|\tilde{\boldsymbol{F}}^{(i)}\|_1$
8: $\quad$ Solve problem (5.6) to obtain $\tilde{\boldsymbol{R}}^{(i)}$ and $\tilde{\boldsymbol{w}}^{(i)}$
9: **end while**

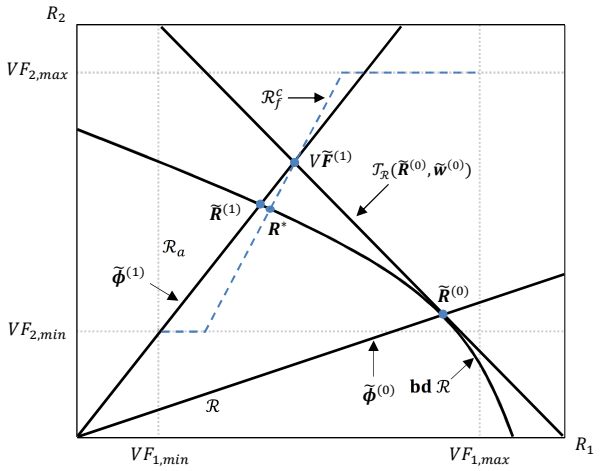


Figure 5.2: An example of first step of the ILA algorithm for a two-user case.

i.e.,

$$\lim_{i\to\infty} V\tilde{\boldsymbol{F}}^{(i)} = \boldsymbol{R}^*. \tag{5.10}$$

Finally, it is worth noting that the optimal solution $\boldsymbol{R}^*$ may not be supported by the available SVC encoding schemes corresponding to a finite set of rate values. To derive the optimal discrete solution, it is a common practice to extract the largest achievable rate which is smaller than $\boldsymbol{R}^*$, at the cost of a minimum waste of bandwidth.

5.4 MAC Layer Subproblem: Resource Allocation

Given the directional vector $\tilde{\boldsymbol{\phi}}$ from the APP layer at each iteration of the ILA algorithm, the MAC layer is able to solve problem (5.6). The optimal solution $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{w}}$ to problem (5.6) can be obtained by using the Lagrangian dual and stochastic subgradient methods, as shown in (Wong and Evans, 2008a). In this section, we will show the procedures of solving problem (5.6), as in (Wong and Evans, 2008a), with an additional assumption, namely, the time-sharing of subcarriers.

Let λ be the Lagrangian multiplier related to the average power constraint implied by (5.6c) and $\tilde{\boldsymbol{w}}$ be the Lagrangian multiplier vector related to proportional rate constraint (5.6b), the Lagrangian associated with (5.6) is then denoted as $L(\boldsymbol{\tau}, \boldsymbol{p}, \lambda, \tilde{\boldsymbol{w}})$. Accordingly, the dual problem is:

$$\min_{\lambda \geq 0, \tilde{\boldsymbol{w}} \succcurlyeq \mathbf{0}} g(\lambda, \tilde{\boldsymbol{w}}) \tag{5.11a}$$

$$s.t. \quad \tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{\phi}} = 1 \tag{5.11b}$$

where $g(\lambda, \tilde{\boldsymbol{w}}) = \max_{(\boldsymbol{\tau}, \boldsymbol{p}) \in S} L(\boldsymbol{\tau}, \boldsymbol{p}, \lambda, \tilde{\boldsymbol{w}})$ is the Lagrangian dual function associated with problem (5.6). Note that the constraint (5.11b) prevents the optimal rate being infinity or zero. As shown in Section 3.2.2, given λ and $\tilde{\boldsymbol{w}}$, the unique solution of the dual function $g(\lambda, \tilde{\boldsymbol{w}})$ is attained when each subcarrier is exclusively assigned to one single user per time slot and the power is allocated per user across subcarriers following a water-filling principle as shown in (3.11) and (3.13), respectively. To derive the solutions to problem (5.11), a subgradient method is used. That is, given an initial

λ^0, the optimal λ^* can be obtained through the following subgradient iterations:

$$\lambda^{s+1} = [\lambda^s - h_\lambda^s g_\lambda^s]^+ \tag{5.12}$$

where $h_\lambda^s \geq 0$ is the sth step size chosen to ensure convergence and g_λ^s is the subgradient defined as:

$$g_\lambda^s = \bar{P} - \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{m=1}^{M} p_{k,m}^*(\lambda^s, \tilde{\boldsymbol{w}}^s) \right]. \tag{5.13}$$

Similarly, given an initial $\tilde{\boldsymbol{w}}^0$, the optimal $\tilde{\boldsymbol{w}}^*$ can be obtained through subgradient iterations as follows:

$$\tilde{\boldsymbol{w}}^{s+1} = \frac{[\tilde{\boldsymbol{w}}^s - h_{\tilde{\boldsymbol{w}}}^s g_{\tilde{\boldsymbol{w}}}^s]^+}{\tilde{\boldsymbol{\phi}}^T [\tilde{\boldsymbol{w}}^s - h_{\tilde{\boldsymbol{w}}}^s g_{\tilde{\boldsymbol{w}}}^s]^+} \tag{5.14}$$

where $h_{\tilde{\boldsymbol{w}}}^s \geq 0$ is the sth step size chosen to ensure convergence and g_λ^s is the subgradient defined as:

$$g_{\tilde{\boldsymbol{w}}}^s = \boldsymbol{r}^* - \tilde{\boldsymbol{\phi}}^T \|\boldsymbol{r}^*\|_1 \tag{5.15}$$

where $\boldsymbol{r}^* = \{r_1^*, r_2^*, \cdots, r_K^*\}$ and $r_k^* = \mathbb{E}_{\boldsymbol{\gamma}} \left[m_{sc} \sum_{m=1}^{M} r_{k,m}(\tau_{k,m}^*(\lambda^s, \tilde{\boldsymbol{w}}^s), p_{k,m}^*(\lambda^s, \tilde{\boldsymbol{w}}^s)) \right]$.

The practical challenge of the aforementioned subgradient method is that a perfect knowledge of the CSI is required to compute the expected values of the power and rates. Moreover, even though there exist methods to estimate the CSI (Ross, 2006), the procedure of computing the expected power and rates is computationally expensive. Taking into consideration of the practical challenge, an alternative stochastic subgradient method as in (Wong and Evans, 2008a) and (Wang and Giannakis, 2011) is applied to derive the solution where the updates of the dual variables, (5.12) and (5.14), are performed, for each time slot, across time and the subgradients (5.13) and

(5.15) are replaced by their stochastic approximation, i.e.,

$$\begin{cases} \lambda[t+1] = [\lambda[t] - h_{\lambda}[t] g_{\lambda}[t]]^{+} \\ \tilde{\boldsymbol{w}}[t+1] = ([\tilde{\boldsymbol{w}}[t] - h_{\tilde{\boldsymbol{w}}}[t] g_{\tilde{\boldsymbol{w}}}[t]]^{+})/(\tilde{\boldsymbol{\phi}}^{T}[\tilde{\boldsymbol{w}}[t] - h_{\tilde{\boldsymbol{w}}}[t] g_{\tilde{\boldsymbol{w}}}[t]]^{+}) \\ g_{\lambda}[t] = \bar{P} - m_{sc} \sum_{m=1}^{M} p_{k,m}^{*}(\lambda[t], \tilde{\boldsymbol{w}}[t]) \\ g_{\tilde{\boldsymbol{w}}}[t] = \boldsymbol{r}^{*}[t] - \tilde{\boldsymbol{\phi}}^{T} \|\boldsymbol{r}^{*}[t]\|_{1} \end{cases} \tag{5.16}$$

where we have $\boldsymbol{r}^{*}[t] = \{r_1^{*}[t], r_2^{*}[t], \cdots, r_K^{*}[t]\}$, $r_k^{*}[t] = m_{sc} \sum_{m=1}^{M} r_{k,m}^{*}[t]$ and $r_{k,m}^{*}[t] = r_{k,m}[t](\tau_{k,m}^{*}(\lambda[t], \tilde{\boldsymbol{w}}[t]), p_{k,m}^{*}(\lambda[t], \tilde{\boldsymbol{w}}[t]))$. Based on the stochastic subgradient method, at the ith iteration of the ILA algorithm, the MAC performs n_i iterations to find the optimal Lagrangian multiplier vector $\tilde{\boldsymbol{w}}^{(i)}$, and the associated optimal rate vector $\tilde{\boldsymbol{R}}^{(i)}$, which will be forwarded to the APP layers.

5.5 APP Layer Subproblem: Source Adaptation

Problem (5.9) can be expressed as the following constraint-satisfaction problem aiming to find $\boldsymbol{F}$ such that:

$$\begin{cases} \tilde{\boldsymbol{w}}^{T}(V\boldsymbol{F} - \tilde{\boldsymbol{R}}) = 0 \\ \boldsymbol{F}_{min} \preceq \boldsymbol{F} \preceq \boldsymbol{F}_{max} \\ \triangle(D_i, D_j) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j. \end{cases} \tag{5.17}$$

In (Cicalò and Tralli, 2014), the authors proposed an algorithm that requires in the worst case a maximum number of $K(K-1)/2$ iterations to find the value of $\boldsymbol{F}$ that solves problem (5.17). At each iteration, one numerical search is performed to

obtain a distortion value D to be assigned to a set of videos, following with at most K rate evaluations base on model (4.3).

Chapter 6

PSNR-fair Cross-layer Resource Allocation - A Faster Application Layer Algorithm

One drawback of the algorithm used to solve the source adaptation problem in (Cicalò and Tralli, 2014) is the need to perform, in the worst case, $K(K-1)/2$ numerical searches. If K becomes large, the algorithm becomes unnecessarily complex since the required complexity can actually be reduced. To reduce the complexity, we propose a faster algorithm where the number of numerical searches does not depend on K. Specifically, the algorithm requires only one numerical search to find the solution. To this end, we first formulate a joint resource allocation and source adaptation (JRASA) problem with the aim of maximizing the sum of the PSNRs while minimizing the PSNR difference among the received videos. The JRASA problem is equivalent to the cross-layer resource allocation problem in (5.1), thus, it can be similarly decomposed into a resource allocation problem at the MAC layer and a source

adaptation problem at the APP layer. Therefore, the ILA algorithm can be used to obtain the global solution. But differently, we will consider the R-D model given in (4.4) and quality fairness in terms of PSNR fairness in the JRASA problem. However, since the relationship between PSNR and distortion can be described by the bijective function (4.1), the methods and algorithms developed here can be easily extended to optimization problems with distortion fairness constraints. Moreover, we express differently the fairness constraints by introducing into the JRASA problem a variable q that determines the target PSNR values to be assigned to all videos. As we will see later, this way of expressing the fairness constraints allows us to develop a faster algorithm to the source adaptation problem.

6.1 The Joint Resource Allocation and Source Adaptation Problem

From a fairness perspective, given an arbitrary $q \in \mathbb{R}^+$ representing a quality level in terms of PSNR, we are interested in assigning to each video a PSNR value determined by q such that the PSNR difference between any two videos is minimized. To this end, let us define for now the following function $\hat{Q}_k(q)$ of a continuous variable q:

$$\hat{Q}_k(q) = \begin{cases} Q_{k,min} & q \leq Q_{k,min} \quad (6.1a) \\ q & Q_{k,min} < q < Q_{k,max} \quad (6.1b) \\ Q_{k,max} & q \geq Q_{k,max} \ . \quad (6.1c) \end{cases}$$

The function $\hat{Q}_k(q)$ maps an arbitrary $q \in \mathbb{R}^+$ to an achievable PSNR of the kth video. If the PSNR difference between two videos, i.e., $\triangle(Q_i, Q_j)$, is defined similarly

to $\triangle(D_i, D_j)$ in (5.2), it is clear from eq. (6.1) that given an arbitrary $q \in \mathbb{R}^+$, the PSNR difference between any two videos is always equal to zero, i.e., $\triangle(Q_i, Q_j) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j$, where $Q_k = \hat{Q}_k(q)$ and $Q_j = \hat{Q}_j(q)$.

The JRASA problem can then be described by the following constrained PSNR maximization:

$$\max_{q \geq 0} \quad q \tag{6.2a}$$

$$s.t. \quad (\boldsymbol{\tau}, \boldsymbol{p}) \in \mathcal{A} \tag{6.2b}$$

$$R_k(\boldsymbol{\tau}, \boldsymbol{p}) = VF_k(\hat{Q}_k(q)), \forall k \in \mathcal{K} \tag{6.2c}$$

$$V\boldsymbol{F}_{min} \preceq \boldsymbol{R}(\boldsymbol{\tau}, \boldsymbol{p}) \preceq V\boldsymbol{F}_{max}. \tag{6.2d}$$

According to (6.2c), the feasible solutions lie on a one-dimensional monotonically increasing manifold in the $\mathbb{R}^K$ space described by the following set:

$$\mathcal{R}_f^c = \{\boldsymbol{R} \mid R_k = VF_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}, \forall q \geq 0\}. \tag{6.3}$$

Since $F_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}$ is a nondecreasing function of q, the sum of the rates will be maximized if q is maximized. Therefore, it is clear that the JRASA problem is equivalent to problem (5.1). Moreover, maximizing q is equivalent to maximizing the sum of the PSNR values. Similar to problem (5.1), the optimal rate solution $\boldsymbol{R}^*$ associated with the optimal allocation policies $(\boldsymbol{\tau}^*, \boldsymbol{p}^*)$ to problem (6.2) is attained on the boundary $\mathbf{bd}\,\mathcal{R}$. The problem (6.2) can be decomposed into a resource allocation problem at the MAC layer as given in (5.6) and a source adaptation problem at the APP layer that aims to find q and $\boldsymbol{F}(q) = [F_1(\hat{Q}_1(q)), F_2(\hat{Q}_2(q)), \cdots, F_K(\hat{Q}_K(q))]^T$

such that :

$$V\boldsymbol{F}(q) \in \mathcal{R}_f^c \cap \mathbf{bd}\,\mathcal{R}. \tag{6.4}$$

Interestingly, there is no explicit fairness constraints in problem (6.4), but the fairness can be always achieved if the rate solution is found through a rate adaptation according to q.

6.2 APP Layer Subproblem: A Faster Application Layer Algorithm

Given the information of the $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ to the rate region $\mathcal{R}$ at $\tilde{\boldsymbol{R}}$, the problem in (6.4) can be approximated by the following constraint-satisfaction problem that aims to find q and $\boldsymbol{F}(q)$ such that:

$$\begin{cases} \tilde{\boldsymbol{w}}^T(V\boldsymbol{F}(q) - \tilde{\boldsymbol{R}}) = 0 \\ \boldsymbol{F}_{min} \preceq \boldsymbol{F}(q) \preceq \boldsymbol{F}_{max}. \end{cases} \tag{6.5}$$

Let $\boldsymbol{Q}_{max} = [Q_{1,max}, Q_{2,max}, \cdots, Q_{K,max}]^T$ and $\boldsymbol{Q}_{min} = [Q_{1,min}, Q_{2,min}, \cdots, Q_{K,min}]^T$ be the vectors of the maximum and minimum PSNRs of the video streams, respectively. Then, we denote with Q_{max}^{all} and Q_{min}^{all} the largest element of $\boldsymbol{Q}_{max}$ and the smallest element of $\boldsymbol{Q}_{min}$, respectively. Furthermore, let us define a function:

$$\Gamma(\boldsymbol{F}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = \sum_{k=1}^{K} V\tilde{w}_k F_k - \sum_{k=1}^{K} \tilde{w}_k \tilde{R}_k. \tag{6.6}$$

According to eq. (4.4) and eq. (6.1), the rate $F_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}$, is a nondecreasing function of q. As a result, the function $\Gamma(\boldsymbol{F}(q), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = \sum_{k=1}^{K} V\tilde{w}_k F_k(\hat{Q}_k(q)) - \sum_{k=1}^{K} \tilde{w}_k \tilde{R}_k$ is also a nondecreasing function of q. Therefore, we can apply the bisection search method to find q^* such that $\Gamma(\boldsymbol{F}(q^*), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$ and obtain the solution $\boldsymbol{F}^*(q^*)$ to the source adaptation problem. We summarize the pseudocode of the bisection search-based source adaptation algorithm in **Algorithm 2** below.

Algorithm 2 Source adaptation algorithm to solve problem (6.5)

if $\Gamma(\boldsymbol{F}_{min}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
  report infeasibility and terminate the ILA algorithm
else if $\Gamma(\boldsymbol{F}_{max}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$ **then**
  report infeasibility, set $\boldsymbol{F}^* = \boldsymbol{F}_{max}$ and terminate the algorithm
else
  Initialize: $low = Q_{min}^{all}$; $high = Q_{max}^{all}$; set tolerance e_{bs};
  while $(high - low)/2 > e_{bs}$ **do**
    $q^* = (high + low)/2$;
    for all $k \in \mathcal{K}$ **do**
      if $q^* \leq Q_{k,min}$ **then**
        $Q_k^* = Q_{k,min}$; $F_k^* = F_{k,min}$;
      else if $q^* \geq Q_{k,max}$ **then**
        $Q_k^* = Q_{k,max}$; $F_k^* = F_{k,max}$;
      else
        $Q_k^* = q^*$; $F_k^* = F_k(Q_k^*)$, based on model (4.4);
      end if
    end for
    if $\Gamma(\boldsymbol{F}^*(q^*), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ **then**
      $low = q^*$;
    else if $\Gamma(\boldsymbol{F}^*(q^*), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
      $high = q^*$;
    else
      break
    end if
  end while
end if

The algorithm first checks two feasibility conditions $\Gamma(\boldsymbol{F}_{min}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ and

$\Gamma(\boldsymbol{F}_{max}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \geq 0$ which are the relaxations of $V\boldsymbol{F}_{min} \in \mathcal{R}$ and $V\boldsymbol{F}_{max} \notin \mathcal{R}$, respectively. For each iteration of the ILA algorithm, if the first condition is violated, the ILA algorithm will be terminated by the APP layer because any source adaptation is impossible. On the other hand, if the second condition is violated, the ILA algorithm will continue, but the source adaptation algorithm terminates unsuccessfully and the source adaptation solution will be $\boldsymbol{F}^* = \boldsymbol{F}_{max}$ at the current iteration. The bisection search procedure from line 7 to line 25 strives to find the optimal q^* and the optimal rate $\boldsymbol{F}^*(q^*)$ such that $\Gamma(\boldsymbol{F}^*(q^*), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$.

To find the optimal solution, **Algorithm 2** requires only one numerical search, i.e., bisection search, and is thus considerably faster in comparison to the source adaptation algorithm proposed in (Cicalò and Tralli, 2014), which requires in the worst case a maximum number of $K(K-1)/2$ numerical searches to obtain the optimal distortion and rate solutions. The low-complexity of **Algorithm 2** comes from the introduction of the function in (6.1). This is because it enables us to express $\Gamma(\boldsymbol{F}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ as a nondecreasing function of the quality variable q. As a result, the solution to $\Gamma(\boldsymbol{F}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$ can be efficiently found through the bisection search method.

The optimal PSNR and rate solutions, given q^*, are given as follow:

$$Q_k^* = \begin{cases} Q_{k,min} & q^* \leq Q_{k,min} \\ q^* & Q_{k,min} < q^* < Q_{k,max} \\ Q_{k,max} & q^* \geq Q_{k,max} \end{cases} \tag{6.7}$$

and

$$F_k^* = \begin{cases} F_{k,min} & q^* \leq Q_{k,min} \\ \frac{\theta_k}{255^2 10^{-Q_k^*/10} + \alpha_k} + \beta_k & Q_{k,min} < q^* < Q_{k,max} \\ F_{k,max} & q^* \geq Q_{k,max}. \end{cases} \tag{6.8}$$

The optimality of solutions (6.7) and (6.8) can be easily proved by noting that the PSNR difference between any two videos is always equal to zero, i.e., $\triangle(Q_i^*, Q_j^*) = 0$, and the optimal rate vector $\boldsymbol{F}^*(q^*)$ satisfies that $\Gamma(\boldsymbol{F}^*(q^*), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$.

Chapter 7

PSNR-fair Cross-layer Resource Allocation - A Benchmark Scheme

As shown in Chapters 5 and 6, the use of R-D models enables us to predict the minimum required rate to achieve a target video quality, and thus facilitates the optimal allocation of system resources among users such that the optimal rates and target qualities are achieved. Therefore, the accuracy of the R-D models has a direct impact on the performance of the video transmission systems that use them. Among the three types of R-D models, empirical R-D models are the most accurate since they are constructed using all the empirical R-D points. The high accuracy provided by empirical models motivates us to use an empirical model in our optimization problem.

In this chapter, we show that the cross-layer optimization method presented in Chapter 6 is still applicable in the case of source adaptation using a discrete empirical R-D model. To this end, the JRASA problem in (6.2) is reconsidered, but the source adaptation is performed based on the empirical R-D model in (4.2). Moreover, we present and discuss an algorithm to solve the source adaptation problem. Due to its

high accuracy, the optimal solution obtained based on the empirical model can be used as a benchmark for comparing the results produced using the semi-analytical model in (4.4).

7.1 The Joint Resource Allocation and Source Adaptation Problem

To facilitate the development of a low-complexity source adaptation algorithm, let us first introduce the following step function with a continuous variable $q \geq 0$ that represents a quality level in terms of PSNR:

$$\hat{Q}_k(q) = \sum_{c=1}^{C_k} q_{k,c} \chi_{I_{k,c}}(q),\ q \in \mathbb{R}^+ \tag{7.1}$$

where $I_{k,c}$ are PSNR intervals defined as:

$$I_{k,c} = \begin{cases} [0, q_{k,c}] & c = 1 \\ (q_{k,c-1}, q_{k,c}] & c = 2, \cdots, C_k - 1 \\ [q_{k,c}, +\infty) & c = C_k \end{cases} \tag{7.2}$$

and $\chi_I(q)$ is the indicator function of interval I:

$$\chi_I(q) = \begin{cases} 1 & q \in I \\ 0 & q \notin I. \end{cases} \tag{7.3}$$

Given an arbitrary $q \in [0, \infty)$, the function (7.1) maps q to the smallest PSNR value in $\mathcal{Q}_k$ that is larger than q or to q_{k,C_k} if $q \geq q_{k,C_k}$. With eq. (7.1), the R-D model (4.2) can be considered as a nondecreasing function of a continuous variable q, i.e., $F_k(\hat{Q}_k(q))$.

Let us denote with $\triangle(Q_i, Q_j), \forall i \neq j$ the achieved PSNR difference between any two received videos. Ideal PSNR fairness among users will require that $\triangle(Q_i, Q_j) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j$. However, it is unlikely to achieve ideal fairness because that (i) the PSNR of each video has maximum and minimum values and (ii) the R-D model in (4.2) is a discrete function. Taking these facts into consideration, the PSNR difference is defined as follows:

$$\triangle(Q_i, Q_j) = \begin{cases} 0 & (Q_i, Q_j) \in \mathbb{Q} \vee (Q_j, Q_i) \in \mathbb{Q} \quad (7.4a) \\ 0 & (Q_i, Q_j) \in \mathbb{F} \vee (Q_j, Q_i) \in \mathbb{F} \quad (7.4b) \\ |Q_i - Q_j| & \text{otherwise} \quad (7.4c) \end{cases}$$

where $\mathbb{Q} = \{(Q_i, Q_j) \mid (Q_i = Q_{i,max} \wedge Q_j > Q_i) \vee (Q_i = Q_{i,min} \wedge Q_j < Q_i)\}$ and $\mathbb{F} = \{(Q_i, Q_j) \mid (Q_j \geq Q_i = q_{i,c}) \wedge (Q_j < q_{i,c+1}), \forall c = 1, 2, \cdots, C_i - 1\}$. The case in (7.4a) takes into consideration the maximum and minimum PSNR constraints. The case in (7.4b) considers the discrete nature of R-D function. Given that $Q_j \geq Q_i$ and $Q_i = q_{i,c}$, if we have $Q_j < q_{i,c+1}$, the difference between Q_i and Q_j has already achieved its minimum value. In this case, the PSNR difference is set to zero.

Based on (4.2) and (7.1), the joint resource allocation and source adaptation

problem is formulated as the following constrained PSNR maximization:

$$\max_{q \geq 0} \quad q \tag{7.5a}$$

$$s.t. \quad (\boldsymbol{\tau}, \boldsymbol{p}) \in \mathcal{A} \tag{7.5b}$$

$$R_k(\boldsymbol{\tau}, \boldsymbol{p}) = VF_k(\hat{Q}_k(q)), \forall k \in \mathcal{K} \tag{7.5c}$$

$$V\boldsymbol{F}_{min} \preceq \boldsymbol{R}(\boldsymbol{\tau}, \boldsymbol{p}) \preceq V\boldsymbol{F}_{max} \tag{7.5d}$$

where $q \in [0, \infty)$ determines the target PSNR values to be assigned to all videos.

The constraint in (7.5c) restricts the feasible solutions to a discrete set of rate vectors that achieve fairness among users, i.e., the set is defined as:

$$\mathcal{R}_f = \{\boldsymbol{R} \mid R_k = VF_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}, \forall q \geq 0\}. \tag{7.6}$$

Clearly, $\forall \boldsymbol{R} \in \mathcal{R}_f$, the achieved PSNR difference between any two videos is zeros, i.e., $(Q_i, Q_j) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j$.

Since $F_k(\hat{Q}_k(q)), \forall k \in \mathcal{K}$ is a nondecreasing function of q, the optimal rate solution $\boldsymbol{R}^*$ to problem (7.5) must be the largest rate vector in the set $\mathcal{R}_f \cap \mathcal{R}_a$. In Fig. 7.1, we show an example of the optimization problem for two users where the optimal solution $\boldsymbol{R}^*$ is the asterisk marked in red. For the following analysis, we will refer to the rate vectors in the set $\mathcal{R}_f$ as fair rate vectors. Moreover, we say that two fair rate vectors $\boldsymbol{R}'$ and $\boldsymbol{R}''$ are adjacent when $\boldsymbol{R}''$ is the smallest fair rate vector that is larger than $\boldsymbol{R}'$ or $\boldsymbol{R}''$ is the largest fair rate vector that is smaller than $\boldsymbol{R}'$.

To exploit the ILA algorithm to solve problem (7.5), let us give the following proposition whose proof is reported in Appendix A.

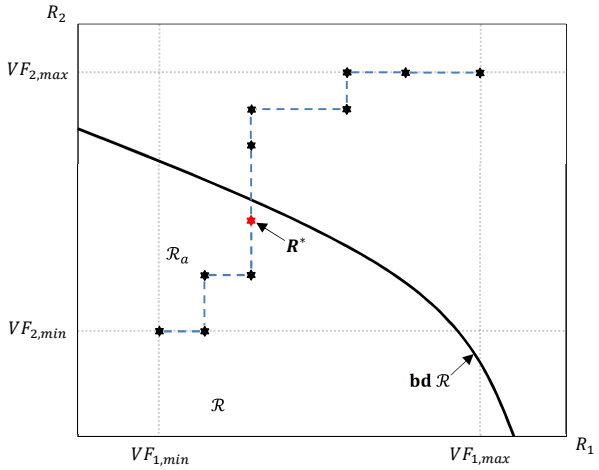


Figure 7.1: An example of the optimization problem for a system with two users. The asterisks represent rate vectors satisfying the fairness constraint $\triangle(Q_1, Q_2) = 0$ and every two adjacent asterisks are connected by a dash line.

Proposition 1. *The piecewise curve $\mathcal{C}_{pw}$, formed by all the line segments bounded by any two adjacent fair rate vectors, is a one-dimensional monotonically increasing manifold with boundary in the $\mathbb{R}^K$ space, i.e.,*

$$\mathcal{C}_{pw} = \bigcup_{(\boldsymbol{R},\boldsymbol{R}')\in\mathcal{R}_{ad}} \{\boldsymbol{R} + t(\boldsymbol{R}' - \boldsymbol{R}) \mid t \in [0,1]\} \tag{7.7}$$

where the set $\mathcal{R}_{ad}$ is defined as:

$$\mathcal{R}_{ad} = \{(\boldsymbol{R}, \boldsymbol{R}') \mid \boldsymbol{R}, \boldsymbol{R}' \in \mathcal{R}_f \wedge \boldsymbol{R} \preccurlyeq \boldsymbol{R}' \wedge (\boldsymbol{R} \text{ and } \boldsymbol{R}' \text{ are adjacent})\}. \tag{7.8}$$

Due to its one-dimensionality and monotonicity, the piecewise curve $\mathcal{C}_{pw}$ intersects with the boundary $\mathbf{bd}\,\mathcal{R}$ of the rate region $\mathcal{R}$ at a unique rate point $\boldsymbol{R}_{int}$. Clearly, the optimal solution $\boldsymbol{R}^*$ to problem (7.5) is the largest fair rate vector that is smaller than $\boldsymbol{R}_{int}$ or $\boldsymbol{R}_{int}$ itself. Therefore, the first step to obtain $\boldsymbol{R}^*$ is to find $\boldsymbol{R}_{int}$. If we have the information about the line where the intersection point lies, which is

identified by $\boldsymbol{R}_{int} = \boldsymbol{\phi} r$, problem (7.5) can be simplified into a resource allocation problem at the MAC layer as shown in (5.6). The optimal solution $\boldsymbol{R}^*$ to problem (7.5) can then be evaluated from the solution $\boldsymbol{R}_{int}$ to the resource allocation problem. On the other hand, if we have the information of the boundary $\mathbf{bd}\,\mathcal{R}$, problem (7.5) can be simplified into a source adaptation problem at the APP layer that aims to find $\boldsymbol{F}$ such that:

$$V\boldsymbol{F} = \mathbf{bd}\,\mathcal{R} \cap \mathcal{C}_{pw} \tag{7.9}$$

and then $\boldsymbol{R}^*$ is evaluated from the $V\boldsymbol{F}$. By exploiting the $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ to the rate region $\mathcal{R}$ at $\tilde{\boldsymbol{R}}$ as the local approximation of the boundary $\mathbf{bd}\,\mathcal{R}$, the ILA algorithm discussed in Section (5.3) can be applied to obtain the intersection point $\boldsymbol{R}_{int}$ and accordingly the optimal solution $\boldsymbol{R}^*$. In Fig. (7.2), we show an example of the first step of the ILA, for two users, used to solve problem (7.5).

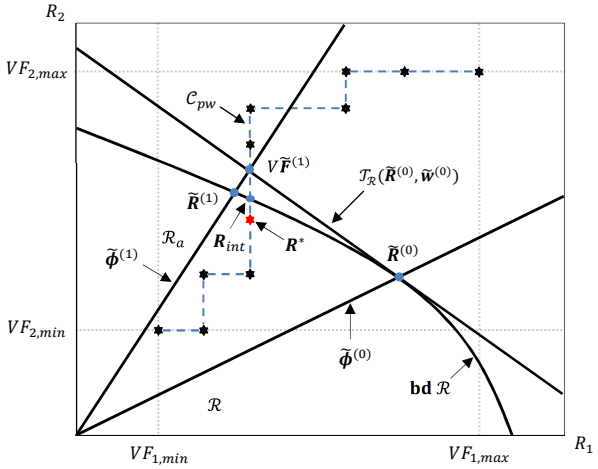


Figure 7.2: An example of first step of the ILA algorithm, for a two-user case, to find $\boldsymbol{R}_{int}$.

7.2 APP Layer Subproblem: Source Adaptation with An Empirical R-D Model

Given the information of the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$, the problem (7.9) at the APP layer is reformulated as the following constraint-satisfaction problem where the objective is to find $\boldsymbol{F}$ such that:

$$\begin{cases} \Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0 \\ V\boldsymbol{F} \in \mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'') \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \\ \boldsymbol{F}_{min} \preceq \boldsymbol{F}' \preccurlyeq \boldsymbol{F}'' \preceq \boldsymbol{F}_{max} \\ (V\boldsymbol{F}', V\boldsymbol{F}'') \in \mathcal{R}_{ad} \end{cases} \tag{7.10}$$

where $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'') = \{V\boldsymbol{F}' + t(V\boldsymbol{F}'' - V\boldsymbol{F}') \mid t \in [0, 1]\}$ is the line segment bounded by $V\boldsymbol{F}'$ and $V\boldsymbol{F}''$.

At each iteration of the ILA algorithm, the APP layer exploits the information of the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ to derive $\tilde{\boldsymbol{F}}$ that solves problem (7.10). The rate vector $V\tilde{\boldsymbol{F}}$ is an intersection point of the tangent space with the line segment $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$ bounded by two adjacent fairness rate vectors $V\boldsymbol{F}'$ and $V\boldsymbol{F}''$ satisfying the relations: $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$ and $\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \geq 0$, respectively. The resulting direction vector $\tilde{\boldsymbol{\phi}}$, computed from $\tilde{\boldsymbol{F}}/\|\tilde{\boldsymbol{F}}\|_1$, is then forwarded to the MAC layer, which projects the solution on the boundary of $\mathcal{R}$ by solving the problem (5.6) to get $\tilde{\boldsymbol{R}}$ and the corresponding weight vector $\tilde{\boldsymbol{w}}$.

Based on the bisection search method, we develop an algorithm to find $\boldsymbol{F}$ that solves problem (7.10). The pseudocode of the algorithm is reported in **Algorithm 3** below, whereas the details are discussed in Appendix B.

Algorithm 3 Source adaptation algorithm to solve problem (7.10)

if $\Gamma(\boldsymbol{F}_{min}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
  report infeasibility and terminate the ILA algorithm
else if $\Gamma(\boldsymbol{F}_{max}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$ **then**
  report infeasibility, set $\boldsymbol{F} = \boldsymbol{F}_{max}$ and terminate the algorithm
else
  Initialize: $\boldsymbol{F}' = \boldsymbol{F}_{min}$; $\boldsymbol{F}'' = \boldsymbol{F}_{max}$; $low = Q_{min}^{all}$; $high = Q_{max}^{all}$; $cond_{EQ}$ = false; $cond_{END}$ = false; set tolerance e_{bs}
  while $(high - low)/2 > e_{bs}$ **and** $cond_{END} ==$ false **do**
    $q = (high + low)/2$;
    Compute a fairness rate vector: $V\boldsymbol{F}(q)$ [i];
    if $\Gamma(\boldsymbol{F}(q), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ **then**
      Find $\boldsymbol{F}_r$ such that $(V\boldsymbol{F}(q), V\boldsymbol{F}_r) \in \mathcal{R}_{ad}$ [ii];
      if $\Gamma(\boldsymbol{F}_r, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ **then**
        $low = q$; $\boldsymbol{F}' = \boldsymbol{F}_r$;
      else if $\Gamma(\boldsymbol{F}_r, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
        $\boldsymbol{F}' = \boldsymbol{F}(q)$; $\boldsymbol{F}'' = \boldsymbol{F}_r$; $cond_{END}$ = true;
      else
        $\boldsymbol{F} = \boldsymbol{F}_r$; $cond_{EQ} = cond_{END}$ = true;
      end if
    else if $\Gamma(\boldsymbol{F}(q), \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
      Find $\boldsymbol{F}_l$ such that $(V\boldsymbol{F}_l, V\boldsymbol{F}(q)) \in \mathcal{R}_{ad}$ [iii];
      if $\Gamma(\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
        $high = q$; $\boldsymbol{F}'' = \boldsymbol{F}_l$;
      else if $\Gamma(\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ **then**
        $\boldsymbol{F}' = \boldsymbol{F}_l$; $\boldsymbol{F}'' = \boldsymbol{F}(q)$; $cond_{END}$ = true;
      else
        $\boldsymbol{F} = \boldsymbol{F}_l$; $cond_{EQ} = cond_{END}$ = true;
      end if
    else
      $\boldsymbol{F} = \boldsymbol{F}(q)$; $cond_{EQ}$ = true;
    end if
  end while
  if $cond_{EQ} ==$ false **and** $cond_{END} ==$ true **then**
    $V\boldsymbol{F} = \mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'') \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ [iv];
  end if
  if $cond_{EQ} ==$ false **and** $cond_{END} ==$ false **then**
    Linear Search: Find $\boldsymbol{F}'$ and $\boldsymbol{F}''$ such that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ [v];
    $V\boldsymbol{F} = \mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'') \cap \mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$;
  end if
end if

It is worth noting that 5 simple auxiliary algorithms, reported and explained in Appendix B, are derived to (i) compute a fair rate vector $V\boldsymbol{F}(q)$ (line 9), (ii) find a larger adjacent fair rate vector of $V\boldsymbol{F}(q)$ (line 11), (iii) find a small adjacent fair rate vector of $V\boldsymbol{F}(q)$ (line 20), (iv) find an intersection of the line segment $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$ and the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$ (line 33), and (v) carry out a linear search procedure to find adjacent $V\boldsymbol{F}'$ and $V\boldsymbol{F}''$ such that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ (line 36) in **Algorithm 3**.

For the case of error-free transmission, **Algorithm 3**, in the worst case, runs in $O\left(KC_{LN} + K\log_2[(Q_{max}^{all} - Q_{min}^{all})/e_{bs} + C_{max}^{all}]\right)$ time where $C_{max}^{all} = \max_k(C_k)$ and C_{LN} is the number of searches of the linear search procedure. For the case of error-prone transmission, the worst-case running time complexity of **Algorithm 3** is $O(\gamma KLIC_{LN} + \gamma KLI\log_2[(Q_{max}^{all} - Q_{min}^{all})/e_{bs} + L_{max}])$.

Chapter 8

Adjustable PSNR-fair Cross-layer Resource Allocation

In Chapters 5, 6 and 7, the objective of the discussed optimization frameworks for SVC video transmission is to maximize the sum of the average rates while minimizing the PSNR difference among the received videos. Such objective usually requires that most of the available resources be assigned to the users that experience bad channel conditions and request high-complexity videos such that the goal of keeping all the received video qualities at the same level is achieved. However, such objective also leaves many users without a chance to fully utilize the system resources and adversely affect the system efficiency in terms of the overall received video quality. This conflict between fairness and efficiency motivates us to develop an optimization framework to address the trade-offs between fairness and efficiency. In this chapter, we propose a cross-layer optimization framework for SVC video transmission in OFDMA wireless networks. The objective of the optimization is to maximize the overall received video PSNR while limiting the PSNR difference among the received videos within an

acceptable and adjustable range. To limit the PSNR difference within an acceptable range, a common target PSNR value is chosen and it is required that the absolute value of the relative difference between the target PSNR and achieved PSNR of each video be bounded from above by a nonnegative scalar. The confinement of the PSNR difference within an adjustable range can be achieved by varying the value of the scalar. The larger the scalar is, the larger the range and the looser the fairness constraints will be. As a result, varying the value of the scalar, according to application requirements, allows us to achieve the trade-offs between fairness and efficiency. In comparison to the framework proposed in (Su *et al.*, 2006) where only a limited number of trade-off points can be achieved, our framework supports an infinite number of trade-off points.

8.1 The Optimization Problem

The starting point for formulating the optimization problem is to find a reasonable common target PSNR value $\bar{Q}$. Even though the framework we are going to develop holds for a wide range of common target PSNR values, for the following analysis, $\bar{Q}$ is given as the optimal quality level obtained in the JRASA problem (6.2), i.e., $\bar{Q} = q^*$. Given the common target PSNR value $\bar{Q}$ and a scalar σ that controls the maximum absolute value of the relative difference between $\bar{Q}$ and the achieved PSNR of each video, one possible formulation of the optimization problem is the following

fairness-constrained sum-PSNR maximization:

$$\max_{\boldsymbol{R}\in\mathcal{R}_a} \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.1a}$$

$$s.t. \quad \left|\frac{Q_k(R_k/V)-\bar{Q}}{\bar{Q}}\right| \leq \sigma, \forall k \in \mathcal{K} \tag{8.1b}$$

$$Q_{k,min} \leq Q_k(R_k/V) \leq Q_{k,max}, \forall k \in \mathcal{K} \tag{8.1c}$$

where $Q_k(R_k/V) = F_k^{-1}(R_k/V)$ represents the achieved PSNR and is given as the the inverse function of (4.4):

$$Q_k(R_k/V) = F_k^{-1}(R_k/V) = 20\log_{10}(255) - 10\log_{10}\left(\frac{\theta_k}{R_k/V-\beta_k} - \alpha_k\right). \tag{8.2}$$

The constraints in (8.1b) limit the PSNR difference between any two videos within $2\sigma\bar{Q}$. The achieved PSNR of each video is constrained to be within the range bounded by its minimum and maximum values according to constraint (8.1c). According to constraints (8.1b) and (8.1c), the problem is feasible if and only if $\bar{Q}(1+\sigma) \geq Q_{k,min}$ and $\bar{Q}(1-\sigma) \leq Q_{k,max}, \forall k \in \mathcal{K}$, are satisfied. To keep the feasibility of the problem even when the conditions $\bar{Q}(1+\sigma) \leq Q_{k,min}$ and $\bar{Q}(1-\sigma) \leq Q_{k,max}, \forall k \in \mathcal{K}$, are

violated, we rewrite the constraints, leading to the following optimization problem:

$$\max_{\boldsymbol{R}\in\mathcal{R}_a} \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.3a}$$

$$s.t. \quad Q_k(R_k/V) = Q_{k,min}, \forall k \in \mathcal{K}_1 \tag{8.3b}$$

$$Q_k(R_k/V) = Q_{k,max}, \forall k \in \mathcal{K}_2 \tag{8.3c}$$

$$|Q_k(R_k/V) - \bar{Q}| \leq \bar{Q}\sigma, \forall k \in \mathcal{K} \setminus (\mathcal{K}_1 \cup \mathcal{K}_2) \tag{8.3d}$$

$$Q_{k,min} \leq Q_k(R_k/V) \leq Q_{k,max}, \forall k \in \mathcal{K} \setminus (\mathcal{K}_1 \cup \mathcal{K}_2) \tag{8.3e}$$

where $\mathcal{K}_1 = \{k \in \mathcal{K} \mid \bar{Q}(1+\sigma) < Q_{k,min}\}$ and $\mathcal{K}_2 = \{k \in \mathcal{K} \mid \bar{Q}(1-\sigma) > Q_{k,max}\}$. The equality constraints in (8.3b) and (8.3c) are motivated by following considerations. Ideally, the fairness constraints would require that the PSNR difference between any two videos be within $2\sigma\bar{Q}$. However, if for the k-th video, we have $\bar{Q}(1+\sigma) \leq Q_{k,min}$ or $\bar{Q}(1-\sigma) \geq Q_{k,max}$, its PSNR value will be set to $Q_{k,min}$ (with rate $F_{k,min}$) or $Q_{k,max}$ (with rate $F_{k,max}$). The optimization is then performed over the set of other videos. In this way, the problem is feasible, and the optimization seeks to maximize the sum of the PSNR while guaranteeing that the absolute PSNR difference between any other two videos is within $2\sigma\bar{Q}$ and the PSNR differences between the k-th video and other videos are minimized.

Note that if $\sigma = 0$, problem (8.3) is equivalent to:

$$\max_{\boldsymbol{R}\in\mathcal{R}_a} \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.4a}$$

$$s.t. \quad Q_k(R_k/V) = \hat{Q}_k(q^*), \forall k \in \mathcal{K} \tag{8.4b}$$

where $\hat{Q}_k(q), \forall k \in \mathcal{K}$ is the function in (6.1). Interestingly, the optimal solution to

problem (8.4) is the same as that to problem (6.2) where the PSNR difference between any two videos is zero, i.e., $\triangle(Q_i, Q_j) = 0, \forall i, j \in \mathcal{K} \wedge i \neq j$.

On the other hand, if σ is sufficiently large, problem (8.3) is equivalent to:

$$\max_{\boldsymbol{R}\in\mathcal{R}_a} \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.5a}$$

$$s.t. \quad Q_{k,min} \leq Q_k(R_k/V) \leq Q_{k,max}, \forall k \in \mathcal{K}. \tag{8.5b}$$

The optimization problem in (8.5) seeks to maximize the overall video PSNR under a set of maximum and minimum PSNR constraints. Without considering fairness, such a system efficiency driven optimization could lead to large quality variations among the received videos.

Clearly, the proposed optimization framework enables us to achieve any fairness level ranging from the fairest solution with $\bar{Q} = q^*$ and $\sigma = 0$ to the most unfair but efficiency-maximizing solution with $\sigma \to \infty$. In the next subsection, we will present a method to solve problem (8.3) with any σ and $\bar{Q}$.

8.2 Problem Solution

Translating the fairness constraints in (8.3b) - (8.3e) into rate constraints, the problem (8.3) can be rewritten as:

$$\max \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.6a}$$

$$s.t. \quad \boldsymbol{f}_{min} \preceq \boldsymbol{R}/V \preceq \boldsymbol{f}_{max} \tag{8.6b}$$

$$\boldsymbol{R} \in \mathcal{R} \tag{8.6c}$$

where $\boldsymbol{f}_{min} = [f_{1,min}, f_{2,min}, \cdots, f_{K,min}]^T$ and $\boldsymbol{f}_{max} = [f_{1,max}, f_{2,max}, \cdots, f_{K,max}]^T$, with

$$\begin{cases} f_{k,min} \in \{\max\left(F_{k,min}, F_k(\bar{Q}(1-\sigma))\right), F_{k,min}, F_{k,max}\} \\ f_{k,max} \in \{\min\left(F_{k,max}, F_k(\bar{Q}(1+\sigma))\right), F_{k,min}, F_{k,max}\}. \end{cases}$$

According to the constraints (8.6b) and (8.6c), any feasible solution to problem (8.6) belongs to $\mathcal{R}_{adj} = \{\boldsymbol{R} \in \mathcal{R} \mid V\boldsymbol{f}_{min} \preceq \boldsymbol{R} \preceq V\boldsymbol{f}_{max}\}$, if it is not empty. This can be guaranteed if and only if $V\boldsymbol{f}_{min} \in \mathcal{R}$. Moreover, the problem has a trivial solution if $V\boldsymbol{f}_{max} \in \mathcal{R}$. Since the objective (8.6) is concave (Boyd and Vandenberghe, 2004, Section 3.2.4) and increasing, $\forall \boldsymbol{R} \in \mathcal{R}$, the optimal solution $\boldsymbol{R}^*_{adj}$ is clearly attained at the boundary $\mathbf{bd}\,\mathcal{R}$ under the assumptions $V\boldsymbol{f}_{min} \in \mathcal{R}$ and $V\boldsymbol{f}_{max} \notin \mathcal{R}$. In Fig. 8.1, we draw an example of the optimization problem for a two-user case.

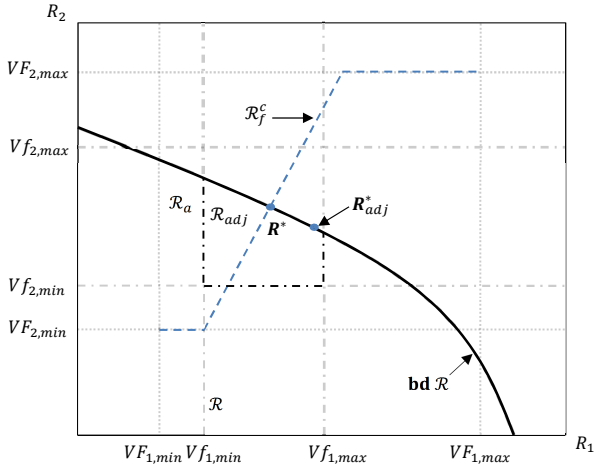


Figure 8.1: An example of two-user optimization problem (8.6). $\boldsymbol{R}^*$ is the optimal solution to the problem when $\sigma = 0$ and $\bar{Q} = q^*$ where q^* is the optimal quality level of problem (6.2), whereas $\boldsymbol{R}^*_{adj}$ is the optimal solution for a general σ and $\bar{Q} = q^*$.

According to the definition of $\mathcal{R}$, the problem (8.6) can be reformulated as:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S},\boldsymbol{f}_{min}\preceq\boldsymbol{R}/V\preceq\boldsymbol{f}_{max}} \quad \sum_{k=1}^{K} Q_k(R_k/V) \tag{8.7a}$$

$$s.t. \quad R_k \leq \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{m=1}^{M} m_{sc} r_{k,m}\left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right)\right], \forall k \in \mathcal{K} \tag{8.7b}$$

$$\mathbb{E}_{\boldsymbol{\gamma}}\left[m_{sc}\sum_{k=1}^{K}\sum_{m=1}^{M} p_{k,m}(\boldsymbol{\gamma})\right] \leq \bar{P}. \tag{8.7c}$$

The problem (8.7) is a strictly feasible convex optimization problem because of the concavity of the objective in (8.7a) and $r_{k,m}$. The solution to problem (8.7) can be found by the Lagrangian dual method, as in (Wang and Giannakis, 2011). We retrace here the main results of (Wang and Giannakis, 2011).

Let $\boldsymbol{\mu}$ be the Lagrangian multiplier vector related to the constraint (8.7b) and λ be the Lagrangian multiplier related to the average power constraint (8.7c), then, the Lagrangian associated with (8.7) is given as:

$$\begin{aligned} L(\boldsymbol{\tau},\boldsymbol{p},\boldsymbol{R},\lambda,\boldsymbol{\mu}) &= \sum_{k=1}^{K} Q_k(R_k/V) + \lambda\left\{\bar{P} - \mathbb{E}_{\boldsymbol{\gamma}}\left[m_{sc}\sum_{k=1}^{K}\sum_{m=1}^{M} p_{k,m}(\boldsymbol{\gamma})\right]\right\} \\ &\quad + \sum_{k=1}^{K}\mu_k\left\{\mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{m=1}^{M} m_{sc} r_{k,m}\left(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})\right)\right] - R_k\right\} \\ &= \sum_{k=1}^{K} Q_k(R_k/V) + \boldsymbol{\mu}^T\boldsymbol{R} \\ &\quad + \lambda\bar{P} + m_{sc}\mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k=1}^{K}\sum_{m=1}^{M}\mu_k r_{k,m}(\tau_{k,m}(\boldsymbol{\gamma}), p_{k,m}(\boldsymbol{\gamma})) - \lambda p_{k,m}(\boldsymbol{\gamma})\right]. \end{aligned} \tag{8.8}$$

The related Lagrangian dual function is given as:

$$\Theta(\lambda,\boldsymbol{\mu}) = \max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S},\boldsymbol{f}_{min}\preceq\boldsymbol{R}/V\preceq\boldsymbol{f}_{max}} \quad L(\boldsymbol{\tau},\boldsymbol{p},\boldsymbol{R},\lambda,\boldsymbol{\mu}) \tag{8.9}$$

and the dual problem associated with (8.7) is $\min_{\lambda>0,\boldsymbol{\mu}\succeq\mathbf{0}} \Theta(\lambda,\boldsymbol{\mu})$. Given λ and $\boldsymbol{\mu}$, $\Theta(\lambda,\boldsymbol{\mu})$ can be derived by solving two decoupled subproblems across $\boldsymbol{R}$ and $(\lambda,\boldsymbol{\mu})$, respectively. The first subproblem is associated with $\boldsymbol{R}$, i.e.,

$$\max_{\boldsymbol{f}_{min}\preceq\boldsymbol{R}/V\preceq\boldsymbol{f}_{max}} \quad Q_k(R_k/V)+\boldsymbol{\mu}^T\boldsymbol{R} \tag{8.10}$$

which is a convex optimization problem where efficient algorithms are available to find the solution $\boldsymbol{R}^*(\boldsymbol{\mu})$. The second subproblem is related to $(\boldsymbol{\tau},\boldsymbol{p})$ and given as:

$$\max_{(\boldsymbol{\tau},\boldsymbol{p})\in\mathcal{S}} \lambda\bar{P}+m_{sc}\mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{k=1}^{K}\sum_{m=1}^{M}\mu_k r_{k,m}(\tau_{k,m}(\boldsymbol{\gamma}),p_{k,m}(\boldsymbol{\gamma}))-\lambda p_{k,m}(\boldsymbol{\gamma})\right] \tag{8.11}$$

which is the same as the dual function associated with (3.8) with $\boldsymbol{\mu}\equiv\boldsymbol{w}$. Therefore, the solution $\boldsymbol{\tau}^*(\lambda,\boldsymbol{\mu},\boldsymbol{\gamma})$ and $\boldsymbol{p}^*(\lambda,\boldsymbol{\mu},\boldsymbol{\gamma})$ are given by (3.11) and (3.13).

Since (8.7) is convex and Slater's condition holds, the duality gap between the primal and dual problems is zero. Therefore, replacing λ and $\boldsymbol{\mu}$ with the optimal dual variables λ^* and $\boldsymbol{\mu}^*$ provides the almost surely optimal resource allocation policy $\boldsymbol{\tau}^*(\lambda^*,\boldsymbol{\mu}^*,\boldsymbol{\gamma})$ and $\boldsymbol{p}^*(\lambda^*,\boldsymbol{\mu}^*,\boldsymbol{\gamma})$ and the corresponding optimal rate vector $\boldsymbol{R}^*(\boldsymbol{\mu}^*)$ which is a boundary point of the rate region $\mathcal{R}$. The optimal λ^* and $\boldsymbol{\mu}^*$ can be obtained through the method of stochastic subgradient iterations, as shown in (8.12),

$$\begin{cases}\lambda[t+1]=\lambda[t]+\delta[t]\left(m_{sc}\sum_{m=1}^{M}p^*_{k,m}(\lambda[t],\boldsymbol{\mu}[t],\boldsymbol{\gamma}[t])-\bar{P}\right)\\ \mu_k[t+1]=\mu_k[t]+\delta[t]\left(R^*_k(\boldsymbol{\mu}[t])-m_{sc}\sum_{m=1}^{M}r^*_{k,m}(\lambda[t],\boldsymbol{\mu}[t],\boldsymbol{\gamma}[t])\right)\end{cases} \tag{8.12}$$

where $r^*_{k,m}(\lambda[t],\boldsymbol{\mu}[t],\boldsymbol{\gamma}[t]) = r_{k,m}\left(\tau^*_{k,m}(\lambda[t],\boldsymbol{\mu}[t],\boldsymbol{\gamma}[t]),p^*_{k,m}(\lambda[t],\boldsymbol{\mu}[t],\boldsymbol{\gamma}[t])\right)$. Starting from any intial $\lambda>0$ and $\boldsymbol{\mu}\succ\mathbf{0}$, the iterations in (8.12) converge to the optimal λ^*

and $\boldsymbol{\mu}^*$.

Finally, it should be pointed out that the optimal solution $\boldsymbol{R}^*_{adj} = \boldsymbol{R}^*(\boldsymbol{\mu}^*)$ may not be achievable since the available SVC encoding schemes support only a discrete set of rate values. Following the common practice, the optimal discrete solution is obtained by extracting the largest achievable rate which is smaller than $\boldsymbol{R}^*_{adj}$.

Chapter 9

Numerical Results

In this chapter, we evaluate the performance of the proposed optimization frameworks. We consider an OFDMA WLAN with $K = 6$ users, $M = 112$ subchannels unless otherwise stated, a time slot duration $t_{slot} = 0.5$ ms and a total average power $\bar{P} = 1$ W. Supposing without loss of generality that each subchannel consists of only one subcarrier, i.e., $m_{sc} = 1$, and the bandwidth of the subcarrier is 15 kHz. The Rayleigh fading channels between the BS and each user are simulated using the ITU Vehicular Channel A model (SMG, 1997) which has a root mean square delay spread $\tau_{rms} = 0.37$ μs and 50% coherence bandwidth of $B_c = 1/(5\tau_{rms}) \approx 540$ kHz. The average normalized SNRs for all users are assumed to be 25 dBW. The modulation and coding scheme adopted at the PHY layer is characterized by a *rate adjustment* $a_1 = 0.905$ and *SNR gap* $a_2 = 1.34$ (Mazzotti *et al.*, 2012). We encode six 160-frame videos, one for each user, with different spatial-temporal complexities, i.e., Foreman, Ice, Soccer, Crew, Football and Mobile, in CIF resolution with a frame-rate of 30 frames per second. Each sequence is encoded IDR-period-by-IDR-period by the JSVM reference software (Reichel *et al.*, 2007) with the GOP size and IDR

period set to 8 and 16 frames, respectively. The encoded stream consists of one base layer and two enhancement layers, and the basis quantization parameters of the three layers are set to 40, 34 and 28, respectively. Each enhancement layer is further split into five MGS layers with MGS vector [3 2 4 2 5]. Then, the post-processing priority level assignment is carried out. Without loss of generality, the overhead constant V is set to 1. The estimate of the three parameters of model (4.4) is performed every IDR period. The duration of the *application period* is set to an IDR period, which leads to an *application period* window $N_{slot} = 1066$.

To assess individual received video quality, we use the PSNR calculated using the luminance MSE aveMSE$_k$ averaged over all the 160 frames if not specified otherwise:

$$PSNR_k = 10 \log_{10} \left(\frac{255^2}{aveMSE_k} \right). \tag{9.1}$$

To measure the system efficiency, we average the PSNR$_k$ for all user received videos, i.e.,

$$avePSNR = (1/K) \sum_{k=1}^{K} PSNR_k. \tag{9.2}$$

The higher avePSNR is, the higher system efficiency we have. The performance of fairness is evaluated through the standard deviation of the PSNRs, i.e.,

$$stdPSNR = \sqrt{(1/K) \sum_{k=1}^{K} (PSNR_k - avePSNR)^2} \tag{9.3}$$

and the average of the absolute value of the PSNR difference between any two videos, i.e.,

$$\triangle_{ave} = (1/G) \sum_{i=1}^{K} \sum_{j=i+1}^{K} |PSNR_i - PSNR_j| \tag{9.4}$$

where $G = K(K-1)/2$. The lower stdPSNR or $\triangle_{ave}$ is, the fairer service each user receives.

Let us denote by F-ILA the ILA algorithm with a faster APP layer algorithm and D-ILA the ILA algorithm where source adaptation is performed using a discrete R-D model, respectively. We compare the performance of the F-ILA and D-ILA algorithms with an equal-rate adaptation scheme, denoted by ERA, which provides fairness among users in terms of allocated video rate without violating the maximum and minimum rate constraints. The equal-rate adaptation problem can be formulated from problem (6.2) or (7.5) by replacing the fairness constraints in (6.2c) or (7.5c) with new rate-fair constraints, i.e., $R_k(\boldsymbol{\tau}, \boldsymbol{p}) = VF_k(\hat{F}_k(f)), \forall k \in \mathcal{K}$, where $f \geq 0$ and $\hat{F}_k(f)$ is defined similarly to (6.1) or (7.1). The solution to it can be obtained by using the ILA algorithm where the APP layer algorithm aims to find an optimal rate-fair solution rather than a quality-fair solution. In the following section, we first compare the performance of the different optimization frameworks without considering packet losses and UXP. Section 9.2 presents the simulation results for the case of transmission with packet losses and UXP.

9.1 Performance Evaluation with Error-free Transmission

Fig. 9.1 shows the received PSNR of each video resulting from the D-ILA, F-ILA and ERA algorithms. We first note that D-ILA outperforms F-ILA in terms of individual video quality. For all videos, D-ILA achieves a PSNR gain ranging from 0.1 to 0.4 dB due to the accuracy of the empirical R-D model used in the source adaptation process.

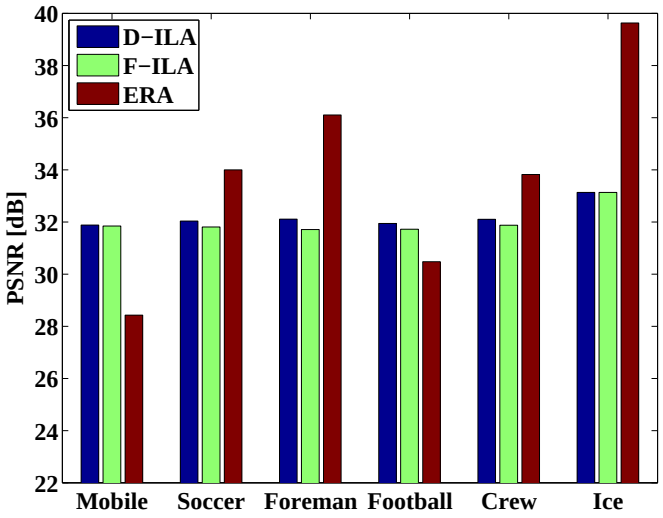


Figure 9.1: The PSNR of each video obtained from the ILA algorithm with a faster APP layer algorithm (F-ILA), ILA algorithm based on a discrete R-D model (D-ILA) and equal rate adaptation (ERA) algorithm.

The relatively small gain, on the other hand, shows that the semi-analytical model is a good approximation to the empirical R-D data and that the solution provided by F-ILA is a good approximation to the benchmark solution provided by D-ILA. As expected, the ERA algorithm results in large quality variations among the videos because it blindly assigns the same rate to all videos without considering the R-D relationship of individual video. By contrast, both D-ILA and F-ILA are able to achieve approximately the same quality among all videos.

The superiority of F-ILA and D-ILA over ERA in terms of quality fairness is more clear in Table 9.1 where the average absolute value of PSNR difference $\triangle_{ave}$ and the standard deviation of the PSNRs stdPSNR in each IDR period are given. Note that the individual PSNR is calculated using the MSE averaged over all the frames in an IDR period. We first note the significant improvement of D-ILA and F-ILA over ERA. Specifically, both $\triangle_{ave}$ and stdPSNR are significantly reduced up to ten times.

Table 9.1: The average absolute value of PSNR difference $\triangle_{ave}$ and standard deviation of PSNRs stdPSNR in each IDR period for D-ILA, F-ILA and ERA.

IDR Index	$\triangle_{ave}$ [dB]			stdPSNR [dB]		
	D-ILA	F-ILA	ERA	D-ILA	F-ILA	ERA
1	0.57	0.57	5.17	0.42	0.45	3.99
2	0.47	0.74	5.06	0.40	0.57	3.84
3	0.65	0.75	5.38	0.62	0.69	4.17
4	0.59	0.70	5.12	0.55	0.60	3.90
5	0.76	0.96	5.39	0.70	0.86	3.98
6	0.46	0.54	4.55	0.43	0.47	3.53
7	0.28	0.65	5.05	0.23	0.50	3.80
8	0.30	0.39	4.81	0.26	0.34	3.63
9	0.37	0.57	5.01	0.37	0.46	3.73
10	0.43	0.68	5.0	0.39	0.55	3.71
Average	0.49	0.66	5.05	0.44	0.55	3.83

It is also worth noting that D-ILA slightly outperforms F-ILA in terms of quality fairness.

In Fig. 9.2(a) and 9.2(b), we show the per-IDR PSNRs, obtaining from D-ILA and ERA, respectively, of all the six videos. Note that the results obtained from F-ILA are similar to those of D-ILA, and thus are omitted here. Fig. 9.2 further shows the advantage of D-ILA and F-ILA over ERA in providing quality fairness among users. While the received PSNRs, resulting from ERA, for the videos are considerably different at each IDR period, D-ILA provides approximately uniform PSNR to each video, over all the simulated IDR periods, except Ice whose minimum rate constraint is active for most of the time.

In Table 9.2, we give the minimum and maximum received PSNRs overall all IDR periods for D-ILA, F-ILA and ERA. Table 9.2 shows that D-ILA and F-ILA not only provide quality fairness, but also improve the quality of the most demanding

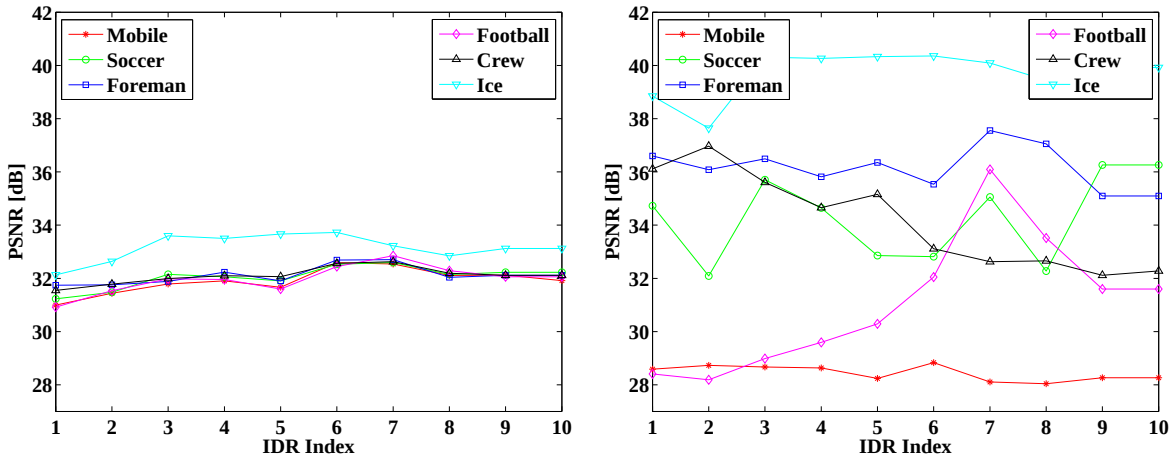


Figure 9.2: Per-IDR PSNRs of all the six videos obtained from the D-ILA algorithm (a) and ERA algorithm (b).

videos, e.g., Mobile and Football, by achieving a minimum PSNR of 31 dB, which is approximately 3 dB higher than that provided by the ERA algorithm.

Table 9.2: The minimum and maximum received PSNRs over all IDR periods for D-ILA, F-ILA and ERA.

min-max PSNR [dB]	D-ILA	F-ILA	ERA
Mobile	31.0 - 32.6	31.0 - 32.5	28.0 - 28.8
Soccer	31.2 - 32.6	31.2 - 32.8	32.1 - 36.3
Foreman	31.7 - 32.7	30.8 - 32.7	35.1 - 37.6
Football	31.0 - 32.9	31.0 - 32.3	28.2 - 36.1
Crew	31.5 - 32.6	30.9 - 32.7	32.1 - 37.0
Ice	32.1 - 33.7	32.1 - 33.7	37.6 - 40.4

To demonstrate the visual quality advantage of the quality-fair algorithms, i.e., D-ILA and F-ILA, for the high-complexity videos, we give a visual comparison of the reconstructed sample video frames of Football and Mobile resulting from ERA and D-ILA in Fig. 9.3 and 9.4, respectively. Fig. 9.3(a) and 9.3(b) are the reconstructed

frames by applying ERA and D-ILA, respectively, from the IDR period, i.e., the first IDR period as we can see in Fig. 9.2(b), where D-ILA has the worst PSNR performance for Football. Similarly, Fig. 9.4(a) and 9.4(b) are the reconstructed frames resulting from ERA and D-ILA, respectively, from the IDR period where D-ILA achieves the lowest PSNR for Mobile. It is clear that the frames are sharper by applying D-ILA. The D-ILA is able to improve the performance of the high-complexity videos, even in the worst case simulated here.

(a) (b)

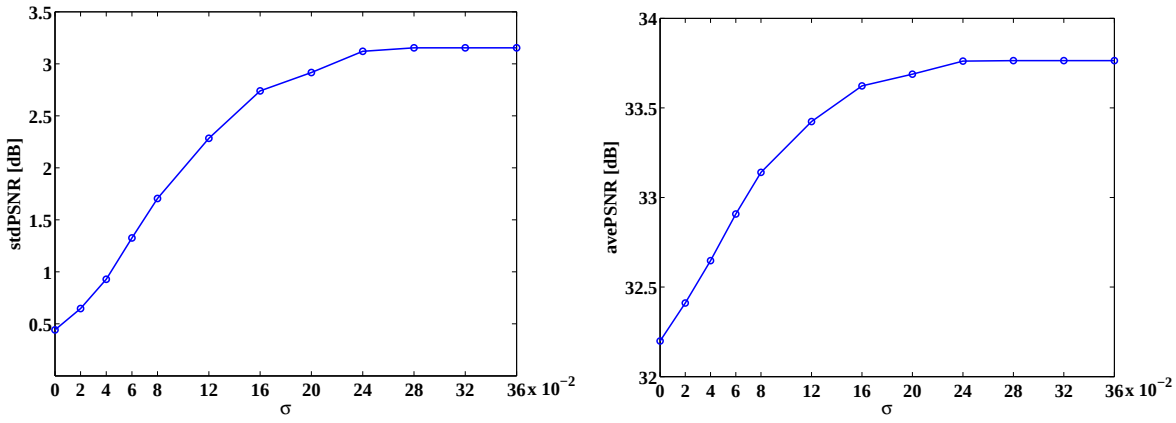
Figure 9.3: Reconstructed sample video frames of Football, by applying (a) ERA and (b) D-ILA, respectively, from the IDR period where D-ILA has the worst performance.

The trade-off between the fairness and system efficiency is illustrated in Fig. 9.5(a) and 9.5(b) where the standard deviation of the PSNRs stdPSNR and average PSNR avePSNR, resulting from the optimization problem (8.6) where $\bar{Q} = q^*$, are plotted for different values of σ, respectively. We see that the totally fair solution (corresponding to $\sigma = 0$) achieves the lowest PSNR deviation but has the lowest average PSNR. The solution corresponding to a larger σ has higher PSNR deviation but higher average PSNR than the solution related to a smaller σ. The fairness is traded off against the system efficiency as σ increases, due to the increasingly looser fairness constraints in

(a) (b)

Figure 9.4: Reconstructed sample video frames of Mobile, by applying (a) ERA and (b) D-ILA, respectively, from the IDR period where D-ILA achieves the lowest PSNR.

problem (8.6). As revealed from Fig. 9.5, when σ is larger than a certain value, e.g., 0.28, in the scenario simulated here, the avePSNR and stdPSNR do not increase as σ increases. This is because as σ becomes larger than a certain value, the intersection of $\mathcal{R}_{adj}$ and the $\mathbf{bd}\,\mathcal{R}$ is no longer changed, and an optimal solution to problem (8.6) with higher avePSNR and stdPSNR becomes unavailable.



(a) (b)

Figure 9.5: Fairness in terms of standard deviation of the PSNRs, stdPSNR, (a) and system efficiency in terms of the average PSNR, avePSNR, (b) for different values of σ.

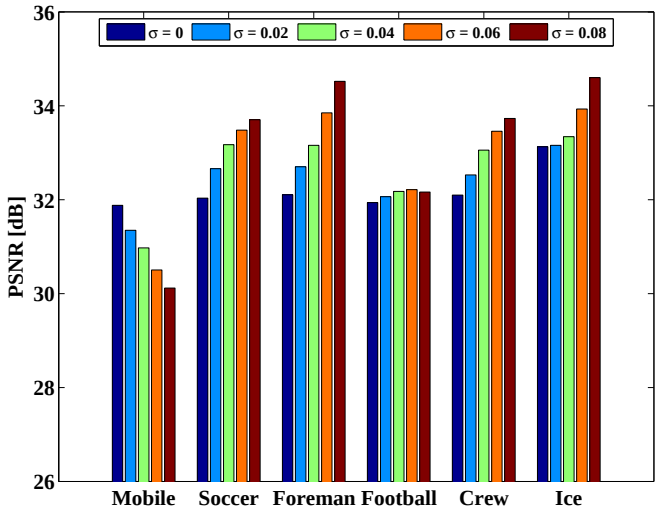


Figure 9.6: The PSNR of each video resulting from the optimization problem (8.6) where $\bar{Q} = q^*$, for a set of values of σ.

Finally, we show the individual PSNR of each video resulting from the optimization problem (8.6) where $\bar{Q} = q^*$, for a set of values of σ in Fig. 9.6. We first note that as σ increases, the PSNR of the most demanding video, i.e., Mobile, will decrease, while the PSNR of the other videos will increase, with the exception of Football whose PSNR first increases, then decrease slightly. Together with Fig. 9.5, it shows that the increase of the system efficiency as σ increases comes at the cost of decreasing the quality of the most demanding videos. This is because the optimization becomes more efficiency driven as σ increases. As a result, more rate is allocated to the low-complexity videos in order to improve the overall video quality. In the scenario simulated here, with a proper selection of σ, e.g., $\sigma = 0.08$, the average PSNR, avePSNR, can be increased by approximate 1 dB compared to the totally fair solution while the quality of the most demanding video, i.e., Mobile, is maintained at an acceptable level by achieving an individual PSNR of 30 dB.

9.2 Performance Evaluation with Error-prone Transmission

Table 9.3: The average received PSNR PSNR_{rec} and average expected PSNR PSNR_{exp}, resulting from D-ILA, F-ILA and ERA, of each video sequence for a packet loss rate $r_{rtp} = 0.05$.

Video	D-ILA		F-ILA		ERA	
	PSNR_{rec}	PSNR_{exp}	PSNR_{rec}	PSNR_{exp}	PSNR_{rec}	PSNR_{exp}
Mobile	31.79	31.78	31.78	31.77	28.35	28.34
Soccer	31.96	31.92	31.73	31.70	33.58	33.56
Foreman	32.07	32.0	31.63	31.58	36.03	36.02
Football	31.82	31.80	31.67	31.67	30.36	30.33
Crew	31.95	31.93	31.87	31.72	33.74	33.73
Ice	33.16	32.67	33.14	32.42	39.45	39.43

In this section, we evaluate the performance of the different optimization frameworks in the scenario of error-prone transmission where packet losses and UXP are considered. We assume that the number of available subcarriers is 144, which allows the system to support the transmission of the base layers of all videos with UXP. Moreover, we set the size of a RTP packet to 600 bytes and simulate a RTP packet loss rate r_{rtp} of 5% as in (Mansour *et al.*, 2008) and (Cicalò and Tralli, 2014). The maximum number of bytes per RS codeword is set to 255. The minimum number of bytes per RS codeword is dependent on the video content and r_{rtp}.

In Table 9.3, we show the average PSNR (in dB) of each video sequence, resulting from D-ILA, F-ILA and ERA, where the averages are calculated over ten IDR periods. Here, PSNR_{rec} is the average received PSNR, defined in (9.1), at the receiver side and PSNR_{exp} is the average expected PSNR, which is the optimal solution of the

cross-layer optimization problem. The UXP scheme ensures that all data of each transmitted video stream can be correctly recovered from packet erasures with very high probility, in the scenario simulated here. We first note that the expected PSNRs are approximately equal to the received PSNRs, for the three algorithms. Such results show the applicability and goodness of discussed cross-layer optimization frameworks when packet losses and UXP are considered. The results also justify that the R-D models in (4.2) and (4.4) can be extended to cover the case of error-prone transmission with UXP.

Table 9.4: The average absolute value of PSNR difference $\triangle_{ave}$ and standard deviation of PSNRs stdPSNR in each IDR period for D-ILA, F-ILA and ERA.

IDR Index	$\triangle_{ave}$ [dB]			stdPSNR [dB]		
	D-ILA	F-ILA	ERA	D-ILA	F-ILA	ERA
1	0.75	0.73	5.01	0.57	0.63	3.80
2	0.54	0.63	5.07	0.48	0.49	3.90
3	0.66	0.79	5.43	0.62	0.69	4.18
4	0.66	0.80	5.12	0.59	0.65	3.90
5	0.78	0.92	5.38	0.74	0.85	3.97
6	0.53	0.61	4.56	0.48	0.53	3.58
7	0.28	0.62	5.09	0.23	0.50	3.84
8	0.36	0.44	4.71	0.32	0.38	3.56
9	0.44	0.67	5.07	0.40	0.51	3.77
10	0.47	0.60	4.50	0.43	0.53	3.43
Average	0.55	0.68	4.99	0.49	0.57	3.79

In Table 9.4, we give the average absolute value of PSNR difference $\triangle_{ave}$ and standard deviation of the PSNRs stdPSNR, obtained from D-ILA, F-ILA and ERA, in each IDR period. Together with Table 9.3, it shows that D-ILA and F-ILA significantly outperform ERA in terms of quality fairness, even in the presence of packet losses. Moreover, D-ILA achieves slightly better individual video quality and quality

fairness than F-ILA, as expected, because of the use of the accurate empirical R-D model for source adaptation.

The benefits of the D-ILA and F-ILA algorithms in terms of quality fairness can be further shown in Fig. 9.7(a) and 9.7(b) where the per-IDR PSNRs, obtained from D-ILA and ERA, respectively, of all the six videos are given.

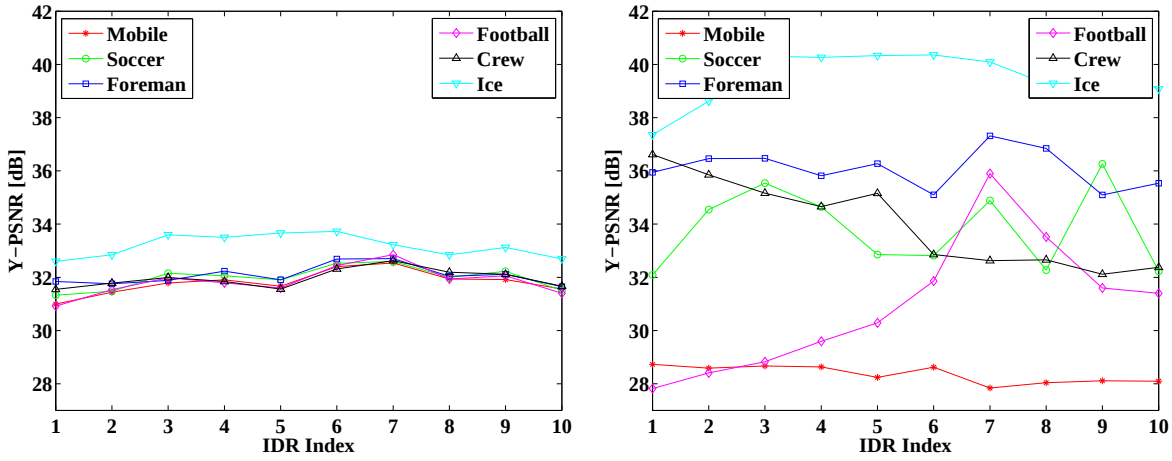


Figure 9.7: Per-IDR PSNRs of all the six videos obtained from the D-ILA algorithm (a) and ERA algorithm (b).

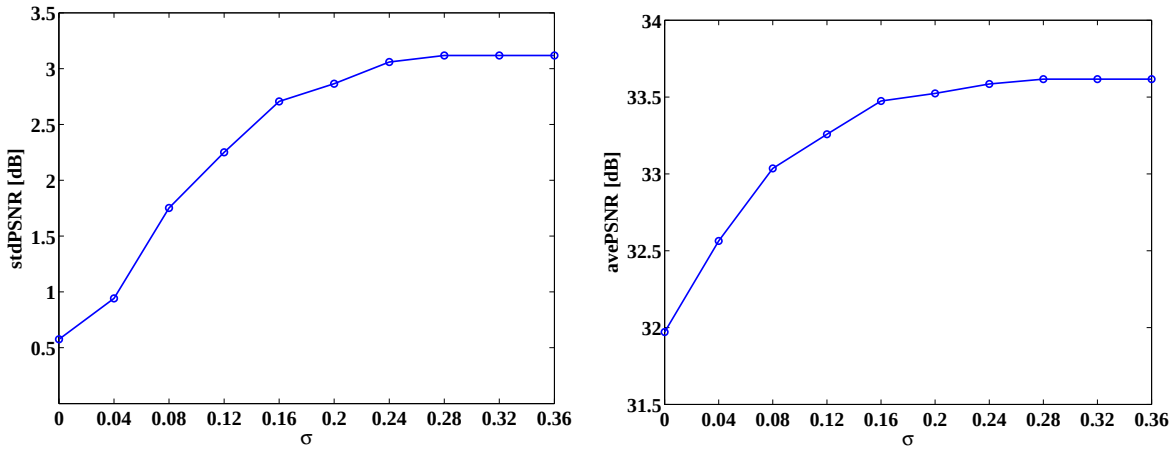
Table 9.5: The minimum and maximum received PSNRs over all IDR periods for D-ILA, F-ILA and ERA.

min-max PSNR [dB]	D-ILA	F-ILA	ERA
Mobile	31.0 - 32.5	31.0 - 32.4	27.8 - 28.7
Soccer	31.3 - 32.6	30.9 - 32.6	32.1 - 36.3
Foreman	31.6 - 32.7	30.8 - 32.7	35.1 - 37.3
Football	31.0 - 32.9	31.0 - 32.3	27.8 - 35.9
Crew	31.5 - 32.6	31.4 - 32.3	32.1 - 36.6
Ice	32.6 - 33.7	32.6 - 33.7	37.4 - 40.3

Table 9.5 summarizes the minimum and maximum received PSNRs over all IDR

periods for D-ILA, F-ILA and ERA. The D-ILA and F-ILA algorithms, while achieving quality fairness, provide the most demanding videos, e.g., Mobile and Football, with a minimum received PSNR of 31 dB, whereas ERA algorithm achieves a considerably lower minimum received PSNR of 27.8 dB for Mobile and Football.

Fig. 9.8(a) and 9.8(b) show the fairness and system efficiency results for different values of σ, respectively. The results are obtained from solving the optimization problem (8.6) where $\bar{Q} = q^*$. We see a similar trend as in the case of error-free transmission. The fairness is traded off against the system efficiency as σ increases until σ becomes larger than a certain value, e.g., 0.28.



(a) (b)

Figure 9.8: Fairness in terms of standard deviation of the PSNRs, stdPSNR, (a) and system efficiency in terms of the average PSNR, avePSNR, (b) for different values of σ.

In Fig. 9.9, the individual PSNR of each video resulting from the optimization problem (8.6) where $\bar{Q} = q^*$, for a set of values of σ are shown. As in the case of error-free transmission, with the increase of σ, the PSNR of the most demanding video, i.e., Mobile, decreases, whereas the PSNR of the other videos will increase, with the exception of Football whose PSNR first increases, then decrease slightly. In

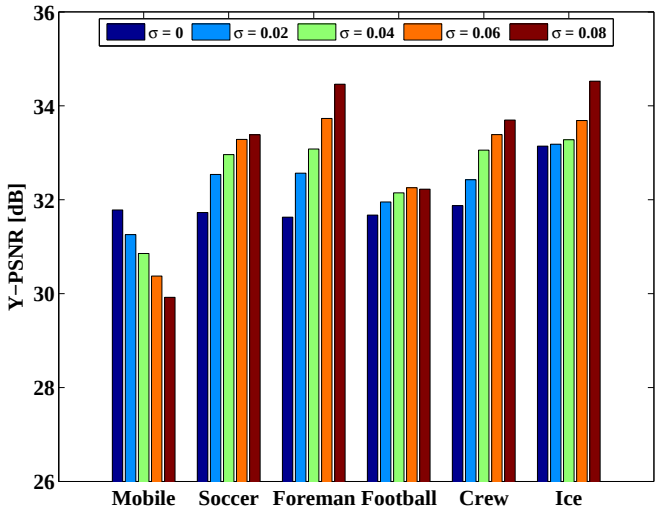


Figure 9.9: The PSNR of each video resulting from the optimization problem (8.6) where $\bar{Q} = q^*$, for a set of values of σ.

the scenario simulated here, with a proper selection of σ, e.g., $\sigma = 0.08$, the average PSNR, avePSNR, can be increased by approximate 1 dB compared to the totally fair solution while the quality of the most demanding video, i.e., Mobile, is maintained at an acceptable level by achieving an individual PSNR of 29.9 dB.

Chapter 10

Conclusion and Future Work

10.1 Conclusion

In this thesis, we have tackled the problem of sending scalable videos to multiple users over OFDMA wireless networks where quality fairness and system efficiency are jointly considered. This problem has been recently addressed by Cicalò and Tralli who proposed a cross-layer optimization framework for the maximization of the sum of the ergodic rates under totally quality-fair constraints. The constrained sum-rate maximization was decomposed into two subproblems and the ILA algorithm has been proposed to achieve the optimal solution.

We have first proposed a quality fairness-oriented cross-layer optimization framework that solves the JRASA problem where the objective is to maximize the sum of the PSNRs while minimizing the PSNR difference among the received videos. We have proved that the JRASA problem is equivalent to the aforementioned constrained sum-rate maximization and can be solved by the ILA algorithm. A considerably faster algorithm has been designed to solve the source adaptation problem at the APP layer.

Moreover, we have shown that the above optimization framework can be extended to solve efficiently the JRASA problem based on accurate empirical R-D models where the resulting solution can be used as a benchmark to assess the performance of solutions based on the semi-analytical R-D models.

To achieve the trade-offs between quality fairness and system efficiency, we have then proposed an adjustable quality-fair cross-layer optimization framework that seeks to maximize the sum of the PSNRs while limiting the absolute value of the relative difference between the PSNR of each video and a predefined common PSNR value. We have shown that the optimization problem is a general utility-based resource allocation problem, for which efficient algorithms are available to obtain an almost surely optimal solution.

In both error-free and error-prone scenarios, the numerical results have shown that the proposed quality fairness-oriented optimization framework provides significantly better performance in terms of quality fairness and the provision of better quality to high-complexity videos with respect to the equal-rate adaptation scheme. Moreover, a desired trade-off between fairness and system efficiency can be achieved using the adjustable quality-fair cross-layer optimization framework.

10.2 Future Work

The work presented in this thesis provides two lines of research which could be pursued in the future.

Firstly, we consider in this thesis a single-input single-output antenna configuration, it is possible to extend the frameworks, methods and algorithms developed here to cover the scenario of multiple-input and multiple-output OFDMA systems.

The second line of research is looking for different ways to find a reasonable common target PSNR value $\bar{Q}$ for problem (8.3). At this point, $\bar{Q}$ is selected as the optimal quality level q^*, which is obtained only after solving the JRASA problem (6.2).

Appendix A

Proof of Proposition 1

The first step to prove **Proposition 1** is the following proposition:

Proposition 2. *For any given K-tuples $\boldsymbol{r}$, $\boldsymbol{r}' \in \mathcal{R}_f$, if $r_k < r'_k$, then*

$$[r_1, \cdots, r_{k-1}, r_{k+1}, \cdots, r_K] \preceq [r'_1, \cdots, r'_{k-1}, r'_{k+1}, \cdots, r'_K]. \tag{A.1}$$

Proof: For any $\boldsymbol{r} \in \mathcal{R}_f$, if its k-th element $r_k \in \mathcal{F}_k$ and the associated PSNR $Q_k = F_k^{-1}(r_k/V)$ are fixed, the PSNRs of the other $K-1$ videos are given, according to the definition of $\mathcal{R}_f$ in (7.6), as:

$$Q_l = \begin{cases} Q_{l,max} & Q_k \geq Q_{l,max} \\ q_{l,m+1} & q_{l,m} < Q_k \leq q_{l,m+1} \\ Q_{l,min} & Q_k \leq Q_{l,min} \end{cases} , \forall l \in \mathcal{K} \setminus \{k\}. \tag{A.2}$$

If $r_k < r'_k$, we have $Q_k < Q'_k$ because of the strictly increasing monotonicity of F_k^{-1}. According to (A.2), we should have $Q_l \leq Q'_l, \forall l \in \mathcal{K} \setminus \{k\}$, which together with the

strictly increasing monotonicity of F_k, proves **Proposition 2**.

To prove the one-dimensionality of $\mathcal{C}_{pw}$, it is sufficient to prove that for any K-tuple $\boldsymbol{r} \in \mathcal{C}_{pw}$, its coordinates in $\mathbb{R}^{K-1}$ can be determined by one coordinate $\mathbb{R}$, i.e., $\mathbb{R} \xrightarrow{\mathcal{C}_{pw}} \mathbb{R}^{K-1}$. According to (A.2), any fairness rate vector $\boldsymbol{R} \in \mathcal{R}_f$, which is the endpoint of one of the line segments constituting $\mathcal{C}_{pw}$, satisfy the condition. Let $\boldsymbol{R}_a$ be any rate point belonging to $\mathcal{C}_{pw}$ and $\mathcal{L}(\boldsymbol{R}_l, \boldsymbol{R}_r) = \{\boldsymbol{R}_l + t(\boldsymbol{R}_r - \boldsymbol{R}_l) \mid t \in [0,1]\}$ be the line segment that comprises $\boldsymbol{R}_a$, if one coordinate R_a^i of $\boldsymbol{R}_a$ is fixed, the other $K-1$ coordinates $[R_a^1, \cdots, R_a^{i-1}, R_a^{i+1}, R_a^K]$ can be expressed as $K-1$ functions of R_a^i. That is,

$$R_a^j = \left(\frac{R_a^i - R_l^i}{R_r^k - R_l^k}\right) \cdot (R_r^j - R_l^j) + R_l^j, \forall j \in \mathcal{K} \setminus \{i\} \tag{A.3}$$

where $\boldsymbol{R}_l = [R_l^1, R_l^2, \cdots, R_l^K]$ and $\boldsymbol{R}_r = [R_r^1, R_r^2, \cdots, R_r^K]$. Therefore, (A.2) and (A.3) prove the one dimensionality of $\mathcal{C}_{pw}$.

To prove the monotonically increasing property, it is equivalent to prove that for any given K-tuples $\boldsymbol{r}, \boldsymbol{r'} \in \mathcal{C}_{pw}$, if $r_k < r'_k$, then $[r_1, \cdots, r_{k-1}, r_{k+1}, \cdots, r_K] \preceq [r'_1, \cdots, r'_{k-1}, r'_{k+1}, \cdots, r'_K]$. According to **Proposition 1**, any fairness rate vector satisfies this condition. Let $\boldsymbol{R}_a$ and $\boldsymbol{R}_b$ be any two different rate points belonging to $\mathcal{C}_{pw}$ and $\mathcal{L}(\boldsymbol{R}_l, \boldsymbol{R}_r) = \{\boldsymbol{R}_l + t(\boldsymbol{R}_r - \boldsymbol{R}_l) \mid t \in [0,1]\}$ be the line segment that comprises $\boldsymbol{R}_a$ and $\boldsymbol{R}_b$. According to (A.3), if $R_a^i < R_b^i$, we will have $R_a^j \leq R_b^j, \forall j \in \mathcal{K} \setminus \{i\}$ where the equality is achieved when $R_r^j = R_l^j$. On the other hand, if $\boldsymbol{R}_a$ and $\boldsymbol{R}_b$ belong to different line segments, it is clear that $\boldsymbol{R}_a \preccurlyeq \boldsymbol{R}_b$ holds if there exists at least one k such that $R_a^k < R_b^k$. Therefore, the piecewise curve $\mathcal{C}_{pw}$ is a monotonically increasing manifold.

Propositions 1 and **2** justify the uniqueness of the optimal solution $\boldsymbol{R}^*$ and the intersection rate point $\boldsymbol{R}_{int}$ that given by the intersection between the boundary $\mathbf{bd}\mathcal{R}$

of the rate region $\mathcal{R}$ with the monotonically increasing one-dimension piecewise curve $\mathcal{C}_{pw}$ ($\mathcal{C}_{pw}$ comprises the line segment $\mathcal{L}(\boldsymbol{R}^*, \boldsymbol{R}')$ bounded by the optimal solution $\boldsymbol{R}^*$ and it adjacent fairness rate vector $\boldsymbol{R}'$). Based on the method of proof by contradiction, let us first assume that both $\boldsymbol{R}_{int1}$ and $\boldsymbol{R}_{int2}$ are intersection rate points of the boundary $\mathbf{bd}\,\mathcal{R}$ and $\mathcal{C}_{pw}$. Then, we will have $\boldsymbol{R}_{int1} \preccurlyeq \boldsymbol{R}_{int2}$ if there exists at least one k such that $R^k_{int1} < R^k_{int2}$. However, if $\boldsymbol{R}_{int1} \preccurlyeq \boldsymbol{R}_{int2}$, according to (5.4), $\boldsymbol{R}_{int1}$ cannot be a boundary point. This contradicts with the fact that $\boldsymbol{R}_{int1} \in \mathbf{bd}\,\mathcal{R} \cap \mathcal{C}_{pw}$.

Appendix B

The Algorithms Used at the APP Layer

Algorithm 3 works as follows. The algorithm first checks two feasibility conditions, $\Gamma(\boldsymbol{F}_{min}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ and $\Gamma(\boldsymbol{F}_{max}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \geq 0$, which are the relaxations of $V\boldsymbol{F}_{min} \in \mathcal{R}$ and $V\boldsymbol{F}_{max} \notin \mathcal{R}$, respectively. The procedure from line 7 to line 31 strives to find $(V\boldsymbol{F}', V\boldsymbol{F}'') \in \mathcal{R}_{ad}$ such that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ and $\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$. If the search for the desired $\boldsymbol{F}'$ and $\boldsymbol{F}''$ is unsuccessful using the aforementioned procedure, the search will be switched to a linear search procedure in line 36. Finally, the solution $\boldsymbol{F}$ is evaluated from $V\boldsymbol{F}$, which is computed as the intersection of the line segment $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$ and the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$.

The 5 auxiliary algorithms used in **Algorithm 3** are given as follows. The first algorithm shown in **Algorithm 4** computes a fair rate vector $V\boldsymbol{F}(q)$. For the case of error-free transmission, given an arbitrary q, **Algorithm 4** first evaluates the target PSNR Q_k for each video using the step function $\hat{Q}_k(q)$ in (7.1). The index i_k records the information about which PSNR interval q belongs to. Then, the minimum rate

Algorithm 4 Pseudo code to compute a fair rate vector $V\boldsymbol{F}(q)$

1: **Input:** q, functions $\hat{Q}_k(q)$ in (7.1) and $F_k(Q)$ in (4.2), $\forall k \in \mathcal{K}$
2: **Output:** The fairness rate vector $V\boldsymbol{F}(q)$ and an index vector $\boldsymbol{i}$
3: **for all** $k \in \mathcal{K}$ **do**
4: $\quad Q_k = \hat{Q}_k(q)$;
5: $\quad i_k = c$ when $q \in I_{k,c}$;
6: $\quad F_k = F_k(Q_k)$;
7: **end for**

F_k required to achieved the target PSNR for each video is computed according to the R-D function in (4.2). In this way, the PSNR difference between any two videos is guaranteed to be zero according to the definition in (7.4), and thus the fair rate vector is obtained. Three examples of computing a fair vector are shown in Fig. B.1 where R-D relationships (not based on real data, but only for the purpose of demonstration) of three videos are also depicted. If q is chosen to be q_{c1}, the target PSNR values for the

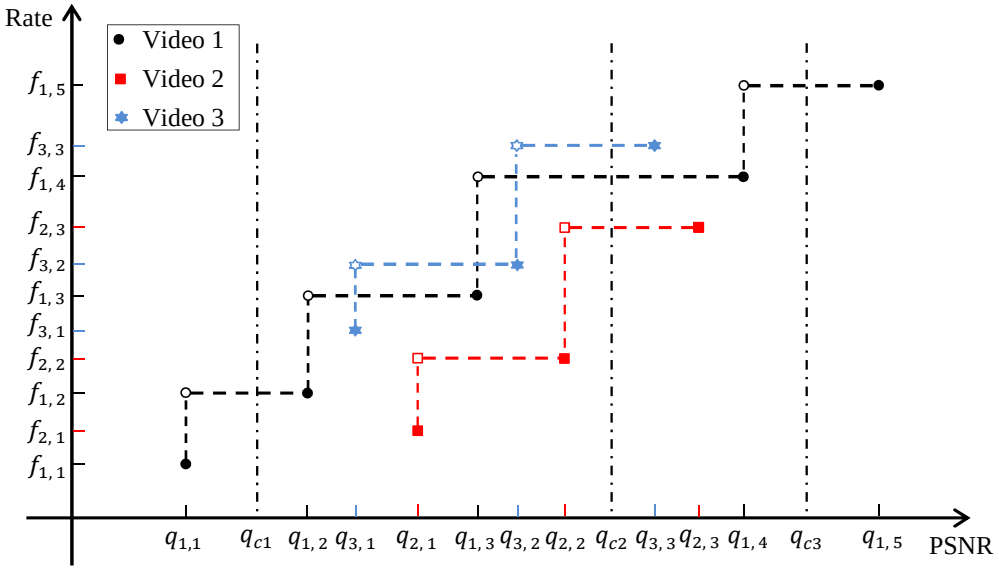


Figure B.1: An illustration that shows how to identify a fair rate vector.

three videos will be $q_{1,2}$, $q_{2,1}$ and $q_{3,1}$. Then, the corresponding rates are $f_{1,2}$, $f_{2,1}$ and $f_{3,1}$. Therefore, the fair rate vector is $V\boldsymbol{F} = \{f_{1,2}, f_{2,1}, f_{3,1}\}$. Similarly, if q is equal

to q_{c2} or q_{c3}, the corresponding fair rate vector is $\{f_{1,4}, f_{2,3}, f_{3,3}\}$ or $\{f_{1,5}, f_{2,3}, f_{3,3}\}$. The worst-case running time complexity of **Algorithm 4** is $O(K \log_2 C_{max}^{all})$ where $C_{max}^{all} = \max_k(C_k)$. For the case of error-prone transmission, evaluating the target PSNR for each video requires, in the worst case, computing $\log_2 L_{max}$ expected R-D values on-the-fly. Therefore, the worst-case running time complexity of **Algorithm 4** is $O(\gamma KLI \log_2 L_{max})$.

Algorithm 5 Pseudo code to find $V\boldsymbol{F}'$ such that $(V\boldsymbol{F}(q), V\boldsymbol{F}') \in \mathcal{R}_{ad}$

Input: $\boldsymbol{F}(q)$, $\boldsymbol{Q}$, $\boldsymbol{i}$, $\boldsymbol{Q}_{max}$ and $\mathcal{F}_k, \forall k \in \mathcal{K}$
Output: $\boldsymbol{F}'$ and its related PSNR vector $\boldsymbol{Q}'$ and index vector $\boldsymbol{i}'$
$\mathcal{V} = \{k \in \mathcal{K} \mid Q_k \neq Q_{k,max}\}$;
$\mathcal{V}^* = \arg\min_{k \in \mathcal{V}} q_{k,i_k}$;
for all $k \in \mathcal{V}^*$ **do**
    $VF'_k = f_{k,i_k+1}$; $Q'_k = q_{k,i_k+1}$; $i'_k = i_k + 1$;
end for
for all $k \in \mathcal{K} \setminus \mathcal{V}^*$ **do**
    $VF'_k = VF_k(q)$; $Q'_k = Q_k$; $i'_k = i_k$;
end for

Next, let us consider **Algorithm 5** that computes a larger adjacent fair rate vector $V\boldsymbol{F}'$ of a given fair rate vector $V\boldsymbol{F}(q)$. For the case of error-free transmission, for any video whose PSNR value Q_k is not equal to its maximum PSNR value, i.e. $Q_{k,max}$, the algorithm puts it into a candidate set $\mathcal{V}$. The algorithm then finds the set $\mathcal{V}^*$ of videos whose PSNR values are the smallest among $Q_k, \forall k \in \mathcal{V}$. The number of elements in $\mathcal{V}^*$ ranges from 1 to $K-1$. Then, the larger adjacent fair rate vector $V\boldsymbol{F}'$ is obtained in a way that its k-th element, $\forall k \in \mathcal{V}^*$, is set to f_{k,i_k+1}, which is the smallest rate, larger than $VF_k(q)$, in $\mathcal{F}_k$, and its other elements are set to the co-located elements of $V\boldsymbol{F}(q)$. **Algorithm 5**, in the worst case, runs in $O(K)$ time. For the case of error-prone transmission, since we do not pre-compute and store all the expected R-D information, we do not have the information about $\mathcal{F}_k, \forall k \in \mathcal{K}$. In addition, the

index i_k will record the information about the transmission budget, i.e., the value of L. After we determine the set $\mathcal{V}^*$, given that $VF_k(q) = IL_k, \forall k \in \mathcal{V}^*$, $V\boldsymbol{F}'$ is obtained in the following way: $VF'_k = I(L_k + 1), \forall k \in \mathcal{V}^*$ and $VF'_k = IL_k, \forall k \in \mathcal{K} \setminus \mathcal{V}^*$. The individual expected PSNR Q'_k is computed based on F'_k and the index i'_k is set to either $i_k + 1$ or i_k. In this case, the worst-case running time complexity of **Algorithm 5** is $O(\gamma KLI)$.

The pseudo code used to compute a smaller adjacent fair rate vector $\boldsymbol{F}'$ of $\boldsymbol{F}(q)$ is summarized in **Algorithm 6** as follow. **Algorithm 6** is similar to **Algorithm 5**. For the case of error-free transmission, the worst-case running time is $O(K)$. For the case of error-prone transmission, the worst-case running time is $O(\gamma KLI)$.

Algorithm 6 Pseudo code to find $\boldsymbol{F}'$ such that $(V\boldsymbol{F}', V\boldsymbol{F}(q)) \in \mathcal{R}_{ad}$

Input: $\boldsymbol{F}(q)$, $\boldsymbol{Q}$, $\boldsymbol{i}$, $\boldsymbol{Q}_{min}$ and $\mathcal{F}_k, \forall k \in \mathcal{K}$
Output: $\boldsymbol{F}'$ and its related PSNR vector $\boldsymbol{Q}'$ and index vector $\boldsymbol{i}'$
$\mathcal{V} = \{k \in \mathcal{K} \mid Q_k \neq Q_{k,min}\}$;
$\mathcal{V}^* = \arg\max_{k \in \mathcal{V}} q_{k,i_k-1}$;
for all $k \in \mathcal{V}^*$ **do**
    $F'_k = f_{k,i_k-1}$; $Q'_k = q_{k,i_k-1}$; $i'_k = i_k - 1$;
end for
for all $k \in \mathcal{K} \setminus \mathcal{V}^*$ **do**
    $F'_k = F_k(q)$; $Q'_k = Q_k$; $i'_k = i_k$;
end for

Algorithm 7 Pseudo code finding an intersection of $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$ and $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$

Input: $\boldsymbol{F}'$, $\boldsymbol{F}''$, $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{w}}$
Output: An intersection rate vector $V\boldsymbol{F}$
$F_k = (F''_k - F'_k)t + F'_k, \forall k \in \mathcal{K}$ where t is a real number;
Solve $\Gamma(\boldsymbol{F}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$ to obtain t and thus $V\boldsymbol{F}$;

We now consider **Algorithm 7** that finds an intersection of $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$ and the tangent space $\mathcal{T}_{\mathcal{R}}(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})$. Since the intersection point is on $\mathcal{L}(\boldsymbol{F}', \boldsymbol{F}'')$, we have $t =$

$(F_k - F'_k)/(F''_k - F'_k), \forall k \in \mathcal{K}$. Therefore, the elements of the intersection rate vector can be expressed as $VF_k = V(F''_k - F'_k)t + VF'_k, \forall k \in \mathcal{K}$. Since the intersection rate vector is also on the tangent space, i.e., $\Gamma(\boldsymbol{F}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) = 0$, we have $t = \sum_{k=1}^{K}[w_k(R_k - VF'_k)/(VF''_k - VF'_k)]$. Then, $VF_k, \forall k \in \mathcal{K}$ can be evaluated using t. The running time of **Algorithm 7** is $O(K)$.

Algorithm 8 Pseudo code executing a linear search procedure to find $\boldsymbol{F}'$ and $\boldsymbol{F}''$ such that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$

1: **Input:** $\tilde{\boldsymbol{R}}$, $\tilde{\boldsymbol{w}}$, $\boldsymbol{F}''$, $\boldsymbol{Q}''$, $\boldsymbol{Q}_{max}$ and $\mathcal{F}_k, \forall k \in \mathcal{K}$
2: **Output:** $\boldsymbol{F}'$ and $\boldsymbol{F}''$ satisfying $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$
3: **Initialize:** $cond_{suc}$ = false;
4: **repeat**
5: Find $V\boldsymbol{F}_l$ such that $(V\boldsymbol{F}_l, V\boldsymbol{F}'') \in \mathcal{R}_{ad}$;
6: **if** $\Gamma(\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ **then**
7: $\boldsymbol{F}'' = \boldsymbol{F}_l$;
8: **else**
9: $\boldsymbol{F}' = \boldsymbol{F}_l$; $cond_{suc}$ = true;
10: **end if**
11: **until** $cond_{suc}$ == true

The pseudo code that carries out a linear search procedure to find $\boldsymbol{F}'$ and $\boldsymbol{F}''$ is reported in **Algorithm 8** below. Since the R-D function $F_k(\hat{Q}_k(q))$ maps an infinite number of nonnegative values to a set of discrete rate values, updating q (line 8 of **Algorithm 3**) may result in no update for the fair rate vector $\boldsymbol{F}(q)$, and accordingly for $\boldsymbol{F}'$ and $\boldsymbol{F}''$. Consequently, we may not obtain two adjacent rate vectors $V\boldsymbol{F}'$ and $V\boldsymbol{F}''$ such that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ when $(high - low)/2$ become less than e_{bs}. In such case, however, it is guaranteed that $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) < 0$ and $\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$ when the procedure from line 7 to line 31 is terminated. Therefore, starting from $V\boldsymbol{F}''$, the algorithm finds a smaller adjacent fair rate vector $V\boldsymbol{F}_l$ of $V\boldsymbol{F}''$ and checks whether or not $\Gamma(\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$. If $\Gamma(\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) > 0$, $\boldsymbol{F}''$ is updated by setting

$\boldsymbol{F}'' = \boldsymbol{F}_l$. This procedure is repeated until $\boldsymbol{F}_l$ satisfies $\Gamma(V\boldsymbol{F}_l, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$. Then, $\boldsymbol{F}'$ is updated by setting $\boldsymbol{F}' = \boldsymbol{F}_l$. In this way, we have $(V\boldsymbol{F}', V\boldsymbol{F}'') \in \mathcal{R}_{ad}$ and $\Gamma(\boldsymbol{F}', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}})\Gamma(\boldsymbol{F}'', \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{w}}) \leq 0$. For the case of error-free transmission, **Algorithm 8**, in the worst case, runs in $O(KC_{LN})$ time where C_{LN} is the number of searches, which is expected to be small. For the case of error-prone transmission, the running time complexity of **Algorithm 8** is $O(\gamma KLIC_{LN})$.

Bibliography

(2006). 3rd generation partnership project, technical specification group radio access network; physical layer aspects for evolved universalterrestrial radio access (utra). *3GPP Std. TR 25.814 v.7.0.0*.

Amonou, I., Cammas, N., Kervadec, S., and Pateux, S. (2007). Optimized rate-distortion extraction with quality layers in the scalable extension of h. 264/avc. *Circuits and Systems for Video Technology, IEEE Transactions on*, **17**(9), 1186–1193.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Chen, Z., Li, M., and Tan, Y.-P. (2010). Perception-aware multiple scalable video streaming over wlans. *Signal Processing Letters, IEEE*, **17**(7), 675–678.

Cicalò, S. and Tralli, V. (2014). Distortion-fair cross-layer resource allocation for scalable video transmission in ofdma wireless networks. *Multimedia, IEEE Transactions on*, **16**(3), 848–863.

Cicalò, S., Haseeb, A., and Tralli, V. (2012). Fairness-oriented multi-stream rate

adaptation using scalable video coding. *Signal Processing: Image Communication*, **27**(8), 800–813.

Cisco, C. V. N. I. (2015). Global mobile data traffic forecast update, 2014–2019. *White Paper*.

Committee, I. L. S. *et al.* (2006). Air interface for fixed and mobile broadband wireless access systems. *IEEE Std. 802.16e-2005*.

Dai, M., Loguinov, D., and Radha, H. M. (2006). Rate-distortion analysis and quality control in scalable internet streaming. *Multimedia, IEEE Transactions on*, **8**(6), 1135–1146.

De Cock, J., Notebaert, S., Lambert, P., and Van De Walle, R. (2009). Architectures for fast transcoding of h. 264/avc to quality-scalable svc streams. *Multimedia, IEEE Transactions on*, **11**(7), 1209–1224.

Dumitrescu, S., Wu, X., and Wang, Z. (2007). Efficient algorithms for optimal uneven protection of single and multiple scalable code streams against packet erasures. *Multimedia, IEEE Transactions on*, **9**(7), 1466–1474.

Guan, Z., Yuan, D., and Zhang, H. (2009). Optimal and fair resource allocation for multiuser wireless multimedia transmissions. *EURASIP Journal on Wireless Communications and Networking*, **2009**, 17.

Ha, H. and Yim, C. (2008). Layer-weighted unequal error protection for scalable video coding extension of h. 264/avc. *Consumer Electronics, IEEE Transactions on*, **54**(2), 736–744.

Ha, H., Yim, C., and Kim, Y. Y. (2008). Cross-layer multiuser resource allocation for video communication over ofdm networks. *Computer Communications*, **31**(15), 3553–3563.

Haseeb, A., Martini, M. G., Cicalò, S., and Tralli, V. (2012). Rate and distortion modeling for real-time mgs coding and adaptation. In *Wireless Advanced (WiAd), 2012*, pages 85–89. IEEE.

He, L. and Liu, G. (2014). Quality-driven cross-layer design for h. 264/avc video transmission over ofdma system. *Wireless Communications, IEEE Transactions on*, **13**(12), 6768–6782.

Hsu, C.-H. and Hefeeda, M. (2008). On the accuracy and complexity of rate-distortion models for fine-grained scalable video sequences. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **4**(2), 15.

Huang, J., Li, Z., Chiang, M., and Katsaggelos, A. K. (2008). Joint source adaptation and resource allocation for multi-user wireless video streaming. *Circuits and Systems for Video Technology, IEEE Transactions on*, **18**(5), 582–595.

Khalek, A. A., Caramanis, C., and Heath, R. W. (2015). Delay-constrained video transmission: Quality-driven resource allocation and scheduling. *Selected Topics in Signal Processing, IEEE Journal of*, **9**(1), 60–75.

Khan, N., Martini, M. G., and Bharucha, Z. (2012). Quality-aware fair downlink scheduling for scalable video transmission over lte systems. In *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, pages 334–338. IEEE.

Khan, N., Martini, M. G., and Staehle, D. (2013). Opportunistic proportional fair downlink scheduling for scalable video transmission over lte systems. In *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, pages 1–6. IEEE.

Kwon, D.-K., Shen, M.-Y., and Kuo, C. J. (2007). Rate control for h. 264 video with enhanced rate and distortion models. *Circuits and Systems for Video Technology, IEEE Transactions on*, **17**(5), 517–529.

Li, L. and Goldsmith, A. J. (2001). Capacity and optimal resource allocation for fading broadcast channels. i. ergodic capacity. *Information Theory, IEEE Transactions on*, **47**(3), 1083–1102.

Li, Y., Li, Z., Chiang, M., *et al.* (2009). Content-aware distortion-fair video streaming in congested networks. *Multimedia, IEEE Transactions on*, **11**(6), 1182–1193.

Maani, E. and Katsaggelos, A. K. (2010). Unequal error protection for robust streaming of scalable video over packet lossy networks. *Circuits and Systems for Video Technology, IEEE Transactions on*, **20**(3), 407–416.

Maani, E., Pahalawatta, P. V., Berry, R., Pappas, T. N., and Katsaggelos, A. K. (2008). Resource allocation for downlink multiuser video transmission over wireless lossy networks. *Image Processing, IEEE Transactions on*, **17**(9), 1663–1671.

Mansour, H., Krishnamurthy, V., and Nasiopoulos, P. (2008). Channel aware multiuser scalable video streaming over lossy under-provisioned channels: Modeling and analysis. *Multimedia, IEEE Transactions on*, **10**(7), 1366–1381.

Mazzotti, M., Moretti, S., and Chiani, M. (2012). Multiuser resource allocation with

adaptive modulation and ldpc coding for heterogeneous traffic in ofdma downlink. *Communications, IEEE Transactions on*, **60**(10), 2915–2925.

Mohr, A. E., Ladner, R. E., Riskin, E., *et al.* (2000). Approximately optimal assignment for unequal loss protection. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 1, pages 367–370. IEEE.

Munaretto, D. and Zorzi, M. (2012). Robust opportunistic broadcast scheduling for scalable video streaming. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 2134–2139. IEEE.

Reichel, J., Schwarz, H., and Wien, M. (2007). Joint scalable video model 11 (jsvm 11). *Joint Video Team, Doc. JVT-X202*.

Ross, S. M. (2006). *Simulation, 4th Edition*. Academic Press.

Schierl, T., Schwarz, H., Marpe, D., and Wiegand, T. (2005). Wireless broadcasting using the scalable extension of h. 264/avc. In *Multimedia and Expo, 2005. IEEE International Conference on*, pages 884–887. IEEE.

Schwarz, H. and Wiegand, T. (2007). Implementation and performance of fgs, mgs, and cgs. *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-V126, Marrakech*.

Schwarz, H., Hinz, T., Kirchhoffer, H., Marpe, D., and Wiegand, T. (2004). Technical description of the hhi proposal for svc ce1. *ISO/IEC JTC 1/SC 29/WG 11, Doc. M11244*.

Schwarz, H., Marpe, D., Wiegand, T., *et al.* (2005). Hierarchical b pictures. *Joint Video Team, Doc. JVT-P014, Poznan, Poland*.

Schwarz, H., Marpe, D., and Wiegand, T. (2006). Analysis of hierarchical b pictures and mctf. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1929–1932. IEEE.

Schwarz, H., Marpe, D., and Wiegand, T. (2007). Overview of the scalable video coding extension of the h. 264/avc standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, **17**(9), 1103–1120.

Seferoglu, H., Gurbuz, O., Ercetin, O., and Altunbasak, Y. (2007). Rate-distortion based real-time wireless video streaming. *Signal processing: Image communication*, **22**(6), 529–542.

SMG, E. (1997). Universal mobile telecommunications system (umts); selection procedures for the choice of radio transmission technologies of the umts. *ETSI Document TR*, **101**, 112.

Song, G. and Li, Y. G. (2005). Cross-layer optimization for ofdm wireless networks-part ii: algorithm development. *Wireless Communications, IEEE Transactions on*, **4**(2), 625–634.

Stuhlmüller, K., Färber, N., Link, M., and Girod, B. (2000). Analysis of video transmission over lossy channels. *Selected Areas in Communications, IEEE Journal on*, **18**(6), 1012–1032.

Su, G.-M., Han, Z., Wu, M., and Liu, K. (2006). A scalable multiuser framework for video over ofdm networks: fairness and efficiency. *Circuits and Systems for Video Technology, IEEE Transactions on*, **16**(10), 1217–1231.

van Der Schaar, M. and Sai, S. N. (2005). Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms. *Wireless Communications, IEEE*, **12**(4), 50–58.

Wang, X. and Giannakis, G. B. (2011). Resource allocation for wireless multiuser ofdm networks. *Information Theory, IEEE Transactions on*, **57**(7), 4359–4372.

Wenger, S. and Stockhammer, T. (2005). Rtp payload format for h. 264 video.

Wiegand, T., Sullivan, G. J., Bjøntegaard, G., and Luthra, A. (2003a). Overview of the h. 264/avc video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, **13**(7), 560–576.

Wiegand, T., Schwarz, H., Joch, A., Kossentini, F., and Sullivan, G. J. (2003b). Rate-constrained coder control and comparison of video coding standards. *Circuits and Systems for Video Technology, IEEE Transactions on*, **13**(7), 688–703.

Wien, M., Schwarz, H., and Oelbaum, T. (2007). Performance analysis of svc. *Circuits and Systems for Video Technology, IEEE Transactions on*, **17**(9), 1194–1203.

Wong, I. C. and Evans, B. L. (2008a). Adaptive downlink ofdma resource allocation. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 2203–2207. IEEE.

Wong, I. C. and Evans, B. L. (2008b). Optimal downlink ofdma resource allocation with linear complexity to maximize ergodic rates. *Wireless Communications, IEEE Transactions on*, **7**(3), 962–971.

Zyren, J. and McCoy, W. (2007). Overview of the 3gpp long term evolution physical layer. *Freescale Semiconductor, Inc., White Paper*.