# Robust Predictive Resource Allocation for Video Delivery Over Future Wireless Networks

by

Ramy Atawia

A thesis submitted to the

Department of Electrical and Computer Engineering

in conformity with the requirements for

the degree of Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

August 2017

# Dedication

*To my father Tarek, my mother Maha, my sister Reem, and my wife Mariam*

# Abstract

The promising energy saving and quality of service (QoS) gains of Predictive Resource Allocation (PRA) for video streaming have recently been recognized in the wireless network research community. The PRA relies on future channel conditions to strategically deliver the video content of the mobile users. For instance, the whole video is pushed to the users moving towards the cell edge while prebuffering is postponed for others heading to the cell center in order to minimize the transmission energy. The focus of this thesis is to present a Robust Predictive Resource Allocation (R-PRA) framework to tackle practical uncertainties in the predicted information. In essence, the R-PRA adopts stochastic optimization techniques such as chance-constrained and recourse programming to model the uncertainties in the problem constraints and objectives. Although deterministic convex approximations are feasible, guided heuristic algorithms are introduced to provide real-time allocation. Moreover, Bayesian filtering methods (e.g. Kalman Filter) are adopted to continuously learn the degree of uncertainty which decreases the cost of robustness and maintains the prediction gains. Different variants for the robust framework are proposed such as energy-minimization and predictive adaptive streaming under erroneous prediction of channel rate, user demand and network resources. The variants unleash various design challenges for the network operators such as the trade-off between the complexity of uncertainty modelling and the prediction gains. All the variants are evaluated using a standard

compliant simulation environment that comprises a network simulator 3 (ns-3) integrated with commercial solvers to obtain benchmark solutions. The results demonstrated the ability of R-PRA to meet the QoS level while maintaining the prediction gains over the opportunistic schemes employed in current networks. We believe that this framework set the groundwork for future robust predictive content delivery in which time horizon decisions are taken under practical uncertainties.

# Acknowledgment

In the name of God, the Most Gracious, the Most Merciful, and the Most Compassionate. All praises and thanks are due to God for granting us endless blessings, and the ability to envision, reason, learn, and conduct research. All my work is accomplished only with His guidance and blessing.

I would like to thank my family for their unconditional care, help, love and support. Many thanks to my father Eng. Tarek Atawia for being my role model. Special thanks to my mother Eng. Maha Anwar for motivating me and being always by my side. To my sister Dr. Reem Atawia, many thanks for being my best friend who made me love research and studying. May God bless you and your small family, Dr. Wael Eldahshan and our little princess Noor.

I couldn't thank my wife Eng. Mariam Elazab enough for being a true partner and doing her best to support me during MSc and PhD journeys. May God bless you and grant you happiness, you deserve all the best in this life. I dedicate all my achievements to you.

My sincere appreciation are due to my advisors Prof. Hossam S. Hassanein and Prof. Aboelmagd Noureldin. Your guidance, patience, advice and feedback led to the successful completion of this work. It has been a great pleasure working with you, I owe you a lot.

To my great thesis committee: Prof. Terence Todd, Prof. David Rappaport, Prof. Michael J. Korenberg and Prof. Carlos Saavedra, many thanks for your valuable feedback

and advice.

I would like to thank Queen's University for giving me the opportunity to conduct my PhD studies and supporting me in every challenge I faced. Special thanks to the school of graduate studies (SGS) and the department of electrical and computer engineering (ECE). Many thanks to Ms. Debie Fraser for being very supportive and helpful.

I would like to thank my Professors at the German University in Cairo: Prof. Yasser Hegazy and Prof. Hany Hammad for their constant support. Special thanks to my BSc and MSc supervisors Prof. Tallal Elshabrawy and Prof. Mohamed Ashour for their patience and effort in teaching me how research should be done. Many thanks to Prof. Mustafa Amer (may your soul rest in peace), I owe you all the mathematics in this thesis. Many thanks to Dr. Hamed Salah for inspiring me to pursue postgraduate research.

To my great friend, brother and mentor, Dr. Hatem Abou-zeid. This work wouldn't have been done without your guidance, advice and effort. I want to thank you for inspiring and encouraging me to reach beyond my limits. You, Dr. Hafsah and Amelle deserve all the best, may God bless you.

Many thanks to all the people who helped me during my very first days in Canada till the moment I am writing this last page. Special thanks to Prof. Khaled Elgazzar, Prof. Abdelhamid Taha, Dr. Khaled Hayajneh, Dr. Ahmed Medallal and Prof. Ayman Elmanasry.

To my best friend and brother Dr. Hesham Farahat, many thanks for always having my back during the last four years. This research couldn't have been done without your help in ns-3 and Linux, beside all other challenges I faced. Mariam and myself had a pleasant time in Kingston mainly because of you and Rana, may God grant you both, and your little princess Joana happiness and joy.

To my friend and brother, Anas Mahmoud, I enjoyed all our interesting and inspirational discussions. Many thanks for all your help and words of wisdom. Special thanks to my dear friend Prof. Sharief Oteafy for his priceless help and advice. You and Dr. Layan deserve all the best, may God bless you and your little angel Reem. I would like to thank my friend Dr. Mohamed Moussa for being there for me. Wishing you, Mai, Omar and Ziad all the best. Many thanks to Dr. Mohamed Hosny and Dr. Gehan Selim, wishing you and your cutest Lara the best of this life. To my awesome friend Abdalla Abdelrahman, many thanks for your help in moving and printing the thesis. May God bless you, Tasneem and our sweetest Deema.

Many thanks to all my friends in the Telecommunication Research Lab (TRL): Wenjie Li, Mohamed Adel, Dr. Shereen Abd El-Hameed, Dr. Abdallah Alma'aitah and Mohannad Alswailim. Special thanks to Mrs. Basia Palmer for your unconditional help and endless effort to handle all our concerns.

To my lifetime friends and brothers: Mohamed Saeed, Omar Khayrat, Sherif Ashraf and Ahmed Elsahhar, thank you so much for your support and motivation during the 20 years of our friendship. All the rewards of this thesis are dedicated to our friend Seif Eldin Nabil (may your soul rest in peace), till we meet again, I will keep praying for you.

Many thanks to my friends Mohamed Refaat and Mohamed Fathy for all your help. I would like to thank all the people who significantly shaped my personality: Capt. Ahmed Salah, Mr. Ahmed Shams, Eng. Mohamed Ghanem and Eng. Ayman Beiram.

Many thanks to my friends Ahmed Moheeb, Ahmed Nafe, Abdullah Mostafa, Ehab Mohamed, Karim Fathy, Ahmed Magdy, Ahmed Khaled and Maged Wageeh for all the support and help they provided to me.

*Ramy Atawia, Kingston, Ontario*

# Co-Authorship

The work in Chapter 4 resulted in [1, 2], Chapter 5 led to [3, 4]. Similarly, part of Chapter 6 is published in [5–7] while Chapter 7 appeared in [8, 9]


[1] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, pp. 4683–4688, Dec 2014

[2] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Chance-constrained qos satisfaction for predictive video streaming," in *Proc. IEEE LCN*, pp. 253–260, 2015

[3] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE J. Select. Areas Commun.*, vol. 34, pp. 1389–1404, May 2016

[4] R. Atawia, H. S. Hassanein, H. Abou-zeid, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Wireless Commun.*, pp. 1–14, 2017

[5] R. Atawia, H. Hassanein, and A. Noureldin, "Energy-efficient predictive video streaming under demand uncertainties," in *Proc. IEEE ICC*, pp. 1–6, May 2017

[6] R. Atawia, H. S. Hassanein, N. A. Ali, and A. Noureldin, “Robust content delivery and uncertainty tracking in predictive wireless networks,” *IEEE Trans. Green Commun. Netw.*, pp. 1–14, 2017

[7] R. Atawia, H. Hassanein, and A. Noureldin, “Fair robust predictive resource allocation for video streaming under rate uncertainties,” in *Proc. IEEE GLOBECOM*, pp. 1–6, Dec 2016

[8] R. Atawia, H. S. Hassanein, and A. Noureldin, “Robust long-term predictive adaptive video streaming under wireless network uncertainties,” *IEEE Trans. Wireless Commun.*, pp. 1–14, 2017

[9] R. Atawia, H. Hassanein, and A. Noureldin, “Optimal and robust qos-aware predictive adaptive video streaming for future wireless networks,” in *Proc. IEEE GLOBECOM*, pp. 1–6, Dec 2017

# List of Abbreviations

**CDF** Cumulative Density Function

**CCP** Chance Constrained Programming

**JCCP** Joint Chance Constrained Programming

**PA** Power Amplifier

**REM** Radio Environment Map

**GPS** Global Positioning System

**MCS** Modulation and Coding Scheme

**RAN** Radio Access Network

**BS** Base Station

**RP** Recourse Programming

**SINR** Signal to Interference plus Noise Ratio

**CQI** Channel Quality Indicator

**DASH** Dynamic Adaptive Streaming over HTTP

**KF** Kalman Filter

**QoE** Quality of Experience

**RAN** Radio Access Network

**QoS** Quality of Service

**RA** Resource Allocation

**R-PRA** Robust Predictive Resource Allocation

**PRA** Predictive Resource Allocation

**QoS** Quality of Service

**TTI** Time Transmission Interval

**LTE** Long Term Evolution

**BA** Bernstein Approximation

**GA** Gaussian Approximation

**SA** Scenario Approximation

**PF** Particle Filter

**SoCP** Second order Cone Programming

**PDF** Probability Density Function

**ns-3** Network Simulator 3

**MGF** Moment Generating Function

**MIMO** Multiple Input Multiple Output

**MOS** Mean Opinion Score

**P-DASH** *Predictive*-DASH

**RP-DASH** *Robust Predictive*-DASH

**BER** Bit Error Rate

**R-GPRA** *Robust*-Green Predictive Resource Allocation

**SAA** Sample Average Approximation

**IPM** Interior Point Method

# List of Symbols

$\mathcal{T}$ Prediction window

$t \in \mathcal{T}$ Time slot index

$\mathcal{M}$ An active user set

$i \in \mathcal{M}$ An active user index

$\tilde{r}_{i,t}$ Predicted erroneous rate for user $i$ at time slot $t$

$\bar{r}_{i,t}$ Average rate for user $i$ at time slot $t$

$\sigma^2_{i,t}$ Prediction error variance (Calculated Off-line)

$\bar{\sigma}^2_{i,t-1}$ Measured rate error variance by user $i$ during previous time slot $t-1$

${\sigma^2_{i,t+\delta t}}^+$ Estimated rate error variance

$[r^l_{i,t}, r^u_{i,t}]$ lower and upper bounds of predicted erroneous rate

$x_t$ Resource allocation variable at time slot $t$

$x_{i,t}$ Airtime fraction for user $i$ at time slot $t$

$v_{i,t}$ Streaming rate at time slot $t$ requested by user $i$

$D_{i,t}$ Cumulative user $i$ demand at each time slot $t$

$\tilde{D}_{i,t}$ Erroneous cumulative user $i$ demand at each time slot $t$

$\tilde{C}_t$ Uncertain resources for real-time users at each time slot $t$

$Q$ Set of possible segment qualities

$q \in Q$ Video segment quality level

$\kappa_{i,t}^{(q)}$ Binary decision variable, equals $1$ if quality $q$ is selected to user $i$ at time slot $t$

$\beta$ Quality of Service (QoS) satisfaction Level

$\epsilon$ Maximum allowed QoS degradation ($= 1 - \beta$)

$\zeta_{i,t}$ Per slot risk probability

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

### 1.1.1 Evolution of Mobile Video Traffic

Mobile phones and data applications are undergoing a constant development that drives forces for cellular network expansion. As expected, the number of mobile devices has increased exponentially over the last decade and already surpassed the world's population in 2014 with a total of 7.4 billion devices [11]. Such growth is expected to continue in the next few years reaching 11.6 billion by 2021. In addition, the upsurge in multimedia services and social networking applications, among others, will cause an exponential increase in total wireless data traffic of 49 Exabytes per month in 2021. This will put network operators under huge pressure as they strive to manage user experience with minimal capital and operational expenditures.

Concurrently, mobile video traffic is experiencing substantial growth as more than $78\%$ of the global mobile data traffic is expected to be video content in $2021$ [11]. This is attributed to the high bit rates required by video content compared to other data applications.

Ongoing development of mobile devices, streaming servers such as YouTube, and adaptive streaming protocols are supporting the availability and delivery of video content at different quality levels that improve the user experience [12, 13]. This large volume of traffic, however, must be delivered to users at a certain quality of service (QoS) level, e.g. maximum delay and service interruptions, using the available resources. To that end, the cellular operators focus on Resource Allocation (RA) that provides proficient usage of available network resources such as the licensed spectrum and access nodes.

### 1.1.2 Challenges and Ongoing Efforts

Among the network elements, the Radio Access Network (RAN) accounts for more than 50% of the network energy consumption [14]. As such, designing novel energy-efficient RAN frameworks is paramount to reducing the network carbon footprint while satisfying the increasing Telecom market demands. This includes techniques such as efficient Power Amplifier (PA) design [15], cell switch off [16, 17], and traffic-aware scheduling [18], among others.

A more efficient RAN is also beneficial for operators as it can postpone investment in equipment installations and new spectrum. Thus, in addition to minimizing the energy-related operational expenditures (OpEx), the capital expenditures (CapEx) can also be reduced since radio equipment installations can make up to 70% of CapEx [19]. To address these recent developments, *energy-efficient* RA schemes for wireless video streaming are gaining momentum. Such schemes are also important for future wireless paradigms such as Vehicular Ad-hoc Networks (VANETs) in which energy saving remains a challenge [20, 21].

Another advancement in video streaming protocols is the adaptive selection of quality

(i.e. video definition) [13, 22]. Dynamic Adaptive Streaming over HTTP (DASH) refers to one type of these protocols which has been standardized in the 3GPP [23]. Each video file is encoded at multiple bit rates within the server, and thus enables channel aware quality selection. This selection is currently user-driven, yet increases the risk of buffer underrun and video stops when users greedily request high bitrates that require more resources than the amount calculated by the resource allocator. Hence, a shift towards selection becoming network-centric is getting attention in current research thereby to include the decisions of radio resource allocator especially in multi-user scenarios [24]. In essence, DASH schemes aim to maximize the Quality of Service (QoS) by minimizing the number and durations of video stops, and initial buffer delays while maximizing the video quality measured by the bitrate [13].

These stringent requirements on energy consumption and QoS necessitate novel design of RA schemes to optimally calculate the resources and select the video quality. The predictability of user's behaviour and mobility, and wireless channels enabled a new paradigm referred to as Predictive Resource Allocation (PRA) [25–30]. Extensive network measurements demonstrated the predictability of users' behaviour up to 93 % [31], including human mobility and activity [32]. Meanwhile, the radio signal strength and available bandwidth are found to follow repetitive spatio-temporal patterns [33–35]. The availability of navigation systems (e.g. Global Positioning System (GPS)) at current user devices has enabled mobile operators to correlate the radio measurements (e.g. channel rates) with geographical locations, and constructs the Radio Environment Map (REM) [36].

PRA that exploits these patterns of signal strength and mobility prediction over a time horizon has recently been recognized as a promising approach to improve video streaming QoS [26,37,38], and minimize transmission energy [25,27,29]. In essence, the PRA avoids

allocating resources to users during poor radio conditions, that consume more airtime per byte, while maximizes the allocation during peak conditions by leveraging the content availability and prebuffering capabilities at the Base Station (BS) and user devices. To derive performance gains over non-predictive schemes, the PRA literature [25–29] assumed perfect prediction of future information. However, real-world uncertainty should be taken into consideration to support the implementation of PRA in practice. Prediction techniques typically rely on real-time channel measurements characterized by spatio-temporal variations [33]. This is in addition to adopting low-cost and low-power filters at user devices which decreases the prediction accuracy over the time horizon. Nevertheless, dynamics in the environment will result in changes of user behaviour, mobility and demands which make perfect prediction infeasible. All these sources of uncertainties prompt a change in the PRA design to achieve a *robust* solution that guarantees QoS satisfaction and maintains the reported prediction gains.

## 1.2 Objective and Thesis Contribution

In this thesis, we address the problem of imperfect predictions and handle the resultant uncertainties to limit their impact on the PRA performance. The main focus is on the following research questions:

What is the impact of information uncertainties on the prediction gains?
How to develop a *robust*-PRA scheme to model and handle these uncertainties?
What is the cost of *robustness*?

The first question aims to quantify and analyze the impact of uncertainties on the user satisfaction and prediction gains while adopting existing PRA under typical error models.

The second question is related to introducing a novel PRA design that is *robust* to errors in the predicted information. Finally, the third question will assess whether the reported performance gains in the PRA literature are still attainable by the *robust* forms. We believe this work provides a practical direction towards the development of deployable PRAs in future generation networks.

We summarize the contributions of this thesis, to tackle the above questions, as follows:

- We propose, for the first time in literature, a *Robust* Predictive Resource Allocation (R-PRA) framework that handles prediction uncertainties over a time horizon through *probabilistic* modelling, *stochastic* optimization, *Bayesian* learning, and guided heuristic search. The framework comprises the following main stages:
  - Modelling the future information as random variables in order to capture the impact of prediction errors. This is unlike the existing PRA approaches that adopts the average values of predicted information and ignored their variations and uncertainties.
  - We adopt *stochastic* optimization techniques such as Chance Constrained Programming (CCP) and Recourse Programming (RP) to limit the degree of violation in QoS constraints and minimize the expected loss in network gains, respectively. Such probabilistic modelling allows the framework to strike a balance between providing high network gains when predictions are accurate, and minimizing the risks associated with erroneous predictions during periods of uncertainty. Unlike traditional *robust* optimization, new models are proposed here to capture the interdependency between the time slot decisions and guarantee *joint* QoS satisfaction over the time horizon.

  - The main challenges in such probabilistic model is the lack of non-closed form solution. A deterministic equivalent formulation is therefore derived using the statistics of predicted information such as variance and Probability Density Function (PDF). Thus, a tractable solution can be obtained and solved by commercial solvers.
  - Although the statistics of random variables can be calculated off-line, radio measurement studies reveal that the degree of predictability varies significantly with geographical location and time of day [33]. Therefore, a mechanism to *track* the uncertainty level in predicted information is proposed for a practical solution. This is as opposed to the *stochastic* literature in which the uncertainty level was constant and thus provided suboptimal or non-robust decisions when the degree of predictability varies over time.
  - We propose a low-complexity guided heuristic search algorithm to obtain real-time solutions for the deterministic equivalent formulation. Although the formulated model is convex and can be optimized by commercial solvers, real-time solutions are not attainable by conventional numerical methods whose complexity increases with the time horizon length and number of users.

- We propose four variants of the R-PRA framework for video streaming under different network objectives and sources of uncertainties summarized as follows:
  - **Energy-efficiency under Gaussian uncertainty**: We introduce a novel model for video streaming QoS over a time horizon that accounts for uncertainty in the predicted user rates. Herein, the objective is to minimize BS energy consumption while guaranteeing a long-term QoS. As recent practical and theoretical

findings indicate that the variations in predicted rates can be modeled as multivariate Gaussian random numbers [34], we employ probabilistic Joint Chance Constrained Programming (JCCP) to formulate the problem mathematically. We then show that the resultant formulation is non-convex and apply proportional risk allocation for joint chance constraints. The problem is decomposed into two convex sub-problems, where the first stage optimizes the *individual* risk levels at each time slot, which are subsequently used by the second stage to solve the robust RA problem. By applying such a *non-uniform* risk allocation, we generalize the solution to achieve less conservative (i.e, energy-efficient) and more practical QoS aware RA decisions. We develop an efficient low complexity guided search heuristic that guarantees the satisfaction of joint QoS levels. Due to the inconsistency in the rate variance over time and location, we adopt Kalman Filter (KF) to accurately track such variations, providing an additional degree of robustness to the statistical parameters. With such a framework, QoS guarantees can be ensured during high variance while energy minimization is achieved during low varying cases.

– **Energy-efficiency under Generic uncertainty**: Unlike the first variant, this one provides a solution that is not dependent on a particular error Probability Density Function (PDF) in order to save the cost of error modelling. We adopt the Bernstein Approximation (BA) which only requires error bounds to satisfy the QoS constraint. Under such uncertainty model, we also demonstrate how a Particle Filter (PF) can be adopted to effectively achieve the channel tracking functionality, and adapt the BA rate bounds. Finally, we present a guided heuristic algorithm based on local search to provide a real-time solution for the

BA formulation.

– **Energy-efficiency under Demand and Resource uncertainty**: While the first two variants tackle errors in predicted rate, we capture here uncertainties in both the demand and radio resources. The model relies on Recourse Programming (RP) to consider the risk of wasting resources due to users terminating the video session before watching the prebuffered content [39, 40]. Similarly, a CCP is adopted to control the QoS degradations under resources fluctuations due to the random arrival of real-time traffic. The deterministic equivalent is derived using the PDF of video watching durations to quantify both the possibility of energy-saving and the risk of wasting resources. Similarly, the PDF of users arrival and their traffic load are used to obtain a deterministic form for the CCP model. The proposed guided heuristic algorithm allows the network to prebuffer future demands with high likelihood of watching, and delay the delivery of upcoming uncertain content, while accounting for the fluctuations in the network resources. In addition, the trade-off between energy-savings and the risk of QoS violation during resources uncertainty is modelled and ensures that the QoS degradations does not surpass predefined level in CCP.

– **QoS-Aware DASH under Rate Uncertainty**: Unlike the previous variants that solve only for resources at a fixed quality, to save energy in low load scenario, this last approach seeks joint optimization of radio resources and video quality selection to maintain prediction gains in high load scenarios. The main objective is to achieve *long-term* quality fairness among users over the time horizon while avoiding the video stops due to buffer underrun. Unlike non-predictive counterparts, the proposed approach allows the network to prebuffer upcoming

video content in high quality to users with poor future rates. The deterministic equivalent of CCP is based on Scenario Approximation (SA) that adopts the discrete PDF of predicted rates. As the decision is taken over a time horizon and for both resources and quality, conventional SA results in a combinatorial complexity. As such, we introduce a linear approximation to aggregate the dependency between the time horizon constraints which reduces the formulation to a polynomial model. While SA provides benchmark solutions for the robust approach, mobile operators strive to minimize the effort of obtaining the discrete PDF. Hence, we propose a second deterministic model based on Gaussian Approximation (GA) that only require the variance and the inverse Cumulative Density Function (CDF) of predicted rate. We also propose a low-complexity guided heuristic search algorithm to obtain real-time solutions for the deterministic GA formulation.

- We evaluate the performance of all proposed variants to unleash the impact of uncertainties and robustness on the reported prediction gains. The evaluation framework is summarized as:

  – We modify the scheduling module in a Long Term Evolution (LTE) standard compliant network simulator (ns-3) and integrate it with optimal solvers such as MATLAB and Gurobi to evaluate the proposed algorithms and state of the art solutions.

  – Typical error models, reported in the literature based on measurement campaigns, are adopted by the simulator to perform sensitivity analysis and assess the performance gains.

– New performance metrics are defined to quantify and model the trade-off between the network and QoS gains. In particular, *Cost of Robustness*, *prediction gains*, optimality gaps and complexity, among others, are examples of such metrics that help operators in measuring the rewards of Robust Predictive Resource Allocation (R-PRA).

## 1.3 Organization of Thesis

The thesis is organized as follows:

In Chapter 2, we provide a background on PRA and review the state of the art. In addition, we discuss the sources of prediction uncertainties which were overlooked in PRA literature and review the resulting limitations.

In Chapter 3, a background on both *Robust optimization* and *uncertainty tracking* techniques is provided. The focus is on *Stochastic* optimization and the deterministic equivalent forms. In addition, Bayesian inference techniques used in this thesis will be also reviewed.

In Chapter 4, our general *Robust*-PRA framework is proposed and the main building blocks are summarized. This is in addition to the system model and a Monte-Carlo framework for estimating the statistics of prediction errors.

In Chapter 5, Chapter 6 and Chapter 7 we propose the different variants of the robust framework. Each variant contains a problem description, system model, mathematical formulation, heuristic search, and simulation results.

Finally, in Chapter 8, we summarize and conclude the thesis, and highlight the main findings of this work. Future directions are then recommended to support the momentum of implementing the PRA in next generation wireless networks.

# Chapter 2

# Existing Predictive Resource Allocation (PRA)

## 2.1 Conceptual Overview

Today's wireless networks adopt opportunistic resource allocation schemes based on reported measurements from the user devices [41–43]. The channel conditions at each user device are reported periodically in the form of Channel Quality Indicator (CQI) which guides the network to select the appropriate Modulation and Coding Scheme (MCS). For instance, users experiencing poor channel conditions, i.e. low Signal to Interference plus Noise Ratio (SINR), due to low signal strength or high interference will report low CQI values. As a consequence, the network will select a low order MCS that is robust to such low SINR values and thus user can receive and decode his content at a target Bit Error Rate (BER) value. Although such adaptive transmission provides optimal utilization of radio resources, it poses new challenges to network operators while designing the strategy of opportunistic resource allocation. Each user experiences different radio conditions and

served by an MCS that differs from other users. Thus, optimizing network resources such as minimizing energy consumption or maximizing bandwidth utilization while achieving fair QoS among users is not attainable by existing opportunistic resource allocation schemes. Achieving fair QoS would typically result in allocating more resources to users with poor conditions (i.e. low MCS), yet this increases the energy consumption and minimizes the bandwidth utilization. Future wireless networks as such should employ a new paradigm that handles the conflict between network objectives and QoS requirements.

*Predictive* resource allocation (PRA) has recently been recognized as a promising approach to improve the resource utilization for video content delivery [25–27, 37, 38, 44, 45]. This is accomplished by leveraging the knowledge of the future link capacity users are expected to experience, and then performing *long-term* predictive RA plans over several seconds. By doing so, BSs can prioritize users heading to poor channel conditions (i.e. low MCS), or delay transmission until a user reaches better channel conditions (i.e. high MCS). Prioritizing users allows the BS to prebuffer the future content and thus maintains the target QoS, while delaying the transmission results in optimal bandwidth utilization. Stored video content such as YouTube and Netflix is well suited for such approaches as it can be strategically prebuffered and cached locally at the mobile device. In-network caching enables the content availability at the BS under user mobility [46–50].

## 2.2 Mobility and Channel Prediction

Radio signal measurement studies indicate that cellular network users moving along the same path will typically experience similar signal strength variations as reported in [33, 44]. The PRA relies on long-term prediction of future channel conditions over a duration that extends to some tens of seconds and may be minutes. In that period, user locations

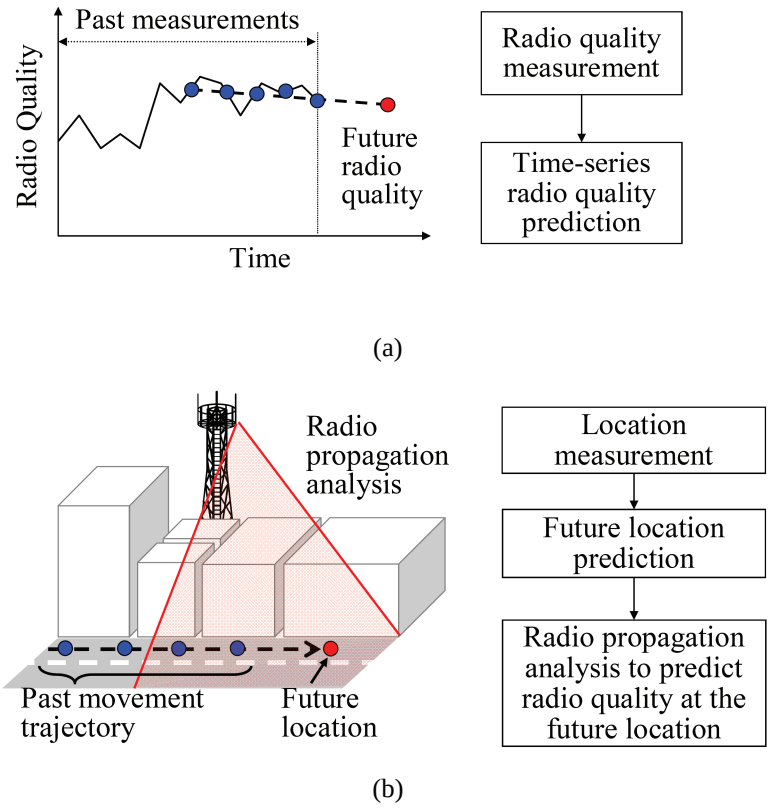



Figure 2.1: Illustrative example for the a) regression and b) location based predictions [10]

significantly change the channel conditions which remove the correlation between future and past signal samples. This makes traditional regression methods [51], Fig. 2.1 (a), unsuitable for PRA, as such the location based prediction in Fig. 2.1 (b) is typically used [10]. This technique relies on the user's future location, mobility and motion behaviour for modelling the upcoming large scale shadowing. This enabled the estimation of radio conditions in urban, suburban or rural areas which are characterized with time disjointed measurements and high user mobility as illustrated in Fig. 2.1.

The Radio Environment Map (REM) has been introduced as a main building block of the location based channel prediction. In essence, REM is a database that stores the users'

reported channel measurements at different locations in the network. These measurements will be used afterwards to retrieve either the received power or rate values at given location. The REM was firstly developed in cognitive radio networks to store medium access information and statistics of users to control the spectrum usage [52]. In cellular networks, the REM is also exploited to detect coverage holes [53] and construct automatic neighbour relations [54] without the need of manual drive test. The REM is frequently updated by the users' measurement reports according to the 3GPP Drive Test Minimization Standard [55] which enables autonomous construction of REM and its application in PRA.

**REM Construction**

The REM construction undergoes two main stages: user location estimation and RAN measurement collection [36].

- User location estimation

  Current 3GPP LTE standard adopts the Evolved Serving Mobile Location Centre (E-SMLC) to localize the user device upon request from RAN [55]. The E-SMLC calculates the position using one or a combination of the available localization systems at the user device.

  In [56], cellular network positioning was adopted based on: 1) observed time difference (OTDs) between consecutive messages at the BSs , and 2) relative time difference (RTD) between BSs. The scheme has four main inputs: 1) measurement report message (MRM) that contains channel measurements (e.g. signal power) along with time stamp, 2) cell configuration which comprises the base station physical parameters (e.g. location), 3) Round trip time (RTT) and 4) Elevation data and thus only 2-D positioning is done using RTT or the difference between the trigger and arrival

of the MRM. The RTD is then calculated based on both the OTD received at two BSs and the propagation delays. The resulted position is further filtered using KF to remove incorrect positions relative to the average pedestrian or vehicular speed. The filter prevents a sudden increase or decrease in the velocity and thus eliminates disruptions in the estimated user location over short time interval.

Other approaches used satellite signal to calculate more accurate user locations by leveraging the GPS in today's user devices [36, 55]. Moreover, the GPS can be also integrated with other localization systems that have complimentary features to improve the positioning accuracy. Integration can take place in a loosely coupled form [57, 58] in which two positions are obtained, one from GPS and another from LTE reference positioning signal, and then aggregated to one final position. Aggregation is done by weighting both positions based on the trustiness of each localization system. Another example of the tight integration is the assisted GPS (A-GPS) which is currently implemented in user devices. In particular, the device can communicate with the cellular network to acquire the available satellite information at the BS. This speeds up the satellite signal acquisition at the mobile device, thus saves energy and decreases latency.

- RAN measurement collection

  Network information such as load and interference between different BSs are collected and stored in the REM. Such measurements are used to create and update the REM in one of two ways: Pixel Update or Propagation Model Tuning. In the former, the REM is represented as a geographical area, divided into square grids, and the reported user device measurements are mapped to the nearest grid. In the second type, the reported measurements are used to tune a selected empirical path

loss model. Tuning is done periodically through calculating correction variables that minimize the difference between the measurements and the model based values. The tuned model can then be used to estimate the received power and interference in all the geographical locations.

### Mobility Prediction

After constructing the REM, the network will predict the future user locations based on both current user location, velocity and routine.

The user mobility behaviour is classified into two types: macroscopic and microscopic. The former includes the daily activities such as going from home to office or from desk to a meeting room. In the microscopic behaviour, the motion is restricted to certain locations such as the office locations or corridors, in case of indoor, or defined routes in case of outdoor or road network [59]. Moreover, the human velocity is highly predictable either as a pedestrian or a driver. The velocity is probably 2 m/s in the former case, while in the second case it depends on the road information. Such motions follow a pattern that can be used to estimate the future mobility traces.

The vehicle trajectory can be mainly predicted using information about: vehicle, environment and driver [60]. The vehicle's velocity, acceleration and angular speed can be used for providing a short term prediction of user's location. On the other hand, the environment information can provide a longer term prediction for the user's trajectory.

### REM Based Channel Prediction

After calculating the anticipated user locations, the corresponding future channel rates can be retrieved from the REM either directly or through geographical/spatial interpolation

techniques. These techniques are fitting methods used to complete a 2-D surface by interpolating the missing points using the stored values in the REM. Different interpolation functions are used to model the relation between the points comprising the same curve. Surveys on different methods, their accuracy and complexity can be found in [61, 62]. One possible classification is in [63] which proposes three interpolation categories: Local Neighbourhood, Geostatistical, and Variational.

In the first category, the interpolated data is a weighted sum of the surrounding neighbourhood measurements. Among its types are: Inverse Distance Weighted, Natural Neighbour Interpolation [64] and Triangular Irregular Network [65]. Inverse Distance Weighted assumes that near points are more correlated than the far ones. Accordingly, the location with missing measurements is predicted (interpolated) as a weighted sum of the surrounding measured points, each one is weighted by the inverse of its distance. In Triangulated Irregular Network, triangles are formed such that the circumcirlce of each triangle should contain a maximum of one measurement point. The three vertices of the same triangle are chosen such that the smallest angle in all triangles is maximized [66]. The geostatistical interpolation technique is based on the channel statistics that model the randomness and uncertainties in the measurements. The most commonly known method is called Kriging [66, 67] that guarantees the minimum mean square error. The method constructs an empirical semivariogram that uses the semivariance to reflect the spatial correlation between the different points. A theoretical semivariogram model (e.g. exponential, Gaussian or spherical) is then selected to approximate the empirical model using appropriate fitting technique such as least square method [68]. The variational interpolation introduces a smooth small varying function called Splin. The most well-known technique of this class is the Thin Plate Splin (TPS) which uses a radial basis function centred at every measurement

and the point to be predicted [67].

## 2.3 PRA Schemes and Potential Gains

Under perfect knowledge of the future network conditions (i.e. error-free REM), the PRA techniques in [25–29, 38] demonstrated how the total BS energy can be significantly reduced, compared to opportunistic allocation, without any buffer underrun at the user device. In [26,69], the PRA achieved long-term QoS fairness over the time horizon resulting in uniform user experience. Moreover, the PRA was extended to *Predictive*-DASH (P-DASH) in order to jointly select the video quality and resources devoted to users over the time horizon. Thus, maximizes the total quality for each user during the streaming session and minimizes the total BS energy [28].

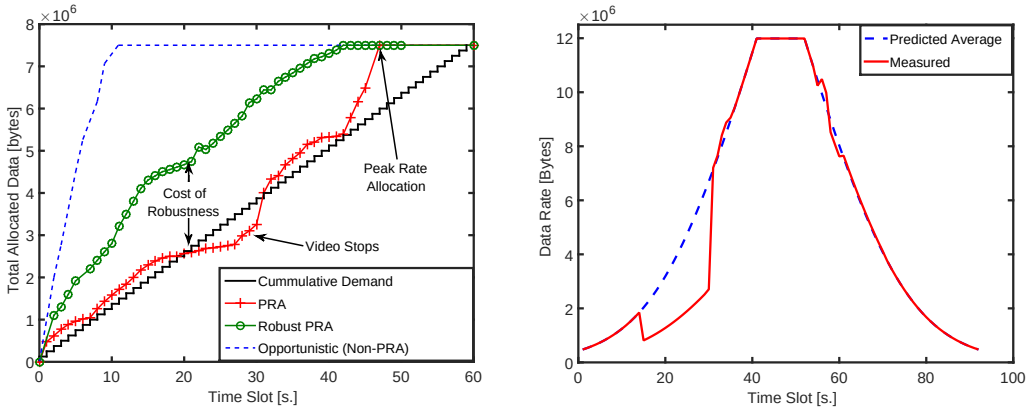
### 2.3.1 Energy Savings

The first gain achieved by PRA is the minimization of total energy consumed by both the BS and user device in transmitting and receiving the video content, respectively. The PRA work in [25, 27, 29, 37, 38] has focused on energy minimization under QoS constraints. In particular, the QoS level is said to be satisfied when the video is played back without stops. Quantitatively, this is achieved when total amount of data delivered to the user at a certain time slot is not less than the *cumulative* demanded data at a fixed streaming rate to avoid buffer underrun. In order to achieve energy savings, the total airtime allocated to the users over the time horizon has to be minimized.

When the user experiences poor radio conditions, e.g. near the cell edge, the BS will adopt low order MCS. This results in low transmission rate that consumes more resources

per bit and increases the energy consumption. As such, the PRA will devote the bare minimum amount of resources to the user such that the video does not freeze. The BS can go into sleep mode, to minimize energy, or allocate the remaining amount of resources to other users. On the contrary, the BS waits until this user reaches his peak radio conditions, e.g. near the cell center, to leverage the high order MCS. The attained peak transmission rates motivate the BS to allocate large amount of resources and exploit the storage capabilities in user devices by transmitting large portion of the video. Thus the whole video can be delivered before the user experiences poor conditions in the future. The BS can also go into sleep mode while the user plays back the prebuffered content in the future. Such strategy allows the PRA to transmit the video content with fewer resources compared to the traditional opportunistic RA technique. The latter overlooks the future radio conditions and thus neither delays prebuffering, for cell edge user, nor prioritizes users at the cell center experiencing peak radio conditions.

An example of such an energy-efficient PRA is illustrated in Fig. 2.2(a). In that example, the user started moving from the cell edge at t = 0 experiencing the lowest channel rate as shown in Fig. 2.2(b). This user is also expected to move towards the cell center reaching the peak channel rate at t = 40. With these future rates in mind, the PRA will strategically serve the user with the minimum airtime to barely satisfy his demand. This allocation will guarantee an optimal balance between QoS satisfaction and energy consumption. Allocating less airtime will result in video stops, while more airtime increases the energy consumption. The PRA adopts this strategy until the user reaches the peak channel conditions at t =40 where the video is prebuffered by maximizing airtime allocated to that user. The main aim of this greedy allocation is to download the whole video before the user leaves the cell center and reaches the poor radio conditions again at t > 50. As




(a) Examples of resource allocation strategies. (b) Examples of predicted rate variations.

Figure 2.2: Illustration of PRA for energy saving under QoS satisfaction

opposed to PRA, opportunistic (Non-PRA) that is unaware of the future peak rates will greedily allocate the whole airtime to the user at the early time slots (near t = 0) resulting in more energy consumption.

### 2.3.2 Long-Term QoS Fairness

The second gain achievable by PRA is the long-term fairness for video streaming users. While energy-saving can be attained during low load scenarios, fair QoS among users can be accomplished during high load scenarios. This includes strategic allocation of video freezes [26, 69] and selection of video quality over a time horizon [37, 70].

Similar to the energy saving PRA techniques, the predictive fair resource allocation was introduced in [26, 69] in which the future rates are exploited to prioritize users. In particular, the predictive proportional fair (PPF) scheduler in [26] considers distributing all the available resources (i.e. airtime) among the users proportional to their anticipated channel rates. Thus, a user experiencing his peak rates and moving towards location with poor radio conditions shall be prioritized. More resources are allocated to that user to

prebuffer a large amount of video which can be watched during future poor conditions. Avoiding allocation during poor conditions will save resources that can be utilized by other users. In particular, lower priority is assigned to users located in bad radio conditions (might experience video stops), yet will be prioritized later when they reach their peak radio conditions.

Such allocation is similar to the opportunistic non-predictive proportional fair scheduler. However, the gain of PPF is emphasized when users are experiencing similar data rates at the same time but their future rates are different. Thus, users with low rates in the future will have a higher priority than the other users with high rates, although both are currently experiencing the same peak rates. As a result, optimal resource utilization and fairness are achieved by the PPF compared to the non-predictive scheme. Moreover, other objective functions that consider fairness such as max-min, $\alpha - fair$ and Jain's index [69] can still be applied to achieve similar gains as the above-mentioned PPF.

### 2.3.3 Maximizing DASH Quality

Dynamic Adaptive Streaming over HTTP (DASH) was essentially introduced to improve the user experience and resource utilization under wireless channel fluctuations [12]. The video file is split and delivered in the form of small segments where the quality of each segment is adapted proportionally to the user's channel condition. In particular, low-quality segments are selected when the user is experiencing low channel rates (e.g. user at the cell edge) in order to avoid video freezes. On the contrary, high-quality segments are delivered when peak channel rates are observed (e.g. user at the cell center) to exploit the available radio resources and improve the user's experience.

The original DASH protocol relies on user device, aware of available video bitrates,

and the channel conditions to select the segment quality and request it from the streaming server. Such user-centric approach, however, is unaware of the total network load and other users demands which are considered by the resource allocator. Therefore, a user might select a high-quality level, due to the measured high channel rate, although the network resource allocator will not necessarily devote the whole radio resources to that user in the next time slot. Such limited resources, selected by resource allocator, might not be sufficient to deliver the high-quality segment, requested by the user, and thus increases the risk of video freeze [24].

Research efforts are currently concerned with shifting the DASH from a user-centric decision to a network driven decision in order to bridge the gap between the decisions of individual users and the resource allocator [71]. To that end, the network jointly optimizes the segments qualities and the resource sharing among the users. Thus, avoids the greedy quality selection by the users when they overestimate the available radio resources. Different implementations of network-centric DASH, with minimal changes to the current user-centric strategy, were proposed in [24, 71]. At the BS, the resource allocator overwrites the user's requested quality before forwarding it to the server [71]. Recent BS storage capability provides another implementation flexibility where the video is locally cached with different quality representations, and the segments are sent at the resource allocator's quality.

Conventional network-centric DASH [72–74] adopts recent channel measurements, reported by the user's device, to opportunistically optimize the network resources (e.g. resource utilization) and QoS (e.g. quality and interruptions). Each user reports the current channel conditions to the network which in turn calculates both the segment's quality and the amount of resource share for each user at a certain time slot. These reactive decisions

only achieve local optimal network performance without QoS guarantees due to overlooking the users' future radio conditions. *Predictive*-DASH (P-DASH) [25, 26, 28, 44, 70], oppositely, relies on future radio conditions to derive long-term policy while allocating the current resources. For example, two users at the cell center (i.e. good radio conditions), one is heading towards the cell edge while the other will stay for longer time. As the former experiences poor radio conditions in the future, the resource allocator must prioritize this user by allocating more resources during peak radio conditions. Long-term fairness with regards to quality and stalls can be achieved by either prebuffering the future content or increasing the current video quality. This strategy allows the user to stream the prebuffered high quality content during poor radio conditions resulting in higher fairness. On the contrary, users in poor conditions and moving towards high channel rates will be allocated a small amount of resources until reaching their peak conditions.

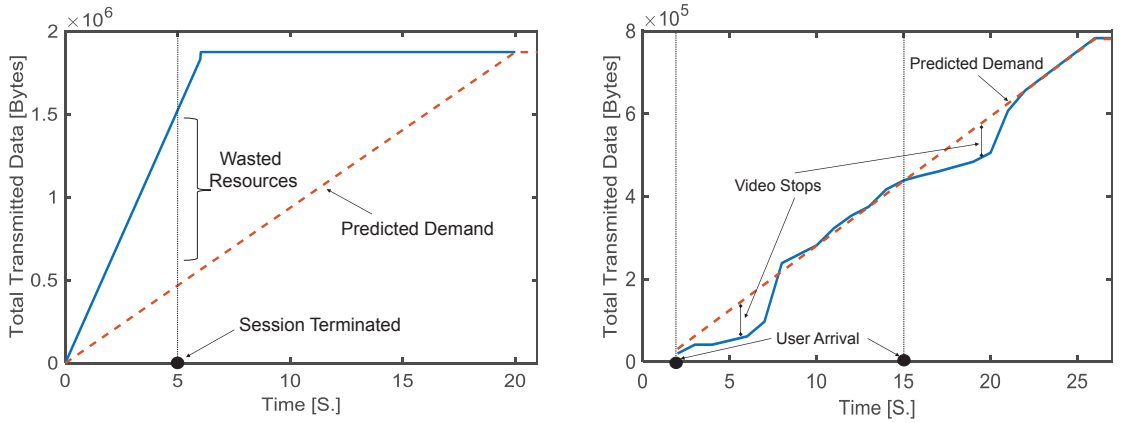
## 2.4 Sources and Limitations of Prediction Uncertainty

In wireless medium, channel rates and network resources can not be perfectly predicted and thus typically modelled as random variables. Similarly, users demands are subjected to variations according to the user experience and behaviour, and type of content. Although users future locations can be accurately predicted by advanced positioning techniques [75–79], other sources of uncertainties such as users skipping the video session and arrival of real-time traffic are envisioned in future networks irrespective of mobility. Existing PRA strategies in [25–27, 37, 38] modelled each of these uncertain components by the expected (average) value to obtain a deterministic formulation. However, this approach results in *non-robust* and suboptimal allocations when network conditions deviate from the expected value. Despite these reported gains in the literature, the following practical challenges

related to prediction uncertainty must be addressed:

- **Channel Rate Fluctuations:** The first parameter used in PRA is the future channel rate of mobile users based on their trajectory. In practice, channel predictions are typically associated with uncertainties due to the low-power filters used in the mobile devices [34] and the random behaviour of the received signal level as shown in Fig. 2.2(b). Deterministic decisions by existing PRA [25–28] do not guarantee QoS satisfaction when predicted future rates fall below the expected values. In this case, the minimal airtime fraction allocated to the cell edge users will not be sufficient to meet their demand and buffer underrun occurs causing video stalls as depicted in Fig. 2.2(a). In addition, when peak rates exhibit lower values than expected, energy savings will be suboptimal as the large allocated airtime will deliver a small amount of video content.

- **Demand Uncertainty:** The user demand is represented by both the streaming bitrate (i.e. video quality) and the watching duration. Users can frequently change the quality of video, skip some frames or terminate the session without watching the entire video [40]. Fig. 2.3(a) depicts an example of energy wastage under the PRA literature, which assumed perfect prediction of streaming duration, however the user terminates the session at t=5. The risk of wasting resources increases as PRA maximizes prebuffering for users at the cell center (i.e. experiencing peak rates). Existing robust non-PRA techniques [80, 81] decide when to be prebuffer the video at the current slot, to save the tail energy, or postpone the delivery. The PRA, however, requires further efforts to consider the trade-off over the time horizon since postponing full video delivery requires more resources to transmit the remaining content during




(a) Existing PRA under uncertain demand (b) Existing PRA under uncertain network resources

Figure 2.3: Illustration of high energy consumption and QoS degradations

future poor channel conditions. The impact of demand uncertainty is thus more severe in case of PRA, to strike a balance between both the risk of wasting resources if the prebuffered content is skipped and the likelihood of energy consumption if prebuffering is delayed till poor conditions.

- **Radio Resources Variation:** The stochastic arrival of users with stringent service delay requirements, such as voice calls, will decrease the total available resources for streaming users. Such random arrival will increase the risk of violating QoS requirements of video users at poor conditions who are allocated a small portion of the available resources. Fig. 2.3(b) depicts this scenario where the network follows a minimal allocation strategy for a cell-edge user to minimize the energy consumption. The risk of violating the demand, when the user does not receive the minimum amount of data, has to be modelled by the R-PRA. Thus, minimal allocation can be only followed during resources stability while an opportunistic risk-aware strategy is adopted in uncertain conditions.

*Robust* PRA frameworks are therefore paramount to unleash the gains of predictions under real-life constraints. This involves 1) modeling the rate, resources and demand uncertainty, 2) developing models to provide probabilistic QoS guarantees, and 3) efficiently tracking the prediction uncertainty in real-time. Integrating these functionalities should enable PRAs to strike a balance between providing network gains such as energy savings when predictions are accurate, and minimizing the risks associated with erroneous predictions during periods of uncertainty.

# Chapter 3

# Stochastic Optimization and Uncertainty Tracking

In this chapter, we provide a background on the robust optimization techniques that will be used in our R-PRA framework. Robust optimization refers to a class of decision making problems in which input information are erroneous. In essence, a certain level of constraint satisfaction has to be met by the decision maker while solving a problem accommodating uncertain information. Mathematically, the coefficients and bounds of objective function and constraints are modelled as uncertain variables rather than constants in the deterministic optimization problems.

## 3.1 Robust Optimization

*Robust* non-predictive RA techniques have been discussed in the literature in the context of handling both uncertainties and delays in the user reported measurements [43, 82, 83]. Two fundamental optimization techniques namely *Fuzzy* and *Stochastic* are used to provide a

robust formulation of the RA problem. In the former, the varying predicted information is represented as fuzzy numbers associated by a membership function [84]. On the other hand, *Stochastic* optimization represents the uncertain values as random variables characterized by their probability density functions [85]. Commonly, these two techniques provide a closed form representation of the robust formulation referred to as *deterministic equivalent* or *robust counterpart*. Although the *Fuzzy* results in a deterministic form that does not change the order of complexity of the original non-robust formulation [84], an unsustainable conservatism is attained, resulting in suboptimal RA decisions [85, 86]. Conservatism means over-satisfying the constraints at the expense of the objective function optimality. *Stochastic* optimization, which is less conservative, was thus extensively adopted in non-predictive RA schemes. The main drawback compared to fuzzy approach is the increased complexity. Hence we adopt the stochastic optimization to avoid the effect of conservatism on resource allocation and prediction gains, while the complexity is handled through convex decompositions or linear approximations, and supported by guided heuristic search to obtain real-time solutions.

## 3.2 Stochastic Optimization

*Stochastic* optimization utilizes two main techniques: Chance Constrained Programming (CCP) and Recourse Programming (RP) to handle the uncertainty in constraints and objective functions coefficients, respectively [85].

### 3.2.1 Recourse Programming

#### Two-Stage Recourse Programming

In Recourse Programming (RP), resource allocator takes some actions as a first stage, after that a random event is observed and impacts the optimality of the first-stage decision. A recourse decision is thus needed in the second stage to compensates for any suboptimal effects experienced by the network as a result of the first-stage decision [85]. The RP model consists of both first-stage decision variables and recourse decision variables (i.e. second-stage variables).

A standard formulation of stochastic two-stage RP is depicted as:

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \left\{ \sum_{\forall t \in \mathcal{T}} \big( F(x_t) + \mathbb{E}\big[ H(y_t, \eta_t) \big] \big) \right\} \tag{3.1}$$

where $\mathbb{E}$ is the expectation with respect to the random vector $\eta_t$ that represents uncertain resources or demands. $\mathbf{x}$ and $\mathbf{y}$ are vectors of the first and second stage decision variables, respectively. $H(.)$ is the recourse function that calculates the second-stage actions after the random component is unveiled.

#### Deterministic Equivalent

The first approach to obtain a closed form solution for Eq. 3.1 is the continuous PDF where integration is adopted to calculate the expectation operator over the probability space. Although the model size remains the same (i.e. no extra variables are defined), non-linearities are introduced which increases the computational effort. The integration of PDF in some cases is very challenging which make this approach intractable. Another approach is the SA in which the event space is considered to be discrete. The main challenge of such an approach is the additional decision variables in the problem. Planning over a time horizon

will make such an approach more difficult as a tree of events is created from the discrete variables of each time slot. As the PDF integration is very challenging, in the first approach, while the problem size increases exponentially with the SA, searching and simulation based methods can be applied such as Sample Average Approximation (SAA) [87].

### 3.2.2 Chance Constraint Programming

CCP was initially introduced in [88] to handle uncertainties and randomness in the constraints, and used in one of the two forms described below.

#### Individual Chance Constraint Programming (ICCP)

The individual chance constraint can be formulated as:

$$Pr\left\{F(x_t, \eta_t) \geq D_t\right\} \geq \beta, \qquad \forall t \in \mathcal{T}, \tag{3.2}$$

where $x_t$ is the resource allocation variable at time slot $t$, and $\eta_t$ denotes the random information (e.g. channel rate). The function $F(x_t, \eta_t)$ models the relation between $x_t$, $\eta_t$ and the demand $D_t$ for each time slot $t$ in the time horizon $\mathcal{T}$. The above formulation guarantees that the allocation at each time slot satisfies the corresponding demand with at least probability $\beta$. This represents the QoS level, where a higher value results in allocating more resources (e.g., more energy consumption) to ensure demand satisfaction. The above form of CCP has been applied in several applications of non-predictive resource allocation such as OFDM scheduling [89, 90], channel assignment [91], and power assignment in wireless networks [92].

### Joint Chance Constraint Programming (JCCP)

The aforementioned form of chance constraints can only guarantee the QoS satisfaction level during each time slot, and does not model the satisfaction over the *time horizon*. In particular, allocating less resources in one time slot will result in the demand violation of both the current and the future instances. Thus, satisfying $\beta\%$ of the demand of one time slot will not guarantee the same satisfaction degree in the coming time slot, since each time slot does not account for the partial satisfaction of the preceding slot demands. This is because the demand across the time slots is cumulative and allocation should be able to recover from outages in the previous slots. To avoid the propagation of such outages, allocation of all the time slots in the horizon should be jointly considered. This is typically done using Joint Chance Constrained Programming (JCCP) [93] and expressed mathematically as follows

$$Pr\left\{F(x_t, \eta_t) \geq D_t, \quad \forall t \in \mathcal{T}\right\} \geq \beta. \tag{3.3}$$

JCCP has been successfully adopted in the literature to solve numerous networking problems where the decision made on one constraint affects the satisfaction of the others. Among these, application to routing and bandwidth assignment is discussed in [94], and uplink resource allocation in OFDM networks in [95] where the QoS satisfaction of one user might affect the others. In such models, the chance constraints are found to be independent and their joint probability is simply the product of their individual probabilities. However, such an independence is not applicable in PRA since the constraints are no longer independent due to the cumulative demand at each time slot.

Due to the difficulty of obtaining the pairs of joint probabilities, Boole's inequality [96]

can be used to bound this joint probability. However, applying such a bound is very conservative and can result in suboptimal allocations that deteriorate the network optimization objective. Therefore, the individual probabilities of each constraint should be optimized to result in less conservative solutions. Example of applications that apply time dependent JCCP are model predictive control [97, 98] and the unit commitment in power generation systems [99, 100] in which the demand is cumulative among the time slots and therefore joint satisfaction is needed. Individual probabilities of chance constraints can be either determined optimally if the RA problem with unknown individual probabilities remains affine or convex, as in [101]. Otherwise, both individual probabilities and RA decisions are jointly determined using simulation based or iterative search techniques as in [99]. In summary, the joint chance constraint solves for two decision vectors: 1) the individual probabilities of each time slot QoS constraint, and 2) the resource allocation among the users. The former is subjected to Boole's inequality while the latter is subjected to user QoS satisfaction at each time slot in order to satisfy the overall QoS level over the time horizon.

**Deterministic Equivalent**

The common challenge in both types of CCP is that the problem does not have a closed form solution Eq. 3.2 or Eq. 3.3. As such, the problem is either solved using simulation based approaches or analytical methods. In the former type, realizations of the random component are generated [85] and allocation is decided to satisfy $\beta^{th}$ percentile of the scenarios [89]. On the other hand, analytical methods replace the chance constraints either with its CDF, PDF or Moment Generating Function (MGF) [89, 90]. These methods are found to provide better accuracy [102] when the CDF is invertible, unimodal and results in

affine or convex optimization.

We focus on analytical methods in which a deterministic equivalent form is derived to obtain a closed form RA formulation, and provide a solution in real-time. Such deterministic form should handle three main challenges: conservatism, safety and complexity. The first ensures that the constraints should not be over satisfied to avoid suboptimal network gains. The second challenge, safety, refers to the ability of capping the maximum violation probability by a certain degree denoted by $\epsilon = 1 - \beta$. With regards to complexity, the robustness typically converts the linear RA formulation to a non-linear or a discrete form. Hence, only convex continuous or linear formulations should be considered to obtain optimal robust solutions.

To derive the CCP deterministic form, robust stochastic work utilizes different techniques such as Scenario Approximation (SA), Gaussian Approximation (GA), Bernstein Approximation (BA) and Markov inequality [43, 103], among others. The GA assumes that all the random variables, in the formulation, follow a normal distribution. Their summation results in a multivariate random variable whose mean and covariance is a function of the statistical parameters of each single random variable. This derives a Second order Cone Programming (SoCP) formulation which also incorporates the inverse of the Gaussian Cumulative Density Function (CDF) and the QoS degradation level $\epsilon = 1 - \beta$. Similarly, the BA adopts the MGF to develop a SoCP deterministic form that only depends on the support of random variables and the QoS degradation level $\epsilon$ as well. The *Markov* inequality [83] on the other hand provides a linear empirical approximation. However, the optimal coefficients for such approximation are not easily attainable and do not model the trade-off between optimality and degree of constraint satisfaction. The SA utilizes the discrete PDF of the random variables to create a scenario tree using all the combinations. The allocator

has to ensure that the calculated resources satisfy the scenarios with total probability more than $\beta$.

In general, the GA and BA deterministic forms will have a higher complexity order than the non-robust form. For instance, the BA will transform a linear CCP into a SoCP which increases the computational burden [104]; due to the typically used convex optimization techniques such as Interior Point Method (IPM) [105, 106]. The robust non-predictive RA in [83] adopted the Markov inequality to approximate the CCP using a linear formulation. Previous approaches in [43] and [82] tackled the complexity of both GA and BA's SoCP by adopting either the first or the infinite order norms to obtain linear low-complexity deterministic forms for uplink non-predictive RA. However, both norms resulted in conservative solutions that are acceptable only for single time slot allocations (i.e. non predictive RA) to maximize the bandwidth efficiency.

## 3.3 Uncertainty Tracking

Both the feasibility and optimality of the obtained resource allocation solution are highly sensitive to the parameters of random variables such as variance. Applying the deterministic equivalent form with low error variance results in *unsafe* solution that does not guarantee the constraint satisfaction since less resources will be allocated to the user (e.g. when the channel rate falls below the average value). On the other hand, using a large variance results in a conservative solution that allocates too many resources especially in relatively high data rate time slots. Due to the fluctuation of prediction error variance with the user location and time of the day as reported in [33], a fixed variance becomes suboptimal. We therefore propose to adaptively track the variance based on the user's previous measurements. The tracking procedure is implemented using Bayesian based inference such as

Kalman Filter (KF) and Particle Filter (PF).

### 3.3.1 Kalman Filter

Kalman Filter (KF) is known to be the optimal linear estimator in the mean square error sense in case of Gaussian noise. In essence, KF is composed of two stages as summarized below [107]:

**Prediction Phase:**

$$\mathcal{X}_t^- = \Phi_t \mathcal{X}_{t-1}^+. \tag{3.4}$$

$$\mathcal{P}_t^- = \Phi_t \mathcal{P}_{t-1}^+ \Phi_t^{'} + \mathcal{Q}. \tag{3.5}$$

**Measurement Phase:**

$$\mathcal{K}_t = \mathcal{P}_t^- H_t^{'} (H_t \mathcal{P}_t^- H_t^{'} + \mathcal{R})^{-1}. \tag{3.6}$$

$$\mathcal{X}_t^+ = \mathcal{X}_t^- + \mathcal{K}_t (z_t - H_t \mathcal{X}_t^-). \tag{3.7}$$

$$\mathcal{P}_t^+ = \mathcal{P}_t^- - K_t H_t \mathcal{P}_t^-. \tag{3.8}$$

where $\mathcal{X}_t^-$ and $\mathcal{X}_t^+$ are the priori and posterior state vectors respectively. $\mathcal{P}_t^-$ and $\mathcal{P}_t^+$ are the priori and posterior state estimation covariance matrices respectively. $H$ and $\Phi$ are the observation (design) and state transition matrices respectively, while $\mathcal{K}$ is the KF gain. $\mathcal{Q}$ and $\mathcal{R}$ are the process and the measurement noise covariance matrices respectively.

The Kalman filter performs state vector estimation using two phases: Prediction and Measurement. In first phase, the predicted state value $\mathcal{X}_t^-$ is calculated using the previously estimated value $\mathcal{X}_{t-1}^+$ in time slot $t-1$ as indicated in Eq. 3.4. In the measurement phase,

the new state is calculated using a weighted difference between the observed measurements $z_t$ and the predicted state Eq. 3.7. This weighting is done using Kalman gain $\mathcal{K}_t$ calculated in Eq. 3.6, that is dependent on both the measurement noise covariance $\mathcal{R}$ and the predicted state estimation covariance $\mathcal{P}_t^-$ in Eq. 3.6.

### 3.3.2 Particle Filter

The Particle Filter (PF) is typically adopted when the system noise is non-Gaussian. Initially, the PF generates a set of values (i.e., particles) following a proposed distribution and assigns them equal weights. These weights are then tuned based on the reported user measurements according to a predefined likelihood function. A final estimate of the PF state is a weighted sum of the particles' values. The measurements represent the reported deviation between the predicted and the measured channel rates.

$p(y_{t+1}|Z_t)$ denotes the unknown posterior distribution of the state variable $y$ given a set of previous measurements/observations $Z$ at time $t$. This probability distribution is calculated based on a Bayesian method called Chapman-Kolmogorov defined as [108]

$$p(y_{t+1}|Z_t) = \int p(y_{t+1}|y_t)p(y_t|Z_t)dy_t \tag{3.9}$$

where $p(y_{t+1}|y_t)$ is used to calculate the evolution of state $y$ over the time horizon, while $p(y_t|Z_t)$ is an initial estimate of the posteriori probability at the current time slot and calculated as follows using Bayes' rule

$$p(y_t|Z_t) = \frac{p(Z_t|y_t)p(y_t|Z_{t-1})}{\int p(Z_{t+1}|y_{t+1})p(y_{t+1}|Z_t)dy_t} \tag{3.10}$$

where $p(Z_t|y_t)$ represents the likelihood probability of receiving measurements as $Z_t$ while

assuming state $y_t$. The denominator in Eq. 3.10 ensures that the estimated posteriori PDF will sum up to 1 over the time horizon.

The best estimate of the state $y_t$ in the mean square error sense is denoted by $\bar{y}_t$ and calculated as

$$\bar{y}_t = \int y_t p(y_t|Z_t) dy_t \tag{3.11}$$

In order to provide a tractable solution for the above equations, different techniques can be applied such as *Sequential Importance Sampling (SIS)* technique [109].

# Chapter 4

# Problem Statement and Proposed Robust-PRA Framework

In this chapter, we introduce the preliminaries, system model and the main building blocks of the proposed *Robust*-Predictive Resource Allocation (R-PRA) framework.

## 4.1 Preliminaries

We use the following notational conventions throughout the thesis: $\mathcal{X}$ denotes a set and its cardinality is denoted by $X$. Matrices are denoted with subscripts, e.g. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $\tilde{r}$ represents a random variable (r.v.) and its expectation is denoted by $\mathbb{E}[\cdot]$. $Pr\left(\bigcap_{\forall \mathcal{S}} s_i\right)$ and $Pr\left(\bigcup_{\forall \mathcal{S}} s_i\right)$ denote the joint and disjoint probabilities of all events in set $\mathcal{S}$. The gradient and Hessian of function $\mathbf{f}(\cdot)$ are denoted by $\nabla \mathbf{f}(\cdot)$ and $\nabla^2 \mathbf{f}(\cdot)$ in order. $\tilde{r}$ represents a random variable, whose probability density function follows normal distribution, while its cumulative density function is the Q-function denoted as $Q$. The $n^{th}$ percentile of a zero mean and unit variance normal random variable is denoted by $Q^{-1}_{1-n}$.

The $log(\cdot)$ denotes the natural logarithmic function and $\mathbb{1}_y$ is an indicator function which equals 1 if $y$ is satisfied and 0 otherwise.

## 4.2 System Model

Each BS serves an active user set $\mathcal{M}$ where the user index is denoted by $i \in \mathcal{M}$. At every time slot $t$, each mobile user requests video segment with a streaming rate $v_{i,t}$ that corresponds to a certain quality level.

### 4.2.1 Resource Allocation Model

#### Radio Resources

The active users can share the BS resources (airtime fractions) at each time slot $t$. The resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0,1] : i \in \mathcal{M}, t \in \mathcal{T})$ gives the fraction of time slot $t$ during which BS's bandwidth is assigned to user $i$.

#### Video Quality Selection

Each video segment can be transmitted and streamed by quality level $q \in Q$, where $Q$ is the set of possible segment qualities. The binary decision variable $\kappa_{i,t}^{(q)}$ is equal to $1$ if the video segment transmitted to user $i$ at time slot $t$ is encoded in quality $q$, and $0$ otherwise. Each segment consists of $v_q$ bytes of data, which depends on the selected quality level $q$.

### 4.2.2 Future Information

#### Predicted Channel Rate and Radio Resources

We assume that user's mobility trace is known for the next $T$ seconds, called the prediction window $\mathcal{T}$, and at a per second granularity where $\mathcal{T} = \{1, 2, \cdots, T\}$. Future rate prediction is obtained by mapping the user's trace to the REM available at the service provider. The REM contains the average rate for user $i$ at time slot $t$ and denoted as $\bar{r}_{i,t}$ [110].

#### Predicted Demand

The average demand of user $i$ at time slot $t$ is denoted by $v_{i,t}$ which corresponds to the data content played back with fixed quality during the time slot. The cumulative user demand at each time slot is denoted by $D_{i,t} = \sum_{t'=1}^{t} v_{i,t'}$. Although current streaming standards are user driven, the network can access the file between the user and streaming server to overwrite the video quality selected by the user device [24, 71].

### 4.2.3 Prediction Uncertainty

At each time slot, the resources are shared among both the streaming users (considered by the R-PRA) and other real-time users. The traffic of the latter is modeled using their arrival rate and demanded resources. Accordingly, we model the uncertainty associated with network resources as the total load of users requesting real-time service. This load depends on both the per user demand and the total number of users whose probability is calculated using the PDF of users arrival denoted by $P^A$. Similarly, the channel rates are subjected to uncertainties and thus modelled as random variables that can take a value according to the available MCSs at the BS, and the PDF of random rates, denoted by $P^R$.

Herein, we assume that the demand is uncertain as the user can terminate the video at

any time slot. Accordingly, the per slot demand is modeled as a random variable $\tilde{v}_{i,t}$ that is equal to 0 (user terminated the video) or $v_{i,t}$ (user streaming the video). The probability of terminating the video at each time slot will be determined by the PDF of the watching time denoted by $P^W$. Thus, the cumulative demand is also denoted as a random variable $\tilde{D}_{i,t} = \sum_{t'=0}^{t} \tilde{v}_{i,t'}$.

## 4.3 Problem Statement

The problem is to solve the resource allocation matrix $\mathbf{x} = (x_{i,t})$ and select the quality matrix $\kappa = \kappa_{i,t}^{(q)}$ to achieve a certain network metric such as minimizing energy or fair allocation of quality among the user. The QoS is said to be satisfied when the cumulative data allocated to the user $R_{i,t} = \sum_{t'=0}^{t} x_{i,t} r_{i,t}$ is not less than the cumulative demand $D_{i,t}$ at each time slot $t$. Both matrices are calculated under the uncertainty of all three predicted information, future rate, demand and radio resources. The R-PRA variants in Chapter 5 and Chapter 6 solve only for the resource allocation matrix while assuming a predefined quality level that minimizes energy consumption. This is unlike the DASH variant in Chapter 7 that solves for both decision variables to achieve fair QoS among the users.

## 4.4 Framework Overview

The proposed R-PRA framework aims to provide a real-time adaptive robust predictive allocation, and consists of four main blocks (see Fig. 4.1):

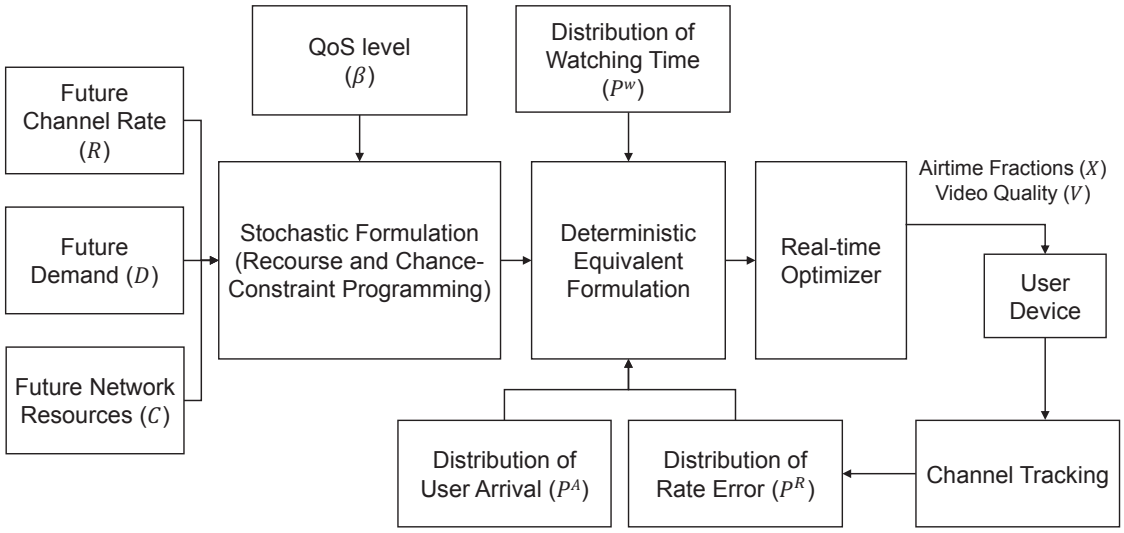



Figure 4.1: Schematic Diagram of Proposed R-PRA Framework

### 4.4.1 Stochastic Formulation

The first block provides a mathematical representation comprising of the resource allocation variables (i.e. airtime fractions and quality) and the future information represented as random variables to account for the prediction uncertainty. In essence, the formulation should model the trade-off between the network gains (e.g. energy) and the user satisfaction, which is governed by the QoS level $\beta$. The network resources and QoS constraints appear in a probabilistic form, i.e. CCP, and are bounded by predefined violation levels. The network gains are typically captured by objective function whose optimality can be impacted by the prediction uncertainties. Thus, RP will be used in this block to maintain the prediction gains over the time horizon. This first block typically consists of two types of input: the predicted information and the QoS level $\beta$. The former represents the future channel rates, capacity (i.e. radio resources) and demands of the video streaming users. The user's QoS satisfaction level is represented by $\beta$ over the considered time horizon.

The operator has the flexibility to assign different QoS levels to the users reflecting their priorities in the network. Moreover, this value can be changed over the time horizon to strike a balance between network objectives (i.e. prediction gains) and the degree of user satisfaction. The variants in Chapter 5 and Chapter 7 adopted the CCP to satisfy the QoS constraint at a certain probabilistic level under uncertain channel rates. In Chapter 6, both CCP and RP are adopted to handle uncertainties in network resources and users demands, respectively.

### 4.4.2 Deterministic Equivalent

In the second block, the formulated probabilistic model is transformed into a deterministic representation using the properties of random variables to capture uncertainties in predicted information (i.e. rate, resources and demand). In particular, these variations can be represented by the random variables properties such as PDF, support (i.e. limits) and the variance. Such properties are typically obtained either from extensive measurements or using Monte-Carlo simulations while adopting typical analytical error models. The main challenge in such module is how to choose the best approximation that handles the trade-off between conservatism, safety and complexity, as highlighted in Chapter 3, and the error modelling cost which depends on the type of random variables. For example, using the exact PDF will have a higher modelling cost than adopting only the variance. Another challenge is to specify the properties of random variables according to spatio-temporal changes and environments. For instance, a user moving in urban areas can suffer from rate variations characterized by larger error variance compared to another user in rural area. Similarly, the variance in both channel rate and network capacity during rush hours (e.g. afternoon) is very high compared to the evening of the same day [33].

Moreover, this module has to consider the joint uncertainties of predicted information over consecutive time slots. In particular, errors in both the mobility trace and variations in the wireless channel have to be jointly considered while modelling the uncertainness in the future rates. Handling these challenges will allow the framework to obtain a closed form model that can be solved by the optimizer in the next stage, and satisfies the QoS level.

The proposed two variants in Chapter 5 adopted GA and BA to obtain a deterministic equivalent in the case of normally distributed or unknown error model, respectively. In Chapter 6, SA is used due to the small dimension of the network resource constraint. In Chapter 7, both linearized SA and GA are proposed to obtain a closed form representation with non-polynomial complexity. Different conclusions on each equivalent form are drawn in the variants as their performance vary with the network objectives and type of constraint.

### 4.4.3 Real-time Optimizer

Although the deterministic form is convex, optimal gradient search methods cannot be adopted due to their high complexity. This module implements a low complexity local search guided algorithm that starts by satisfying the constraints and then moves on for optimizing the objective. The outcome is a real-time solution provided to schedulers and channel assignment modules in the access network.

In particular, the optimizer solves for the airtime fractions and video quality, and sometimes also solves for the QoS level. The main challenge of the optimizer is to obtain such optimal solutions in real-time (e.g. within 1 ms, which is the scheduling interval) that are also scalable with the system load and length of prediction window. Thus, this module will adopt guided heuristic algorithms that exploit the problem structure to generate feasible solutions and further enhance them to reach near-optimal values within the scheduling

interval. A near-optimal solution refers to an allocation decision whose objective function value is close to the value obtained by commercial solvers, while a feasible solution is the one that satisfies all the constraints. Moreover, this optimizer has to be adapted according to the considered constraints and the objective with a stable performance for different QoS levels, statistical parameters and problem dimensions.

All the R-PRA variants in the next chapters will develop a problem specific guided heuristic technique that initially satisfies the QoS constraints and then sequentially improves the value of objective function without changing the satisfaction of resource constraints.

### 4.4.4 Channel Tracking

The optimality of the robust deterministic form depends to a great extent on on the accuracy of rate deviations which differ with time and location [33]. This module uses Bayesian inference techniques to track the degree of uncertainty and adapt the statistical parameters such as variances based on the reported user measurements without prior knowledge of the channel statistics. In addition, it also allows cooperative uncertainty tracking among users and thus provides real-time updates for new arriving users to the network. The two variants in Chapter 5 adopt KF and PF to track the degree of uncertainty in predicted rates under Gaussian and generic error models, respectively.

## 4.5 Monte Carlo Framework for Statistical Parameters Estimation

The optimality of resultant allocation depends on the accurate calculation of random variable parameters. In this section we show one variant of determining the statistical measures of the rate (i.e., variance ${\sigma^r_{i,t}}^2$ and maximum deviation $\hat{r}_{i,t}$). Lower values of ${\sigma^r_{i,t}}^2$ or $\hat{r}_{i,t}$ than

the actual measurements will result in low level of robustness which increases the risk of violating the QoS constraint, and the converse is true. To address this, off-line Monte Carlo simulations are adopted prior to solving the RA problem. The simulation generates all the possible channel rates and adds random errors to them to build the rate distribution function.

Different values of the signal to interference plus noise ratio (SINR) are generated. For each value, the corresponding rate is calculated and denoted as $R$. Concurrently, $N$ random samples are generated and added to the current SINR, resulting in erroneous SINR denoted as $SINR_e$. Then, $N$ rates are constructed from $SINR_e$ and denoted as $R_e$. These rates are used to construct the probability distribution $\mathbb{P}$ of rate $R$. The simulation continues to generate a new value of SINR and repeats the above procedure until the maximum rate is generated. Finally, the bounds of each distribution and the variance are calculated while considering $R$ to be the mean value. It is worth noting that the SINR is mapped to the corresponding CQI level using formulas in [111]. The latter is then converted to the channel

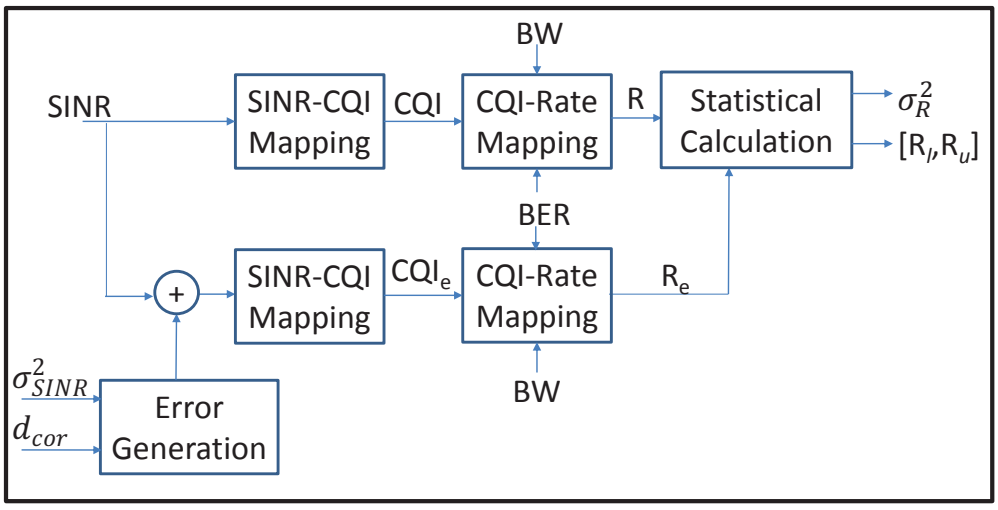



Figure 4.2: Block diagram for generating statistical parameters of the predicted rates using offline Monte-Carlo simulations

rate using the bandwidth (BW) and bit error rate (BER) values according to [112], and the generated error follows the 3GPP correlated fading model in [113]. All the above steps are summarized in Fig. 4.2. The main advantage of performing the above estimation off-line is to generate large samples of both the SINR and the added random variables. This results in accurate statistical estimation of the parameters used in the robust PRA.

# Chapter 5

# Green Robust-PRA under Rate Uncertainty

In this chapter we propose the design details of two variants of the Robust Predictive Resource Allocation (R-PRA) framework. Both variants tackle energy savings under rate uncertainty. In particular, the only source of uncertainty is assumed to be the channel rate, which impacts the QoS constraint satisfaction and thus Chance Constrained Programming (CCP) is adopted. The schemes solve only for the radio resource sharing (i.e. airtime fractions) at a predefined video quality level. The general block diagram of the two schemes is depicted in Fig. 5.1. The only main difference between the two schemes is the assumption of the rate error model. In the first scheme, we assume Gaussian distribution which will be handled by GA based deterministic equivalent and adopts KF for tracking the error variance. For generic or unknown rate error models, the second scheme is proposed and adopts BA which only requires the error bounds providing a solution at less modelling cost. The first section provides the system model, the two schemes are proposed in the second and third sections, while the last section is devoted for discussion and comparison between

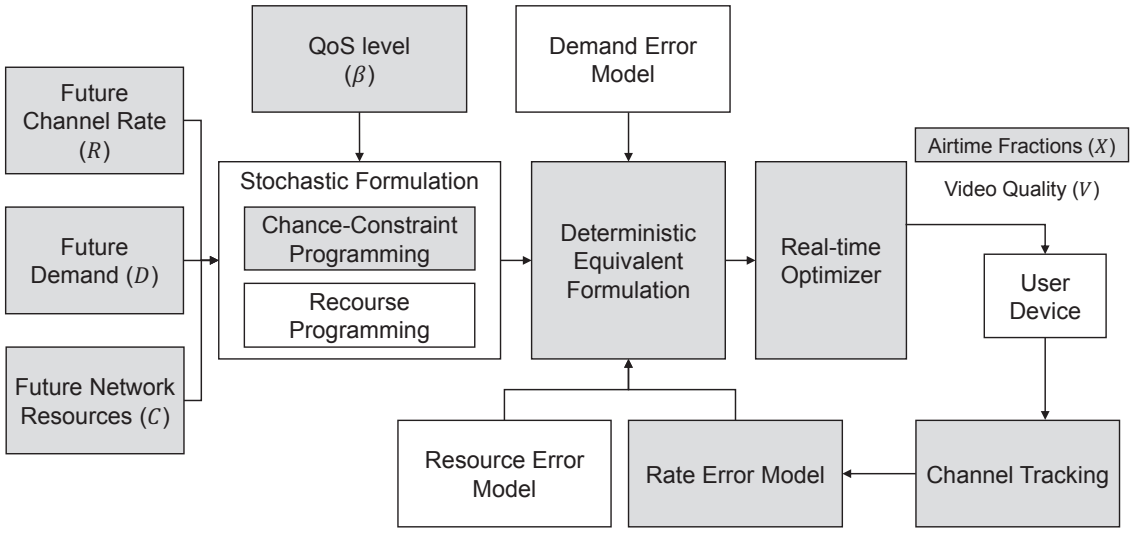



Figure 5.1: Block diagram of energy-saving R-PRA schemes under rate uncertainty

both GA and BA in the light of energy-saving problem at hand.

## 5.1 System Model

### 5.1.1 Predicted Mobility and Demand

In both variants, we assume that the users' mobility is known for the coming T time slots and the average rate is predicted and denoted by $\bar{r}_{i,t}$ for each user $i$ at time slot $t$. At each time slot, the demand of the user is assumed to be fixed at a certain streaming rate which requires a specific amount of bits per second, denoted by $v_{i,t}$, that achieves a compromise between energy minimization and user satisfaction.

### 5.1.2 BS Energy Model

Studies on BS energy consumption and sleeping strategies [17, 112, 114], reveal that the energy consumption $E$ is approximately linearly proportional to the airtime fraction of the BS [27, 115]. This is commonly referred to as time duty-cycling. In essence, $E = P \times \Delta T$ where $P$ is the total transmitted power by the BS and $\Delta T$ is the time during which the BS was switched ON. The dominant part of the power is that transmitted over the wireless channel, which is largely constant as downlink power control is not employed in the current 3GPP standards [112, 116, 117]. Accordingly, the energy consumption can be expressed in terms of the airtime $\Delta T$ to avoid dependencies on the constant power fraction that varies with BS type [115]. Therefore, as in [27, 38], we minimize the energy consumption by minimizing the total time air fractions $x_{i,t}$ allocated to all the users.

### 5.1.3 QoS Satisfaction

To achieve energy savings under QoS satisfaction, the BS should use the minimum resources needed to guarantee the video delivery at the target user quality over a time horizon. Existing energy-efficient RA approaches reveal that playback interruptions, due to buffer underrun, are among the primary sources of user dissatisfaction with video delivery services [25, 118–120]. In essence, video freezing occurs when the allocated airtime up to time slot $t$ results in delivering a total amount of video less than the corresponding cumulative streaming demand. This demand is denoted as $D_{i,t} = \sum_{t'=1}^{t} v_{i,t'}$. The number of video stops can therefore provide a sound QoS metric when modeling RA to optimize the trade-off between energy-minimization and QoS satisfaction.

## 5.2 Robust Model for Gaussian Uncertainty

### 5.2.1 Rate Uncertainty Model

In this first variant, we adopt the Gaussian distribution error model for the predicted rate introduced in [103], and used in recent robust non-PRA works [121]. In particular, predicting the future rates using autoregressive filters, resulted in a Gaussian distributed error model compared to the actual set of collected data as reported in measurement campaigns [103]. This is supported by the same distribution attained while applying the 3GPP correlated shadowing on the average value of predicted rates [113]. In our model, the rate is predicted at a 1 s granularity, which is generally deduced from a large number of samples due to the small feedback interval (1 ms) of the users participating in channel prediction [112]. Such a scenario supports the Central limit theorem (CLT) which approximates the PDF of users' predicted rate as a Gaussian distribution [121]. Nevertheless, resultant formulations are applicable for other error models with closed form and invertible CDF.[1]

### 5.2.2 Problem Formulation

We first model the robust PRA framework for video streaming using traditional *individual* chance constraints which is found to be a convex optimization problem. Thereafter, the problem is extended to the non-convex *joint* chance constraint model to enable QoS satisfaction of the cumulative demand over the time horizon. To provide a tractable solution, the problem is then decomposed into two convex stages that can be optimally solved individually.

[1] It has to be noted that the total probability of negative realizations for the normally distributed random rate has a non-significant value ($\approx 0$). This is attributed to the high average rate values that maintain a positive distribution under typical variances in the 3GPP models and standards [112, 113, 122].

**Individual Chance Constraint Programming**

The robust energy-efficient form is attained by representing the QoS constraint with the individual chance constraint, where predicted rates are replaced by random variables, and a probabilistic constraint is developed as follows

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{M} x_{i,t} \tag{5.1}$$

$$\begin{aligned}
\text{subject to:} \quad & \text{C1:} \quad Pr\left\{\sum_{t'=0}^{t} \tilde{r}_{i,t'} x_{i,t'} \geq D_{i,t}\right\} \geq \beta, \forall\, i \in \mathcal{M}, t \in \mathcal{T}, \\
& \text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, \forall t \in \mathcal{T}, \\
& \text{C3:} \quad x_{i,t} \geq 0 \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.
\end{aligned}$$

Herein, the predicted data rate $\tilde{r}_{i,t'}$ is modeled as a random variable following a normal distribution: $\tilde{r}_{i,t'} \sim N(\bar{r}_{i,t}, \sigma^2_{i,t})$, and $\beta \in [0,1]$ is the QoS satisfaction level.

Accordingly, the summation of the normally distributed random data rates in C1 of Eq. 5.1 is a multivariate normal distribution whose mean is the summation of means of all single random variables, which we denote as $\mu$. The corresponding variance is the covariance matrix denoted by $\Sigma$, and can be evaluated as follows

$$\mu = \sum_{t'=0}^{t} \bar{r}_{i,t}, \qquad \Sigma = \begin{bmatrix} \sigma^2_{i,0} & \dots & \sigma_{i,0,t} \\ \dots & \sigma^2_{i,1} & \dots \\ \sigma_{i,t,0} & \dots & \sigma^2_{i,t} \end{bmatrix}, \tag{5.2}$$

where $\sigma_{i,t,h} = E[(\tilde{r}_{i,t} - \bar{r}_{i,t})(\tilde{r}_{i,h} - \bar{r}_{i,h})]$ and $\sigma^2_{i,t} = \sigma_{i,t,h}, \forall t = h$.

The deterministic closed form of Eq. 5.1 can be expressed using the multivariate random variables and normal cumulative distribution function as shown below.

$$Q\left(\frac{D_{i,t}-\sum_{t'=0}^{t}\bar{r}_{i,t'}x_{i,t'}}{\sqrt{\sum_{t'=0}^{t}\sum_{h=0}^{t}x_{i,t'}^{2}\sigma_{i,t',h}}}\right)\geq\beta,\forall\, i\in\mathcal{M},t\in\mathcal{T}, \tag{5.3}$$

$$\sum_{t'=0}^{t}\bar{r}_{i,t'}x_{i,t'}+Q_{\beta}^{-1}\sqrt{\sum_{t'=0}^{t}\sum_{h=0}^{t}x_{i,t'}^{2}\sigma_{i,t',h}}\geq D_{i,t},\forall\, i\in\mathcal{M},t\in\mathcal{T}.$$

The independence between the realizations of random predicted channel rate at each time slot implies that $\sigma_{i,t',h}=0,\forall t'\neq h$. Accordingly, the chance constraint is represented as follows

$$\sum_{t'=0}^{t}\bar{r}_{i,t'}x_{i,t'}+Q_{\beta}^{-1}\sqrt{\sum_{t'=0}^{t}x_{i,t'}^{2}\sigma_{i,t'}^{2}}\geq D_{i,t}.\forall\, i\in\mathcal{M},t\in\mathcal{T}. \tag{5.4}$$

The above constraint representation is a second order cone programming (SOCP) model which is convex [123] for $\beta>0.5$ and results in a negative value for the inverse of the Q-function. Finally, the deterministic closed form of Eq. 5.1 using individual chance constraint with the preceding assumptions can be summarized below

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{M}x_{i,t} \tag{5.5}$$

$$\text{subject to: C1:} \quad \sum_{t'=0}^{t}\bar{r}_{i,t'}x_{i,t'}+Q_{\beta}^{-1}\sqrt{\sum_{t'=0}^{t}x_{i,t'}^{2}\sigma_{i,t'}^{2}}\geq D_{i,t},\forall\, i\in\mathcal{M},t\in\mathcal{T},$$

$$\text{C2:} \quad \sum_{i=1}^{M}x_{i,t}\leq 1,\forall t\in\mathcal{T},$$

$$\text{C3:} \quad x_{i,t}\geq 0 \quad \forall\, i\in\mathcal{M},t\in\mathcal{T}.$$

As mentioned in Chapter 3, this type of chance constraint formulation ensures that the

QoS is satisfied at each time slot at a certain level $\beta$. However, it does not model the joint satisfaction for each user over the time horizon in which the per slot demand satisfaction is dependent on the total data delivered in the preceding time slots. In order to avoid future buffer starvation, the allocation in each time slot should compensate the unsatisfied previous demands. This is why the joint chance constraint model is needed.

**Joint Chance Constraint Programming**

The joint chance constraint form for the QoS constraint can be expressed as follows

$$Pr\left\{\bigcap_{\forall t\in\mathcal{T}}\sum_{t'=0}^{t}\tilde{r}_{i,t'}x_{i,t'}\geq D_{i,t}\right\}\geq\beta,\forall\, i\in\mathcal{M}. \tag{5.6}$$

We denote the event of individual QoS satisfaction by $S_{i,t}\triangleq\left\{\sum_{t'=0}^{t}\tilde{r}_{i,t'}x_{i,t'}\geq D_{i,t}\right\}$. Similarly, the event of individual QoS dissatisfaction is denoted by $S_{i,t}^{c}$. The probability of joint satisfaction of event $S_{i,t}$ is represented as the complement of disjoint probability of the dissatisfaction event as in Eq. 5.7

$$Pr\left\{\bigcap_{\forall t\in\mathcal{T}}S_{i,t}\right\}=1-Pr\left\{\bigcup_{\forall t\in\mathcal{T}}S_{i,t}^{c}\right\},\forall\, i\in\mathcal{M}. \tag{5.7}$$

According to Boole's inequality, the disjoint probability is tightly bounded from above by the total probability of all individual events [96] as follows

$$Pr\left\{\bigcup_{\forall t\in\mathcal{T}}S_{i,t}^{c}\right\}\leq\sum_{\forall t\in\mathcal{T}}Pr\left\{S_{i,t}^{c}\right\},\forall\, i\in\mathcal{M}. \tag{5.8}$$

The joint probability of the QoS satisfaction event is therefore bounded as below

$$\begin{aligned} Pr\left\{\bigcap_{\forall t\in\mathcal{T}} S_{i,t}\right\} \geq & 1-\sum_{\forall t\in\mathcal{T}} Pr\left\{S_{i,t}^{c}\right\}, \forall\, i\in\mathcal{M}, \\ Pr\left\{\bigcap_{\forall t\in\mathcal{T}} S_{i,t}\right\} \geq & \beta, \forall\, i\in\mathcal{M}, \\ \sum_{\forall t\in\mathcal{T}} Pr\left\{S_{i,t}^{c}\right\} \leq & 1-\beta, \forall\, i\in\mathcal{M}. \end{aligned} \tag{5.9}$$

The above equation implies that the joint probability is satisfied if the summation of individual probabilities of the compliment event is kept below the probability of QoS dissatisfaction (i.e., $1-\beta$). Accordingly, the joint chance constraint in Eq. 5.6 can be replaced by the two constraints in Eq. 5.10 and Eq. 5.11

$$Pr\left\{\sum_{t'=0}^{t} \tilde{r}_{i,t'} x_{i,t'} < D_{i,t}\right\} < \zeta_{i,t}, \forall\, i\in\mathcal{M}, t\in\mathcal{T}. \tag{5.10}$$

$$\sum_{\forall t\in\mathcal{T}} \zeta_{i,t} \leq 1-\beta, \forall\, i\in\mathcal{M}. \tag{5.11}$$

where $\zeta_{i,t}$ is denoted as the probability for not satisfying the individual QoS constraint (i.e., $Pr\left\{S_{i,t}^{c}\right\}$) and is called the probability of *risk* [97].

Each probabilistic constraint in Eq. 5.10 will have the same deterministic equivalent form as the individual chance constraint but with $\beta$ replaced by $\zeta_{i,t}$. After incorporating Eq. 5.10 and Eq. 5.11, this JCCP formulation becomes a function of both variables: $\zeta_{i,t}$ and $x_{i,t}$ as summarized below

$$\underset{\mathbf{x},\zeta}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{M} x_{i,t} \tag{5.12}$$

$$\text{subject to:} \quad \text{C1:} \quad \sum_{t'=0}^{t} \bar{r}_{i,t'} x_{i,t'} + Q^{-1}_{1-\zeta_{i,t}} \sqrt{\sum_{t'=0}^{t} x^2_{i,t'} \sigma^2_{i,t'}} \geq D_{i,t}, \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, \forall t \in \mathcal{T},$$

$$\text{C3:} \quad x_{i,t} \geq 0 \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C4:} \quad \sum_{\forall t \in \mathcal{T}} \zeta_{i,t} \leq 1 - \beta, \forall\, i \in \mathcal{M}.$$

Indeed the above formulation is no longer convex and thus the optimal solution can not be guaranteed by traditional optimization techniques. A proof of its non-convexity is provided in Appendix A. Therefore, to provide a tractable solution, the above formulation is split into two stages: *Risk Allocation* and *Robust PRA*. The first stage determines the optimal values for each risk level (i.e., solves for $\zeta_{i,t}$), while the second stage solves the PRA problem given the calculated QoS satisfaction levels in the prior stage (i.e., solves for $x_{i,t}$).

**Stage A: Risk Allocation**

In this stage, the value of risk probabilities for each constraint is determined such that Boole's inequality Eq. 5.11 is satisfied to guarantee the joint probability satisfaction of Eq. 5.6. An initial feasible solution is to uniformly distribute the probability of risk $(1 - \beta)$ over all the time horizon. In other words, assign an equal risk probability $\zeta_{i,t}$ among all the time slots of each user as below

$$\zeta_{i,t} = \frac{1 - \beta}{T}, \;\; \forall i \in \mathcal{M}. \tag{5.13}$$

However, such equal risk allocation was proven to be very conservative [97] and results in suboptimal resource allocation that compromises the energy savings of the PRA obtained

in the second stage. Hence, optimal risk allocation is applied to consider the optimality of the second stage in addition to the Boole's inequality constraint C4 in Eq. 5.12.

Note that lower risk probability $\zeta_{i,t}$ results in higher airtime $x_{i,t}$ and that $x_{i,t}$ is inversely proportional to its corresponding average rate $\bar{r}_{i,t}$ as depicted in Eq. 5.5. Therefore, the risk of each time slot is allocated proportionally to the corresponding average rate $\bar{r}_{i,t}$ in order to minimize the energy consumption during the resource allocation stage. In other words, time slots with low average data rate will suffer from high airtime for QoS satisfaction. Thus, assigning low risk probability to these slots will result in additional airtime. To that end, the following risk allocation optimization is introduced in Eq. 5.14 to achieve the optimality of the second stage as well

$$\underset{\mathbf{y}}{\text{minimize}} \quad \sum_{t=1}^{T} \left(\frac{\hat{r}_i}{\bar{r}_{i,t}}\right)^n y_{i,t} \;\; \forall\, i \in \mathcal{M}, \tag{5.14}$$

$$\text{subject to:} \quad \sum_{\forall t \in \mathcal{T}} Q(y_{i,t}) \leq 1-\beta, \forall\, i \in \mathcal{M}.$$

where: $y_{i,t} = Q^{-1}_{\zeta_{i,t}}$ to represent the constraint in a differentiable form, $\hat{r_i} = \max_t \bar{r}_{i,t}$ and $n$ is the risk proportionality parameter whose value is positive. The value of $n$ captures the trade-off between the risk of not satisfying the QoS at a certain time slot and the energy savings. For very small values of $n$, the risk is fairly distributed among the time slots and the user will not suffer from successive video degradations. On the other hand, more energy savings are obtained when the value of $n$ increases since high risk is allowed at low data rate values. The mobile operator then may tune $n$ based on the maximum allowable consecutive degradation, or the desired energy savings. The above problem is convex given that $\beta \geq 0.5$, which is valid for practical considerations. A proof of this convexity is provided in Appendix B.

**Stage B: Robust PRA**

After solving the first stage in Eq. 5.14, and determining the risk probabilities $\zeta_{i,t}$ for each constraint, the problem in Eq. 5.12 can be solved without constraint C4. The resulting formulation preserves the form of SoCP, which is still convex due to the positiveness of the calculated risk probabilities.

### 5.2.3 Gradient Based and Guided Heuristic Solution Methods

After decomposing the joint chance constraint programming into two convex optimization stages, the solution methods for each stage are introduced in this section.

**Risk Allocation Solution**

The constrained proportional risk allocation in Eq. 5.14 is solved by calculating the Lagrange formulation and then using Newton's method to search for the saddle points that satisfy the Karush–Kuhn–Tucker (KKT) optimality conditions as follows

$$\mathcal{L}(y,\lambda) = \sum_{t=1}^{T} \big(\frac{\hat{r}_i}{\bar{r}_{i,t}}\big)^n y_{i,t} - \lambda\big(\sum_{\forall t \in \mathcal{T}} Q(y_{i,t}) - (1-\beta)\big) \qquad \forall\, i \in \mathcal{M}, \tag{5.15}$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the constraint in Eq. 5.14.

Since the above problem is optimized for each user separately and performed only once at the beginning of the time horizon, optimal path searching methods provide acceptable performance. We therefore apply Newton's method as summarized in Algorithm 1. The algorithm starts with the uniform risk allocation and then iteratively searches for the saddle points along the gradient while the step size is guided by the Hessian matrix. The calculated step value $\Delta\mathcal{L}$ contains the change in both the decision vector $y_i$ and the Lagrange multiplier $\lambda$ which are denoted as $\Delta y_i$ and $\Delta\lambda$, respectively. In each

iteration, both decision vectors are updated using the calculated step, and the algorithm stops when the iterations no longer result in a significant enhancement, denoted by $\epsilon$.

**Algorithm 1:** Newton's Method for Proportional Risk Allocation

**Input** : Time Horizon: $\mathcal{T}_i$, Average Predicted Rates: $\bar{r}_i$, QoS Level: $\beta$ and Risk Proportionality Factor: $n$

**Output** : $y_i$;

**Initialization**: : $\zeta_{i,t} = \frac{1-\beta}{\mathcal{T}_i}, y_{i,t} = Q^{-1}_{\zeta_{i,t}}, \forall t \in \mathcal{T}$, $\lambda = \lambda_0$, $\epsilon = 0.001$, $\Delta y_i = \Delta y_0$ and $\mathcal{L} = [y_i \ \lambda]^T$

**while** $\Delta y_i \geq \epsilon$ **do**

- $\frac{\partial \mathcal{L}(y_{i,t}, \lambda)}{\partial y_{i,t}} = \big(\frac{\hat{r_i}}{\bar{r}_{i,t}}\big)^n + \lambda \frac{1}{\sqrt{2\pi}} e^{\frac{-y_{i,t}^2}{2}}$;
- $\frac{\partial \mathcal{L}(y_{i,t}, \lambda)}{\partial \lambda} = -\big(\sum_{\forall t' \in \mathcal{T}} Q(y_{i,t}) - (1-\beta)\big)$;
- $\frac{\partial^2 \mathcal{L}(y_{i,t}, \lambda)}{\partial {y_{i,t}}^2} = -\lambda \frac{1}{\sqrt{2\pi}} y_{i,t} e^{\frac{-y_{i,t}^2}{2}}$;
- $\frac{\partial^2 \mathcal{L}(y_{i,t}, \lambda)}{\partial y_{i,t} \partial \lambda} = \frac{1}{\sqrt{2\pi}} e^{\frac{-y_{i,t}^2}{2}}$;
- *Construct:* $\nabla \mathcal{L}(y_i, \lambda)$ and $\nabla^2 \mathcal{L}(y_i, \lambda)$;
- *Calculate* $(\nabla^2 \mathcal{L}(y_i, \lambda))^{-1}$;
- $\Delta \mathcal{L} = -(\nabla^2 \mathcal{L}(y_i, \lambda))^{-1} \nabla \mathcal{L}(y_i, \lambda)$;
- $\mathcal{L} = \mathcal{L} + \Delta \mathcal{L}$ ;
- $\Delta y_i = \Delta \mathcal{L}(1:T)$;
- $\Delta \lambda = \Delta \mathcal{L}(T+1)$ ;
- $y_i = y_i + \Delta y_i$;
- $\lambda = \lambda + \Delta \lambda$;

**end**

return $y_i$

**Robust Real-time optimizer**

The calculated risk probabilities for each user at every time slot are now readily available to the robust PRA stage from the risk allocation solution. The objective of this stage is to solve for the airtime allocation formulated in Eq. 5.12. The solution of this stage is much more complex compared to the risk allocation since here the airtime is determined jointly for all the users over the total time horizon. Based on the users' feedback, this stage is recomputed every $\tau$ seconds according to the received amount of data. To address the resulting impractical complexity, a guided heuristic is also introduced to provide a real-time resource allocation solution, while the derivative based and line search methods are used to provide benchmark solutions.

The formulation in Eq. 5.12 is a SoCP, thus convex and continuous [123]. Its optimal solution can be obtained using Interior Point Method (IPM) [105] which is efficiently implemented in many commercial solvers such as Gurobi [124]. In particular, IPM searches within the set of feasible solutions for the optimal value where the latter is recognized due to its zero (or very small) duality gap. Although the IPM was proved to reach the optimality conditions in fixed number of iterations [106], the complexity per iteration hinders real-time solutions and still depends on the number of constraints. As seen from Eq. 5.12, the dimension of constraints increases with both the number of users M and the length of the time horizon T. In addition, the resource limitation constraint (C3) might cause the dissatisfaction of the QoS constraint (C2) especially at small values of $\epsilon$. In this case, the QoS constraint has to be relaxed which requires extra computations. Our framework hence relies on a suboptimal heuristic algorithm to provide a real-time solution, while optimal techniques (e.g. IPM) are used for benchmarking only.

The introduced guided heuristic search algorithm exploits the problem's features rather

than the direct gradient based iterative search. The algorithm first calculates the minimum allocation for the users to ensure constraint satisfaction (i.e., satisfy C1 in Eq. 5.12) given the calculated risk probabilities and the requested demands. In case of radio resource limit violations (i.e., C2 in Eq. 5.12), airtime reallocation of users is done by granting the excess user requirement in other time slots. In order to achieve energy minimization, users are allocated the residual airtime when they reach the peak average rate location. Residual airtime is the remaining airtime after satisfying the QoS constraints (first step) for all users. The heuristic is summarized in Algorithm 2

**Minimal airtime allocation:** To ensure the satisfaction of QoS constraint, C1 in Eq. 5.12 is turned to equality in the quadratic form $ax^2 + bx + c = 0$ and solved using Eq. 5.16 (Lines 4-11) of Algorithm 2. This is achieved as follows

$$x_{i,t'} = \frac{-b_{i,t'} + \sqrt{b_{i,t'}^2 - 4a_{i,t'}c_{i,t'}}}{2a_{i,t'}}, \tag{5.16}$$

Where: $a_{i,t'} = \bar{r}_{i,t'}^2 - (y_{i,t'}\sigma_{i,t'})^2$,

$b_{i,t'} = -2K_{i,t'}\bar{r}_{i,t'}^2$,

$c_{i,t'} = K_{i,t'}^2 - (y_{i,t'}L_{i,t'})^2$,

$K_{i,t'} = D_{i,t'} - \sum_{h=0}^{t'-1} x_{i,t'}\bar{r}_{i,t'}$,

$L_{i,t'} = \sum_{h=0}^{t'-1} x_{i,t'}^2\bar{r}_{i,t'}^2$.

**Allocation Repair:** The total allocated airtime to all users in each time slot is calculated and the radio resource limitation constraint, C2 in Eq. 5.12, is checked. In case of any violations, the excess airtime is allocated in other time slots with unused resources. Particularly, the heuristic compensates (recovers) any time slot $t \in T$ with a total allocated

**Algorithm 2:** Guided Heuristic Robust Green Allocation

**Input** : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}_i$, Mean of Predicted Rates: $\bar{R}$, Rate Variances: $\Sigma$, Risk Levels: $Y$ and Demand: $D$

**Output** : $X$;

**Initialization**: $x = \emptyset, t_i^{(p)} = \underset{t \in \mathcal{T}}{\text{argmax}}\{\bar{R}_i\}, \forall i \in \mathcal{M}$

/* time slot with maximum average rate (cell center) */ ;
**forall the** $t \in \mathcal{T}$ **do**
  $\tau_t = 0$ /* total airtime fraction allocated in time slot $t$ */ ;
  **forall the** $i \in \mathcal{M}$ **do**
    **if** $t < t_i^{(p)}$ **then**
      Calculate $x_{i,t}$ using Eq. 5.16 /* minimal airtime allocation*/ ;
      $\tau_t = \tau_t + x_{i,t}$ ;
    **end**
    **else**
      $M := M \setminus i$ /* remove user from minimal allocation after reaching cell center*/ ;
    **end**
  **end**
  **if** $\tau_t > 1$ **then**
    $i^{(*)} := \underset{i \in \mathcal{M}}{\text{argmax}}\{x_{i,t}\}$, /*choose the user with maximum airtime violating the constraint*/ ;
    $\delta x_{i^*,t} = \tau_t + x_{i^*,t} - 1$ /*violating airtime excess fraction*/ ;
    **for** $n := t - 1$ *to* $0$ **do**
      **if** $\tau_n + \delta x_{i^*,t} < 1$ **then**
        $x_{i^*,n} := x_{i^*,n} + \delta x_{i^*,t}$ /*Repair the solution*/ ;
        $\tau_n := \tau_n + \delta x_{i^*,t}$ ;
        $\tau_t = 1$ ;
      **end**
    **end**
  **end**
**end**
**forall the** $i \in \mathcal{M}$ **do**
  *AllocatePeaks* $(\tau_t, t_i^p)$ ;
**end**
return $X$ ;

airtime fractions (i.e., $\tau_t = \sum_{\forall i \in I} x_{i,t} \forall t$) more than the slot duration (1 sec.) which occurs due to 1) an increased number of users, 2) high traffic per user or, 3) high QoS level ($\beta$). The heuristic solves this case by iteratively picking the user with the maximum airtime fraction in this time slot and prebuffering his video content in advance to ensure airtime minimization under demand satisfaction (Lines 12-21) in Algorithm 2.

**Peak Average Rate Allocation:** The above allocation strategy guarantees the satisfaction of both QoS and resource constraints. Thus, it continues until the peak data rate time slot is reached. The allocation strategy is then changed (Line 24) to allocate the demand of the future time slots in advance, to minimize the airtime. This follows the following steps for each user $i$

- Calculate the residual demand for user *i*: $\delta D_{i,t'} = D_{i,T} - \sum_{t=0}^{t=t'} D_{i,t}$
- Repeat the allocation strategy in step 1 until either the total residual demand is allocated or the peak rate time slot is full.
- In case of remaining demand while the peak rate time slot is fully loaded, the second peak average rate with remaining airtime is selected and the above procedure continues.
- In each iteration, the residual demand is decremented by $x_{i,t'} \times (\bar{r}_{i,t'} - y_{i,t'}\sigma_{i,t'})$, which is a conservative method since it assumes the worst case channel capacity of the current rate.
- The algorithm terminates when all users received their total demand denoted as $D_{i,T}$.

Both the feasibility and optimality of the obtained resource allocation solution are highly sensitive to the variance $\sigma^2$. Applying the second stage with low variance does

not guarantee the constraint satisfaction since less airtime will be allocated to the user according to Eq. 5.16, especially during low data rates when high risk probability is allowed.

On the other hand, using a large variance $\sigma^2$ results in a conservative solution that allocates too much airtime especially in relatively high data rate time slots when low risk is applied. Due to the fluctuation of $\sigma^2$ with the user location and time of the day as previously mentioned, a fixed value of $\sigma^2$ becomes suboptimal. We therefore propose to adaptively track the variance $\sigma^2$ based on the user's previous measurements. The tracking procedure is implemented using Kalman Filter (KF) described in detail in the following section.

### 5.2.4 Kalman Filter Based Variance Estimation

The variances of the random predicted rates are updated using the channel measurements by the user in the previous time slot. The measured rate variance by user $i$ during the previous time slot $t-1$ is denoted as $\bar{\sigma}^2_{i,t-1}$ and calculated as follows

$$\bar{\sigma}^2_{i,t-1} = (\bar{r}_{i,t-1} - \bar{\mathbf{r}}_{i,t-1})^2, \tag{5.17}$$

where $\bar{\mathbf{r}}_{i,t-1}$ is the average measured data rate by user $i$ during the previous time slot $t-1$. $\bar{\delta}\sigma^2_{i,t}$ is the ratio between the measured and the initial theoretical variances denoted as $\bar{\sigma}^2_{i,t-1}$ and $\sigma^2_{i,t-1}$, respectively, and calculated using the Monte-Carlo framework. Although the variance ratio represents the actual deviations from the initial variance, the former still varies from one time slot to another. Accordingly, the change in the variance over time is modelled as a Gaussian process and thus can be accurately estimated using Kalman Filter, which is known to be the optimal linear estimator in the mean square error sense.

In our problem, the priori state $\mathcal{X}^-_t$ represents the variance ratio $\delta\sigma^2_{i,t}$ and equals the corrected state of the previous time epoch $\mathcal{X}^+_{t-1}$ by setting the state transition to unity.

The observation $z_t$ represents the measured variance ratio $\bar{\delta}\sigma^2_{i,t}$ shown in Eq. 5.17. The observed measurements $z_t$ and the predicted state $\mathcal{X}_t^-$ represent different values for the same quantity (i.e., variance ratio), and therefore the state observation matrix $H$ is set to unity. In summary, our KF model for variance ratio estimation is represented as follows

**Prediction Phase:**

$$\delta\sigma^{2}_{i,t}{}^{-} = \delta\sigma^{2}_{i,t}{}^{+}. \tag{5.18}$$

$$\mathcal{P}_t^- = \mathcal{P}_{t-1}^+ + \mathcal{Q}. \tag{5.19}$$

**Measurement Phase:**

$$\mathcal{K}_t = \mathcal{P}_t^-(\mathcal{P}_t^- + \mathcal{R})^{-1}. \tag{5.20}$$

$$\delta\sigma^{2}_{i,t}{}^{+} = \delta\sigma^{2}_{i,t}{}^{-} + K_t(\bar{\delta\sigma}^{2}_{i,t} - \delta\sigma^{2}_{i,t}{}^{-}). \tag{5.21}$$

$$\mathcal{P}_t^+ = \mathcal{P}_t^- - \mathcal{K}_t\mathcal{P}_t^-. \tag{5.22}$$

The updated ratio $\delta\sigma^{2}_{i,t}{}^{+}$ will be then used to update the predicted variances in the remaining time slots, denoted as $\sigma^{2}_{i,t+\delta t}{}^{+}$, while simultaneously considering their correlation with the current measurement as follows

$$\sigma^{2}_{i,t+\delta t}{}^{+} = \big(1 + \rho_{t,t+\delta t}(\delta\sigma^{2}_{i,t+\delta t}{}^{+} - 1)\big)\sigma^2_{i,t}, \qquad \forall\, \delta t \in [1, T-t], \tag{5.23}$$

where $\rho_{t,t+\delta t}$ is the channel correlation coefficient between the channel fading at time $t$ and $t+\delta t$.

According to Eq. 5.23, in case of high correlation (i.e., $\rho_{t,t+\delta t} \approx 1$), the future variance

will be multiplied by the value of current updated ratio and the term in the brackets becomes 1. On the other hand, very low correlation results in no updates of the future variance. In our model, we calculate the correlation coefficient using an exponentially decaying function with the correlation distance $d_{cor}$ according to the 3GPP slow fading model [113].

### 5.2.5 Performance Evaluation

#### Simulation Set-up

The presented robust PRA techniques are simulated for an LTE network using Network Simulator (ns-3) which is a standard compliant simulator [125], with model parameters and initial values of KF (i.e., $P_0$, $Q$, $R$ and $\delta\sigma_0$) as indicated in Table 5.1. The Gurobi optimization solver is integrated in ns-3 [126] and used to solve the SoCP in Eq. 5.5 and Eq. 5.12 with an efficiently implemented barrier and Interior Point Method (IPM) [127]. The solver exits when it reaches a duality gap less than $0.01\%$. The 3GPP correlated slow fading model and its parameters [113] are incorporated in the received UE power and thus provide predicted rate variations. Simulation results are averaged over 50 runs for statistical validation. Users follow different predefined paths within the cell at varying velocities from 25 to 60 Km/h and request a video stream at a fixed quality. Although the allocation is done at each base station separately, neighbouring BSs are considered at an inter-cell distance of 600 m for practical calculation of SINR and channel rates.

#### Evaluation Metrics and Scheme Notations

In order to assess the introduced Robust Predictive Resource Allocation (R-PRA) framework, we use the two metrics previously discussed in Section 5.1. The first is the percentage of videos stops which reflects the user QoS level. Mathematically, it is calculated as the

percentage of time slots in which the QoS constraint is violated. Existing predictive RA approaches revealed that playback interruptions, due to buffer under-run, are among the primary sources of user dissatisfaction with video delivery services [25, 118]. Thus video stops metric perfectly models the ability of RA to optimize the trade-off between energy-minimization and QoS satisfaction. The percentage of video stops, denoted as VD, is used to quantify the QoS degradation and calculated as the percentage of time slots in which the cumulative transmitted content ($R_{i,t}$) is less than the demand ($D_{i,t}$) per Eq. 5.24.

$$VD = \frac{\sum_{i=1}^{M}\sum_{t=0}^{T} \mathbb{1}_{R_{i,t}<D_{i,t}}}{M \times T} \times 100, \tag{5.24}$$

where $R_{i,t} = \sum_{t'=0}^{t} r_{i,t'} x_{i,t'}$ is the cumulative video content received by user $i$ till time slot $t$ while $r_{i,t}$ is the experienced channel rate by user $i$ at timeslot $t$. A maximum allowable degradation level is defined as the boundary for the metric, and is equal to $(1-\beta) \times 100\%$. The second metric is the average BS airtime which is used to measure the energy consumption in the network. During resource allocation, both the BS and UE consume energy in transmission and reception of data. Therefore, minimizing airtime reduces the energy consumption proportionally [115]. The objective function in Eq. 5.1 is used to quantitatively measure this metric.

In this evaluation study, we denote the proposed optimal ICCP and JCCP, and their corresponding heuristics with the following abbreviations:

- **Optimal-ICCP:** refers to formulation in Eq. 5.5 whose solution is obtained using the IPM implemented in Gurobi.
- **Heuristic-ICCP:** refers to formulation in Eq. 5.5 whose solution is obtained using the guided heuristic in Algorithm 2.

- **Optimal-JCCP:** based on the original non-convex JCCP formulation in Eq. 5.12 and solved using the sequential quadratic programming in MATLAB for a global optimal risk and airtime allocations [128].
- **Optimal-ERA-JCCP:** uses the two stage JCCP in which the first stage solution is obtained with equal risk values Eq. 5.13 and the second stage Eq. 5.5 is solved using the IPM implemented in Gurobi.
- **Heuristic-ERA-JCCP:** similar to the Optimal-ERA-JCCP but the second stage is solved using the guided heuristic in Algorithm 2.
- **Optimal-PRA-JCCP:** similar to the Optimal-ERA-JCCP with first stage formulated as in Eq. 5.14 and solved with Lagrangian Newton in Algorithm 1.
- **Heuristic-PRA-JCCP:** similar to the Heuristic-ERA-JCCP with the first stage formulated as in Eq. 5.14 and solved with Lagrangian Newton in Algorithm 1.

The optimal techniques are used to 1) evaluate the robustness of the introduced framework, and 2) assess the developed real-time guided heuristic in Algorithm 2. The non-convex Optimal-JCCP is used to evaluate the feasibility of the decomposed two-stage JCCP.

### Simulation Results

#### Comparison with Existing Non-Predictive and Non-Robust RA

The first simulated scenario is for one user moving across the cell from one edge to the other. Both the predicted average and the actual experienced rates are shown in Fig. 5.2(a). We consider three typical classes of RA:

Table 5.1: Summary of Model Parameters in the First Variant

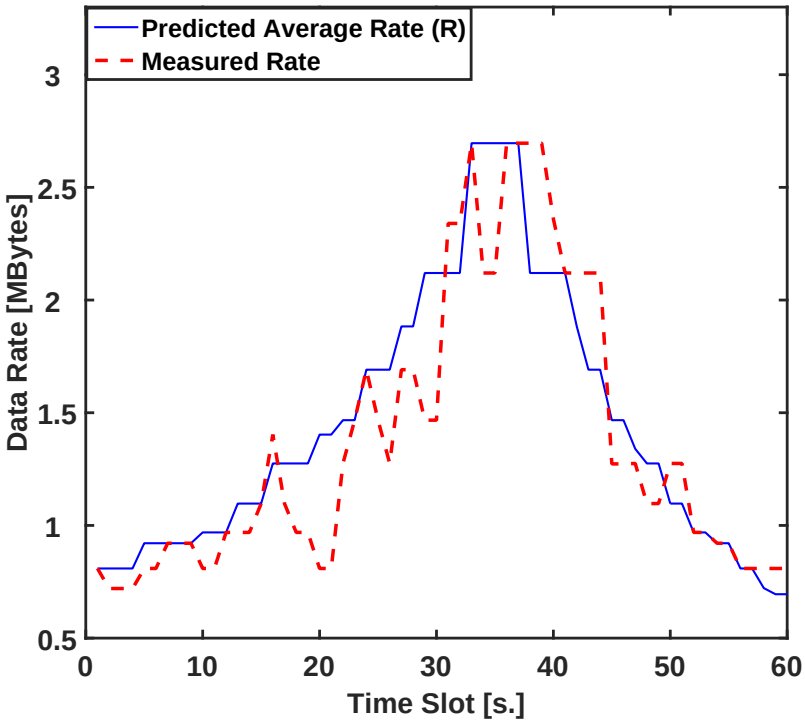
| Parameter | Value |
|---|---|
| BS transmit power | $43$ dBm |
| Bandwidth | $5$ MHz |
| Time Horizon $T$ | $60$ s |
| Streaming rate V | $0.5, 1, 1.5$ [Mbps] |
| Bit Error Rate | $5 \times 10^{-5}$ |
| Shadow correlation distance($d_{cor}$) [113] | $50m$ |
| Shadow standard deviation($\sigma$) [113] | $6dB$ |
| Velocity | From $25$ km/h to $60$ km/h |
| $P_0$ | $1$ |
| $Q$ | $0.1$ |
| $R$ | $1$ |
| $\delta\sigma_0$ | $1$ |
| Risk Proportionality Factor $n$ | $4$ |
| Feedback interval $\tau$ | $5s.$ |
| Packet size | $10^3$ [bytes] |
| Packet rate (from core network to BS) | $10^3 s^{-1}$ |
| Total number of packets | $7.5 \times 10^3$ |
| Buffer size | $10^9$ [bits] |

- **NP-RA:** refers to opportunistic Non-predictive Resource Allocation and the widely used Proportional Fairness [129] will be adopted as a type of this class.

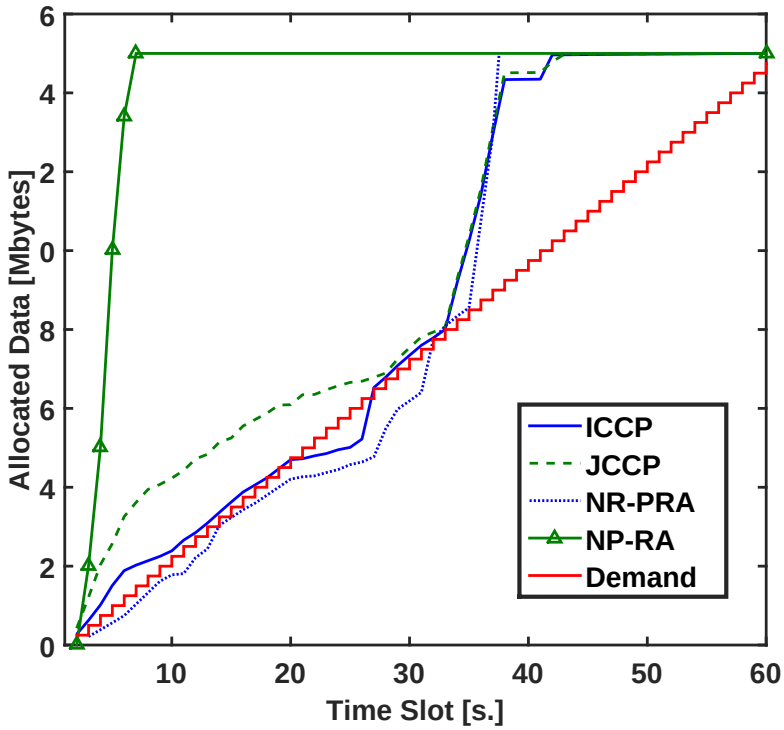
- **NR-PRA:** refers to the existing energy-efficient Non-Robust Predictive Resource Allocation in [27], which assumed perfect prediction and represented the future rate by its average value.

- **R-PRA:** refers to the energy-efficient Robust Predictive Resource Allocation introduced in this work in its two main forms (ICCP and JCCP).

The NR-PRA assumes perfect prediction of the future channel rates and results in the minimum energy consumption compared to both the NP-RA and the R-PRA as illustrated in Fig. 5.3(a). This is because, NR-PRA strategically allocates the minimal airtime that satisfies the demand based on the average predicted rate until the user reaches the cell

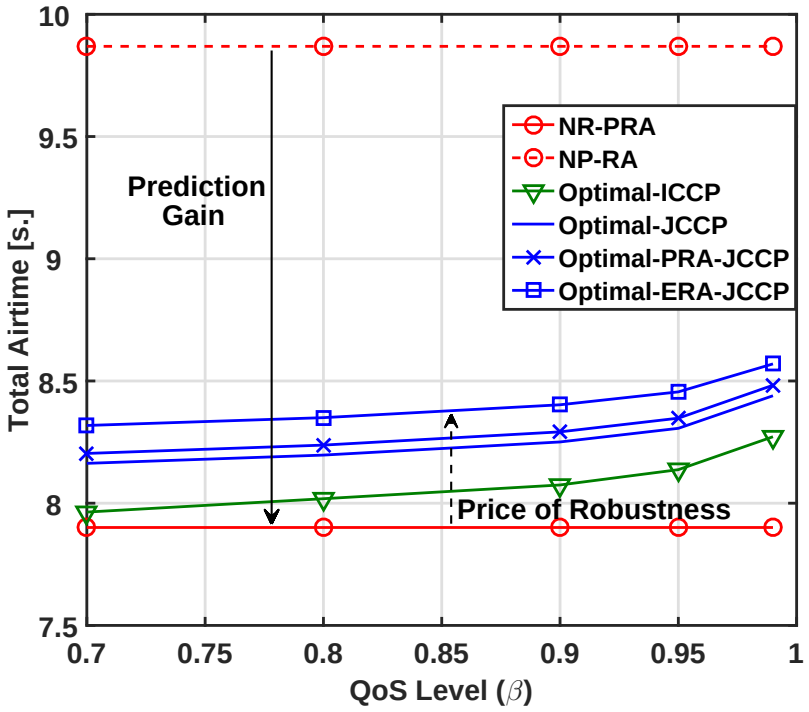



(a) Rate variations.

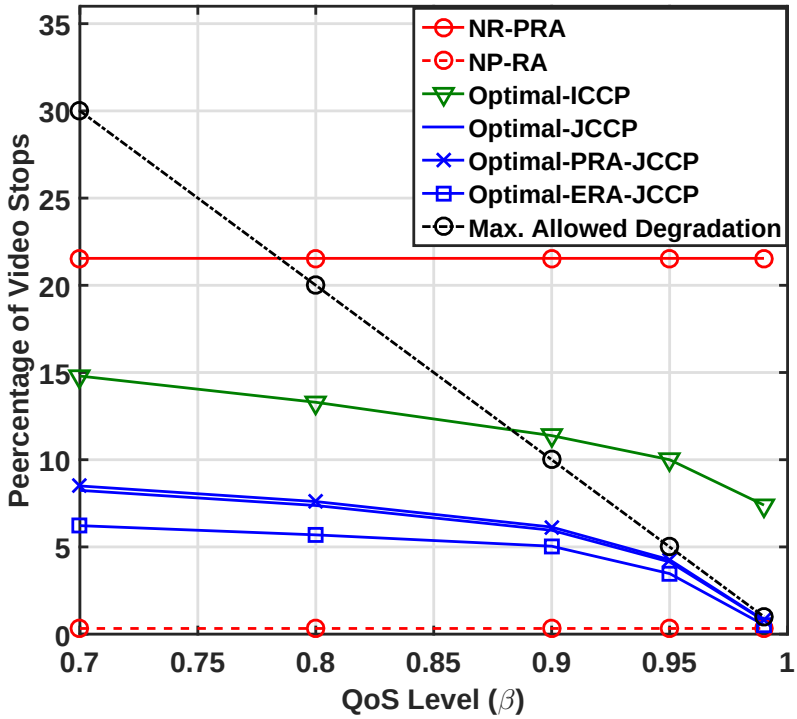



(b) Airtime allocations.

Figure 5.2: Illustrative allocation and rate variations examples for the considered techniques




(a) BS airtime.

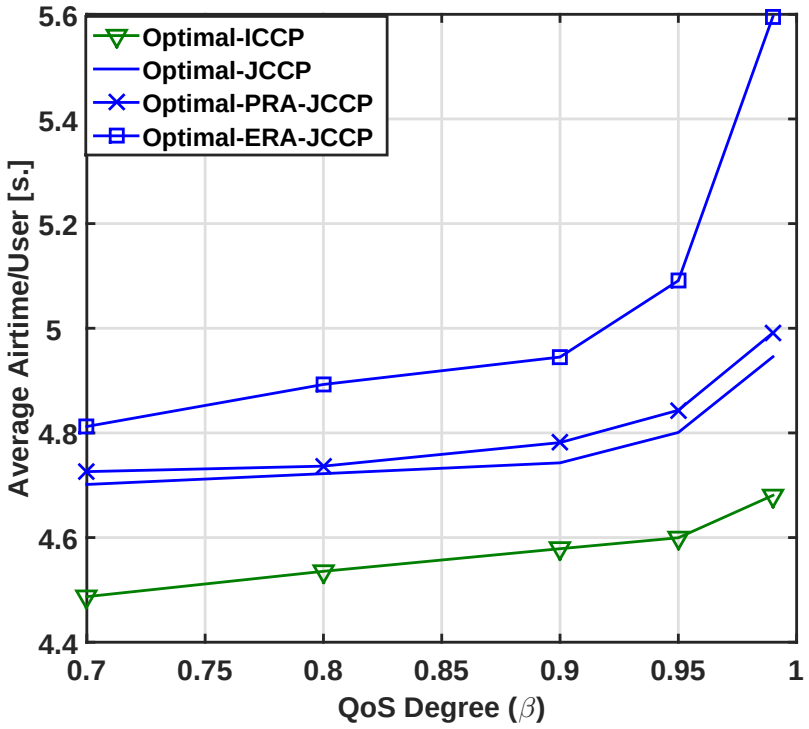



(b) Average Percentage of video stops.

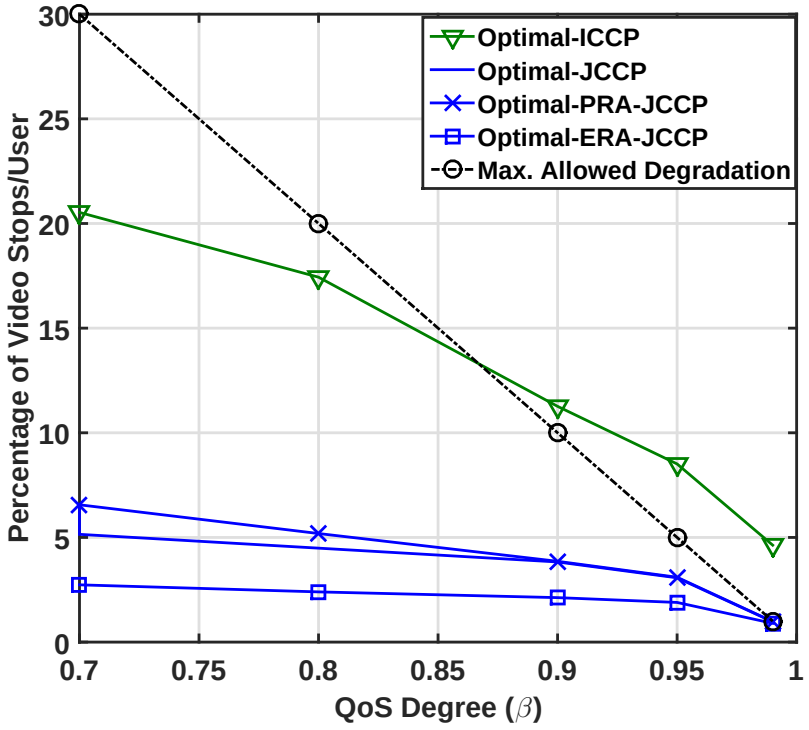
Figure 5.3: Percentage of video stops and average BS airtime for varying QoS degrees $\beta$ for 1 user experiencing rate variations

center. On the other hand, the introduced R-PRA conservatively allocates more airtime than the NR-PRA to guarantee QoS satisfaction under rate variations. The NP-RA, however, greedily assigns all the available resources and thus delivers the video to the user during the initial low rates regardless of the future high rates as shown in Fig. 5.2(b). On the other hand, Fig. 5.3(b) shows that the low-energy NR-PRA failed to satisfy the QoS demand as we can see that the user suffered from a large percentage of video stops. On the contrary, the proposed R-PRA (ICCP and JCCP) was able to compensate the variations by strategically allocating more airtime and the result is much fewer video stops. The traditional non-energy aware NP-RA filled the buffer of the user in the first few seconds, resulting in the highest QoS satisfaction with a negligible number of stops, but at the cost of high energy consumption.

To summarize, the NR-PRA previously introduced in [27] provides large energy savings, denoted as the *Prediction Gain*, compared to the NP-RA. However, this gain was achieved with unacceptable QoS violations under imperfect predictions. To overcome this limitation, the introduced R-PRA is designed to simultaneously satisfy the QoS requirements and energy minimization. This comes at the cost of slightly decreasing the prediction gain by an amount referred to as the *Price/Cost of Robustness* that accounts for rate variations. The above conclusions can also be drawn from the higher load scenario in Fig. 5.5, and indicate that robust PRA can provide significant gains under practical considerations of imperfect predictions. These results are obtained for the optimal forms of the introduced R-PRA (i.e., Optimal-ICCP and Optimal-JCCP) to assess their performance bounds, and the developed real-time heuristic which will be assessed separately. We first compare the performance of the optimal ICCP and JCCP.




(a) Average BS airtime.




(b) Average Percentage of video stops.

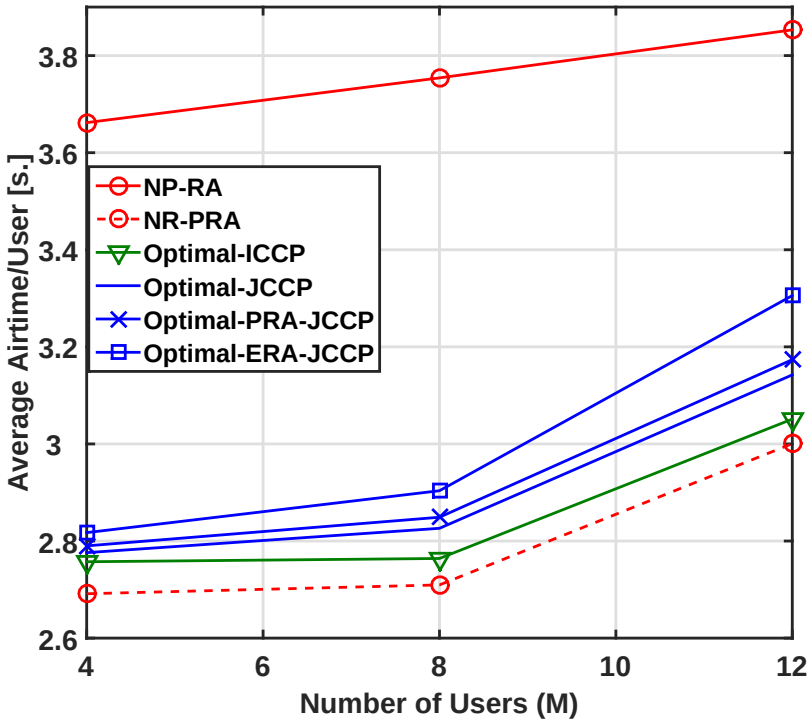
Figure 5.4: Percentage of video stops and average BS airtime for varying QoS degrees $\beta$ for $4$ Users experiencing slow fading with imperfect predictions

**Performance of R-PRA: ICCP and JCCP**

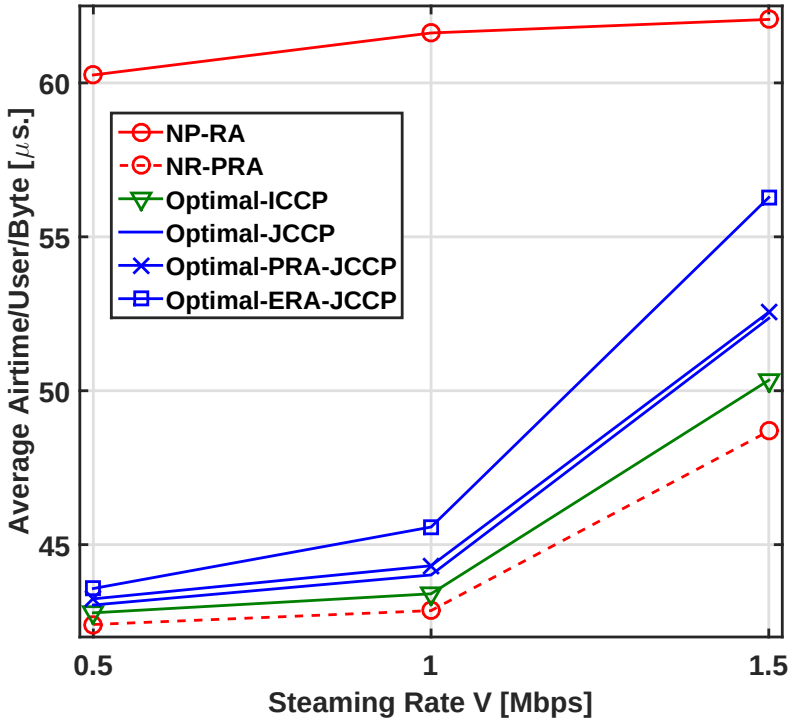
Under the aforementioned low load scenario, the Optimal-ICCP violates the maximum allowable video degradation in case of large QoS levels (i.e., $\beta \geq 0.9$) as shown in Fig. 5.3(b). This is attributed to the ignored dependency between the allocations in the time slots. More specifically, the demand violation occurred at $t = 20\ s$ in Fig. 5.2(b) due to the low rate (shown in Fig. 5.2(a)), resulting in cumulative degradations in the following time slots. This is because the potential outage was not accounted for beforehand. We can see that the buffer occupancy remained below the demand from $t = 20\ s$ to $t = 25\ s$ in Fig. 5.2(b) until the reallocation is done and the unmet demand is compensated. This violation was avoided for lower values of $\beta$ due to the continuous feedback from the user every $\tau$ seconds that enabled the network to recover video outages.

On the other hand, all the JCCP forms: Optimal-JCCP, Optimal-ERA-JCCP and Optimal-PRA-JCCP were able to avoid the above propagation of video stops and thus did not violate the maximum allowed degradation at all QoS levels as shown in Fig. 5.3(b). This was done at the expense of energy savings (i.e., a higher price of robustness) compared to ICCP as depicted in Fig. 5.3(a). The results also demonstrate the ability of the decomposed convex forms of JCCP (Optimal-ERA-JCCP and Optimal-PRA-JCCP) to obtain a solution that satisfies the QoS level. However, compared to the global optimal solution, the Optimal-PRA-JCCP was able to satisfy the QoS level with less energy compared to the Optimal-ERA-JCCP. This result emphasizes the importance of optimizing the risk values over the time horizon to control the conservatism of JCCP, especially when the user is located near the cell edge.

The performance results also indicate that the energy saving gap between the Optimal-PRA-JCCP and Optimal-ERA-JCCP increases with higher QoS levels ($\beta$), number of users




(a) Users experiencing imperfect predictions, $\beta = 0.95$ and $V = 0.5Mbps$




(b) 8 users experiencing imperfect predictions and $\beta = 0.95$

Figure 5.5: Performance of Robust PRA for different simulation scenarios

and higher streaming rates as shown in Fig. 5.4(a), Fig. 5.5(a) and Fig. 5.5(b), respectively. In particular, as $\beta$ increases, lower risk values are attained and the value of the inverse Q-function decreases exponentially which results in more airtime to satisfy C1 in Eq. 5.12. Similarly, increasing the number of users or streaming rate will result in more conservative RA for the cell edge users which decreases the BS airtime available for the cell center users to pre-buffer the video. It should be noted that the range of airtime varies across the scenarios since users follow different paths and velocities in each case.

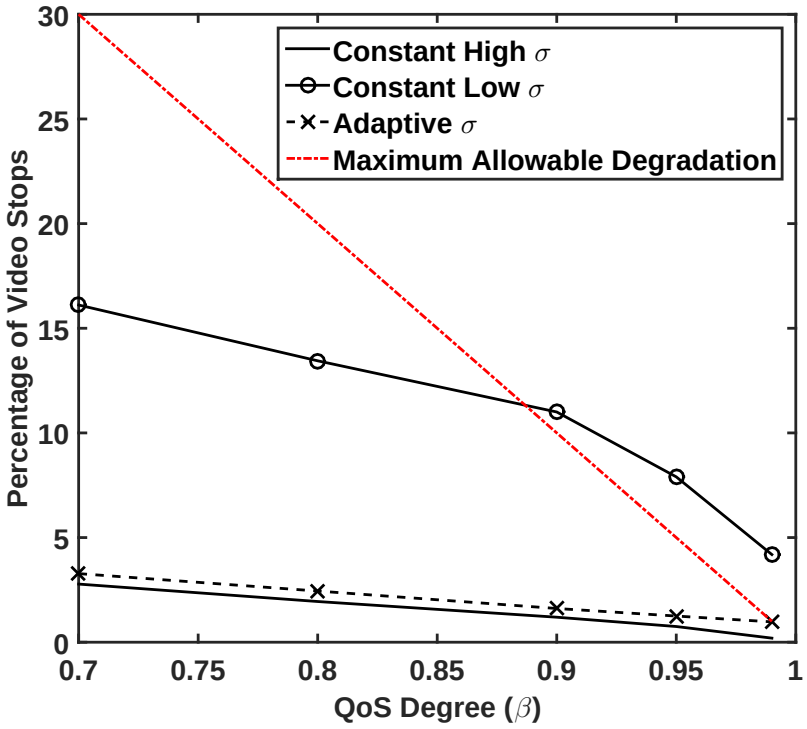
**Optimality and Complexity Analysis**

In order to evaluate the introduced guided heuristic, the optimality gap $Z$ is measured between the heuristic based solutions and the optimal results as $\mathtt{Z} = \frac{M(\mathtt{x})-M(\mathtt{x}^*)}{M(\mathtt{x}^*)} \times 100$, where $M(\mathtt{x})$ and $M(\mathtt{x}^*)$ are the values of objective functions corresponding to the heuristic and optimal solutions, respectively. A small optimality gap indicates that the heuristic solution is very close to the optimal one.

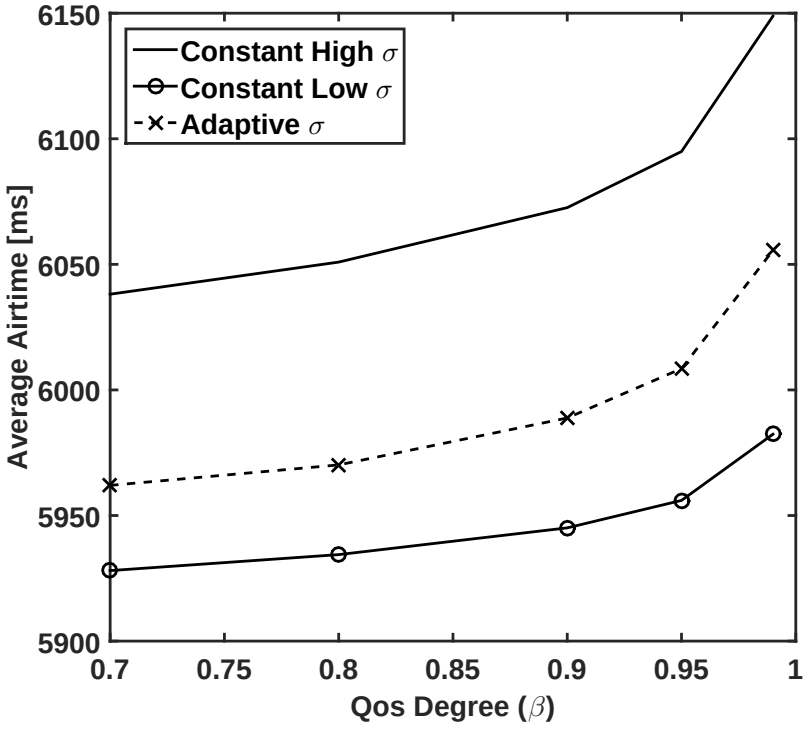
From Table 5.2 we observe that the heuristic solutions can provide the energy savings with small optimality gaps. This performance degrades with an increased competition at the cell center due to either a large number of users located in the cell peak during the same slot or few residual airtime due to conservative allocation of cell edge users (the case of ERA-JCCP). In particular, increasing the number of users at the cell peak will increase the optimality gap since the residual resources (after allocating the cell edge users) need

Table 5.2: Optimality Gap of Heuristic Algorithms

| Technique | Optimality Gap | | | |
|---|---|---|---|---|
| | 1 User | 4 Users | 8 Users | 12 Users |
| Heuristic-ICCP | 0.1 % | 0.15 % | 0.25 % | 0.3 % |
| Heuristic-ERA-JCCP | 0.1 % | 0.2 % | 0.5 % | 1.2 % |
| Heuristic-PRA-JCCP | 0.1 % | 0.15 % | 0.32 % | 0.45 % |




(a) Average Percentage of video stops.




(b) Average BS airtime.

Figure 5.6: Percentage of video stops and average BS airtime for varying QoS degrees $\beta$ for $4$ Users rate variations. Allocation is done using Heuristic-PRA-JCCP.

Table 5.3: Complexity Measures for Introduced Robust Techniques

| Technique | Order of Magnitude | Execution Time | |
|---|---|---|---|
| | | 1 User | 12 Users |
| Optimal-ICCP | $O(\sqrt{MT}(M^3T^4))$ | 90 $s$. | 980 $s$. |
| Heuristic-ICCP | $O(MT + T^2)$ | $< 1ms$. | $< 1ms$. |
| Optimal JCCP | $O(\sqrt{MT}(M^3T^4))$ | $140s$. | $1560s$. |
| Optimal-ERA-JCCP | $O(\sqrt{MT}(M^3T^4))$ | $90s$. | $980s$. |
| Heuristic-ERA-JCCP | $O(MT + T^2)$ | $< 1ms$. | $< 1ms$. |
| Optimal-PRA-JCCP | $O(\sqrt{MT}(M^3T^4))$ | $90s$. | $980s$. |
| Heuristic-PRA-JCCP | $O(M(T)^3)$ | $< 1ms$. | $< 1ms$. |

to be proportionally allocated while considering the future rates. This was not handled by the heuristic algorithm to maintain its low complexity. Instead, the heuristic performs a greedy allocation to the users with the maximum rates. As for QoS satisfaction, the guided heuristic solutions follow the same performance trends as their corresponding optimal counterparts, i.e., the ICCP forms fail to satisfy the maximum degradation at high QoS levels while the JCCP forms succeed for all values.

We next analyze the computational complexity of the different allocation strategies. For SOCP formulations, the optimal solution techniques (e.g., interior point method) require a maximum of $O(\sqrt{K})$ iterations [123] where $K$ is the number of constraints. Each iteration has a complexity of $O(m^2 \sum_{i=1}^{K} n_i)$ [106], where $m$ denotes the total number of decision variables and $n_i$ is the dimension of the $i^{th}$ constraint. For the Newton's method, the main complexity lies in the calculation of the Hessian matrix inverse with a dimension $m \times m$. This gives a complexity of $O(m^3)$ for each step in Newton's method. Table 5.3 summarizes the two complexity measures for all the considered techniques as a function of the problem dimensions, i.e., number users $M$ and time slots $T$. For the heuristic in Algorithm 2, the QoS satisfaction has a complexity of $O(MT)$. The peak allocations and solution repairing

have complexities of $O(M(T - t_p))$ and $O(MT)$, respectively. We also report the execution time measured within the simulation environment on a Quad Core i7-Processor, 3.2 GHz machine. These results highlight the incapability of the optimal solution methods to facilitate real-time implementation. It should be noted that increasing the number of users does not result in a proportional increase in execution time since the algorithms can be executed on multiple threads when there are multiple users. Moreover, the complexity of Newton's method which was executed for each user individually completes in less than 1 ms.

**Adaptive Variance Estimation**

The simulations were extended to test the robustness of the PRA framework to the *variations* in the channel variance. Such variations in the rate variance are typically observed in practical measurements due to the different landscapes and degrees of urbanization [33]. A conservative approach to tackle such variabilities is to optimize with a constantly large value (highest value revealed in simulations) for the rate variance. This will ensure meeting the QoS satisfaction level using JCCP as in Fig. 5.6(a). However, it compromises the energy efficiency as shown in Fig. 5.6(b). On the other hand, starting with a fixed lower value (smallest value revealed in simulations) of variance will result in less energy consumption but at the expense of QoS degradation even when JCCP is applied. The KF based tracking algorithm starts either with an arbitrary value of variance, and then continuously adapts its value based on the error between the channel measurements and initial values. It is therefore able to satisfy the QoS for all values of $\beta$, and with a lower airtime compared to the high variance case. In this scenario, the evaluation is based on the Heuristic-PRA-JCCP since it has a practical complexity and results in more energy savings compared to

the Heuristic-ERA-JCCP as highlighted previously.

## 5.3 Robust Model for Generic Uncertainty

### 5.3.1 Rate Uncertainty Model

In this second variant, the predicted uncertain rate is modelled as a random variable $\tilde{r}_{i,t} \in [r^l_{i,t}, r^u_{i,t}]$, where $r^l_{i,t}$ and $r^u_{i,t}$ are the lower and upper rate bounds, respectively, and the average value is $\bar{r}_{i,t} = \mathbb{E}[\tilde{r}_{i,t}]$. Herein, we assume that the PDF of such random rate is unavailable and only the bounds are used.

### 5.3.2 Problem Formulation

To obtain a robust deterministic form that is equivalent to Eq. 5.1, irrespective of the $\tilde{r}_{i,t}$ distribution, Bernstein Approximation (BA) is used. In essence, BA utilizes the marginal distribution and the moment generating function of the random variable. Generally, the chance constraint is represented as a linear summation of random variables as follows

$$Pr\left(f_0(\mathbf{x}) + \sum_{t=1}^{t'} \eta_t f_t(\mathbf{x}) \leq 0\right) \geq 1 - \epsilon, \qquad \forall\, t' \in \mathcal{T}. \tag{5.25}$$

Here $\eta_t$ is a random variable with marginal distribution $\mathbb{P}_t$, and $f_t(\mathbf{x})$ is a convex function containing the decision vector $\mathbf{x}$. $\epsilon$ is the maximum allowed level of QoS violation and equals to $1 - \beta$. Assuming that all the random variables $\eta_t$ are independent, $\mathbb{P}_t$ has a bounded support on the interval $[-1, 1]$ $\forall t$ and the function $f_t(\mathbf{x})$ is affine in the decision vector $x$, a convex deterministic equivalent for Eq. 5.25 can be obtained as follows

$$\inf_{\lambda>0}\left[f_0(\mathbf{x}) + \sum_{t=1}^{t} \lambda\Lambda_t(\lambda^{-1} f_t(\mathbf{x})) + \lambda \log\frac{1}{\epsilon}\right] \leq 0, \forall t \in \mathcal{T}. \tag{5.26}$$

Herein, $\Lambda_t(z)$ is the logarithm of the moment generating function $M_t(z)$ for r.v. $z$ as depicted in Eq. 5.27

$$\begin{aligned} \Lambda_t(z) =& log M_t(z) \\ M_t(z) =& \mathbb{E}\big[e^{kz}\big] = \int e^{kz} d\mathbb{P}_t(k) \end{aligned} \tag{5.27}$$

Instead of computing the exact value of the logarithmic moment generating function in Eq. 5.27, in addition to solving for the auxiliary variable $\lambda$, a conservative approximation using the upper bound can be adopted as in Eq. 5.28 [130].

$$\begin{aligned} \Lambda_t(z) \leq max\left\{\mu_t^+ z, \mu_t^- z\right\} + \frac{\sigma_t^2}{2} z^2, \forall t \in \mathcal{T} \\ -1 \leq \mu_t^- \leq \mu_t^+ \leq 1 \end{aligned} \tag{5.28}$$

The variables $\mu_t^+$, $\mu_t^-$ and $\sigma_t$ are used to approximate the bounded support [130]. Therefore, a conservative deterministic equivalent for Eq. 5.26 is attained using Eq. 5.28 and the arithmetic inequality as follows

$$f_0(\mathbf{x}) + \sum_{t=1}^{t'} max\left\{\mu_t^+ f_t(\mathbf{x}), \mu_t^- f_t(\mathbf{x})\right\} + \sqrt{2log(\frac{1}{\epsilon})\bigg(\sum_{t=1}^{t'} \sigma_t^2 f_t(\mathbf{x})^2\bigg)} \leq 0, \qquad \forall t' \in \mathcal{T}. \tag{5.29}$$

Finally, the robust PRA chance constraint C1 in Eq. 5.1 is replaced by Eq. 5.29 as depicted in Eq. 5.30

$$\sum_{t=1}^{t} \bar{r}_{i,t} x_{i,t} + \sum_{t=1}^{t'} \mu_{i,t}^- \hat{r}_{i,t} x_{i,t} - \sqrt{2log(\frac{1}{\epsilon})\bigg(\sum_{t=1}^{t'} (\sigma_{i,t} \hat{r}_{i,t} x_{i,t})^2\bigg)} \geq D_{i,t'}, \quad \forall t' \in \mathcal{T}, \tag{5.30}$$

where the random predicted rate $\tilde{r}_{i,t}$ is assumed bounded in $[r^l_{i,t}, r^u_{i,t}]$. To satisfy the assumptions for Eq. 5.26, this rate is normalized in $[-1, 1]$ by using the maximum deviation and the average values denoted by $\hat{r}_{i,t}$ and $\bar{r}_{i,t}$, respectively per

$$\begin{aligned} \hat{r}_{i,t} &= \frac{r^u_{i,t} - r^l_{i,t}}{2}, \quad r^u_{i,t} > r^l_{i,t} \\ \bar{r}_{i,t} &= \frac{r^u_{i,t} + r^l_{i,t}}{2} \end{aligned} \tag{5.31}$$

The constraint in Eq. 5.30 is a SoCP model which is convex for $\epsilon < 0.5$ and $x_{i,t} \in [0, 1]$ [123].

### 5.3.3 Real-time Guided Local Search Heuristic

The guided search algorithm proceeds by allocating the airtime that ensures exact satisfaction of QoS constraint (i.e., solves C1 in Eq. 5.30 as equality) to minimize the airtime. The radio capacity constraint is then checked (i.e., C2 in Eq. 5.30) and reallocation is done in case of violating the maximum time slot duration. Finally, the algorithm pushes all the remaining video content when the user reaches his peak radio conditions (i.e. maximum $\bar{r}$) to avoid allocation in future time slots with lower rates. The second and third steps are very challenging in multi-user scenarios where different users might experience their peak radio conditions simultaneously. The heuristic is summarized in Algorithm 3 and detailed as follows

**QoS satisfaction:** To minimize the energy consumption while guaranteeing QoS satisfaction, C1 in Eq. 5.30 is turned to equality so that the airtime exactly satisfies the demand without violating the maximum degree $\epsilon$. This step is calculated for every time slot for each user until the peak radio conditions are reached (lines 1-8).

**Algorithm 3:** Local-Search Guided Heuristic For Robust Allocation

**Input** : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Average Predicted Rates: $\bar{R}$, Rate Bounds: $\hat{R}$, Maximum Violation: $\epsilon$ and Streaming Demand: $D$;
**Output** : $X$;
**Initialization**: $X = \emptyset$, $N_t = 0 \; \forall t \in T$

**for** $i \in \mathcal{M}$ **do**
  $\hat{t}_i = \underset{t \in \mathcal{T}}{\text{argmax}} \{\bar{R}_i\}, \forall i \in \mathcal{M}$;
  $t = 0$;
  **while** $t < \hat{t}_i$ **do**
    Transform Eq. 5.30 to equality and solve for $x_{i,t}$;
    $N_t = N_t + x_{i,t}$;
  **end**
**end**
**for** $t \in \mathcal{T}$ **do**
  **if** $N_t > 1$ **then**
    $j = \underset{i \in \mathcal{M}}{\text{argmax}} \left\{ \frac{\bar{r}_{i,t}}{\max\limits_{\forall t'} \{\bar{R}_i\}_{t'<t}} \right\}$;
    $\Delta x_{j,t} = N_t - 1$, $k = t - 1$;
    **while** $k > 0$ **do**
      $\Delta x_{j,k} = \Delta x_{j,t} \times \frac{\bar{r}_{j,t}}{\bar{r}_{j,k}}$;
      **if** $N_k + \Delta x_{j,k} \leq 1$ **then**
        $x_{j,k} = x_{j,k} + \Delta x_{j,k}$;
        $N_k = N_k + \Delta x_{j,k}$;$N_t = 1$; $k = 0$;
      **else**
        $k = k - 1$;
      **end**
    **end**
  **end**
**end**
**for** $t \in \mathcal{T}$ **do**
  $\mathcal{L} = \{\mathcal{M} | \hat{t}_i = t \; \forall i \in \mathcal{M}\}$ ;
  **for** $i \in \mathcal{L}$ **do**
    $y_{i,t} = \min \left\{ 1 - N_t, \frac{D_{i,T} - D_{i,t}}{\max \{\bar{R}_i\}} \right\}$;
    $t' = \underset{\mathcal{T} \setminus t}{\text{argmax}} \{\bar{R}_i\}, \forall i \in \mathcal{M}$;
    $y_{i,t'} = \min \left\{ 1 - N'_t, \frac{D_{i,T} - D_{i,t}}{\max\limits_{\mathcal{T} \setminus t} \{\bar{R}_i\}} \right\}$;
    $\delta F = y_{i,t} - y_{i,t'}$;
    **if** $\delta F > \delta \hat{F}$ **then**
      $\delta \hat{F} = \delta F$;
      $\hat{i} = i$;
    **end**
  **end**
**end**
return $X$

**Resource Limitation Satisfaction:** After calculating the airtime fractions for all users in each time slot, the resource constraint, C2 in Eq. 5.1, is checked. In case of violation, the excess airtime is prebuffered in a preceding time slot with vacant resources. To ensure airtime minimization, the user with the highest average predicted rate in a previous vacant time slot is chosen (lines 9-23).

**Peak Local Search Allocation:** The above allocation strategy guarantees the satisfaction of both QoS and resource constraints. Thus, minimal allocation is used until the peak data rate time slot is reached. The challenging part in this stage occurs when more than one user competes on the same time slot. Accordingly, local search is applied to select the user who will result in the highest power consumption if he is not granted this time slot. As such, the local search calculates the difference in airtime between the two scenarios: If he is allocated to this peak time slot or if the second maximum peak is selected (lines 31-34). The user with less airtime in the first scenario is selected to be served in the current slot. The algorithm terminates when all the users' cumulative demands are satisfied.

For the heuristic in Algorithm 3, the QoS satisfaction step has a complexity of $O(MT)$. The peak allocations and solution repairing have complexities of $O(MT)$ and $O(T^2)$, respectively. Thus, the total complexity of the heuristic is $O(MT + T^2)$.

### 5.3.4 Particle Filter Based Rate Deviation Learning

We extend the robustness to scenarios in which the channel variance changes over the time and location [33]. A Particle Filter (PF) is used to tune the rate deviations (initially obtained off-line or theoretically) in order to reflect the channel variance based on the users' measurements. This is done on two steps: Rate deviation update and PF estimation. In particular, the PF estimates the error between the measured variance and its assumed value. This error is then used to update the theoretical variance for the future allocations.

### Rate Deviation Update

We denote the off-line calculated deviations (e.g., using Monte-Carlo in Chapter 4) as $\hat{r}_{i,t}^{(M)}$, while the final tuned deviations using PF are denoted by $\hat{r}_{i,t}^{(P)}$ and calculated as follows

$$\hat{r}_{i,t}^{(P)} = \alpha_{i,t} \times \hat{r}_{i,t}^{(M)}, \tag{5.32}$$

where $\alpha_{i,t} \geq 0$ is the proportionality factor between the off-line and measured rate deviations. As the channel variance changes over time and location, the value of $\alpha$ has to be adapted accordingly using the particle filter as shown in the next subsection.

In multi-user scenarios, cooperative tuning can also be performed where existing users in the network can propagate their estimated value of $\alpha_{i,t}$ to the recent users admitted to the same BS. Such cooperation is done using the channel correlation coefficients between the users based on their distances per Eq. 5.33

$$\begin{aligned} \alpha_{i,t} &= \alpha_{i,t-1} + \max_{j \in \mathcal{M}, j \neq i} \{\rho_{i,j,t}\} \left(\alpha_{j,t-1} - \alpha_{i,t-1}\right), \\ \rho_{i,j,t} &= e^{-\frac{d_{i,j,t}}{d_{cor}}}, \end{aligned} \tag{5.33}$$

where $d_{i,j,t}$ and $\rho_{i,j,t}$ are the distance and distance-dependent channel correlation coefficient between user $i$ and $j$ at time slot $t$, while $d_{cor}$ is the correlation distance. The above formula is adopted from the 3GPP channel fading model [113].

### PF Estimation

The PF initially generates a set of values (i.e., particles) following a proposed distribution and assigns them equal weights. These weights are then tuned based on the reported user measurements according to a predefined likelihood function. A final estimate of the PF state (i.e., $\alpha$) is a weighted sum of the particles' values. The measurements represent the reported deviation between the predicted and the measured channel rates. We apply

the *Sequential Importance Sampling (SIS)* technique [109] to obtain the best estimate of for the PF states. SIS approximates the unknown posteriori distribution by a group of generated particles where each particle is weighted by its conformity to the measurements. Such particles are drawn from a proposed distribution, based on the problem structure, that approximates the original unknown distribution using large number of particles. The particle filter methodology based on SIS is summarized as follows

1 **Initialization**

i Define the proposed distribution $p(Q)$.

ii Generate a set of $N$ particles denoted by $Q_{t=0}$ using the distribution $p(Q_{t=0})$.

iii Initialize equal weights ($\omega^i_{t=0}$) for all particles.

$$\omega^i_{t=0} = 1/N, \forall i = 1, ..., N, \tag{5.34}$$

iv Define the likelihood function $F(Q, Z)$.

2 **Measurement Phase**

i Update the weights of each particle using the measurement $Z_t$ and the likelihood function $F(Q, Z)$:

$$\omega^j_t = \omega^j_{t-1} F(Q, Z), \forall j \in 1, ..., N, \tag{5.35}$$

ii Normalize the weights:

$$\bar{\omega}^j_t = \frac{\omega^j_t}{\sum_{j=1}^{N} \omega^j_t}, \tag{5.36}$$

iii Calculate the best estimate:

$$\bar{y}_t = \sum_{j=1}^{N} z_t \bar{\omega}_t^j, \tag{5.37}$$

3 **Prediction Phase**

i Calculate the gradient:

$$\delta y_t = \frac{\partial z_t}{\partial t}, \tag{5.38}$$

ii Predict the future state:

$$y_{t+1} = A\bar{y}_t + B\delta y_t \delta t : \tag{5.39}$$

4 **Importance Sampling**

i Calculate effective samples:

$$N_{eff} = \frac{1}{\sum_{j=1}^{N} (\omega_t^j)^2}, \tag{5.40}$$

ii Check degeneracy then resample: If $N_{eff} < \hat{N}$ then, resample particles and set $\omega_t^j = 1/N \; \forall j \in 1, ..., N,$

In essence, the calculated weights $\omega_t^j$ in Eq. 5.35 approximate the posteriori PDF in Eq. 3.10, while the priori PDF in Eq. 3.9 is evaluated using the likelihood $F(Q, Z)$ in the initialization phase. In addition, Eq. 5.37 in the measurement phase implements the best estimate of the state (Eq. 3.11). In the prediction phase, the future state $y_{t+1}$ in Eq. 5.39 is a linear weighted combination of both the best estimated state $\bar{y}_t$ and the integral of its rate of change $\delta y_t \delta t$ from the available measurements $z_t$. In Eq. 5.39, A and B are the weights of both the best estimate and integral of the rate of change, respectively.

As the PF updates the weights $\omega_t^j$ in Eq. 5.35 every time slot, their values may converge and few number of particles will have non-zero weights. Such situation is called degeneracy, which has to be avoided as it deviates the weight's distribution from the actual posteriori probability. Thus, the number of effective particles $N_{eff}$ is calculated to check for the degeneracy and in case of dropping below the maximum threshold $\hat{N}$, resampling is done. Each particle contributes, based on its weight, in generating a new particle [109]. The newly generated set of particles will not contain the ones with very low weights. The new weights are equally redistributed similar to the initialization phase.

In our rate deviation tracking, the PF state $y_t$ is the proportionality factor $\alpha_t$ while the measurement $z_t$ is the reported proportionality factor $\bar{\alpha}_t$ calculated as

$$\bar{\alpha}_t = \frac{|\bar{r}_{i,t} - \mathbb{E}[r_{i,t}]|}{\hat{r}_{i,t}^{(M)}} \tag{5.41}$$

where $\mathbb{E}[r_{i,t}]$ is the measured channel rate by user $i$ in the duration from slot $t-1$ to slot $t$.

### 5.3.5 Performance Evaluation

#### Simulation Set-up

We adopt the same simulation set-up as the previous variant, yet with different random mobility traces. All the parameters and their values are presented in Table 5.4.

#### Comparative Schemes and Evaluation Metrics

In this evaluation study, we compare the proposed robust predictive scheme against the existing non-robust PRA and non-predictive RA schemes denoted as follows

- **N-PRA (MT):** refers to a type of non-predictive RA called maximum throughput (MT) [131]. In essence, MT allocates the whole resources to the user with the current maximum channel rate regardless his future channel conditions.

Table 5.4: Summary of Model Parameters in the Second Variant

| Parameter | Value |
|---|---|
| BS transmit power | $43$ dBm |
| Bandwidth | $5$ MHz |
| Time Horizon $T$ | $60$ s |
| Streaming rate V | $0.25, 0.5$ and $1$ [Mbps] |
| Bit Error Rate | $5 \times 10^{-5}$ |
| Shadow correlation distance$(d_{cor})$ [113] | $50m$ |
| Shadow standard deviation [113] | $4$ |
| Velocity | From $25$ km/h to $40$ km/h |
| $p(Q)$ | $\mathcal{U}(0, 4)$ |
| $N$ | $10000$ |
| $\hat{N}$ | $N/3$ |
| $A = B$ | $0.5$ |
| $\mu^-$ | $-0.5$ |
| $\sigma'_t$ | $\frac{1}{\sqrt{12}}$ |
| Feedback interval $\tau$ | $5s$. |
| Packet size | $10^3$ [bytes] |
| Packet rate (from core network to BS) | $10^3 s^{-1}$ |
| Total number of packets | $7.5 \times 10^3$ |
| Buffer size | $10^9$ [bits] |

- **NR-PRA:** refers to the existing non-robust PRA in [27] which only uses the average value of the predicted rate resulting in a deterministic linear formulation. The optimal solution is obtained using the simplex method implemented in Gurobi [124].

- **OR-PRA ($\mathbf{l}_2$):** refers to the introduced BA based robust PRA in this work and formulated in Eq. 5.30. The solution is obtained optimally using the IPM in Gurobi optimizer [124].

- **HR-PRA ($\mathbf{l}_2$):** the same as **OR-PRA ($\mathbf{l}_2$)**, but its solution is obtained using the guided local search heuristic in Algorithm 3.

- **R-PRA ($\mathbf{l}_1$):** refers to the introduced BA robust PRA in this work but linearized similar to [43] and the solution is obtained optimally using the simplex method in

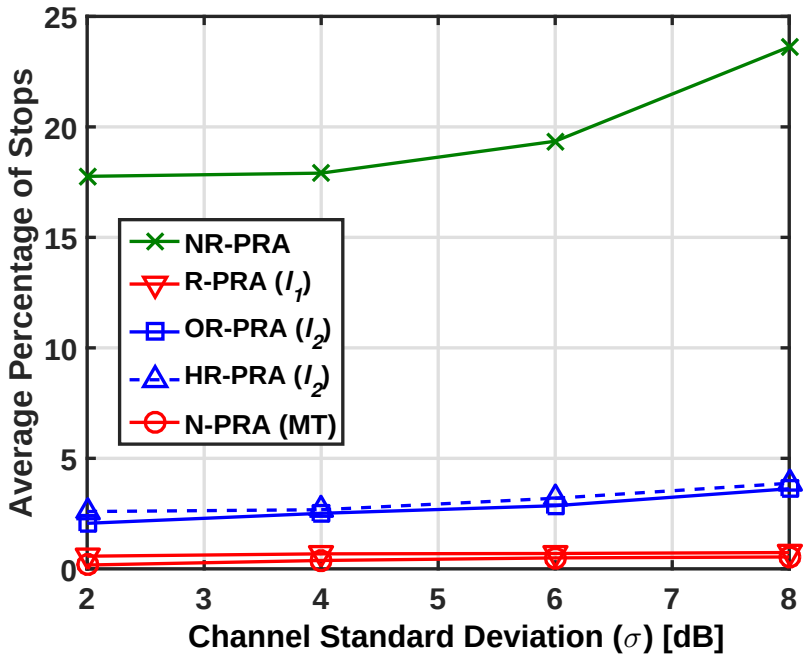
Gurobi optimizer [124].

All the above schemes are assessed using two main metrics: percentage of video stops, referred to as VD, and average airtime to measure the QoS satisfaction and the energy consumption, respectively. The maximum allowed value of VD, calculated per Eq. 5.24, is set to the predefined constraint violation level $(\epsilon) \times 100\%$. The second metric is the average BS airtime which is used to measure the energy consumption in the network, and calculated using the objective function in Eq. 5.1.

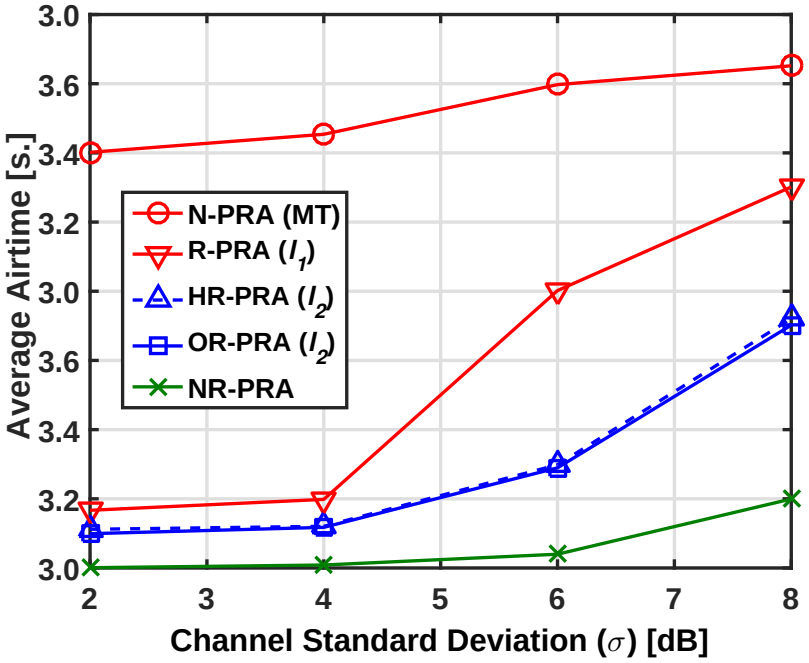
**Comparison with Other Resource Allocators**

We assume that the rate deviation $\hat{r}$ is accurately known and the focus is to show the importance of robust PRA and the heuristic solution. The first scenario considered a high quality video (i.e. $V_i = 1Mbps$) which is a high load scenario relative to the available average channel rate. The non-predictive MT continues to satisfy the QoS level independent on the channel variance as shown in Fig. 5.7(a). This is because the MT schedules the users based on their current reported channel rate irrespective of the variance and the future rates. The non-robust predictive technique [27] fails to satisfy the maximum VD set to 0.1 (i.e. $\epsilon = 0.1$). This QoS performance degrades with the channel variance since the measured rate deviates from the average value. The allocated minimal airtime will not be sufficient to satisfy the demand. Such deterioration is avoided by all the robust forms as the percentage of stops did not pass $\epsilon \times 100\%$ for the considered variances.

Although the non-predictive MT prioritizes users with maximum rates, its energy consumption is higher than the predictive strategies as depicted in Fig. 5.7(b). The MT buffers the video content for the cell peak users, which saves energy, but then turns to push the video for other users located near the cell edge rather than applying minimal allocation. On the other hand, the predictive strategy is able to minimize the energy even in the robust




(a) Average Percentage of video stops.



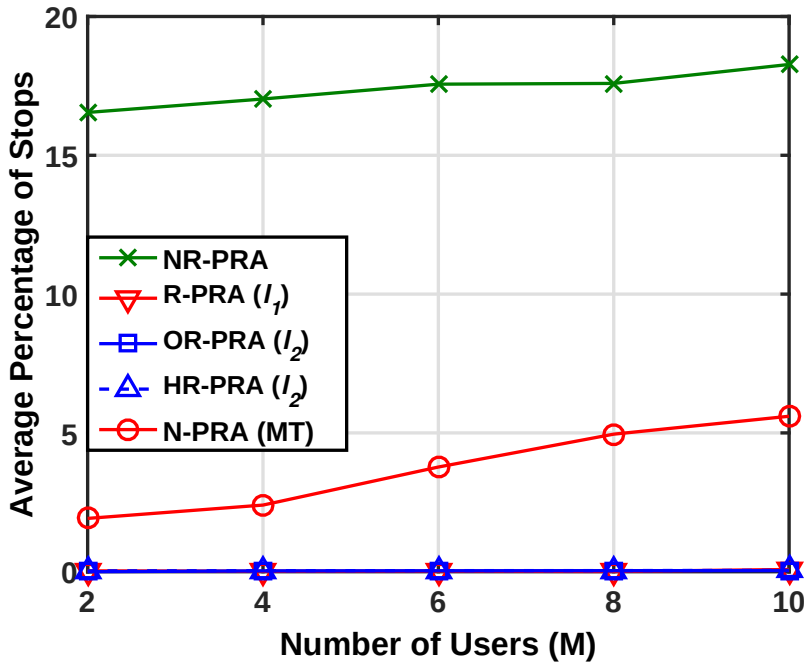

(b) Average BS airtime.

Figure 5.7: Performance evaluation for different channel variances at QoS levels $(1-\epsilon) = 0.9$ and 8 users requesting high quality video

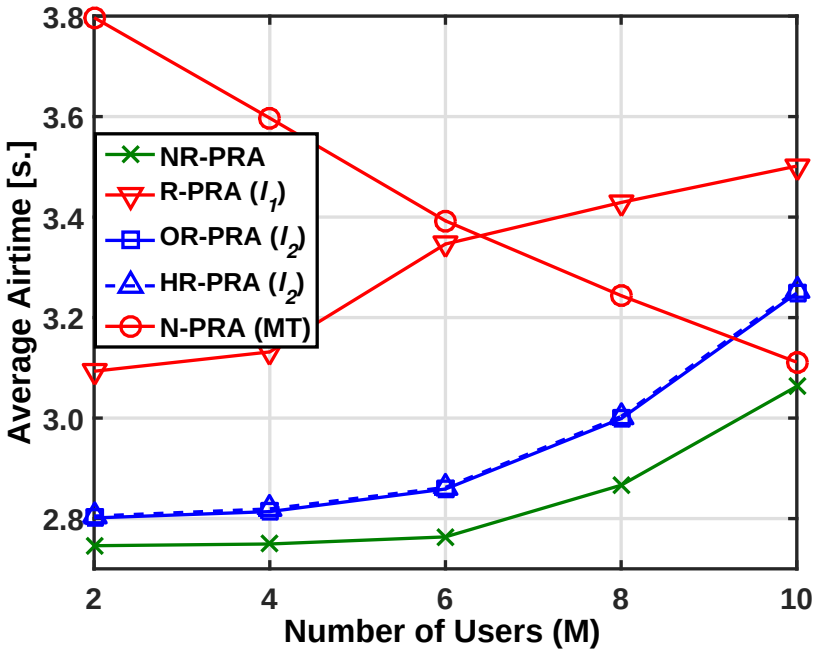
forms. The results also demonstrate the conservatism of the linearized BA used in [43], which decreases the energy saving gain especially at very high channel variances. The energy consumption thus increases and becomes comparable to that of the non-predictive strategy.

Both the load per user and the moving speed are then decreased to medium quality videos (i.e. $V_i = 0.5Mbps$) and 25 Km/h, respectively, to allow more users and higher QoS levels in the simulation scenario. The conservatism of the linearized BA becomes more significant as it consumes more energy than the non-predictive MT at high QoS level (i.e. low $\epsilon$) and high channel variances as in Fig. 5.8(b) and Fig. 5.9(b). The BA in its original SoCP form, however, is able to preserve the prediction gain at these high load conditions. While the energy savings gap between, the predictive and non-predictive schemes decrease for this scenario, the latter fails to meet the QoS level as shown in Fig. 5.8(a) and Fig. 5.9(a). This is because such non-predictive strategy greedily allocated the resources to the cell peak users and ignored serving the cell edge users in order to maximize the total system throughput.

Similar observations are noted for the conservative linearized BA, NR-PRA and MT when the number of users and the QoS level are increased as shown in Fig. 5.9(b). The distributions of QoS satisfaction and degradation are reported in Fig. 5.10(a) and Fig. 5.10(b), respectively. The percentage of users with violated QoS levels mainly depends on their mobility traces and experienced channel rates. In Fig. 5.10(a), the percentage of users with violated QoS levels was around $50\%$ in case of the non-robust PRA. This was found to be the same percentage of users who started the video streaming at the cell edge, and thus were subjected to minimal allocation strategy resulting in buffer underrun. In Fig. 5.10(b),

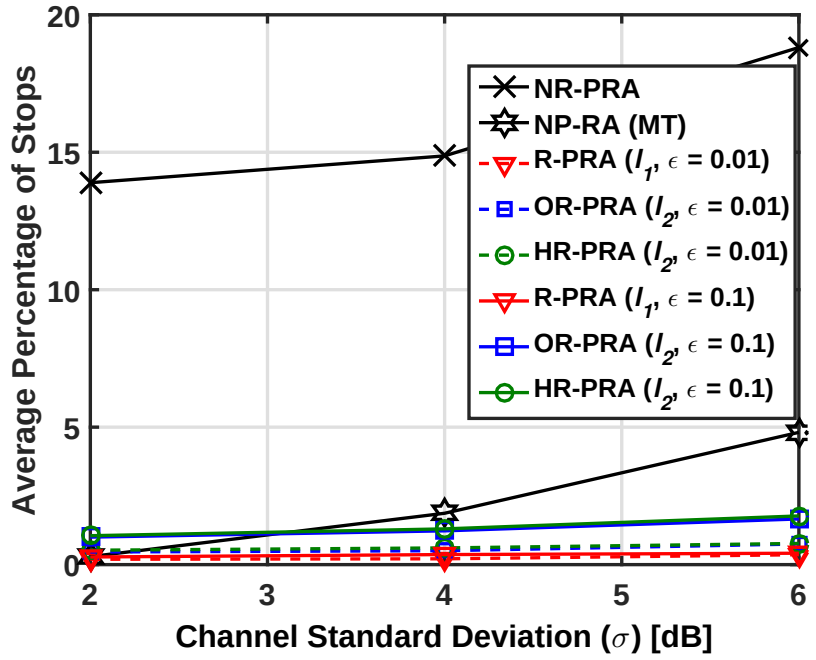



(a) Average Percentage of video stops.

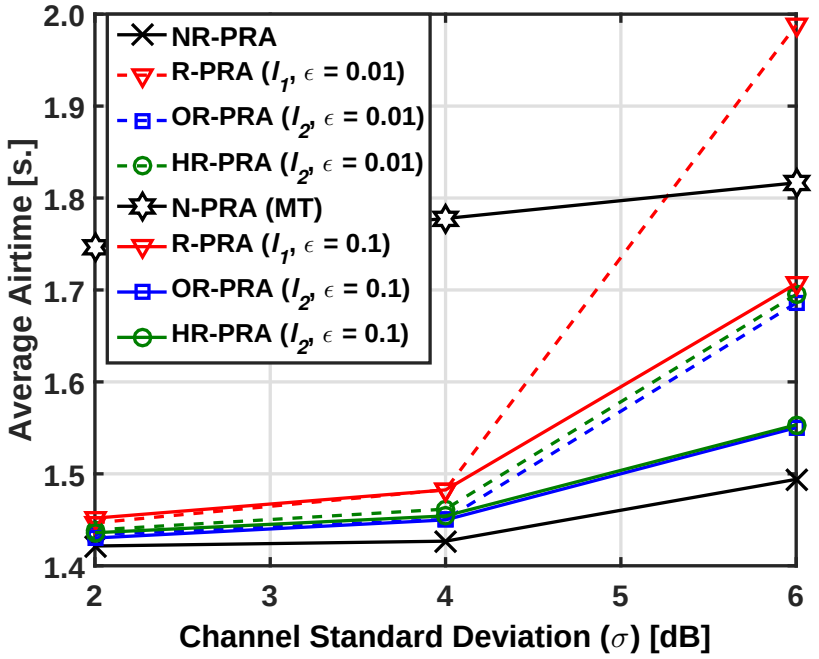



(b) Average BS airtime.

Figure 5.8: Performance evaluation for different number of users requesting HQ at QoS levels $(1-\epsilon) = 0.95$ and experiencing $\sigma = 4$

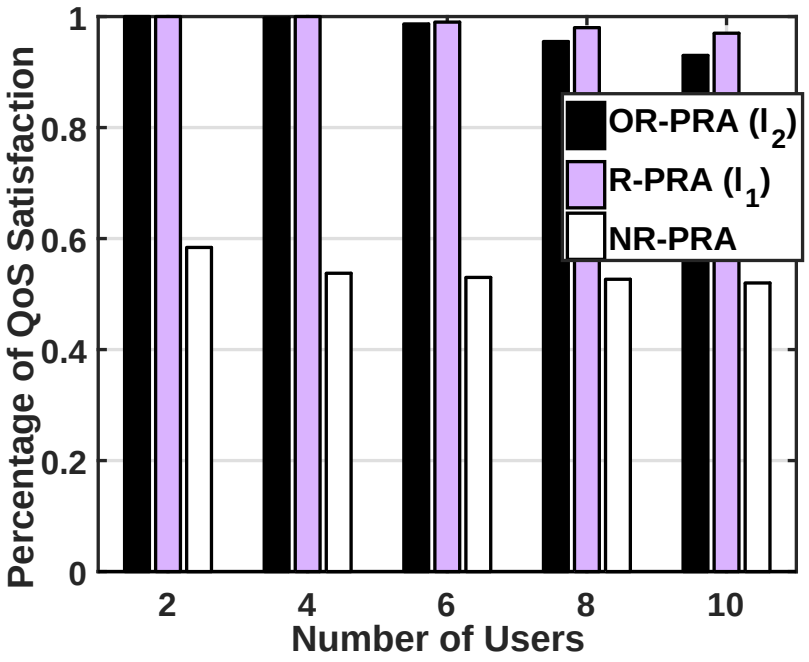



(a) Average Percentage of video stops.

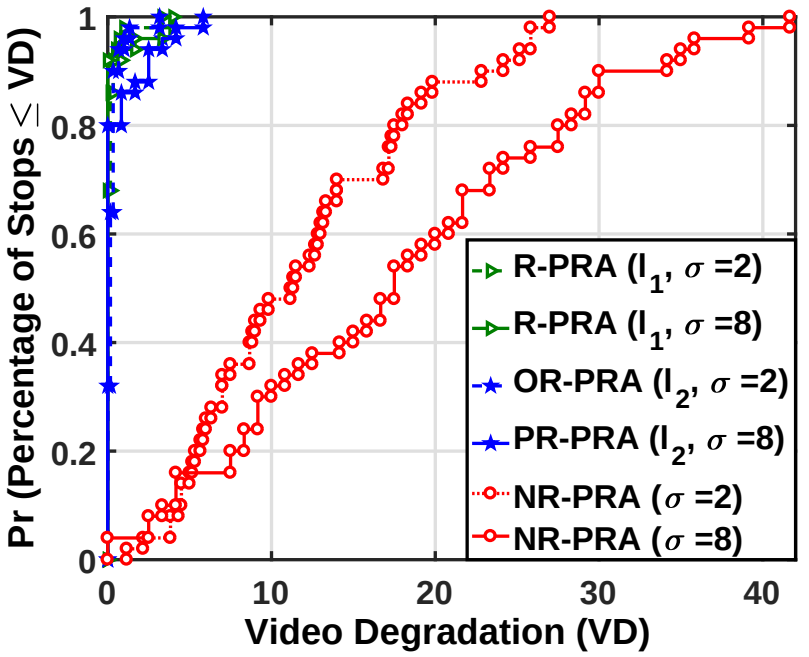



(b) Average BS airtime.

Figure 5.9: Performance evaluation for different channel variances at high QoS levels and 12 users requesting MQ video




(a) Percentage of scenarios with VD below $\epsilon = 0.05$ and $\sigma = 4$




(b) CDF of video degradation for 10 users

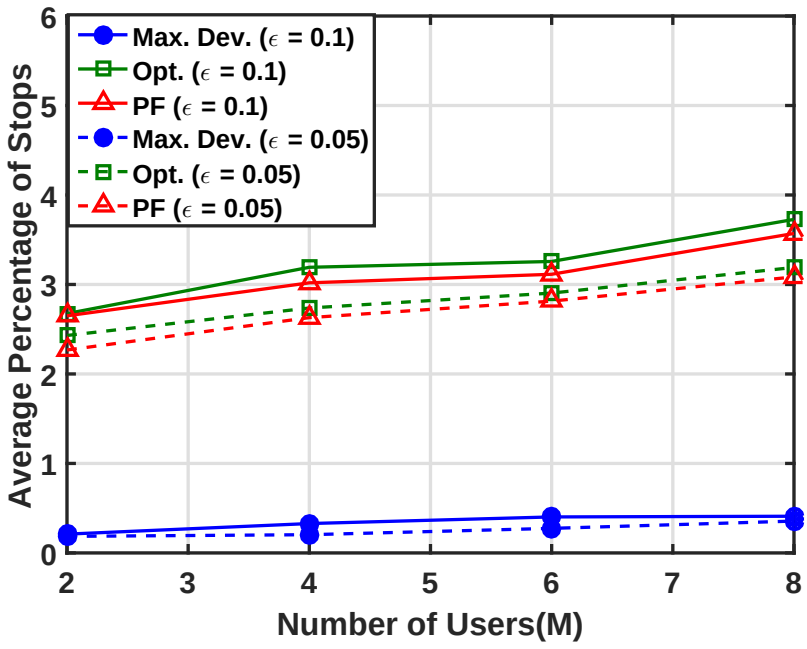
Figure 5.10: Performance evaluation for different channel variances and number of users at QoS levels $(1 - \epsilon) = 0.95$ requesting high quality video

the distribution of video degradation, and its maximum value, illustrate the QoS violation of non-robust PRA. Note that the robust PRA schemes experienced stable QoS performance over the system load and variance. The scenarios above demonstrate that the adopted BA SoCP based PRA formulation: 1) satisfies all QoS levels for different system loads (Fig. 5.8(a)) and 2) preserves the energy-saving gains of the prediction (Fig. 5.8(b)). In addition, the introduced heuristic shows stable performance with a very low optimality gap (< 0.1 %) with respect to the optimal solution's airtime and QoS levels in all considered cases.

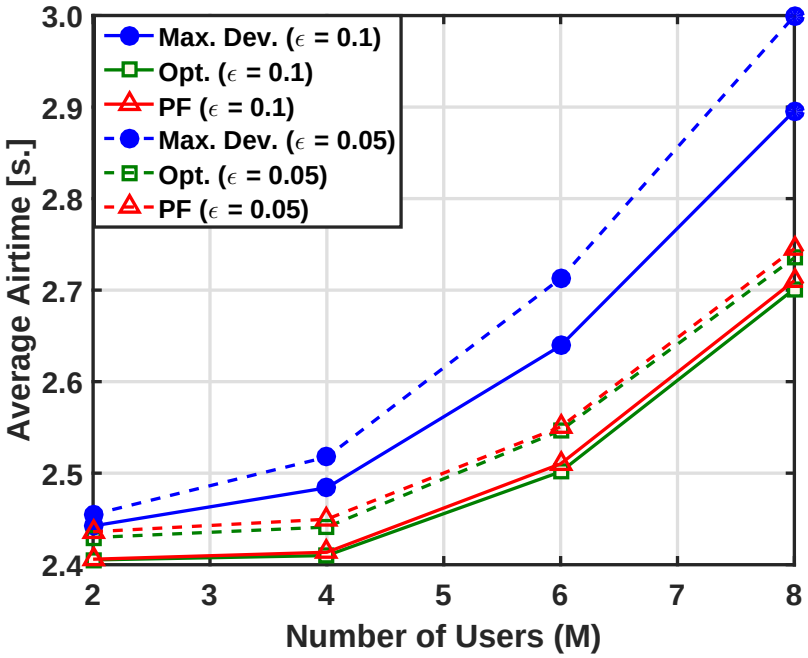
**Performance of Particle Filter**

In this scenario, we assess the ability of the PF to track the rate deviations while adopting the SoCP BA formulation. We compare the PF based variance is compared with both the maximum and optimal theoretical variances denoted by *Max. Dev.* and *Opt.*, respectively. The *Max. Dev.* corresponds to the maximum variance [113] that guarantees the QoS satisfaction under the highest prediction errors. The *Opt.* adopts the exact rate deviation corresponding to the current channel variance. This optimal value satisfies the QoS level without compromising the energy savings. On the other hand, the *PF* initially assumes the highest variance as the *Max. Dev.*, but continuously monitors the channel variance and adapts the rate deviation accordingly.

With regards to QoS satisfaction, the *Max. Dev.* provides a very conservative allocation that greedily satisfies the QoS at the expense of the energy saving as depicted in Fig. 5.11(a) and Fig. 5.11(b), respectively. This is not the case for *PF* which has met the constraint at nearly the exact level as the *Opt.*, resulting in high energy savings. The PF, in essence, decreases the initial maximum rate deviation to reach the lower optimal value and sometimes below. Although going below the optimal rate deviation value increases the risk

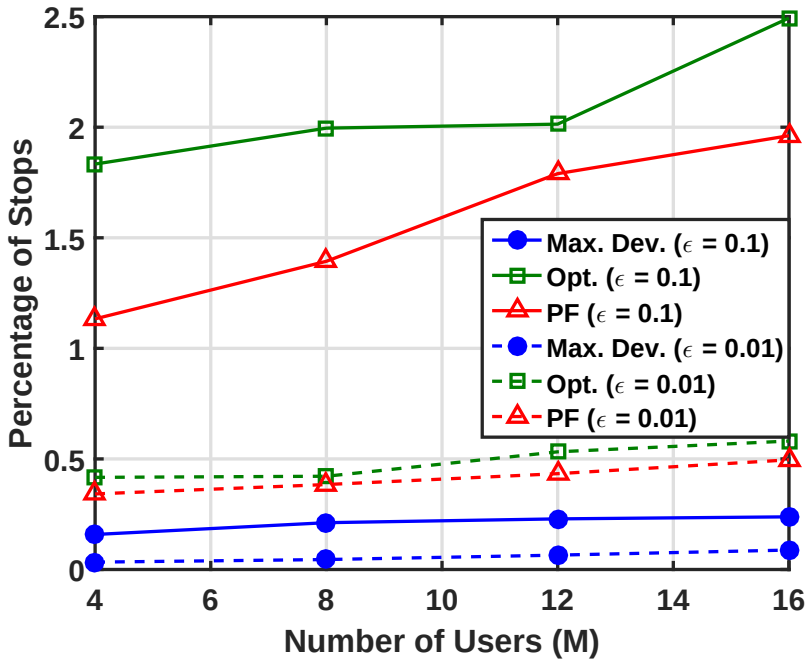



(a) Average Percentage of video stops.

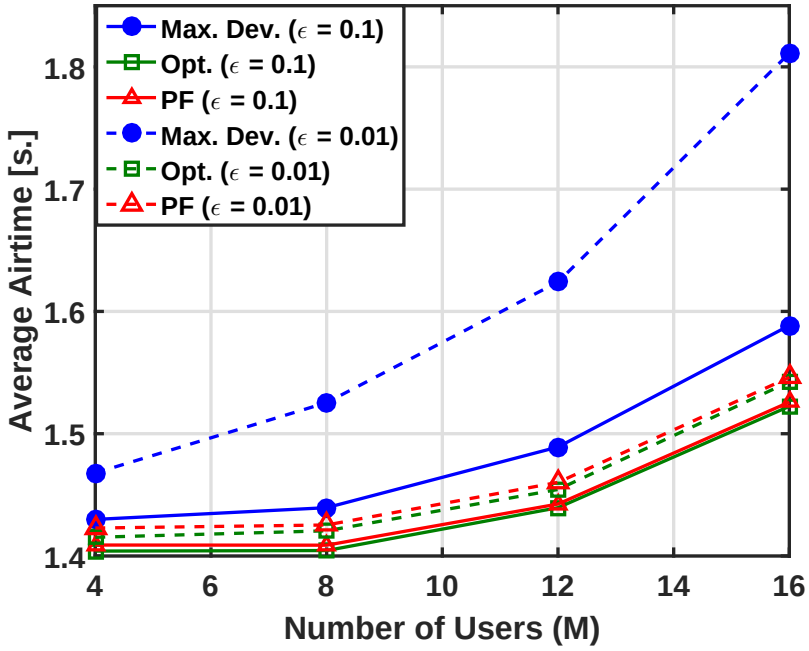



(b) Average BS airtime.

Figure 5.11: Performance evaluation for the robust framework with channel tracking for different number of users experiencing $\sigma = 2$ and requesting MQ video with high QoS level $(1 - \epsilon) = 0.95$




(a) Average Percentage of video stops.




(b) Average BS airtime.

Figure 5.12: Performance evaluation for the robust framework with channel tracking for different number of users experiencing $\sigma = 2$ and requesting LQ video with high QoS level

Table 5.5: Execution Time of the Simulated Schemes

| Technique | Number of Users | | | | Streaming Rate (V) [Mbps] | | |
|---|---|---|---|---|---|---|---|
| | 2 | 4 | 8 | 12 | 0.25 | 0.5 | 1 |
| N-PRA (MT) | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms |
| NR-PRA | 1 s. | 1.5 s. | 2.3 s. | 4 s. | 4 s. | 4 s. | 4 s. |
| OR-PRA ($l_2$) | 50 s. | 80 s. | 150 s. | 250 s. | 200 s. | 250 s. | 290 s. |
| HR-PRA ($l_2$) | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms | <0.1 ms |
| OR-PRA ($l_1$) | 1 s. | 1.5 s. | 2.3 s. | 4 s. | 4 s. | 5 s. | 5.5 s. |

of constraint violation, the conservative BA based allocation in early timeslots avoids such QoS degradation case. The energy gain of the PF-based channel tracking relative to the maximum deviation has increased in the high load scenarios (i.e. more number of users) at high QoS levels and reached up to 15 % as shown in Fig. 5.12(b). This adaptation mechanism results in nearly the same energy savings as the optimal deviation case and with better QoS satisfaction as less video stops have been experienced in the early slots as shown in Fig. 5.12(a) and Fig. 5.12(b).

**Runtime Complexity**

We also report the execution time of all the examined RA schemes in Table 5.5, and measured within the simulation environment on a Quad Core i7-Processor, 3.2 GHz machine. These results highlight the efficiency of the guided heuristic solution methods for providing real-time implementation under different load scenarios. The complexity of the optimal solver increases with *both* the number of users (i.e. the problem dimensions) and the streaming rate (V) since more iterations are required to reach a feasible solution. As opposed to the solver, the guided heuristic resulted in a stable scalable performance regardless

the value of the aforementioned two parameters and with a delay less than the duration of Time Transmission Interval (TTI).

## 5.4 Discussion and Comparison between GA and BA

The Gaussian approximation, in the ICCP form, is theoretically the optimal deterministic representation for the chance constraint where the QoS satisfaction is guaranteed with probability $(1 - \epsilon)$. However, this performance is attained while assuming that the variations in the predicted rates follow the normal distribution entirely. Such assumption does not sustain practically when other imperfect predictions are considered (e.g. user's trajectory). In that case, new rate distributions need to be computed, which neither guarantee to follow the Gaussian nor provide an invertible CDF. Consequently, Bernstein approximation has to be applied to approximate variations by the bounds and thus avoids both the computational effort of deriving new distribution or calculating the inverse of CDF. Nevertheless, this approximation is at the expense of the solution's optimality due to the fact that the logarithmic moment generating function, and its upper bound, are more conservative than the inverse of CDF. Therefore, one might be interested in finding the cost of robustness in applying the Bernstein approximation rather than the Gaussian one. As a result, the trade-off between saving the effort of deriving the inverse CDF for the Gaussian and the extra conservatism cost of the Bernstein has to be compared.

### 5.4.1 Analytical Comparison

Assuming the case when the Gaussian is the optimal theoretical approximate (i.e. predicted rate variations follow a normal distribution), the conservatism cost of the Bernstein is calculated as a function of the difference in their safety terms. In particular, Gaussian and Bernstein safety terms are denoted as $S_G$ and $S_B$, and deduced from Eq. 5.5 and Eq. 5.30

respectively as shown below.

$$
\begin{aligned}
S_G &= Q_\epsilon^{-1}\sqrt{\sum_{t=0}^{t'}(x_{i,t}\sigma_{i,t}^r)^2}, \\
S_B &= -\sum_{t=1}^{t}\mu_{i,t}^-\hat{r}_{i,t}x_{i,t} + \sqrt{2log(\frac{1}{\epsilon})\Big(\sum_{t=1}^{t'}(\sigma_{i,t}\hat{r}_{i,t}x_{i,t})^2\Big)}, \\
&\qquad\qquad \forall\, i \in \mathcal{M},\ t' \in \mathcal{T}.
\end{aligned}
\tag{5.42}
$$

The difference between the above terms is denoted as $S_{B-G}$ and calculated in Eq. 5.43:

$$
\begin{aligned}
S_{B-G} =& S_B - S_G \\
=& -\sum_{t=1}^{t'}\mu_{i,t}^-\hat{r}_{i,t}x_{i,t} + \sqrt{2log(\frac{1}{\epsilon})\Big(\sum_{t=1}^{t'}(\sigma_{i,t}\hat{r}_{i,t}x_{i,t})^2\Big)} \\
&- Q_\epsilon^{-1}\sqrt{\sum_{t=0}^{t'}(x_{i,t}\sigma_{i,t}^r)^2}, \qquad\qquad \forall\, i \in \mathcal{M},\ t' \in \mathcal{T}.
\end{aligned}
\tag{5.43}
$$

From [130], $\mu_{i,t}^-$ and $\sigma_{i,t}$ are set to $-0.5$ and $1/\sqrt{12}$ respectively. For simplicity, the first time slot for the first user is considered and thus the summation is removed as well as the subscripts i and t. Moreover, the bounds of $99.7\%$ of the samples in case of Gaussian can be expressed in terms of the standard deviation $\sigma^r$ as: $\hat{r}_G = 3\sigma$. Accordingly, Eq. 5.43 can be expressed as depicted below:

$$
S_{B-G} = \hat{r}x\left(1 + \sqrt{\frac{1}{6}log(\frac{1}{\epsilon})} - \frac{Q_\epsilon^{-1}}{3}\right) \tag{5.44}
$$

The positivity of Eq. 5.44 indicates that more airtime is assigned by the Bernstein approximation (i.e. high conservatism) than the Gaussian for all practical value of QoS (i.e. $\epsilon < 0.5$). This gap increases with both the QoS level (i.e. $\epsilon$ decreases) and the absolute deviation $\hat{r}$ in the predicted rate. PF based tracking is a potential solution for minimizing

such gap, and so the conservatism cost. Nevertheless, changing the feedback interval to decrease the conservatism is tested in the next subsection.

### 5.4.2 Numerical Comparison

We define the feedback interval $\tau$ in which the solver performs reallocation of all users while considering their total transmitted data. For generating the shadowing based rate variations, the 3GPP slow fading correlated model is used [113]. Simulation results are averaged over 50 runs with different shadowing values. Two mobility scenarios were considered; urban and rural. Users move at a low speed with small inter-vehicle distances in the urban scenario, and thus experience similar average rate values at the same time interval. The rural scenario models high speed moving vehicles with large inter-vehicle distances. Consequently, users experience different data rates from each other at the same time interval. Video content is then requested by all users at a fixed streaming rate over the considered time horizon. The numerical values of all the parameters are summarized in Table 5.1 and Table 5.4, while the variance and bounds of each rate are calculated using the previously discussed Monte-Carlo simulation.

#### Robustness in the Urban Scenario

In urban areas, users start moving from the cell edge towards the centre. In order to decrease computational complexity of the solver, the feedback time $\tau$ was set firstly to a relatively long interval equal to $10\ s$. This is the interval over which the solver recalculates the allocation of all users for the remaining future time slots. In case of GA, the maximum degradation was surpassed for high QoS (i.e., $1 - \epsilon \geq 0.9$) as shown in Fig. 5.13(a). This performance is attributed to the overlooked dependency between the QoS constraints over time. Consequently, demand violation at a certain slot will propagate and affects

the satisfaction in the next slots within the feedback interval. Such violations last until reallocation is done for the next slots. The value of $\tau$ was set to lower values $\tau = 1$ and $5s.$ where less degradation occurs Fig. 5.13(a), but at the expense of both: increased airtime Fig. 5.13(b) and the computational complexity.

The BA approach is very conservative, and thus the percentage of stops was kept below the maximum threshold for all the QoS levels and feedback values of $\tau$ as shown in Fig. 5.13(a). However, the airtime performance with $\tau$ is opposite to that of GA. This is due to the fact that users are moving from a region of low rate towards the cell peak, and BA requires fast feedback to decrease the conservative allocation at the cell edge which consumes more airtime. Large feedback durations continue to allocate large amounts of data at the cell edge.

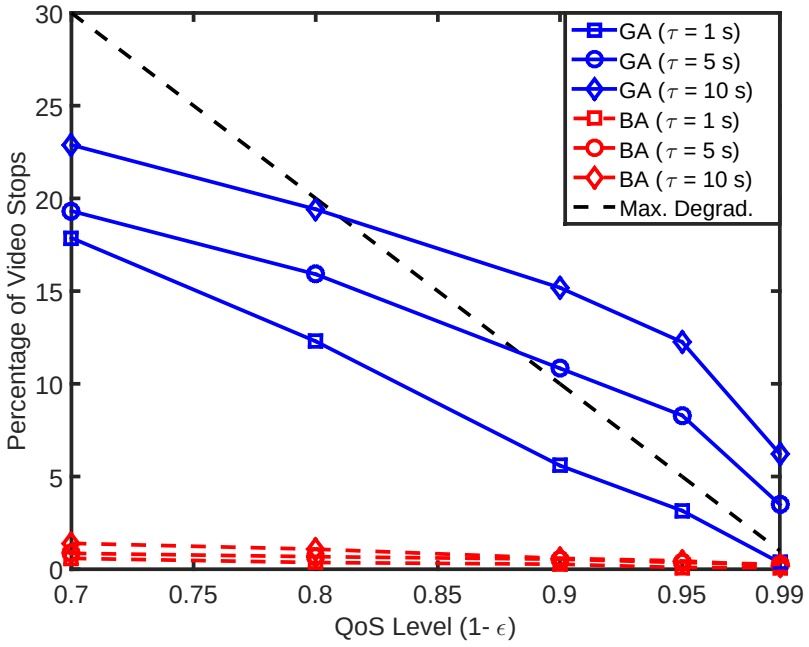
BA requires small feedback durations to correct its conservative allocation. Similarly, GA also requires the same small feedback time but to recover the degradation in any timeslot and prevent it from affecting the coming ones. The allocation for user 1 in Fig. 5.14(a) demonstrates the aforementioned properties. In GA Fig. 5.14(a) where degradation occurs at the first time slot, the small feedback ($\tau = 1\ s$) was able to recover this by recalculating the allocation at the next time slot ($t = 2\ s.$). On the other hand, Bernstein's conservatism avoided the degradation in any of the time slots. However, conservative airtime allocation at early slots (where the rate is minimal) was avoided by frequent feedback, while allocation continues conservatively (large gap above the demand) for the case of $\tau = 10\ s$ as depicted in Fig. 5.14(b).

**Robustness in Rural Scenario**

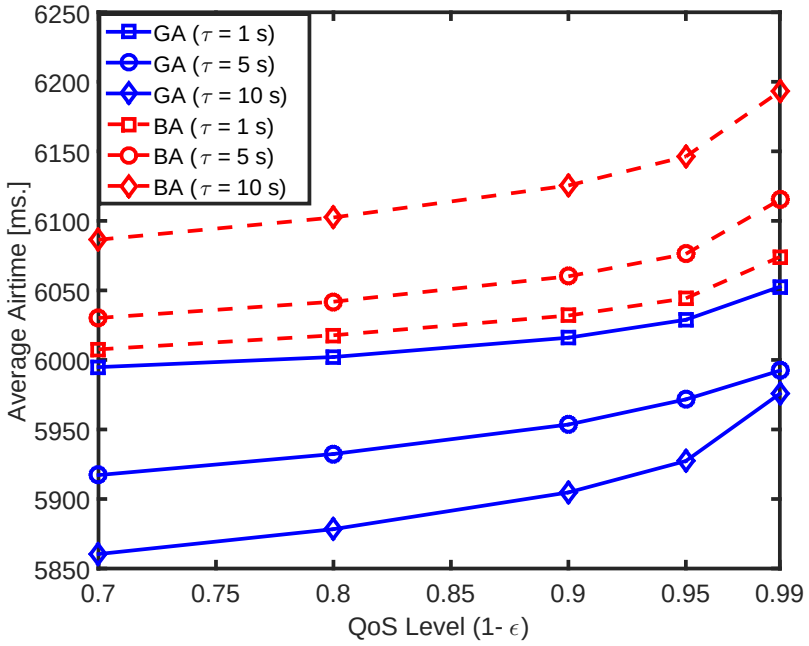
The above conclusions were drawn for the case of users experiencing similar radio conditions at the same time. Thus, very conservative solutions only affects the optimality of

each user individually. We now consider the rural scenario where some users are located as the cell edge while others are at the cell peak and moving towards the edge. Minimal allocation, to satisfy the QoS, is performed for the users at the cell edge while prebuffering is done for the cell peak users to avoid allocation at future low rate locations. In this scenario, the conservatism of cell edge users is more severe and affects the optimality of cell peak users as well due to the provided small airtime for prebufffering. An example of such a case is shown for user 2 (located at cell peak) in Fig. 5.14(b). Due to the conservative allocation of user 1 located at cell edge for $\tau = 10s.$, user 2 was unable to prebuffer in the first 10 seconds while located at the cell peak. Thus, the peak user had to wait until reallocation of the cell edge user at $t = 10s.$ so more airtime is provided for the former to prebuffer at relatively lower rates.

Accordingly, the cost of conservatism in the rural scenario has increased and thus the energy gap expanded between Bernstein at ($\tau = 5$ and $10\ s.$) and the less conservative cases: i.e., Bernstein ($\tau = 1\ s.$) and Gaussian as shown in Fig. 5.15(a). The frequent feedback of Bernstein (i.e. $\tau = 1\ s.$) was able to overcome its expected conservatism and thus results in nearly equal energy consumption compared to the Gaussian case at the same feedback interval. Moreover, the QoS satisfaction of large feedback intervals ($\tau = 5$ and $10\ s.$) is slightly enhanced for the Gaussian case where violation of the maximum degradation occurs only at the highest QoS level for $\tau = 5\ s$, and at the highest two QoS values for $\tau = 10\ s.$ as depicted in Fig. 5.15(b). This is attributed to the prebuffering strategy for the cell peak users and thus their QoS satisfaction never fails resulting in lower average violation.

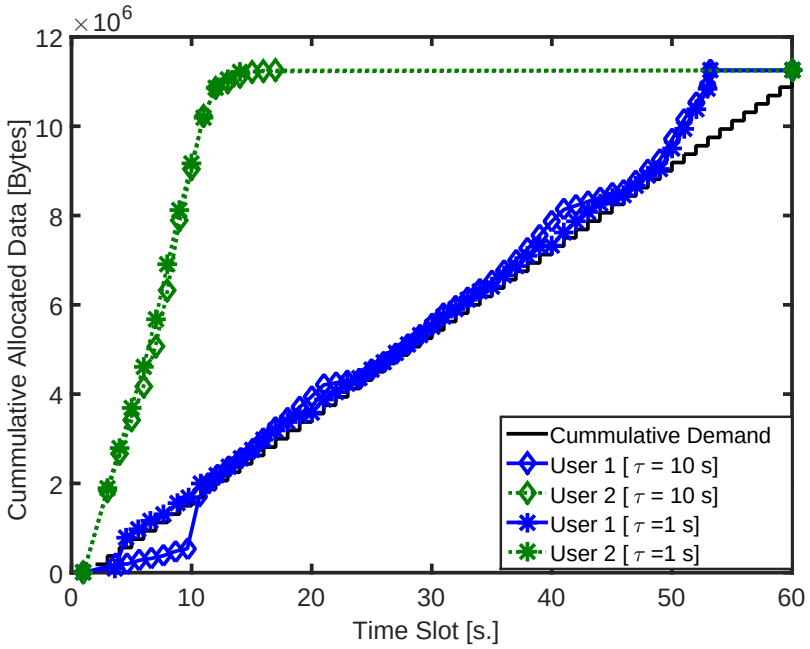



(a) Average Percentage of video stops.

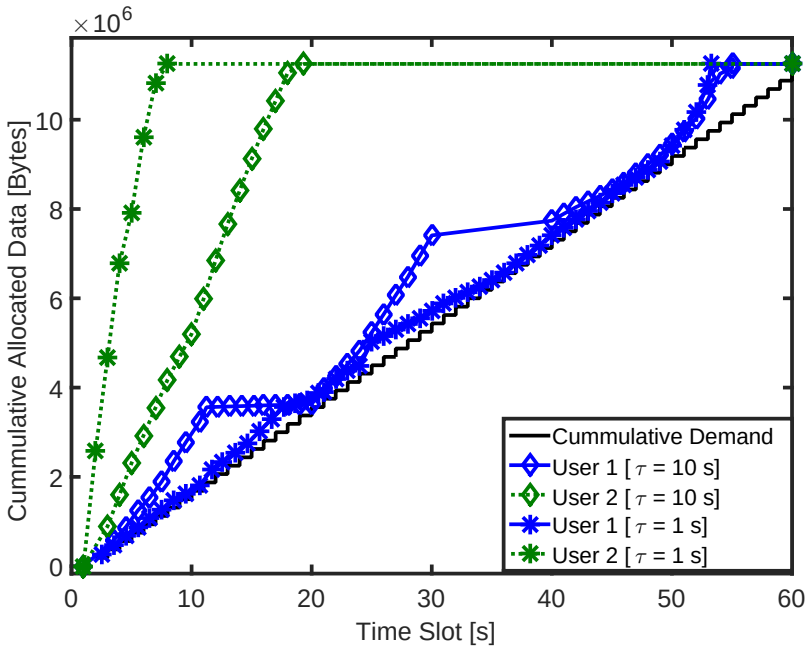



(b) Average BS airtime.

Figure 5.13: Percentage of video stops and average BS airtime for varying QoS levels $(1-\epsilon)$ for 2 users experiencing slow fading with Non Line of Sight (NLoS) variance in urban area

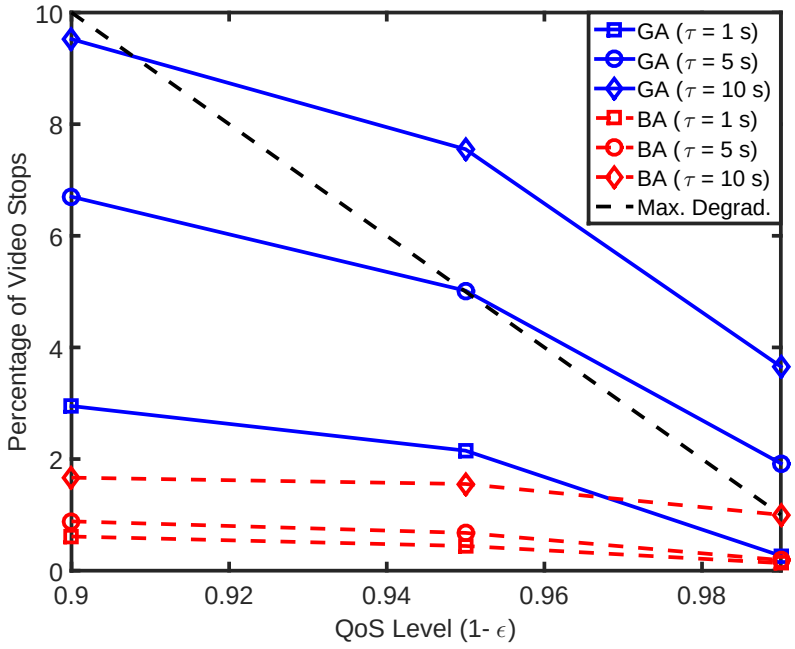



(a) Gaussian approximation

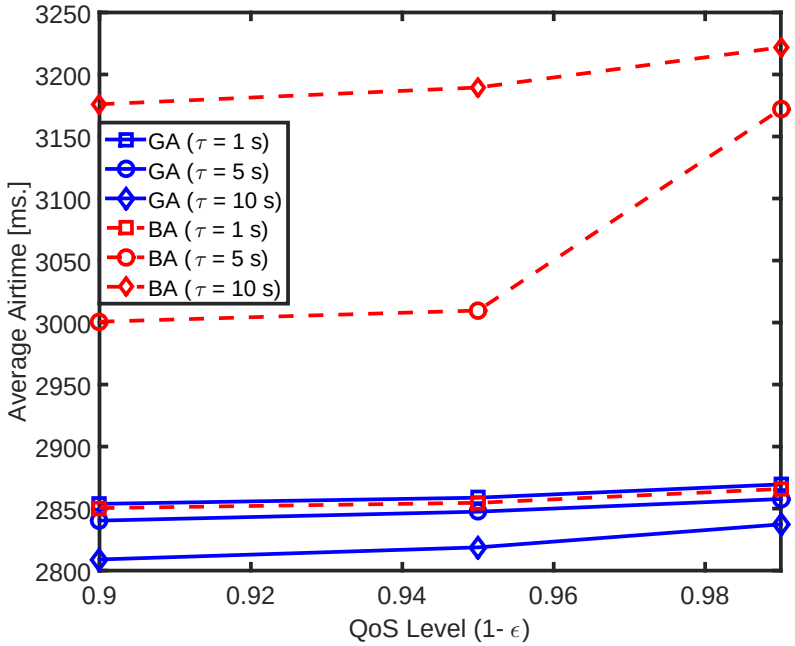



(b) Bernstein approximation

Figure 5.14: Allocation at different feedback intervals for 2 users experiencing slow fading with LoS variance




(a) Average Percentage of video stops.




(b) Average BS airtime.

Figure 5.15: Performance of the robust framework for varying QoS levels $(1-\epsilon)$ for 2 users experiencing LoS variance in rural area.

# Chapter 6

# Robust-Green PRA under Demand and Resources Uncertainty

In the two variants of previous chapter, we showed how the rate uncertainties impact the QoS satisfaction for cell edge users, and demonstrated the importance of *robust* scheme. This chapter introduces the third variant which handles the uncertainties in both the demand and resources. Hence, avoids energy consumption in the case of cell center users terminating the session, and achieves QoS satisfaction to all users when network resources fluctuates due to arrival of real-time traffic. This variant is referred to as *Robust*-Green Predictive Resource Allocation (R-GPRA) and adopts both CCP and RP as illustrated in Fig. 6.1. Similar to the previous two variants, the R-GRPA aims to minimize the total energy and delivers the video at a predefined quality level.

## 6.1 System Model

### 6.1.1 Resource Allocation

The users of the same BS share the available radio resources every time slot $t$, where each user $i$ is allocated a fraction of the slot's airtime denoted by $x_{i,t} \in [0, 1]$. Other real-time users are sharing the same resources, but their allocation is not handled by the R-GPRA.

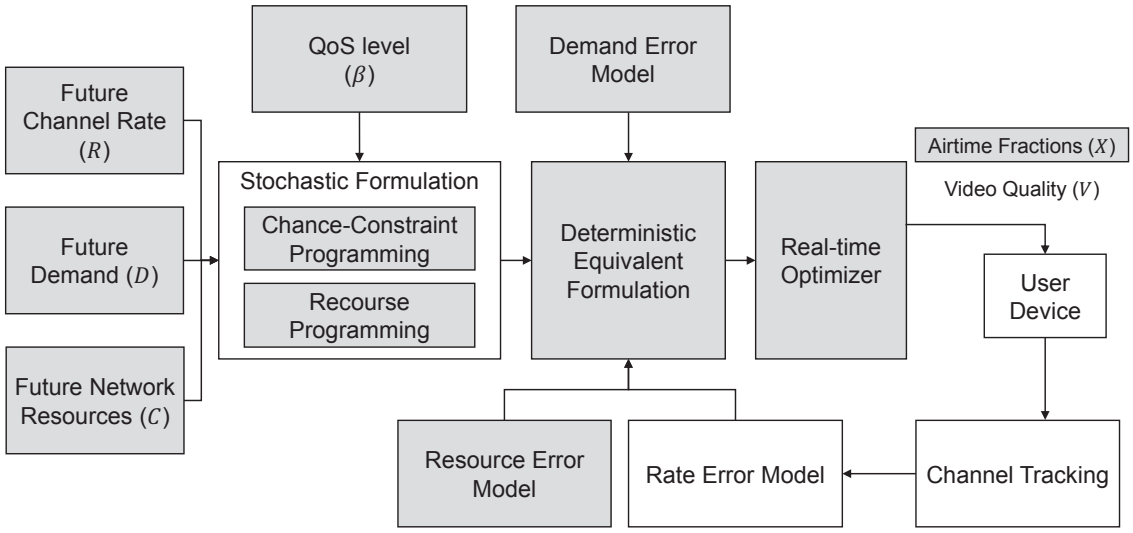



Figure 6.1: Block diagram of energy-saving R-PRA scheme under demand and resource uncertainty

### 6.1.2 Demand Uncertainty Model

The average demand of user $i$ at time slot $t$ is denoted by $v_{i,t}$ which corresponds to the data content played back with fixed quality. Herein, we assume that the demand is uncertain as the user can terminate the video at any time slot. Accordingly, the per slot demand is modeled as a random variable $\tilde{v}_{i,t}$ that is equal to 0 (user terminated the video) or $v_{i,t}$ (user streaming the video). The cumulative demand is thus denoted as a random variable $\tilde{D}_{i,t} = \sum_{t'=0}^{t} \tilde{v}_{i,t'}$.

### 6.1.3 Radio Network Resources Uncertainty Model

At each time slot, the resources are shared among both the streaming users (considered by the R-GPRA) and other real-time users. The traffic of the latter is modeled using their arrival rate and demanded resources. The arrival of real-time users is modeled as a Poisson distribution with mean $\lambda$, and the demand per user is denoted by $C_{i,t}$. The total airtime share

allocated to real-time users at time slot $t$ is denoted by the random variable $\tilde{C}_t = \sum_{i=0}^{\tilde{N}} C_{i,t}$, where $\tilde{N}$ is a random variable representing the number of real-time traffic users at time slot t.

### 6.1.4 Problem Description

The R-GPRA scheme aims to calculate the airtime fractions $x_{i,t}$ for each user at time slot $t$ such that the total allocated resources are minimized to achieve energy-saving or efficient bandwidth utilization. The possibility of terminating the video by the user at a certain time slot is taken into account. By doing so, this prevents the PRA from prebuffering future content to users who might terminate the video at any time slot with a certain probability. Typically, this probabilistic strategy results in more energy savings and optimal bandwidth utilization compared to existing *non-robust* PRA that assumed perfect demand prediction.

As illustrated in Fig. 6.2 (a), the values of predicted rates for three time slots would typically drive a non-robust GPRA to prebuffer the whole content during the first slot to save energy as depicted in Fig. 6.2 (c). However, as shown in Fig. 6.2 (b), the high probability of terminating the video at the third time slot prevents the *robust* GPRA from prebuffering the future content due to the high risk of wasting energy. As such, only the content of the second slot, with low probability of video termination, is prebuffered whereas the delivery of the third slot's content will be postponed as illustrated in Fig. 6.2 (d). To summarize the example, delivering the rest of the video content in the third time slot costs more energy, in case of non-termination, while prebuffering all the contents causes a waste of resources in case of a termination of viewing. The proposed robust GPRA calculates this trade-off based on both the predicted rates and the probability of termination to perform the energy-efficient and QoS-aware allocation.

The uncertainty of future network resources, due to random user arrival, will interfere

with the strategy mentioned earlier. Delaying the transmission in case of high termination probability might be considered suboptimal if the future network resources are scarce. The network, in that case, will miss the chance of exploiting the current channel peaks and vacant resources, thus will not be able to satisfy the user demand with the future anticipated limited resources. As a result, fewer energy-savings are attained in case of future peaks with low resources, while video stops are observed if future low channel rates are further reduced by real-time users arrival.

## 6.2 Problem Formulation

In this section we mathematically formulate the problem of *robust* GPRA (R-GPRA) using stochastic optimization, and then adopt recourse and chance constraint programming to obtain deterministic equivalent forms.

### 6.2.1 Stochastic Model

The introduced *energy-efficient robust* GPRA is formulated using stochastic optimization. In particular, the uncertain demand and future network resources are represented by random variables as follows:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} \right\} \tag{6.1}$$

subject to:

$$\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} \tilde{D}_{i,t}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1 - \tilde{C}_t, \qquad \forall t \in \mathcal{T},$$

$$\text{C3:} \quad x_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

The objective function aims to minimize the total consumed energy represented as a function of the total BS airtime [115]. The QoS constraint in C1 guarantees that the total delivered content to the user satisfies the anticipated cumulative random demand. C2 models the limited resources at each BS by ensuring that the sum of allocated airtime is less than the total available network resources (allocation slot duration) while considering the random resources allocated to the real-time users. The last constraint C3 ensures the non-negativity of the decision variables. The main difference between the proposed *robust* formulation and the existing PRA work is the first and second constraints that now incorporate random demand and network resources. Such randomness has an impact on both objective function value and QoS satisfaction. In particular, when the random demand equals to $v_{i,t}$, the objective function is minimized by prebuffering the future content during slots of peak rates. On the other hand, when the random demand becomes $0$ (due to session termination) the objective function is minimized by avoiding prebuffering of future content. Similarly, allocating more resources than the available capacity, after accounting for the real-time users, will result in video stops since the users will not be able to receive the minimal data amount calculated by the R-GPRA. As such, the network should avoid prebuffering when available resources are low due to periodic arrival of real-time users.

### 6.2.2 Recourse and Chance Constrained Model

To represent the relation mentioned above between constraints C1, C2 and the objective function in a deterministic form, Recourse Programming (RP) and Chance Constrained Programming (CCP) models are used as depicted below:

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} + \mathbb{E}\big[H(\mathbf{y}, \tilde{D})\big] \right\} \tag{6.2}$$

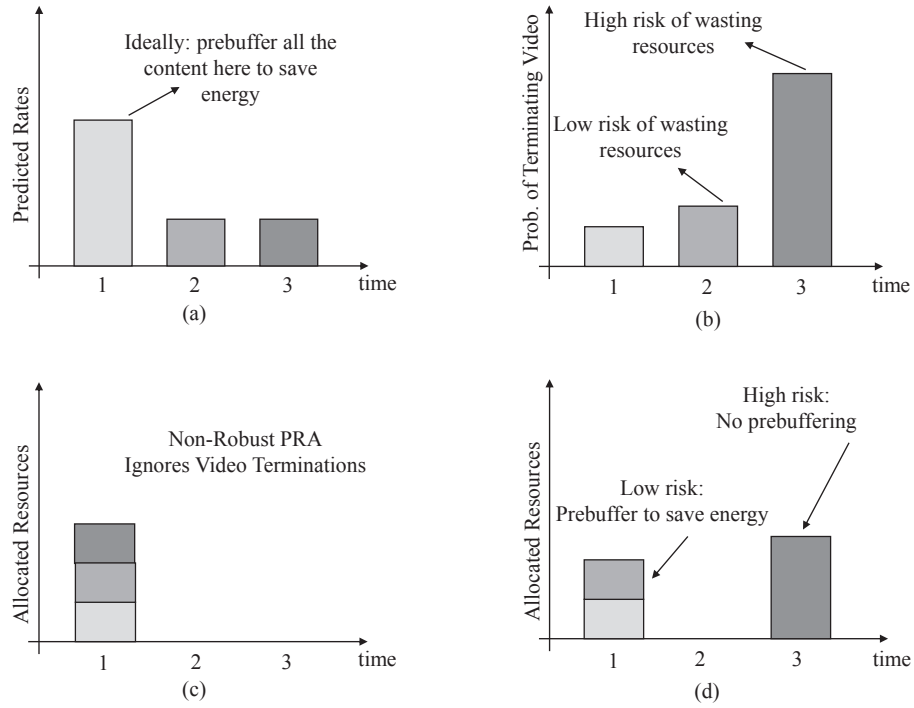



Figure 6.2: Illustration of Robust-GPRA under uncertain video streaming demand

subject to:

$$\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} v_{i,t}, \qquad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C2:} \quad P_r(\sum_{i=1}^{M} x_{i,t} \leq 1 - \tilde{C}_t) \geq \beta, \qquad \forall t \in \mathcal{T},$$

$$\text{C3:} \quad x_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

The objective function herein comprises of two terms whose summation must be minimized. The first term represents the total allocated resources (similar to the non-robust approach) while the second term corresponds to the total amount of wasted resources as a result of terminating the video before watching the prebuffered content. In C2, the probability of satisfying the network resource constraint by the calculated airtime fractions is set

above the QoS level $\beta$. Where $\beta \in [0, 1]$ represents the minimal probability of satisfying the QoS. In the following, we show how to obtain a closed form representation for both the recourse model in the objective function, and the probabilistic constraint in C2.

**Recourse Stage**

The second term of the objective function in Eq. 6.2, i.e. $\mathbb{E}\big[H(\mathbf{y}, \tilde{D})\big]$, is the optimal solution of the recourse stage and formulated as follows:

$$\underset{\mathbf{y},\mathbf{x}}{\text{minimize}} \quad \left\{ \zeta \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} p^{W}_{i,t} y_{i,t} \right\} \tag{6.3}$$

subject to:

$$\text{C4:} \quad r_{i,t-1} y_{i,t-1} + r_{i,t} x_{i,t} - v_{i,t} \leq r_{i,t} y_{i,t}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C5:} \quad y_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

The objective function of the recourse stage in Eq. 6.3 minimizes the expected value of excess allocated resources (i.e. prebuffered) and calculated as a function of both the second stage decision variable $y_{i,t}$ and the probability of terminating the video denoted by $p^{W}_{i,t}$. The variable $\zeta$ is used to model the trade-off between the values of the two stages, and its value is typically less than one. The constraint in C4 is used to calculate the excess resources $r_{i,t} y_{i,t}$ after every time slot $t$. The first two terms on the left hand-side represent the total prebuffered and newly allocated resources in this time slot, respectively. The third term represents the per slot demand. The right hand-side shows the amount of excess resources after slot $t$ which corresponds to the prebuffered future content.

**Deterministic Equivalent**

The probabilistic constraint in C2 is replaced by the following deterministic equivalent form which adopts the probability of arrival of real-time traffic users and their load.

$$\begin{aligned}
&\text{C6:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1 - (C_{t,\omega}\delta_{t,\omega}) \quad \forall t \in \mathcal{T}, \forall \omega \in \Omega, \\
&\text{C7:} \quad \sum_{\forall \omega \in \Omega} \delta_{t,\omega} {p^A}_{t,\omega} \geq \beta \qquad\qquad \forall\, t \in \mathcal{T}. \\
&\text{C8:} \quad \delta_{t,\omega} \in \{0,1\} \qquad\qquad \forall\, t \in \mathcal{T}, \forall \omega \in \Omega,
\end{aligned} \tag{6.4}$$

The binary decision variable $\delta_{t,\omega}$ equals 1 if scenario $\omega$ at time slot $t$ has to be satisfied by the airtime allocation, and equals 0 otherwise. The PDF of user arrival is used to construct the scenarios of network resources at each time slot as a result of real-time traffic user arrival. At each time slot $t$, the scenario $\omega$ represents the existence of $\omega$ real-time traffic users. The constraint in $C6$ demonstrates the scenarios in which the calculated airtime fractions must satisfy the vacant network resources denoted by $1 - C_{t,\omega}$. In $C7$, the total probability of satisfied scenarios must exceed the predefined QoS level $\beta$. The probability of user arrival scenario $\omega$ at time slot $t$ is denoted by ${p^A}_{t,\omega}$. When the scenario is ignored (i.e. $\delta_{t,\omega} = 0$), the right hand-side of $C6$ will be the maximum slot duration (i.e. all network resources are available), and the QoS level $\beta$ will avoid ignoring the most probable scenarios.

### 6.2.3 Deterministic R-GPRA Formulation

The complete deterministic formulation of the proposed R-GPRA can be summarized in the following closed form representation:

$$\underset{\mathbf{x},\mathbf{y},\delta}{\text{minimize}} \quad \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} + \zeta \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} p^W_{i,t} y_{i,t} \right\} \tag{6.5}$$

subject to:

$$\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} v_{i,t}, \qquad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C3:} \quad x_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

$$\text{C4:} \quad r_{i,t-1} y_{i,t-1} + r_{i,t} x_{i,t} - v_{i,t} \leq r_{i,t} y_{i,t}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C5:} \quad y_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

$$\text{C6:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1 - (C_{t,\omega} \delta_{t,\omega}) \qquad \forall t \in \mathcal{T}, \forall \omega \in \Omega,$$

$$\text{C7:} \quad \sum_{\forall \omega \in \Omega} \delta_{t,\omega} {p^A}_{t,\omega} \geq \beta \qquad \forall\, t \in \mathcal{T}.$$

$$\text{C8:} \quad \delta_{t,\omega} \in \{0, 1\} \qquad \forall\, t \in \mathcal{T}, \forall \omega \in \Omega,$$

The above formulation is obtained after combining Eq. 6.3 and Eq. 6.4, resulting in a mixed integer linear programming model. In the next section we explore the possibilities and challenges of solving this NP-complete model, and propose a guided heuristic algorithm for real-time allocation.

## 6.3 Real-time Optimizer

This section reviews the numerical optimization methods that can be used to solve the formulated problem, and introduces the details of heuristic search algorithm followed by analysis of its computational complexity.

### 6.3.1 Optimal Solution

The robust formulation in Eq. 6.5 is a mixed integer linear programming model. As such, an optimal solution, which satisfies all the constraints, can be obtained using branch-and-bound, branch-and-cut or similar techniques in commercial solvers (e.g. Gurobi [124]).

These techniques are capable of reaching the optimal solution with a very small duality gap while satisfying all the constraints. However, the problem at hand requires obtaining an optimal solution in real-time which is unattainable by the commercial solvers or numerical methods that suffer from low scalability and slow convergence. In particular, the complexity of numerical optimization techniques grows exponentially with the number of decision variables [132]. These limitations are due to overlooking the problem structure and exploring a large area of the search space to avoid local optimal solutions. A guided heuristic algorithm is therefore proposed to provide a real-time feasible solution with low optimality gap compared to commercial solvers solutions.

### 6.3.2 Guided Real-time Heuristic

The proposed guided search heuristic algorithm utilizes knowledge about the problem's structure such as the interdependency and conflicts between the constraints, and their impact on the optimality of objective function. In essence, the algorithm starts by satisfying all the QoS constraints using the available radio resources while considering the distribution of user arrival and the predefined QoS level. To achieve energy minimization, resources are allocated to streaming users that have not reached peak channel conditions. Then, the algorithm exploits the prebuffering capabilities of the mobile device for users experiencing peak channel conditions. By doing so, the video content can be pushed in advance to avoid allocation during time slots with low channel rates or high congestion. In the next step, the value of the objective function is further minimized while examining the trade-off between possible energy savings during peak radio conditions, and the risk of wasting resources due to video termination in future time slots. The heuristic is summarized in Algorithm 4 and Algorithm 5, and detailed as follows:

In the first stage, minimal radio resources are calculated (line 2-18) in order to satisfy

the QoS constraint $C1$ in Eq. 6.2 for each slot while considering the network resources uncertainties. The available network resources at each time slot are calculated as follows (lines 2-12):

1. The amount of resources in each scenario are initially sorted in ascending order and probability mass function is sorted accordingly.

2. The scenarios are considered iteratively until the total probability reaches the QoS level $\beta$. Including more scenarios will result in a conservative solution that over-satisfies the QoS and deteriorates the value of the objective function.

3. The resources of the last considered scenario (i.e. the scenario that needs the maximum resources) are selected.

4. The total vacant capacity $C'_t$ remaining for video streaming users is calculated and used in the next stage.

After satisfying constraints $C7 - C8$, the algorithm proceeds to fulfil the per slot demand constraint $C1$. This is accomplished by setting $C1$ to an equality and calculate the resource sharing $x_{i,t}$ that guarantee the satisfaction of demand. Such minimal allocation continues until the user reaches peak radio conditions (line 14). In high load scenarios, due to the large number of users or high streaming rates, the total allocated resources in a certain time slot might violate the airtime constraint $C6$ in Eq. 6.5. Accordingly, the preceding time slots with vacant resources will be used to prebuffer the content of the highly loaded time slots as depicted in lines 19-37 of Algorithm 4. While efficient exploitation of the radio resources is mandatory for these scenarios, the algorithm prebuffers the content of the user with the highest achievable rate. Thus, less airtime is consumed which increases

the chance of satisfying the radio resource constraint $C2$. In case of non-vacant resources, to accommodate the excess demand, the problem is said to be infeasible (lines 34-36).

To further minimize energy consumption, a calculated risk prebuffering strategy is applied by Algorithm 5. In essence, the possibility of prebuffering is checked based on the probability of terminating the video and the difference in channel rates. For each time slot following this peak, the amount of resources in case of prebuffering and non-prebuffering is checked while considering the probability of video termination (lines 3-5) which approximates the objective function in Eq. 6.3. In case of more resource saving (line 7), prebuffering is done (line 8-10). Otherwise, the risk of wasting resources is found to be high and minimal allocation is done for the demand of this slot without prebuffering in the previous slots (lines 13-16).

### 6.3.3 Algorithm Complexity

The first stage of the heuristic consists of sorting the scenarios and calculating the total probability which have complexity of $O(2 \times N^2)$. This stage is repeated for a maximum of $T$ time slots. Thus, the complexity of lines 2-12 is $O(2T \times N^2)$, the minimal allocation in lines 13-18 has complexity of $O(MT)$, while the repairing of resources in lines 19-37 has a complexity of $O(MT^2)$ due to revisiting the preceding time slots to check the possibility of prebuffering. Similarly, the second part of the heuristic has a complexity of $O(MT^2)$ in which previous slots are also revisited for prebuffering any of the future slots with lower rates. Thus, the complexity of the whole proposed heuristic is $O(MT^2)$ which is significantly lower than the mathematical optimization methods whose complexity is non-polynomial and depends on the number of decision variables and constraints.

**Algorithm 4:** QoS Satisfaction under Network Resource Uncertainty

**Input** : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Predicted Rates: $R$, Demand Distribution: $P$, Streaming Rate: $V$ ;
**Output** : $X$;
**Initialization**: $X = \emptyset$, $B = \emptyset$, $Y = \emptyset$, $Z = \emptyset$ $N_t = 0 \; \forall t \in T$;
**Define:** $t'_i = argmax\,\{r_{i,t}, \forall t \in T\}$;
**for** $t \in \mathcal{T}$ **do**
  $\hat{C}'_t = \text{Sort}(P^A_t \forall \omega \in \Omega)$;
  Initialize $S_t = 0$;
  Set minimum capacity $C'_t = 1$ ;
  **while** $S_t \leq \beta$ **do**
    **for** $\omega \in \Omega$ **do**
      Update probability sum: $S_t = S_t + \hat{P}^A_{t,\omega}$;
      Update minimum capacity: $\hat{C'}_t = 1 - \hat{C}_{t,\omega}$;
    **end**
  **end**
**end**
**for** $i \in \mathcal{M}$ **do**
  **for** $t \in \mathcal{T} \,|t \leq t'_i$ **do**
    Calculate minimal airtime $x_{i,t} = v_{i,t}/r_{i,t}$;
    Update used slot fraction $N_t = N_t + x_{i,t}$;
  **end**
**end**
**for** $t \in \mathcal{T}$ **do**
  **if** $N_t > 1$ **then**
    Set $k = t - 1$;
    **while** $k > 0$ & $N_t > C'_t$ **do**
      **if** $x_{i,t} > 0 | i = argmax\,\{r_{i,k}, \forall i \in \mathcal{M}\}$ **then**
        Calculate the violated airtime $\Delta x_{i,t} = N_t - 1$;
        Calculate the demanded airtime $\Delta x_{i,k} = \Delta x_{i,t} \times \frac{r_{i,t}}{r_{i,k}}$;
        **if** $N_k + \Delta x_{i,k} \leq 1$ **then**
          Update $x_{i,k}$, $x_{i,t}$, $N_t$ and $N_k$ ;
          break;
        **end**
      **end**
      $k = k - 1$;
    **end**
  **end**
  **if** $N_t > C'_t$ **then**
    Return Infeasible Problem;
  **end**
**end**

**Algorithm 5:** Calculated Risk Prebuffering for Energy Minimization

**Input** : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Predicted Rates: $R$, Demand Distribution: $P$, Streaming Rate: $V$ ;
**Output** : $X$;
**Initialization**: $X = \emptyset$, $B = \emptyset$, $Y = \emptyset$, $Z = \emptyset$ $N_t = 0 \; \forall t \in T$;
**Define:** $t'_i = argmax \{r_{i,t}, \forall t \in T\}$;
**for** $t \in \mathcal{T} | t > t'_i$ **do**
    Calculate airtime without Prebuffering $x'_{i,t} = v_{i,t}/r_{i,t}$;
    **for** $\tau \in \mathcal{T} \, | \tau < t, r_{i,\tau} > r_{i,t}, B_{i,t} \neq 1$ **do**
        Calculate airtime with prebuffering $z_{i,\tau} = v_{i,t}/r_{i,\tau}$;
        Calculate excess resources $y_{i,\tau} = \gamma \times p^W_{i,t} \times z_{i,t}$;
        **if** $x'_{i,t} > z_{i,\tau} + y_{i,\tau}$ **then**
            Update $x_{i,\tau} = x_{i,\tau} + z_{i,\tau}$;
            Update used slot fraction $N_t = N_t + z_{i,\tau}$;
            Update prebuffering status $B_{i,t} = 1$;
        **end**
    **end**
    **if** $B_{i,t} \neq 1$ **then**
        Update airtime without prebuffering $x_{i,t} = v_{i,t}/r_{i,t}$;
        Update used slot fraction $N_t = N_t + x_{i,t}$;
    **end**
**end**
return $X$

## 6.4 Performance Evaluation

### 6.4.1 Simulation Environment

The proposed R-GPRA is developed in Network Simulator 3 (ns-3) LTE module where Gurobi (a commercial solver) is integrated to obtain benchmark solutions [124]. The probability of terminating the video at any time slot $t$ is calculated using the model in [40]. Users follow random mobility traces within the cell coverage region at a constant velocity typical for suburban areas. The simulation parameters and numerical values are shown in Table 6.1. The simulation is performed 25 times, and the average results of all runs are reported in the next subsections.

The main metric to assess the energy consumption is the total BS airtime [27], and the QoS of video streaming is quantified by the number and duration of video stops [119], denoted by $\eta$ and $\tau$ and calculated as per Eq. 5.24 and Eq. 6.6 respectively.

$$\tau_i = \int_0^T \tau_{i,\kappa} d\kappa / \int_0^T d\kappa. \tag{6.6}$$

where $\tau_{i,\kappa}$ equals to 1 if user $i$ experienced a video stop at time instant $\kappa$ where $\kappa << t$.

While the network performance is calculated by the average of each QoS metric, the resultant Quality of Experience (QoE) is also reported to model the users' perception. QoE, in essence, is a subjective metric that represents the service end-to-end performance level from the user's perspective, and can calculated using the Mean Opinion Score (MOS) formula in [133] and [134] depicted below:

$$MOS_{VS} = \frac{1}{M} \sum_{i=1}^{M} (2.99 * e^{-0.96\eta_i}) + 2.01. \tag{6.7}$$

$$MOS_{VD} = \frac{1}{M} \sum_{i=1}^{M} 4.59 * e^{-3.44\tau_i}. \tag{6.8}$$

Where $MOS_{VS}$ and $MOS_{VD}$ are the MOS values due to number and duration of video stops, respectively. The value of MOS varies from 1 to 5 which represents very poor to excellent service, respectively.

We adopt these metrics to evaluate the proposed R-GPRA, the existing non-robust PRA and the opportunistic RA (i.e. non-predictive). The following abbreviations are used in the next subsection:

- **PF (Non-PRA):** the traditional opportunistic proportional fair scheduler is used to represent the class of non-predictive schemes. It allocates the resources to the users

based on their current channel measurements and cumulative served traffic in previous slots [135].

- **NR-GPRA:** is the existing energy-efficient predictive resource allocation that assumes perfect prediction and adopts the deterministic formulations in [27]. This scheme is simulated by setting the values of $\gamma$ and $C_{i,t}$ to zero in Eq. 6.5, and the resultant formulation is solved using Gurobi optimizer [124].

- **PK-GPRA:** this refers to a hypothetical PRA with perfect knowledge of uncertain demand and network resources. As such it is aware of exact watching duration and amount of available resources. This is achieved by replacing the random variables in Eq. 6.1 by the exact values from the random generator in ns-3.

- **OR-GPRA:** this represents the proposed *robust* green predictive resource allocation as formulated in Eq. 6.5. The probability of video termination follows the distribution in [40]. The optimal solution is obtained by the branch and cut methods in Gurobi optimizer [124].

- **HR-GPRA:** this refers to the heuristic version of **OR-GPRA** in which the solution is obtained by the proposed guided search in Algorithm 4 and Algorithm 5.

### 6.4.2 Simulation Results

**Evaluating Demand Uncertainties**

We initially evaluate the impact of uncertain demand solely on the prediction gains (i.e. energy savings). The system load, in terms of number of users and streaming rates, was configured and set below the available radio resources. Hence, no video stops were observed, and thus the QoS was satisfied by all the schemes, while the main focus remains on

Table 6.1: Summary of Model Parameters in the Third Variant

| Parameter | Value/Definition |
|---|---|
| BS transmit power | $43$ dBm |
| Bandwidth | $5$ MHz |
| Time Horizon $T$ | $60$ s |
| $\zeta$ | $0.99$ |
| Bit Error Rate | $5 \times 10^{-5}$ |
| Velocity | $60$ [kmph] |
| QoS level $\beta$ | $0.95$ |
| Packet size | $10^3$ [bytes] |
| Packet rate (from core network to BS) | $10^3 s^{-1}$ |
| Buffer size | $10^9$ [bits] |
| Probability of watching ratio $p^W_{i,t/T}$ | $2/\sigma\phi(\frac{t-\mu}{\sigma})\Phi(\alpha\frac{t-\mu}{\sigma}), \forall i \in M$ |
| Probability of user arrival $p^A_{\omega,t}$ | $\frac{\lambda^\omega e^{-\lambda}}{\omega!} \forall t \in T$ |
| Standard deviation of watching time ratio $\sigma$ | $0.18$ |
| Skew parameter $\alpha$ | $0.84$ |
| Mean of watching time ratio $\mu$ | $0.27$ |
| User arrival rate $\lambda$ | $0.5$ |
| $\phi(x)$ | PDF of normal distribution |
| $\Phi(x)$ | CDF of normal distribution |

energy consumption. The maximum energy saving gap, referred to as prediction gain, is observed between the opportunistic non-predictive RA and hypothetical perfect knowledge PRA. As reported in the PRA literature, and shown in Fig. 6.3(a), the gain can reach up to $400$ % due to the minimal allocation strategy adopted for cell edge users moving to peak radio conditions. This is in addition to maximizing the allocation for users exiting the cell.

The existing non-robust PRA (NR-GPRA), however, has diminished the gain to 150 % as a result of the greedy prebuffering for cell center users exiting the cell, as yet not watching the full buffered video. On the contrary, the proposed *robust* GPRA has strategically prebuffered the video content to the users exiting the cell region, rather than transmitting their full content. Such risk-aware prebuffering strategy avoids greedy prebuffering of the future content whose delivery can be postponed until the corresponding time slots

are reached, or the user arrives at time slots which have a low probability of terminating the video. This is in addition to following the minimal allocation to users experiencing poor conditions until they reach peak rate values. As such, the robust scheme was able to maintain the prediction gain at $320$ %.

The same impact of uncertainty on the prediction gain was observed while increasing the streaming rate for fewer users Fig. 6.3(b). In this scenario, the maximum prediction gap can reach up to 150 %, however, the uncertainties resulted in a 25 % prediction gap as depicted by the non-robust scheme. The gain was retained to 100 % by adopting the stochastic based robust scheme.

**Evaluating Joint Demand and Resources Uncertainties**

The simulations are extended to incorporate the resources uncertainties, where the QoS and QoE performance are depicted in Fig. 6.4(a)-Fig. 6.4(b) and Fig. 6.5(a)-Fig. 6.5(b), respectively.

The resources uncertainties violated the QoS level under the existing non-robust predictive scheme for different number of users. Due to the arrival of real-time users, the network was unable to deliver the video content with the pre-calculated amount of resources. As such, the demand of cell edge users is not met by the minimal allocated resources that might be occupied by the real-time traffic users. The cell center video streaming users were not impacted due to the prebuffered content that surpasses the demand. Nevertheless, the substantial increase in the normalized number and durations of stops is attributed to the short video segments watched by the streaming users (i.e. demand uncertainty). The corresponding QoS demonstrates the exponential decay of users' experience as a result of encountering a large number and durations of stops.

Unlike the non-robust scheme, the proposed optimal robust technique has satisfied the

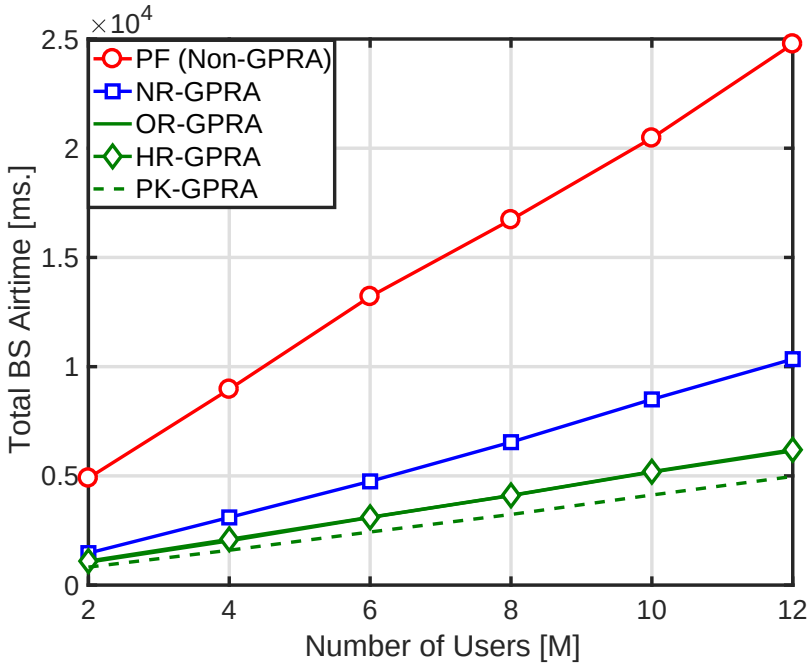
predefined QoS level ($\beta$) for all number of users. The robust scheme balances the amount of allocated resource to the cell edge and cell center users. Prebuffering is minimized for the cell center users and thus reserved more resources for the real-time users. As a result, the amount of allocated resources to cell edge users will be secured during the arrival of real-time users.

The performance of non-robust and robust predictive schemes is compared at different streaming rates and real-time traffic load as shown in Fig. 6.6(a) and Fig. 6.6(b). As the traffic load (streaming or real-time) increases, so does the number of unsatisfied users. With regards to energy savings and the prediction gain, the ability of robust scheme to maintain a high value was observed. Thus, the cost of robustness is said to be very low as the robust scheme avoided generating conservative solutions.

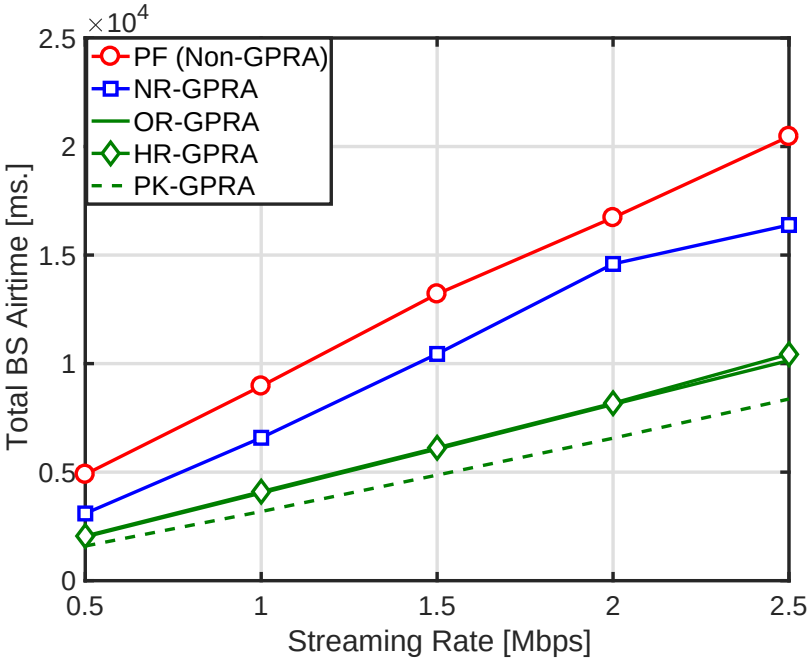
**Performance of Heuristic**

The above-mentioned observations over different system and streaming loads are also reported for the proposed heuristic. In essence, the heuristic was capable of satisfying the QoS level and maintain the prediction gap under demand and network uncertainties. The complexity of both the optimal and heuristic techniques is measured in terms of the computation time of a Quad Core i7-Processor, 3.2 GHz machine. The heuristic algorithm requires less than $0.1ms.$ to solve the robust PRA formulation for all the network configurations (i.e. number of users and streaming rate values). On the other hand, the performance of Gurobi is sensitive to network load and capacity. The execution time varies from $1s.$ to $15s.$ depending on the number of unsatisfied users in previous time slots, streaming rate, and available channel capacity. Requests of high streaming rates during low channel capacity will result in a narrow feasibility region. Such situations are very challenging for the solver that overlooks the problem structure and generates a large number of branches and

nodes to solve the integer programming model.

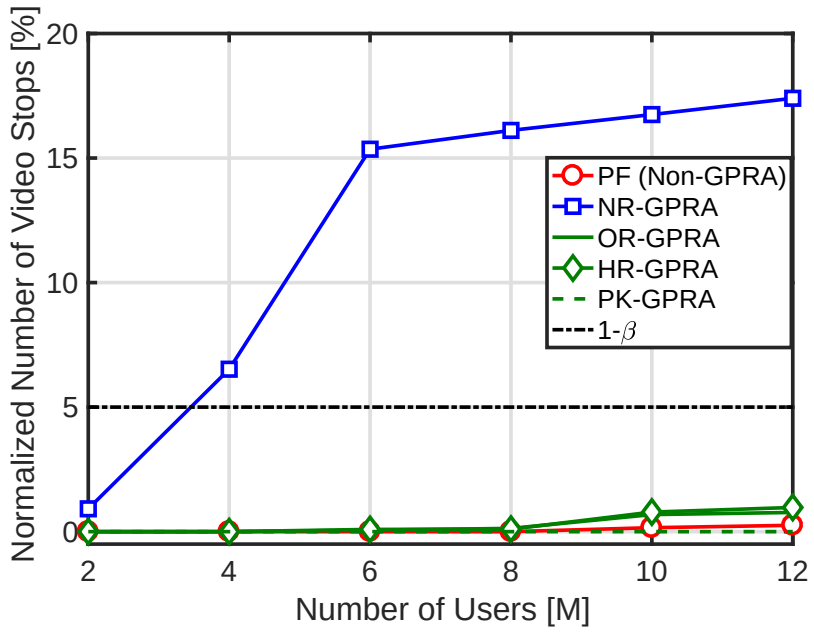



(a) Energy Consumption at V=0.5 Mbps

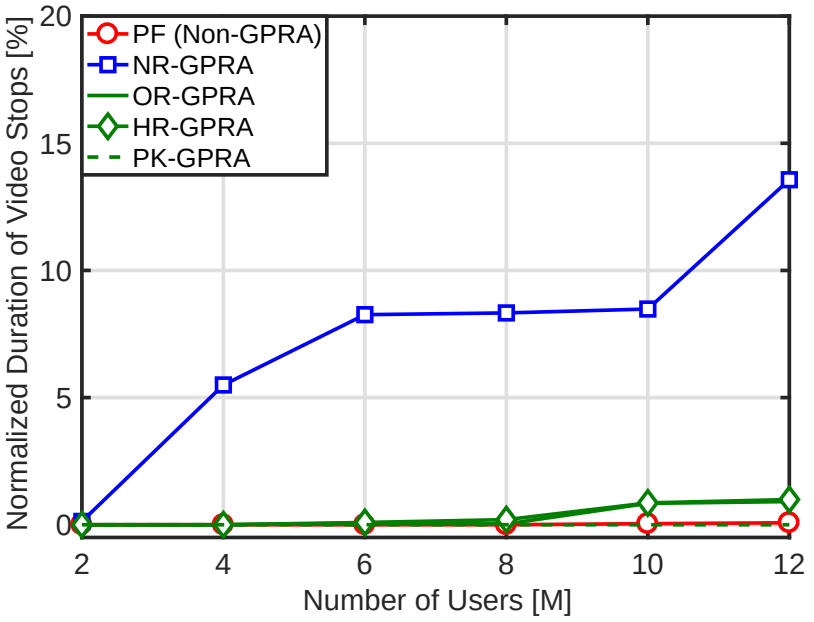



(b) Energy Consumption for 4 Users

Figure 6.3: Airtime-based energy consumption with uncertain demand only

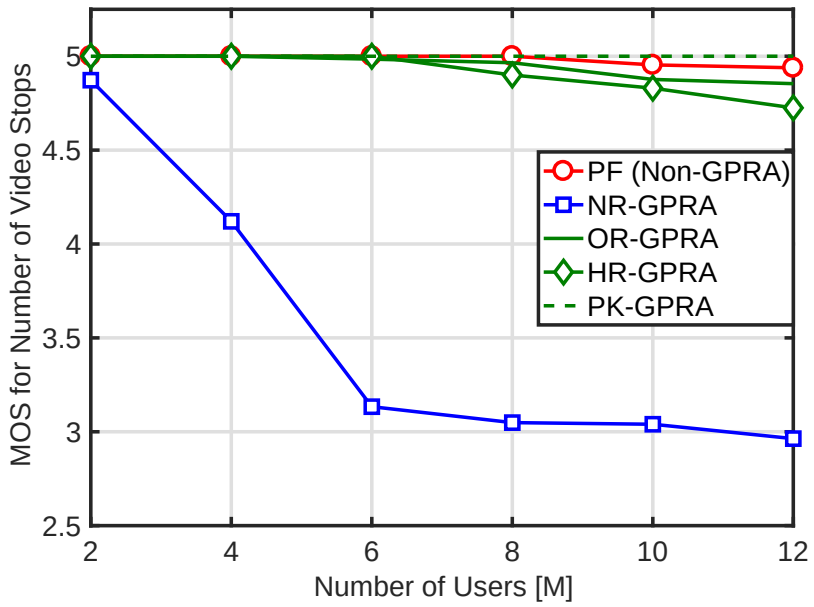



(a) Number of Stops

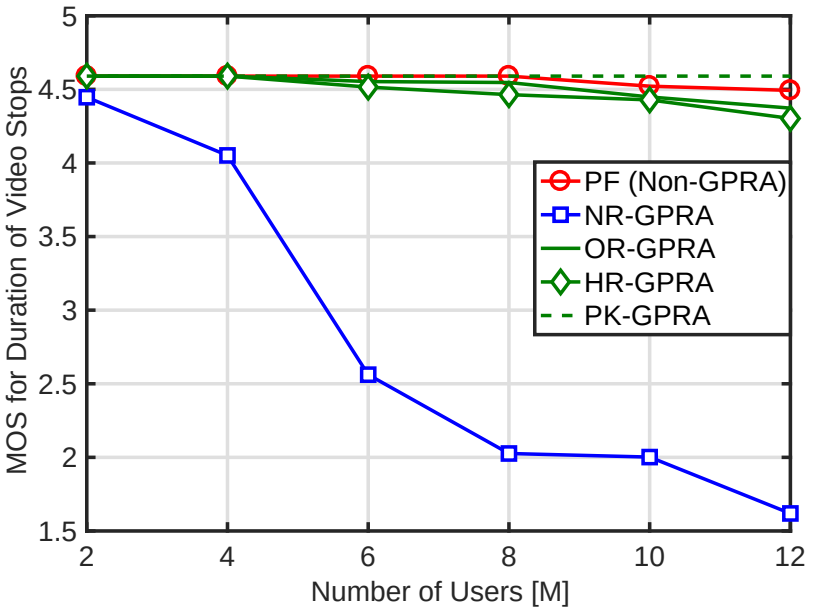



(b) Duration of Stops

Figure 6.4: QoS for number and duration of stops with uncertain demand and network resources at v=0.5 Mbps

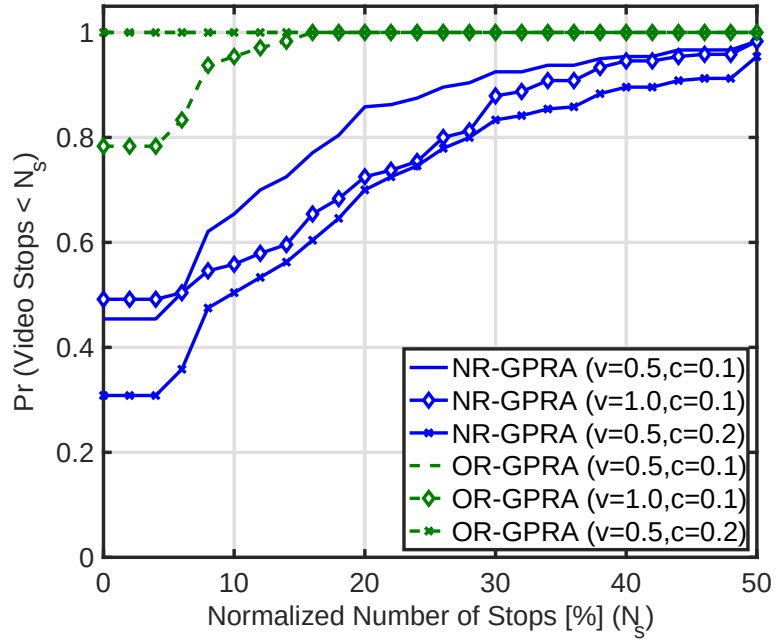



(a) QoE due to Number of Stops

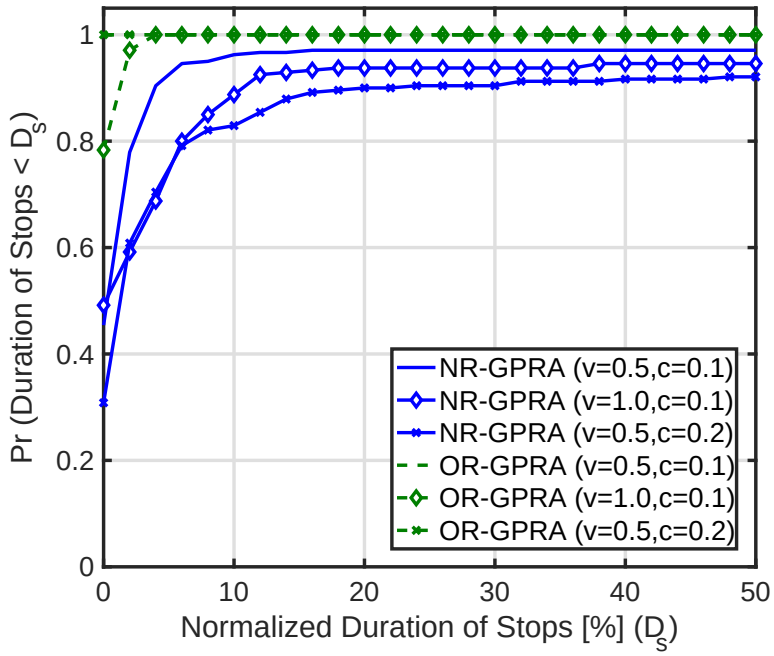



(b) QoE due to Duration of Stops

Figure 6.5: QoE for number and duration of stops with uncertain demand and network resources at v=0.5 Mbps




(a) Distribution of Stop Number




(b) Distribution of Stop Durations

Figure 6.6: Distribution of QoS values for robust and non-robust GPRA

# Chapter 7

# QoS-Aware Robust-DASH under Rate Uncertainty

The three variants in Chapter 5 and Chapter 6 aimed to minimize the energy consumption which is only achievable during low load scenarios as the BS can go into sleep mode. This chapter introduces the fourth variant, in Fig. 7.1, which exploits all the available radio resources (i.e. no energy saving) to achieve *fair* QoS during high load scenarios, and solves for both video quality and airtime fractions over a time-horizon. This is unlike the previous three variants in which the video quality was predefined at each time slot and thus treated as constant in the optimization stage. The scheme in this chapter is an application of network-centric DASH in which the network selects the video quality to achieve fairness among the users, and calculates the corresponding amount of resources (airtime fractions) required to avoid video stops while considering uncertainties in predicted channel rates. The schemes is referred to as *Robust Predictive*-DASH (RP-DASH).

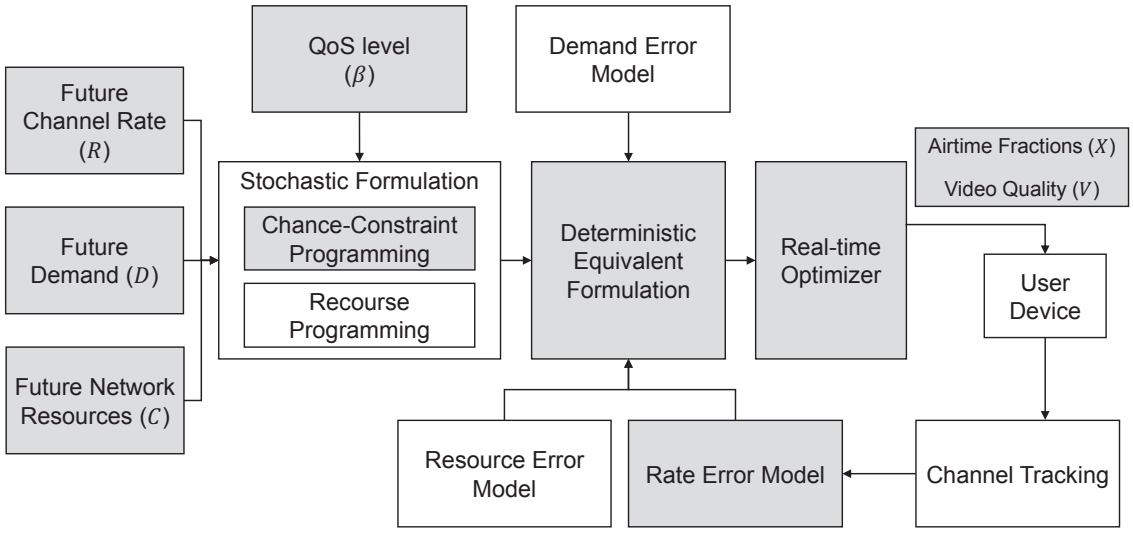



Figure 7.1: Block diagram of RP-DASH scheme under rate uncertainty

## 7.1 System Model

### 7.1.1 Predicted Rate Error Model

In order to model prediction uncertainties, future rate is modelled as random variable denoted by $\tilde{r}_{i,t}$. This random variable is either described by 1) its discrete PDF when the realizations and their probabilities are known or 2) the Gaussian distribution in which the standard deviation is denoted by $\sigma_{i,t}$ and calculated using the framework in Chapter 4. This Gaussian distribution error model is motivated by the findings in [34, 43] and will be used to quantify the trade-off between error PDF modelling and robustness. In both cases, the per slot rate errors are assumed to be independent. Particularly, the error of predicting the rate is function of erroneous rate in REM, variations in the wireless signal (which changes the SINR) and user location uncertainty. These parameters are calculated at each slot based on the independent channel gains [43, 90].

### 7.1.2 Demand Model

In this scheme, we assume that the total video duration the user is going to watch is known (i.e. no demand uncertainty). Yet, the scheme has to decide on the quality of each video segment.

## 7.2 Problem Statement

The RP-DASH scheme aims to calculate both the airtime fractions $x_{i,t}$ and segments quality $\kappa_{i,t}^{(q)}$ for each user $i$ at time slot $t$ such that all users experience fair video qualities while meeting the QoS level. Particularly, QoS is said to be satisfied when users experience video stops, due to buffer underrun, with probability below $\epsilon = (1 - \beta)$.

## 7.3 Problem Formulation

The introduced robust P-DASH and fair quality selection is formulated based on Chance Constrained Programming (CCP) as follows:

$$\underset{\mathbf{x},\kappa}{\text{maximize}} \quad \left\{ \min_{\forall i \in M} \sum_{\forall t \in T} \sum_{\forall q \in Q_i} \kappa_{i,t}^{(q)} v_q \right\} \tag{7.1}$$

subject to:

$$\text{C1:} \quad Pr\left\{\sum_{t'=0}^{t} \tilde{r}_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t'}^{(q)} v_q\right\} \geq 1 - \epsilon_{i,t}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C2:} \quad \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t'}^{(q)} \tau_{i,t'} \geq t, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

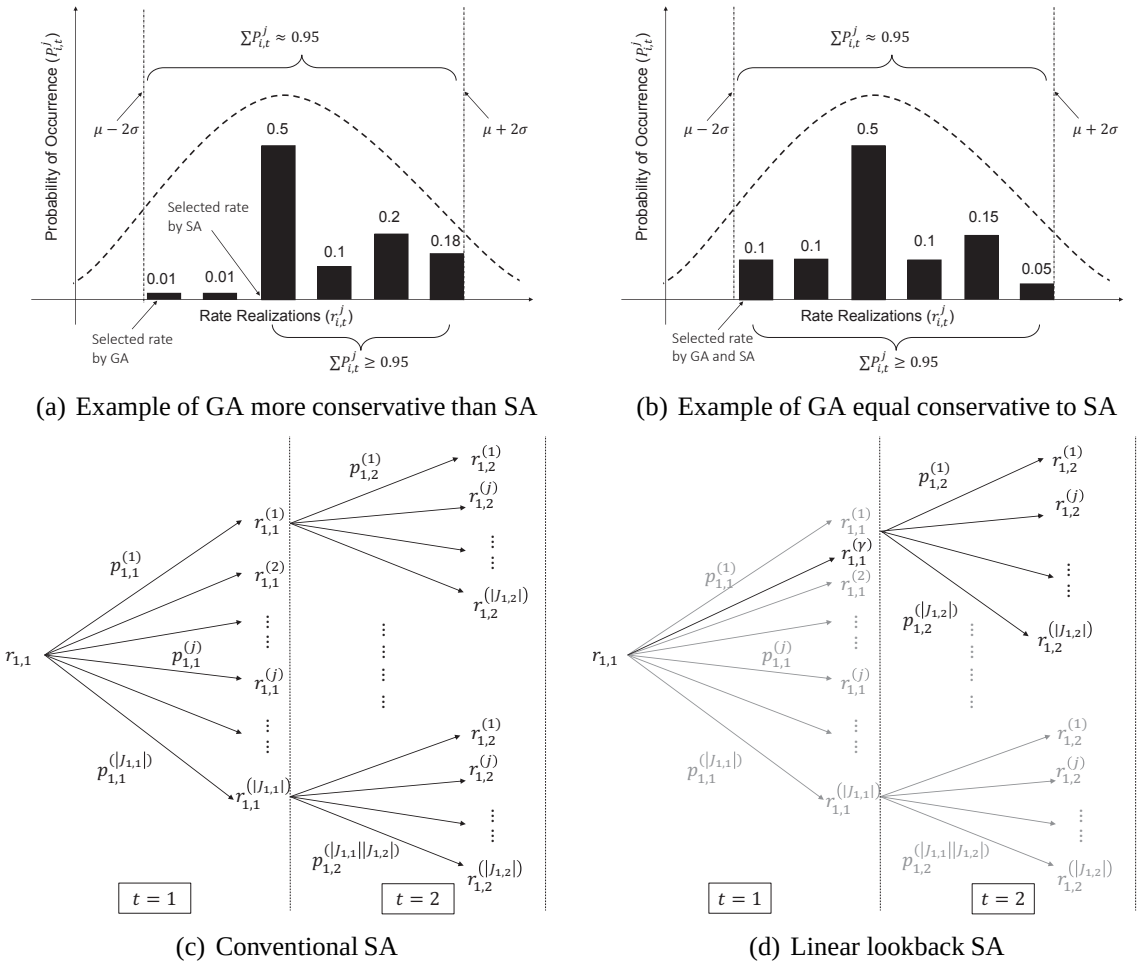
$$\text{C3:} \quad \sum_{q \in Q_i} \kappa_{i,t}^{(q)} \leq 1, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C4:} \quad \kappa_{i,t}^{(q)} \in \{0, 1\}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C5:} \quad \sum_{i=1}^{|\mathcal{M}|} x_{i,t} \leq 1, \quad \forall t \in \mathcal{T},$$

$$\text{C6:} \quad x_{i,t} \geq 0, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

$\epsilon_{i,t} \in [0, 1]$ is the probability that the QoS of user $i$ is unsatisfied at time slot $t$, where $\epsilon_{i,t} = 1$ is the maximum QoS violation. The objective function aims to maximize the minimum total quality of each user to attain the fairness among the users over the time horizon. The QoS chance constraint in C1 guarantees that the total delivered content to the user satisfies the anticipated demand (function of the selected quality) by a minimum probability of $1 - \epsilon$ while considering uncertainties in future rates. The constraint in C2 complements C1 to ensure that the total duration of the selected segments should be greater than the elapsed playback time to avoid video stops. C3 and C4 ensure that, for each user, only one quality level is selected at a given time slot. The fifth constraint C5 models the limited resources at each base station by ensuring that the sum of the airtime fractions is less than 1 second which is the duration of the allocation slot. The last constraint C6 ensures the non-negativity of the decision variable. Indeed the above formulation does not have a closed form solution due to the probabilistic constraint C1. As such, we will initially adopt




(a) Example of GA more conservative than SA

(b) Example of GA equal conservative to SA

(c) Conventional SA

(d) Linear lookback SA

Figure 7.2: Illustration of SA and GA operations

the SA to obtain a deterministic equivalent form in the next section.

## 7.4 Nominal Scenario Approximation Equivalent

The Scenario approximation adopts the discrete Probability Density Function (PDF) of the uncertain rates to derive a deterministic representation for the probabilistic constraint. The PDF of every rate $\tilde{r}_{i,t}$ contains all the realizations $r^{(j)}_{i,t}$ and their probabilities $p^{(j)}_{i,t}$ to construct the scenarios over the time horizon. The approximation ensures that resource allocations

and quality selections satisfy the scenarios whose total probability of occurrence is more than the defined QoS level (i.e. $1-\epsilon$). Each scenario corresponds to one combination of the possible realizations of the uncertain rates in C1. For example, the constraint in the second time slot includes the rates in both the first and second time slot. The scenarios will comprise all the possible combinations of the realizations of these two rates. As illustrated in Fig. 7.2(c), the first scenario consists of $r_{1,1}^{(1)}$ and $r_{1,2}^{(1)}$. Where $r_{1,1}^{(1)}$ represents the first realization of the rate at t=1, and $r_{1,2}^{(1)}$ is the first realization of the rate at t=2, both for the first user. The probability of this scenario will be the product of the individual probabilities (i.e. $s_{1,2}^{(1)} = p_{1,1}^{(1)} \times p_{1,2}^{(1)}$). The deterministic equivalent of C1 in Eq. 7.1 is captured by C7-C9 below

$$\underset{\mathbf{x},\kappa,\delta}{\text{maximize}} \quad \left\{ \min_{\forall i \in M} \sum_{\forall t \in T} \sum_{\forall q \in Q_i} \kappa_{i,t}^{(q)} v_q \right\} \tag{7.2}$$

subject to:

$$\text{C7:} \quad \sum_{t'=0}^{t} r_{i,t'}^{(j)} x_{i,t'} \geq \delta_{i,t}^{(j)} \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t'}^{(q)} v_q, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall j \in \mathcal{J}_{i,t},$$

$$\text{C8:} \quad \sum_{j \in \mathcal{J}_{i,t}} s_{i,t}^{(j)} \delta_{i,t}^{(j)} \geq 1 - \epsilon_{i,t}, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C9:} \quad \delta_{i,t}^{(j)} \in \{0,1\}, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}, j \in \mathcal{J}_{i,t},$$

(C2 - C6)

where $r_{i,t}^{(j)}$ is the $j^{th}$ realization of the uncertain predicted rate at time slot $t$ for user $i$. $s_{i,t}^{(j)}$ is the probability of the $j^{th}$ scenario at time slot $t$ for user $i$. $\delta_{i,t}^{(j)}$ is a binary decision variable which equals to 1 if the $j^{th}$ scenario at slot $t$ must be satisfied by the decision variable and equals 0 otherwise (C9). Constraint C8 guarantees that the total probability of all the satisfied scenarios exceeds the minimal QoS level $1-\epsilon$.

Although the above formulation is deterministic and robust, it poses the following three main challenges to the solver:

1. Non-linearity: due to the joint optimization of quality and airtime fractions, the right hand side in C7 will be non-linear (both decision variables are multiplied). Despite the dimensions of C7, the problem is NP-hard and reaching the optimal solution is not guaranteed.

2. Exponential complexity: the QoS constraint at each time slot is a function of the rate in both the current and preceding slots (C7 in Eq. 7.2 and Fig. 7.2). Thus, at each time slot $t$ the number of considered scenarios will be $\prod_{t'=0}^{t} |J_{i,t'}|$, where $|J_{i,t'}|$ is the number of realizations of the uncertain rate $r_{i,t'}$. Assuming that all the rates have equal number of realizations (i.e. $|J_{i,t'}| = |J_i|$), thus the total number of scenarios for each time slot constraint per user will be $(|J_i|)^{(t)}$.

3. Explicit rate information: the scenario-based approximation requires the exact values of realizations for all the rates and their corresponding probabilities. This requires collecting large number of samples for each achievable channel rate value in order to construct an accurate discrete PDF. Due to the large number of physical layer configurations such as Multiple Input Multiple Output (MIMO) and MCS, more possible rates can be achieved. Hence, increases the burdens of prediction and error modelling.

In the next subsection we address the first two challenges while the third challenge is tackled separately in the next section by the Gaussian based approximation.

## 7.5 Linear Look-Back Scenario Approximation Equivalent

The nonlinearity of QoS constraint C7 is solved by exploiting the problem's structure. The scenario decision variable $\delta_{i,t}^{(j)}$ is governed by the QoS constraint C8, as such a minimal number of scenarios should be satisfied (i.e. $\delta_{i,t}^{(j)} = 1$). For each satisfied scenario (i.e. $\delta_{i,t}^{(j)} = 1$), the corresponding airtime allocation (i.e. left hand side of C7) should guarantee the satisfaction of the selected demand (i.e. video quality) while considering the worst case of the selected scenario. The objective function plays the main role in discarding the scenarios (i.e. $\delta_{i,t}^{(j)} = 0$) whose realizations have very low values. In that case, both sides of C7 are equal to zero, and the scenario is not satisfied by the calculated airtime fractions.

A new linear representation for C7 in Eq. 7.2 is introduced to capture the above strategy and avoid the exponential complexity due to considering the realizations of all previous time slots. Instead, the new formulation considers a linear look-back on the preceding rate realizations to decrease the large number of scenarios at each time slot. Only one conservative realization denoted by $r_{i,t}^{(\gamma)}$ is selected to represent each of the rates in the previous slots. The number of scenarios at slot $t$ will depend only on the realizations in this slot ($|J_{i,t}|$) and the number of previous slots ($t-1$) instead of all the realizations of the latter. In order words, $|J_{i,t}| \times (t-1)$ scenarios are considered instead of $(|J_i|)^{(t)}$. The new linear formulation is represented as follows:

$$\underset{\mathbf{x},\kappa,\delta,\mathbf{Y}}{\text{maximize}} \quad \sum_{i=1}^{M} Y_i \tag{7.3}$$

subject to:

$$\text{C10:} \quad \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t'}^{(q)} v_q - (\sum_{t'=0}^{t-1} r_{i,t'}^{(\gamma)} x_{i,t'} + r_{i,t}^{(j)} x_{i,t}) \leq \mathbb{B}(1 - \delta_{i,t}^{(j)}), \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C11:} \quad \sum_{j \in \mathcal{J}_{i,t}} p_{i,t}^{(j)} \delta_{i,t}^{(j)} \geq 1 - \epsilon_{i,t}, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C12:} \quad \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t}^{(q)} v_q \geq Y_i, \qquad \forall i \in M,$$

(C2 - C6, C9)

The minimum function operator in the objective function of Eq. 7.2 was replaced by introducing auxiliary variable $Y_i$ and the fairness constraint C12 which must be satisfied for all users. C10 represents the linear look-back constraint in which $p_{i,t}^{(j)}$ is the probability of realization $j$ of channel rate $r_{i,t}$ and $\mathbb{B}$ is a very large number that forces the airtime allocation to satisfy the demand when scenario $j$ is considered. $r_{i,t}^{(\gamma)}$ approximates the channel rates of the preceding timeslots and can be calculated as follows

$$\underset{\theta, r_{i,t}^{(\gamma)}}{\text{minimize}} \quad r_{i,t}^{(\gamma)} \tag{7.4}$$

subject to:

$$r_{i,t}^{(\gamma)} \geq \sum_{j \in \mathcal{J}_{i,t}} r_{i,t}^{(j)} \theta_j$$

$$\theta_j \sum_{j'=1}^{j} p_{i,t}^{(j')} \geq \epsilon \qquad , \forall j \in \mathcal{J}_{i,t}$$

$$\theta_j \in \{0, 1\}$$

The objective function in Eq. 7.4 aims to select the optimal value of the aggregated rate $r_{i,t}^{\gamma}$ for the slot realizations such that very low values with conservative solutions and high values with non-robust solutions are ignored. The first constraint ensures that the

calculated value of $r_{i,t}^{\gamma}$ surpasses some realizations due to their low values; where ignoring such realizations avoid conservative solutions. The second constraint guarantees that the sum of probability of the ignored realizations is below the degradation level $\epsilon$, to achieve robustness. The last constraint defines $\theta$ as a binary decision variable. Since the objective function is minimization which is subjected to the second constraint, the decision variable is $\sum_j \theta_j = 1$. Thus only one realization value is selected from the first constraint.

## 7.6 Linearized Gaussian Approximation Equivalent

The third challenge of Scenario Approximation (SA) is tackled by adopting the Gaussian Approximation (GA) which does not require the explicit realizations and their probabilities for all future rates. Instead, GA obtains a deterministic closed form for C1 using the CDF of multivariate random variables denoted by $\Phi$. Thus the probabilistic constraint C1 is replaced by the following deterministic form

$$Pr\left\{\sum_{t'=0}^{t}\tilde{r}_{i,t'}x_{i,t'} \geq \sum_{t'=0}^{t}\sum_{\forall q\in Q_i}\kappa_{i,t'}^{(q)}v_q\right\} = 1 - \int_{-\infty}^{D_{i,t}} N(\mathbf{r},\mu,\Sigma)d\mathbf{r}$$

$$= 1 - \frac{\Phi(\frac{D_{i,t}-\mu_{i,t}}{\Sigma_{i,t}}) - \Phi(\frac{-\mu_{i,t}}{\Sigma_{i,t}})}{S_{i,t}} \geq 1 - \epsilon_{i,t}, \tag{7.5}$$

Using the inverse CDF, the following closed form can be obtained:

$$\mu_{i,t} + S_{i,t}\Phi_{\epsilon_{i,t}}^{-1}\Sigma_{i,t} \geq D_{i,t}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}, \tag{7.6}$$

where:

$$S_{i,t} = \prod_{t'=0}^{t} \left( \Phi(\frac{r_{i,t'}^{(u)} - \bar{r}_{i,t'}}{\sigma_{i,t'}}) - \Phi(\frac{r_{i,t'}^{(l)} - \bar{r}_{i,t'}}{\sigma_{i,t'}}) \right)$$

$$\mu_{i,t} = \sum_{t'=0}^{t} \bar{r}_{i,t'} x_{i,t'},$$

$$\Sigma_{i,t} = \sqrt{\sum_{t'=0}^{t} x_{i,t'}^2 \sigma_{i,t'}^2},$$

$$\sigma_{i,t'}^2 = E[(\tilde{r}_{i,t'} - \bar{r}_{i,t'})^2],$$

$$D_{i,t} = \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa_{i,t'}^{(q)} \upsilon_q$$

$r_{i,t}^{(l)}$ and $r_{i,t}^{(u)}$ are the lower and upper bounds of the realizations of future predicted rate $\tilde{r}_{i,t}$ (i.e. the support). Typical values of the channel rates in the current and future networks are more than the corresponding variance values (i.e. $\mu_{i,t} >> \Sigma_{i,t}$) and thus $\Phi(\frac{-\mu_{i,t}}{\Sigma_{i,t}}) \approx 0$. $S_{i,t}$ is used to normalize the truncated probability distribution of the random rates.

The above deterministic form, however, is a mixed integer quadratic constrained programming which is NP-hard. A linear approximation is adopted, which turns the problem to NP-complete. This is done by the budgeted robust approximation of [82] on Eq. 7.6 as follows. Let $\Sigma_{i,t}^{(L)} = \sum_{t'=0}^{t} |x_{i,t'} \sigma_{i,t'}|$, thus $\Sigma_{i,t} < \Sigma_{i,t}^{(L)}$, and $\Phi(\frac{1}{\Sigma_{i,t}}) > \Phi(\frac{1}{\Sigma_{i,t}^{(L)}})$. This guarantees the satisfaction of C1 by substituting $\Sigma_{i,t}$ with $\Sigma_{i,t}^{(L)}$. Such approximation will result in a linear but conservative formulation compared to the original Gaussian approximation.

The final deterministic mixed integer linear equivalent for the RP-DASH in Eq. 7.1 is summarized as

$$\underset{\mathbf{x},\kappa,\delta,\mathbf{Y}}{\text{maximize}} \quad \sum_{i=1}^{M} Y_i \tag{7.7}$$

subject to:

$$\text{C13:} \quad \mu_{i,t} + S_{i,t}\Phi^{-1}_{\epsilon_{i,t}}\Sigma^{L}_{i,t} \geq D_{i,t}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$$

(C2 - C6, C9)

## 7.7 Real-Time Guided Heuristic

This section introduces the guided heuristic algorithm to obtain real-time solutions for the formulated RP-DASH problem. This is in addition to analyzing its computational complexity.

### 7.7.1 Limitations of Optimal Commercial Solvers

The Scenario-and Gaussian-based robust formulations in Eq. 7.3 and Eq. 7.7 are represented in mixed integer linear programming forms. The main advantage of these forms is that an optimal feasible solution can be obtained using branch and bound or simplex techniques. Such conventional techniques are currently well developed and implemented in many commercial solvers such as Gurobi [124]. These solvers use their own developed heuristic algorithms to calculate an initial feasible solution which satisfies the constraints. Other neighbouring solutions are then explored by means of branch and bound or simplex algorithms, while using the duality gap to evaluate the optimality of each solution. Although zero or low duality gaps (i.e. optimal solutions) can be achieved by commercial solvers, the execution time highly increases with the problem's dimensions (i.e. number of constraints and decision variables). A guided heuristic algorithm is proposed to provide a real-time feasible solution with low optimality gap from commercial solvers solutions.

### 7.7.2 Guided Real-time Heuristic

The introduced guided heuristic search algorithm is aware of the problem's structure that includes the interdependency between the constraints and their impact on the value of objective function. This is in addition to considering the motive of robust and predictive allocation in calculating the airtime fractions and video qualities. In essence, the algorithm starts by satisfying all the QoS constraints using the available radio resources while ignoring the objective function in that stage. This first stage contains two problem specific knowledge: 1) the buffering capabilities of the users, and 2) the direct relation between the QoS and the resource limitation constraints. The former knowledge can be used to push the video content in advance and thus avoids stalling in congested time slots. In the next step, the value of objective function is maximized while exploiting three other problem features: 1) the trade-off between the fairness (i.e. the objective function) and the above-mentioned two constraints, 2) the time horizon and the buffer status of each user, and 3) the competition between the users, experiencing different channels rates, on the radio resources of one time slot.

The heuristic implements two main consecutive stages summarized in Algorithm 6 and Algorithm 7, respectively and are detailed as follows:

#### Satisfaction of Minimal Quality

In this initial stage (Algorithm 1), the lowest video quality is assigned to all users over the time horizon (Algorithm 1, line 3). Then, the amount of airtime that satisfies this quality level is calculated (line 4) and used to update the total amount of allocated resources at each time slot (line 5). Such allocation guarantees the satisfaction of QoS constraint C13 in Eq. 7.7.

However, in high load scenarios, due to high QoS levels (i.e. $1-\epsilon$) or a large number of users, the total allocated resources in a certain time slot might violate the airtime constraint C5 in Eq. 7.1. Accordingly, the preceding time slots with vacant resources will be used to prebuffer the content of the highly loaded time slots as depicted in lines 11-21 of Algorithm 6. While efficient exploitation of the radio resources is mandatory, the algorithm selects the user with the highest achievable rate in this preceding slot and prebuffers the content (lines 13-16). Thus, less airtime is consumed and the chance of satisfying the radio resource constraint C5 is increased. In case of non-vacant resources, the problem is said to be infeasible (lines 23-25). Other bounding and streaming constraints are implicitly satisfied by the above iterative procedure.

**Optimizing Long-term Fairness**

This stage (Algorithm 2) aims to maximize the value of objective function without violating any of the aforementioned satisfied constraints. While the objective is to maximize the long-term quality for each user, the algorithm tries to achieve this on both the current and the future time slots. In each time slot with vacant resources, both the cumulative quality and the required airtime to increase the current slot's quality are calculated for each user (lines 3-7). The user with minimal quality (both cumulative and increased values) is selected as long as the required airtime is less than the available vacant resources (line 9). In case of more than one user with the same quality, the one that requires less airtime is selected (line 10). This procedure is repeated for all users as long as there are vacant resources in the current slot and the video quality is improving.

For low-load scenarios, due to either a small number of users or high achievable rates, the resources at a certain time slot might not be fully utilized. As such, predictive allocation

is performed in order to maximize the quality of the users experiencing their highest channel conditions in the current time slot. This is modelled by calculating the ratio between the achievable rate in this time slot and the minimal future rate. Thus, users with peak radio conditions who are heading towards the cell edge (line 14) will have the highest ratio and thus can use these vacant resources to increase the quality of future video content (lines 13-16). The achievable rate used to calculate all the airtime allocations as a function of the rate average value, variance, CDF and QoS level as derived in C13 of Eq. 7.7.

### 7.7.3 Algorithm Complexity

The first part of the heuristic (i.e. Algorithm 6) consists of two successive loops, the first is in lines 1-9 and has a complexity $O(MT)$. The second loop, however, has a higher complexity of $O(MT^2)$ due to revisiting the preceding time slots in lines 9-27. Similarly, the second part of the heuristic (i.e. Algorithm 7) has a complexity of $O(QMT^2) \approx O(MT^2)$ due to the relatively small number of available quality levels compared to the length of the time horizon. The complexity of the whole proposed heuristic is $O(MT^2)$ which is lower than numerical optimization methods.

## 7.8 Performance Evaluation

### 7.8.1 Simulation Setup

We simulate the proposed RP-DASH using the LTE module in ns-3 [125] which is integrated with Gurobi commercial solver [124] to obtain optimal solutions for all the formulated problems. The fading model of 3GPP defined in [113] is added to the received

**Algorithm 6:** Initialization and QoS Satisfaction Stages of Guided Heuristic

**Input** : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Average Predicted Rates: $\bar{R}$, Rate Variances: $\Sigma$, Maximum Violation: $\epsilon$ and Video Qualities: $Q$;
**Output** : $X$;
**Initialization**: $X = \emptyset$, $\kappa = \emptyset$ $N_t = 0 \; \forall t \in T$
**Define:** $\mathcal{R}_{i,t} = \bar{r}_{i,t} - S_{i,t}\Phi^{-1}_{\epsilon_{i,t}}\sigma_{i,t}$;
**for** $i \in \mathcal{M}$ **do**
  **for** $t \in \mathcal{T}$ **do**
    Set $\kappa^0_{i,t} = 1$ ;
    Set C13 of Eq. 7.7 to an equality and solve for $x_{i,t}$;
    $N_t = N_t + x_{i,t}$;
  **end**
**end**
**for** $t \in \mathcal{T}$ **do**
  **if** $N_t > 1$ **then**
    Set $k = t - 1$;
    **while** $k > 0$ **do**
      Calculate the residual airtime $\Delta x_{i,t} = N_t - 1$;
      Calculate the demanded airtime $\Delta x_{i,k} = \Delta x_{i,t} \times \frac{\mathcal{R}_{i,t}}{\mathcal{R}_{i,k}}$;
      $i* = argmax x_{i,k} \forall i \in \mathcal{M}$;
      **if** $N_k + \Delta x_{i*,k} \leq 1$ **then**
        Update $x_{i*,k}$, $x_{i*,t}$, $N_t$ and $N_k$ ;
        break;
      **end**
      $k = k - 1$;
    **end**
  **end**
  **if** $N_t > 1$ **then**
    Return Infeasible Problem;
  **end**
**end**
return $X$

**Algorithm 7:** Optimization Stages of Guided Heuristic

**Output**: $X$ and $\kappa$;
**Define:** $\mathcal{R}_{i,t} = \bar{r}_{i,t} - S_{i,t}\Phi^{-1}_{\epsilon_{i,t}}\sigma_{i,t}$;
**for** $t \in \mathcal{T}$ **do**
  **while** $N_t < 1$ **do**
    Calculate $\mathcal{V}_{i,t} = \sum_{t'=0}^{t} \sum_{\forall q \in Q} \kappa^{(q)}_{i,t'} v_q$ for all users;
    **for** $i \in \mathcal{M}$ **do**
      Calculate a possible higher quality level $\kappa^{(q')}_{i,t} v_{q'}$;
      Calculate the required airtime $\Delta x_{i,t}$ to satisfy $\kappa^{(q')}_{i,t} v_{q'}$;
      Update $\mathcal{V}_{i,t}$ using $\kappa^{(q')}_{i,t} v_{q'}$ ;
    **end**
    Select the set of users $l$ with minimum $\mathcal{V}_{i,t}$;
    Select user $k$ from $l$ with minimal $\Delta x_{i,t}$;
    Update $N_t$, $x_{k,t}$, and $\kappa^{(q)}_{k,t}$;
    **if** $l$ *is empty* **then**
      **while** $t' < T$ **do**
        Select user $i$ with maximum $(\mathcal{R}_{i,t}) \times (r_{i,t'} - r_{i,t})$;
        Repeat lines $5-6$ and line $11$ for user $i$ with $t = t'$;
      **end**
    **end**
  **end**
**end**
return $X$ and $\kappa$

power at the user device to apply variations in predicted rate. Users follow random predefined paths within the cell coverage region at varying velocities from 25 to 40 km/h, which correspond to typical values in urban areas. All the simulation parameters and values are presented in Table 7.1, and the average of all output results, over 50 simulation runs, is reported in the following subsections. We compare the introduced RP-DASH scheme with an existing non-robust P-DASH technique. The abbreviations, definitions and solution methods of the comparative schemes are summarized in Table 7.2. Existing non-robust P-DASH techniques, referred to as *P-DASH*, are simulated by replacing the random rates in Eq. 7.1

Table 7.1: Summary of Model Parameters in the Fourth Variant

| Parameter | Value |
|---|---|
| BS transmit power | $43$ dBm |
| Bandwidth | $5$ MHz |
| Time Horizon $T$ | $60$ s |
| Streaming rates | $0.5, 1, 1.5, 2, 2.5$ [Mbps] |
| Bit Error Rate | $5 \times 10^{-5}$ |
| Shadow correlation distance$(d_{cor})$ [113] | $50m$ |
| Shadow standard deviation [113] | $4, 6$ |
| Velocity | $25$ - $40$ [km/h] |
| Packet size | $10^3$ [bytes] |
| Packet rate (from core network to BS) | $10^3 s^{-1}$ |
| Buffer size | $10^9$ [bits] |

Table 7.2: Comparative Schemes

| Notation | Definition | Solution Method |
|---|---|---|
| **P-DASH** | Non-robust P-DASH in [26, 136] | Gurobi [124]. |
| **PP-DASH** | P-DASH with perfect channel knowledge | Gurobi [124]. |
| **SRP-DASH** | SA based RP-DASH in Eq. 7.3 | Gurobi [124]. |
| **GRP-DASH** | GA based RP-DASH in Eq. 7.7 | Gurobi [124]. |
| **HRP-DASH** | GA based RP-DASH in Eq. 7.7 | Heuristic in Algorithm 6-7. |

with the average rate values. The performance bounds are obtained by *PP-DASH* which assumes perfect prediction of channel rates (without errors) to replace the random variable in C1 Eq. 7.1.

### 7.8.2 Evaluation Metrics

#### QoS Satisfaction and QoE levels

In order to assess the robustness of the simulated schemes, we measure the QoS satisfaction using the number and duration of video stops denoted by $\eta$ and $\tau$, respectively and calculated as in Eq. 5.24 and Eq. 6.6. Similar to Chapter 6, the resultant QoE is also reported to model the users' perception using the MOS formula in [133] and [134].

**Video Streaming Quality**

A key performance parameter of DASH is the selected quality of all the segments over the time horizon for each user $i$, denoted by $V_i$, and calculated as a function of the segment size as follows

$$V_i = \sum_{\forall t \in T} \sum_{\forall q \in Q_i} \kappa_{i,t}^{(q)} v_q \quad \forall i \in M \tag{7.8}$$

The $V_i$ metric is averaged over all users to assess the conservatism of the schemes, while the optimality of the objective function is measured by the fairness using the Jain's index below

$$J = \frac{(\sum_{i=1}^{M} V_i)^2}{M \sum_{i=1}^{M} {V_i}^2} \tag{7.9}$$

### 7.8.3 Simulation Results

**Comparison with non-Robust P-DASH**

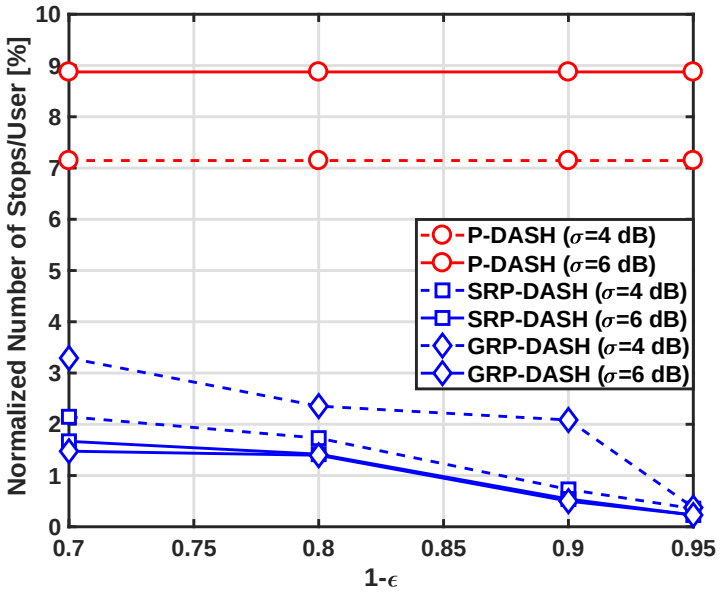
We firstly compare both the SA and GA formulations of the introduced robust P-DASH against the existing *non-robust P-DASH* for different values of QoS degradations and standard deviations. The existing non-robust *P-DASH* suffered from an increased number and durations of video stops with the standard deviations of shadowing as depicted in Fig. 7.3(a) and Fig. 7.3(b), respectively. Although only four users are considered, this QoS degradation resulted in average and poor MOS values due to frequent stops with long durations as shown in Fig. 7.4(a) and Fig. 7.4(b), respectively. This is attributed to the average predicted values of rates adopted by the *P-DASH* which did not account for the rate variations and uncertainties. As such, the highest quality levels were always selected by the non-robust

scheme as depicted in Fig. 7.5(a). This is as opposed to the introduced *GRP-DASH* and *SRP-DASH* formulations which were able to keep the percentage of stops and durations below the QoS degradation level $\epsilon \times 100\%$. An increasing trade-off between the QoS and QoE improvements on one hand and the quality degradation on the other hand is deduced over different $\epsilon$ levels as in Fig. 7.3(a)-Fig. 7.4(b) and Fig. 7.5(a), respectively. The main objective (i.e. quality fairness), did not suffer a significant degradation as reported by the Jain's index in Fig. 7.5(b).

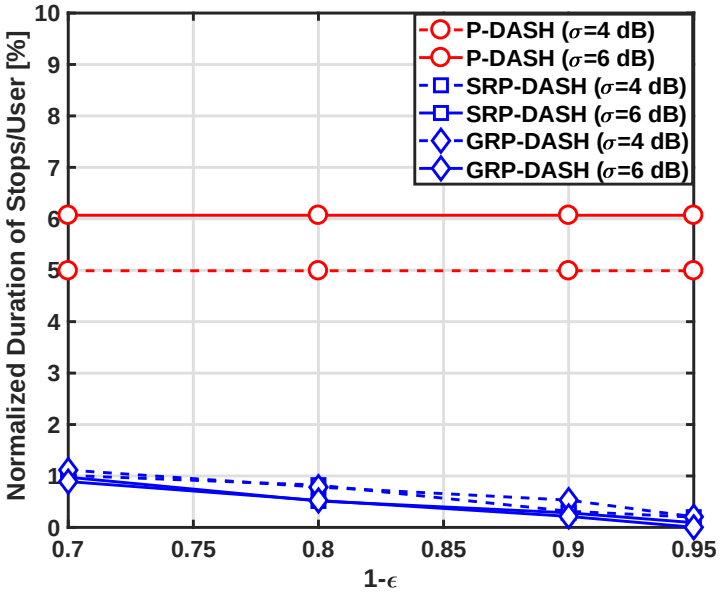
By increasing the number of users, more and longer video stops are observed which resulted thus in low MOS values when using the *P-DASH* as shown in Fig. 7.6(a)-Fig. 7.7(b). This degradation is caused by the optimistic strategy of the *P-DASH* which tries to maximize the quality at the expense of prebuffering and thus increases the chance of stops during channel variations. This was avoided by the *GRP-DASH* which, in essence, allocates more airtime than the *P-DASH* based on the standard deviation and the QoS degradation level $\epsilon$. The optimality gap between the *P-DASH* and *RP-DASH* (GA and SA) also decreases with the increased load as shown in Fig. 7.8(a)-Fig. 7.8(b) since the former has to retroactively allocate extra airtime after detecting the video stops.

**Gaussian and Scenario Based Comparisons**

Comparing the *SRP-DASH* with *GRP-DASH*, the latter is found to be less robust, in terms of average stops, during the low standard deviations and high QoS degradation levels $\epsilon$ as shown in Fig. 7.3(a)-Fig. 7.3(b). However, this is not the case when the MOS is considered which illustrates that GA is equal or more robust than the SA as discussed in Section 7.4 especially at very low values of $\epsilon$. Since the MOS is calculated by an exponential function, it reveals that the GA provides a fair robustness across the users unlike the SA which

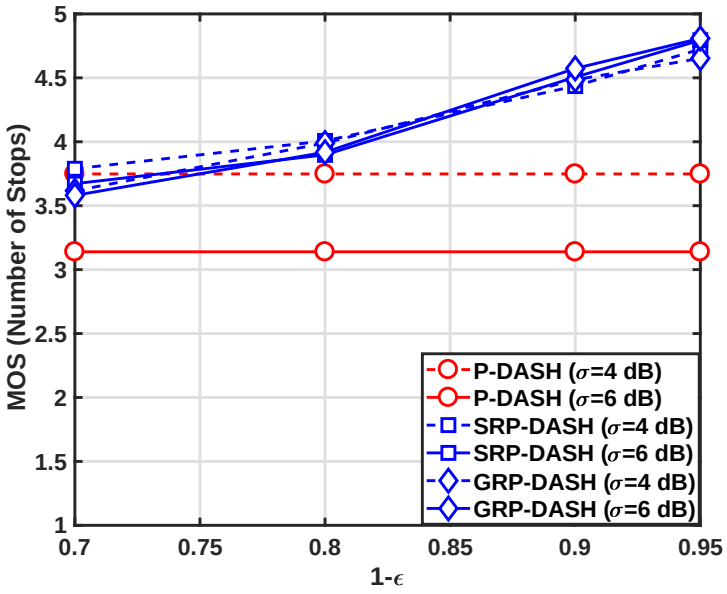



(a) Percentage of video stops

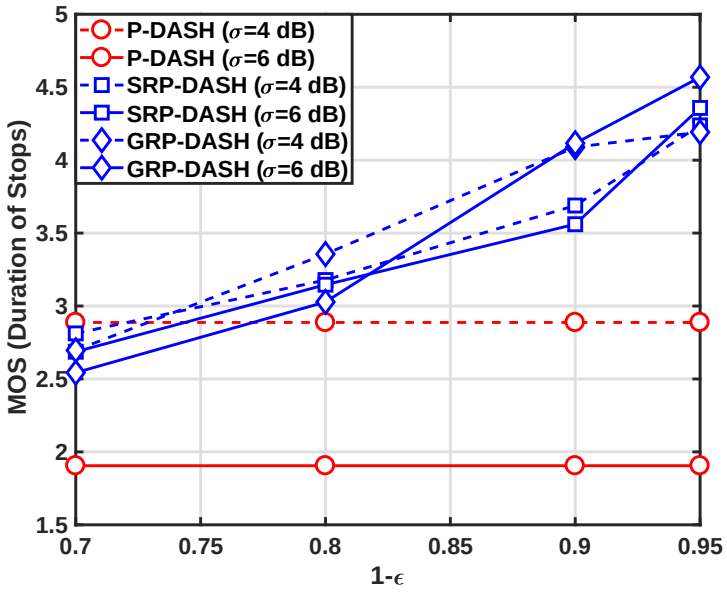



(b) Percentage of video stops duration

Figure 7.3: QoS performance of RP-DASH (SA and GA) for 4 users at different degradation levels

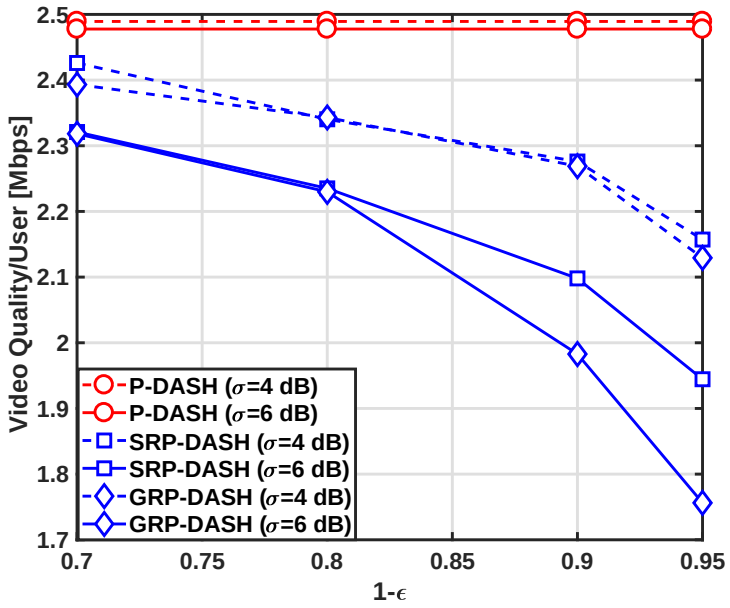



(a) Average MOS for the number of stops

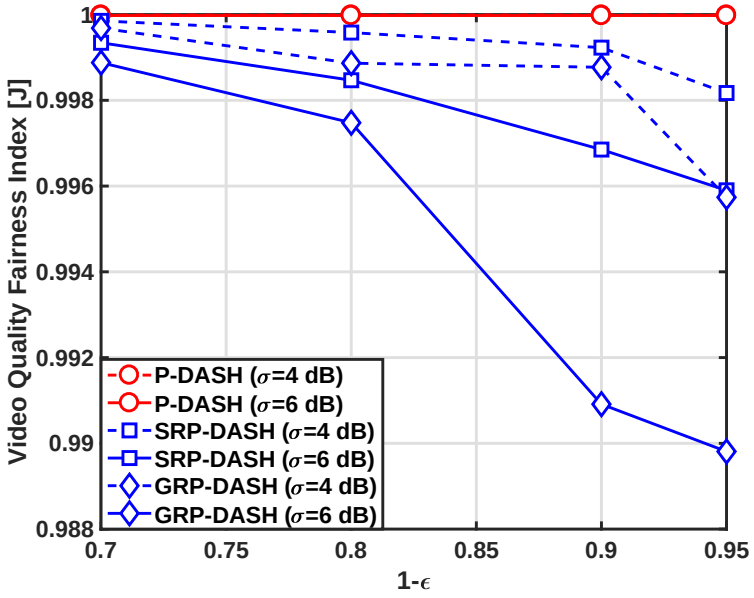



(b) Average MOS for the stop durations

Figure 7.4: QoE performance of RP-DASH (SA and GA) for 4 users at different degradation levels




(a) Avergae video quality



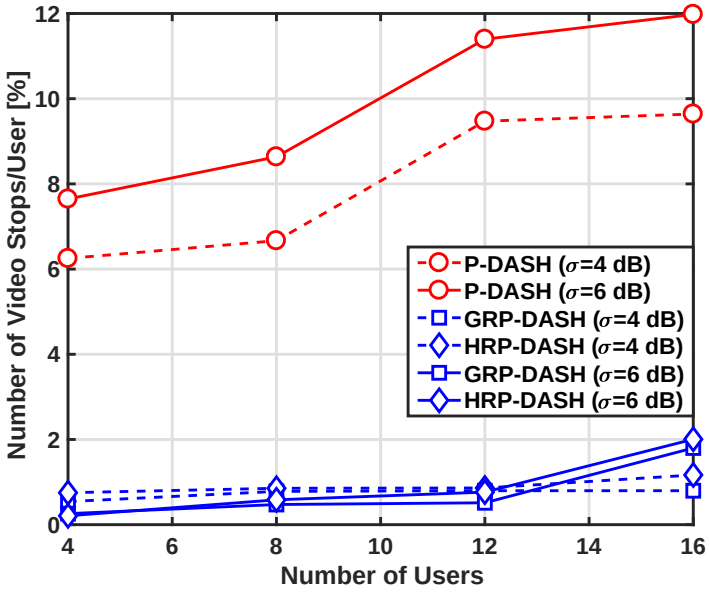

(b) User quality fairness

Figure 7.5: Quality performance of RP-DASH (SA and GA) for 4 users at different degradation levels

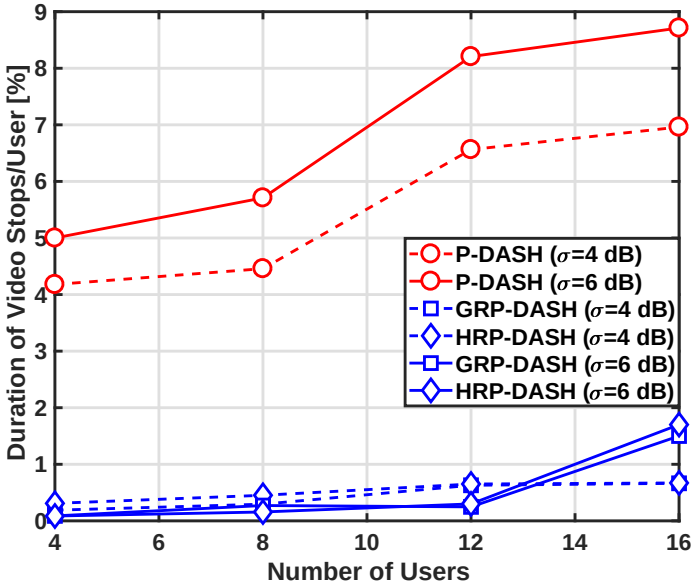
decreases the average degradation and conservatism. The optimality gap in Fig. 7.5(a)-Fig. 7.5(b) reveals another trade-off between the amount of information, required by the SA, and the lower quality obtained by the GA.

**Evaluation of the Heuristic and Complexity**

The performance of the introduced heuristic is reported for different numbers of users in Fig. 7.6(a)-Fig. 7.7(b). Similar to the *GRP-DASH*, the *HRP-DASH* was able to satisfy the maximum QoS degradation level $\epsilon$ and provided a stable QoS performance over the load and the channel standard deviation. It can be also seen that the *HRP-DASH* was slightly more conservative than the *GRP-DASH* and thus reported a smaller optimality gap in Fig. 7.8(a)-Fig. 7.8(b). This demonstrates the ability of the heuristic to exploit the problem structure and obtain near-optimal solutions that also satisfy the defined QoS degradation level $\epsilon$. The complexity of the both optimal and heuristic techniques is measured in terms of the execution time as reported in Table 7.3. The heuristic algorithm only requires less than $0.1ms.$ to solve the RP-DASH formulation irrespective of the network load (i.e. number of users) and the QoS degradation level $\epsilon$. This is unlike the commercial solver which required tens or hundreds of seconds to reach the target duality gap. The execution time increases with both the number of users, due to the larger problem dimension, and the QoS level $(1 - \epsilon)$ due to the tight feasibility region. When the optimal SA is used, more execution time is required compared to the GA due to the added auxiliary decision variables, thus, presenting a new trade-off between the complexity of SA and the conservatism of GA.

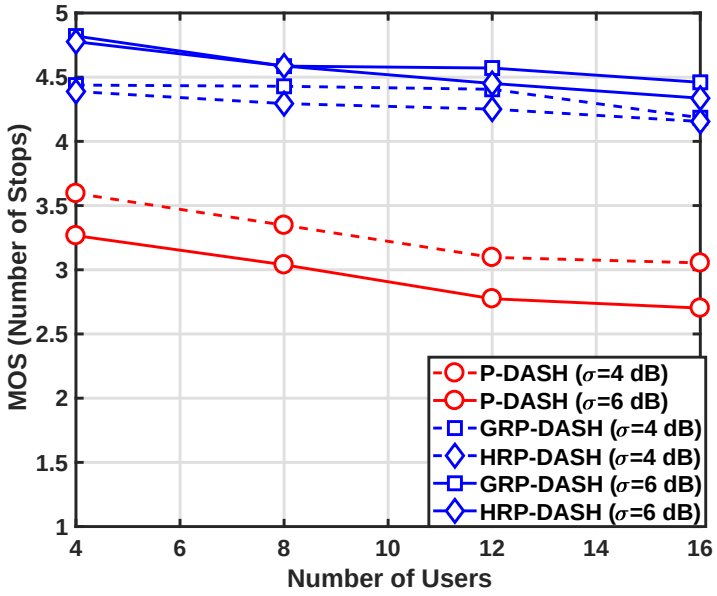



(a) Percentage of video stops

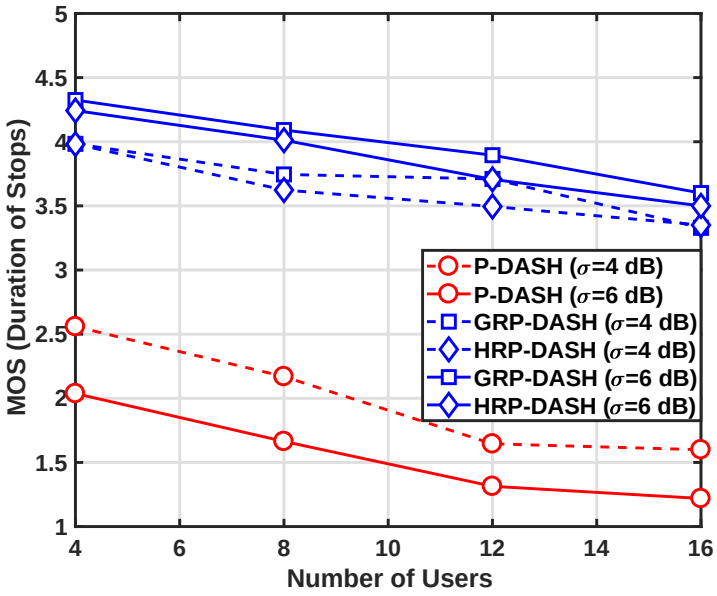



(b) Percentage of video stops duration

Figure 7.6: QoS performance for different number of users at $\epsilon = 0.1$

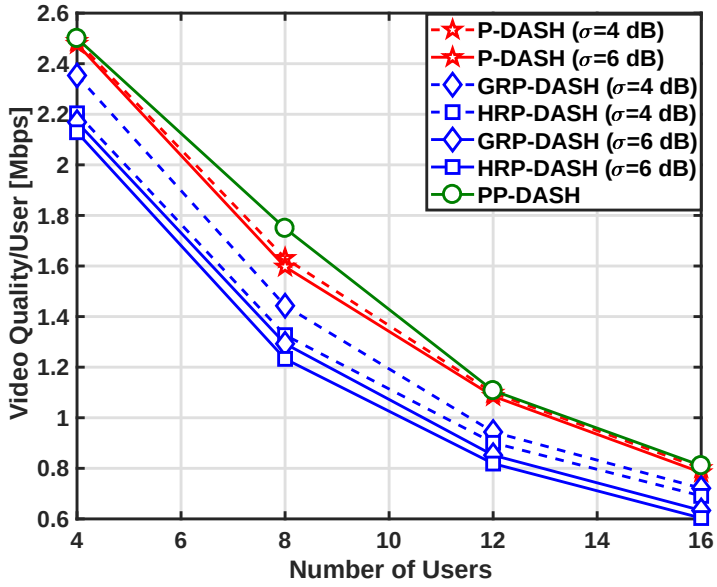



(a) Average MOS for the number of stops

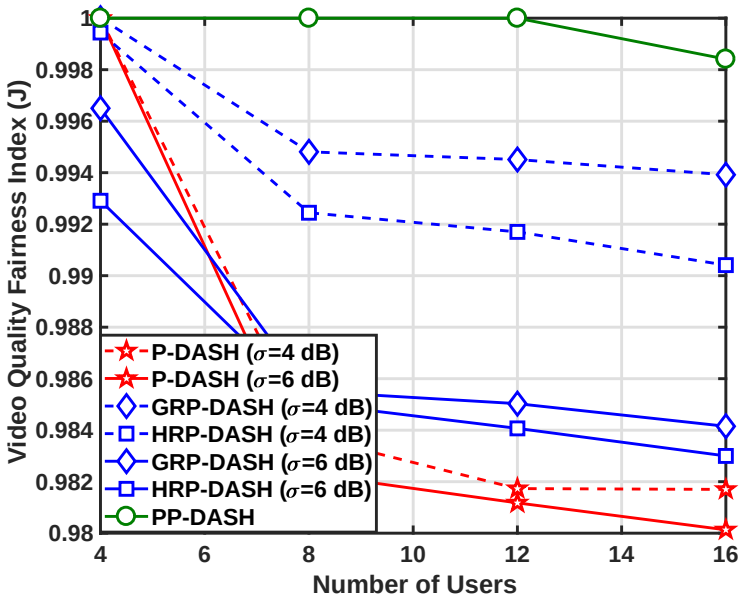



(b) Average MOS for the stop durations

Figure 7.7: QoE performance for different number of users at $\epsilon = 0.1$




(a) Avergae video quality




(b) User quality fairness

Figure 7.8: Video quality performance for different number of users at $\epsilon = 0.1$

Table 7.3: Execution Time of the Simulated Schemes

| Technique | Number of Users | | | | QoS Degrad. Level ($\epsilon$) | | | |
|---|---|---|---|---|---|---|---|---|
| | 4 | 8 | 12 | 16 | 0.3 | 0.2 | 0.1 | 0.05 |
| *P-DASH* | $50s.$ | $120s.$ | $290s.$ | $600s.$ | $30s.$ | $30s.$ | $30s.$ | $30s.$ |
| *SRP-DASH* | $60s.$ | $200s.$ | $320s.$ | $600s.$ | $50s.$ | $55s.$ | $60s.$ | $80s.$ |
| *GRP-DASH* | $50s.$ | $120s.$ | $260s.$ | $560s.$ | $30s.$ | $40s.$ | $50s.$ | $55s.$ |
| *HRP-DASH* | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ | $<0.1ms.$ |

# Chapter 8

# Conclusions and Future Directions

## 8.1 Summary and Conclusions

In this thesis, we addressed the problem of *predictive* resource allocation (PRA) for video streaming under imperfect prediction. In contrast to previous efforts [25–27, 37, 38], we developed a *robust*-PRA framework with uncertainty in mind that provides *joint probabilistic* QoS guarantees and risk-aware prebuffering over a time-horizon. By offering a mechanism to control the probability of constraint satisfaction, operators may strike a balance between network utilities such as energy and the risks associated with erroneous predictions. Furthermore, in order to facilitate practical deployment, near-optimal real-time solutions coupled with a channel variation tracking technique were developed. Different variants of R-PRA framework are introduced for energy-efficiency and QoS-aware adaptive streaming. Stochastic optimization Chance Constrained Programming (CCP) and Recourse Programming (RP) techniques are adopted, and tested under rate, demand and resource uncertainties. Results, obtained by a standard compliant simulator, indicate the resilience of R-PRA framework in meeting QoS constraints, while significantly reducing BS energy and achieve QoS fairness under practical prediction uncertainty. **To summarize, for the first time in literature, a robust framework is introduced for taking decisions over a**

**time horizon while considering the interdependency between the time slot constraints and the demand accumulation. This is further integrated with adaptive tracking of uncertainty degree to control the robustness level.**

The first two variants introduced *energy-efficient robust* schemes under rate uncertainties. The first solution assumed Gaussian distributed rate error model and integrated the Gaussian Approximation (GA) for chance-constraint QoS modeling, a Kalman Filter (KF) for prediction uncertainty tracking, and a guided heuristic that enables real-time implementation. The second variant, on the other hand, adopted Bernstein Approximation (BA) combined with Particle Filter (PF) to handle uncertainty with unknown or complex rate error models in which the Cumulative Density Function (CDF) is non-invertible. Such tracking enables the operator to be greedy during periods of accurate predictions, and thereby maximizes energy savings without compromising QoS. Using a guided heuristic enabled the adoption of the GA and BA in their original less conservative Second order Cone Programming (SoCP) form as opposed to linear approximations in the literature. These results are unlike the existing *non-robust* PRA that rely only on *average* future rates and thus suffered from QoS violations due to increased number of video stops. The results further demonstrated that *non-predictive* RA either consumes excess energy or violates the QoS level under low or high load scenarios, respectively.

Both Gaussian and Bernstein approximations are tested for meeting target QoS level. At small feedback intervals that require frequent optimization via optimal solvers, both approaches were able to meet the QoS level while keeping the energy-saving gain close to the benchmark. Less complex longer feedback intervals showed however different performances in both QoS satisfaction and energy saving. In particular, Gaussian approximation was not robust to avoid the accumulation of video stops over consecutive time slots. This

necessitates either *joint* probabilistic form with optimal risk allocation or short feedback intervals. On the contrary, Bernstein approximation was able to satisfy the QoS level but with high cost of robustness, i.e. more energy consumption, due to using the bounds rather than the CDF of the prediction error.

The third variant focused also on energy-saving, but handled uncertainties in both user demands and network resources over a time horizon. The RP and CCP models adopted the probability of random video termination and arrival of real-time users. The performance evaluation demonstrated the ability of the introduced scheme to maintain the energy-saving gains of PRA while satisfying the QoS levels. An increase in system load underlines the importance of having a robust scheme to avoid excessive allocation for users leaving the cell center and with high probability of terminating the video before viewing the prebuffered content. This is unlike existing PRA schemes that greedily exploit the peak radio conditions by prebuffering the whole future content without taking into consideration the unstable user demand. As such, high energy consumption is observed compared to the non-predictive scheme employed in today's network.

The last variant focused on *Robust Predictive*-DASH (RP-DASH) to jointly calculate the resources and video quality while handling uncertainties in predicted rates and achieving streaming quality fairness among the users. New linear deterministic equivalent forms are then proposed based on GA and SA to provide closed form solutions at a polynomial complexity as opposed to traditional forms. Unlike the robust optimization literature, the allocation over time-horizon will result in GA and SA with non-polynomial complexity and non-convex approximations. As such, linearized and non-conservative yet robust approximations are proposed in this work. The performance showed the ability of probabilistic RP-DASH to satisfy the predefined QoS level. This is unlike the existing non-probabilistic

P-DASH schemes which assume ideal prediction and thus experience high degradation in users' QoS and QoE. The results further revealed a trade-off between the risk of experiencing video stops and maximizing video quality, which increases the need for an explicit modelling of user's preferences. As such, users seeking high video qualities should be assigned low QoS probabilistic levels at the expense of increased number and duration of stops. In addition to satisfying the QoS level, the small optimality gap between the SA and GA promises the adoption of the latter in RP-DASH with quality maximization. This is unlike the existing conclusions on GA that doubted its robustness in long-term energy-efficient predictive video delivery. Adopting the GA in robust predictive DASH will decrease the cost of uncertainty modelling as the network operator will not rely on the exact realizations of future rates. Moreover, near-optimal real-time robust solutions are obtainable for the energy-saving and DASH scheme through a low complexity guided heuristic algorithm that exploits the problem structure. All the above performance improvements and design flexibilities envision the implementation of R-PRA in future wireless networks under practical uncertainties.

**Compared to non-predictive schemes in today's networks, the R-PRA demonstrates that significant prediction gains are still achievable under all kinds of uncertainties.**

## 8.2 Future Directions

The future work considers the following enhancements to the system model, R-PRA framework and the performance evaluation:

1. System Model:
   - Backhaul and Application: The main focus in this thesis is the wireless link.

However, the uncertainties and limitations of the backhaul network have to be taken into consideration. This includes the delivery of the video content from the application server to the BS cache. The delays in the application server response to the user requests, maximum caching capacity in the BS and user device, and the backhaul link capacities have to be explicitly modelled to achieve an optimized end-to-end performance.

- Multi-cell Environment: While each BS individually executes the R-PRA solutions, cooperative scheduling has to be introduced. In particular, neighbouring BSs can jointly exchange future information and calculate resource allocation that controls the inter-cell interference. For instance, the interfering BSs can schedule their sleep interval to void simultaneous transmission and thus increase the total channel capacity.

2. R-PRA Framework:

   - Robust Optimization: the proposed work typically relied on stochastic optimization to handle uncertainties. Other robust techniques such as Fuzzy and decision under uncertainty such as Markov decision process or belief networks can be also introduced. The fuzzy is known for its low complexity but high conservatism which can be handled by real-time tracking of error variance. Other probabilistic decision making techniques such as Markov decision process can provide simple uncertainty modelling as it only requires conditional probability among the system states rather than the error CDF.
   - Real-time Prediction: the future information can be recomputed frequently over

the time horizon to correct the previous predictions while leveraging the unveiled randomness over time. In particular, the KF and PF used to track the degree of uncertainty, can be extended to correct the predicted information using the measurements. This will eventually lead to low uncertainty degree and thus low safety term (i.e. cost of robustness).

3. Performance Evaluation and Optimization:

   – Probabilistic QoE Models: Existing user experience models, i.e. QoE, can be extended to capture the trade-off between video stops and selected quality using the probabilistic metric. Particularly, a new QoE model is needed to consider the user's preference, i.e. both quality and stops, as a function of the QoS level $\epsilon$. Such model would guide the operator while selecting the value of $\epsilon$ jointly with the resources and quality of segments to reflect the user preference.

   – Dynamic Objective: Existing PRA focused on optimizing either the QoS (e.g. video quality), in high load scenarios, or decreasing the energy-consumption in low load cases. A joint optimization model is desirable to autonomously evaluate the network load and select the objective function to optimize (e.g. energy or QoS parameter).

   – Testbed Implementation: An experimental evaluation is needed to assess the R-PRA under real network conditions. Thus, assess the performance gains of both PRA literature and robust techniques while considering practical uncertainness. This is in addition to discovering implementation challenges and verifying the assumptions on system model of the PRA literature and the proposed framework.

# Bibliography

[1] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, pp. 4683–4688, Dec 2014.

[2] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Chance-constrained qos satisfaction for predictive video streaming," in *Proc. IEEE LCN*, pp. 253–260, 2015.

[3] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE J. Select. Areas Commun.*, vol. 34, pp. 1389–1404, May 2016.

[4] R. Atawia, H. S. Hassanein, H. Abou-zeid, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Wireless Commun.*, pp. 1–14, 2017.

[5] R. Atawia, H. Hassanein, and A. Noureldin, "Energy-efficient predictive video streaming under demand uncertainties," in *Proc. IEEE ICC*, pp. 1–6, May 2017.

[6] R. Atawia, H. S. Hassanein, N. A. Ali, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Green Commun. Netw.*, pp. 1–14, 2017.

[7] R. Atawia, H. Hassanein, and A. Noureldin, "Fair robust predictive resource allocation for video streaming under rate uncertainties," in *Proc. IEEE GLOBECOM*, pp. 1–6, Dec 2016.

[8] R. Atawia, H. S. Hassanein, and A. Noureldin, "Robust long-term predictive adaptive video streaming under wireless network uncertainties," *IEEE Trans. Wireless Commun.*, pp. 1–14, 2017.

[9] R. Atawia, H. Hassanein, and A. Noureldin, "Optimal and robust qos-aware predictive adaptive video streaming for future wireless networks," in *Proc. IEEE GLOBECOM*, pp. 1–6, Dec 2017.

[10] K. Kobayashi and Y. Matsunaga, "Radio quality prediction based on user mobility and radio propagation analysis," in *Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pp. 2137–2141, IEEE, 2009.

[11] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021." http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2017. Accessed Jun. 15th, 2017.

[12] T. Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*, pp. 133–144, ACM, 2011.

[13] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[14] L. Correia, D. Zeller, O. Blume, D. Ferling, A. Kangas, I. Godor, G. Auer, and L. Van der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Magazine*, vol. 48, no. 11, pp. 66–72, 2010.

[15] K. Davaslioglu and E. Ayanoglu, "Quantifying potential energy efficiency gain in green cellular wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 16, pp. 2065–2091, Fourthquarter 2014.

[16] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE ICC Comm. Workshops*, pp. 1–5, 2009.

[17] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. IEEE GLOBECOM*, pp. 1–5, 2010.

[18] A. A. Hammad, T. D. Todd, G. Karakostas, and D. Zhao, "Downlink traffic scheduling in green vehicular roadside infrastructure," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1289–1302, 2013.

[19] iGR, "U.S. regional and small operator network infrastructure Capex and Opex forecast, 2012-2017." https://igr-inc.com/, 2013. Accessed Mar. 29th, 2017.

[20] M. Azimifar, T. D. Todd, A. Khezrian, and G. Karakostas, "Vehicle-to-vehicle forwarding in green roadside infrastructure," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 780–795, 2016.

[21] A. Khezrian, T. D. Todd, G. Karakostas, and M. Azimifar, "Energy-efficient scheduling in green vehicular infrastructure with multiple roadside units," *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 1942–1957, 2015.

[22] A. A. Hammad, T. D. Todd, and G. Karakostas, "Variable-bit-rate transmission schedule generation in green vehicular roadside units," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1590–1604, 2016.

[23] 3GPP, "LTE; transparent end-to-end packet-swiched streaming service (pss); progressive download and dynamic adaptive streaming over HTTP (3gp-dash)," Tech. Rep. TS 26.247 v10.7.0, 2012.

[24] A. El Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 6, pp. 988–1001, 2015.

[25] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE INFOCOM*, pp. 2806–2814, 2013.

[26] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, pp. 1339–1347, 2014.

[27] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, 2013.

[28] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013 – 2026, 2014.

[29] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *Proc. IEEE GLOBECOM*, pp. 4714–4719, 2013.

[30] M. S. Zefreh and T. D. Todd, "Energy provisioning in green mesh networks using positional awareness," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 4064–4076, 2014.

[31] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, no. 2923, pp. 1–9, 2013.

[32] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[33] H. Abou-zeid, H. S. Hassanein, Z. Tanveer, and N. A. Ali, "Evaluating mobile signal and location predictability along public transportation routes," in *Proc. IEEE WCNC*, pp. 1195 – 1200, 2015.

[34] N. Bui and J. Widmer, "Modelling throughput prediction errors as Gaussian random walks," in *Poc. KuVS Workshop on Anticipatory Network*, 2014.

[35] J. Yao, S. Kanhere, and M. Hassan, "Improving QoS in high-speed mobility using bandwidth maps," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 603–617, 2012.

[36] M. Neuland, T. Kurner, and M. Amirijoo, "Influence of positioning error on x-map estimation in lte," in *Proc.IEEE VTC (Spring)*, pp. 1–5, 2011.

[37] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013 – 2026, 2014.

[38] H. Abou-zeid and H. S. Hassanein, "Toward green media delivery: location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, 2014.

[39] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing-a study of user behavior in online vod services," in *Proc. IEEE ICCCN*, pp. 1–7, 2013.

[40] Y. Chen, B. Zhang, Y. Liu, and W. Zhu, "Measurement and modeling of video watching time in a large-scale internet video-on-demand system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2087–2098, 2013.

[41] H. Zhou, P. Fan, and J. Li, "Global proportional fair scheduling for networks with multiple base stations," *IEEE Trans. on Veh. Tech.*, vol. 60, pp. 1267–1879, 2011.

[42] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process*, vol. 62, pp. 6184–6193, Dec 2014.

[43] N. Y. Soltani, S.-J. Kim, and G. B. Giannakis, "Chance-constrained optimization of ofdma cognitive radio uplinks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1098–1107, 2013.

[44] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. N. Padmanabhan, "Bartendr: a practical approach to energy-aware cellular data scheduling," in *Proc. ACM Mobicom*, pp. 85–96, 2010.

[45] J. Hajipour and V. C. Leung, "Proportional fair scheduling in multi-carrier networks using channel predictions," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, pp. 1–5, 2010.

[46] H. Farahat and H. S. Hassanein, "Proactive caching for producer mobility management in named data networks," in *Proc. IEEE IWCMC*, pp. 171–176, 2017.

[47] H. Farahat and H. S. Hassanein, "Supporting consumer mobility using proactive caching in named data networks," in *Proc. IEEE GLOBECOM*, pp. 1–6, 2016.

[48] H. Farahat and H. Hassanein, "Optimal caching for producer mobility support in named data networks," in *Proc. IEEE ICC*, pp. 1–6, 2016.

[49] H. Farahat and H. Hassanein, "On the design and evaluation of producer mobility management schemes in named data networks," in *Proc. ACM MSWIM*, pp. 171–178, 2015.

[50] H. Farahat, R. Atawia, and H. S. Hassanein, "Robust proactive mobility management in named data networking under erroneous content prediction," in *Proc. IEEE GLOBECOM*, pp. 1–6, 2017.

[51] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2299–2313, 2007.

[52] Y. Zhao, B. Le, and J. H. Reed, *Network support–The radio environment map*. Elsevier, 2006.

[53] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Automated coverage hole detection for cellular networks using radio environment maps," in *Proc. IEEE WiOpt*, pp. 35–40, 2013.

[54] T. Cai, J. van de Beek, B. Sayrac, S. Grimoud, J. Nasreddine, J. Riihijärvi, and P. Mähönen, "Design of layered radio environment maps for ran optimization in heterogeneous lte systems," in *Proc. IEEE PIMRC*, pp. 172–176, 2011.

[55] J. Johansson, W. A. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3gpp release 11," *IEEE Communications Magazine*, vol. 50, no. 11, 2012.

[56] C. Brunner and D. Flore, "Generation of pathloss and interference maps as son enabler in deployed umts networks," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pp. 1–5, IEEE, 2009.

[57] I. Skog and P. Handel, "In-car positioning and navigation technologies—a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 4–21, 2009.

[58] S. Tao, V. Manolopoulos, S. Rodriguez, M. Ismail, and A. Rusu, "Hybrid vehicle positioning and tracking using mobile phones," in *ITS Telecommunications (ITST), 2011 11th International Conference on*, pp. 315–320, IEEE, 2011.

[59] K. Farkas, T. Hossmann, F. Legendre, B. Plattner, and S. K. Das, "Link quality prediction in mesh networks," *Computer Communications*, vol. 31, no. 8, pp. 1497–1512, 2008.

[60] D. Fernandes Boesel, "Prediction of vehicle trajectories with map data for cooperative systems," 2009.

[61] R. Franke, "A critical comparison of some methods for interpolation of scattered data," tech. rep., DTIC Document, 1979.

[62] R. J. Renka, "Multivariate interpolation of large sets of scattered data," *ACM Transactions on Mathematical Software (TOMS)*, vol. 14, no. 2, pp. 139–148, 1988.

[63] D. Denkovski, V. Atanasovski, L. Gavrilovska, J. Riihijärvi, and P. Mähönen, "Reliability of a radio environment map: Case of spatial interpolation techniques," in *Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 2012 7th International ICST Conference on*, pp. 248–253, IEEE, 2012.

[64] L. Bolea, J. Pérez-Romero, and R. Agustí, "Received signal interpolation for context discovery in cognitive radio," in *Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on*, pp. 1–5, 2011.

[65] W. Fa, F. Xu, and Y. Jin, "Sar imaging simulation for an inhomogeneous undulated lunar surface based on triangulated irregular network," *Science in China Series F: Information Sciences*, vol. 52, no. 4, pp. 559–574, 2009.

[66] S. Üreten, A. Yongaçoğlu, and E. Petriu, "Interference map generation based on delaunay triangulation in cognitive radio networks," in *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, pp. 134–138, 2012.

[67] S. Üreten, A. Yongaçoğlu, and E. Petriu, "A comparison of interference cartography generation techniques in cognitive radio networks," in *Communications (ICC), 2012 IEEE International Conference on*, pp. 1879–1883, 2012.

[68] J. Ojaniemi, J. Kalliovaara, A. Alam, J. Poikonen, and R. Wichman, "Optimal field measurement design for radio environment mapping," in *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pp. 1–6, 2013.

[69] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Long-term fairness in multi-cell networks using rate predictions," in *Proc. IEEE GCC Conf. and Exhibition (GCC)*, pp. 131–135, 2013.

[70] X. K. Zou, J. Erman, V. Gopalakrishnan, E. Halepovic, R. Jana, X. Jin, J. Rexford, and R. K. Sinha, "Can accurate predictions improve video streaming in cellular networks?," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pp. 57–62, ACM, 2015.

[71] S. Cicalo, N. Changuel, V. Tralli, B. Sayadi, F. Faucheux, and S. Kerboeuf, "Improving QoE and fairness in HTTP adaptive streaming over LTE network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2015.

[72] K. J. Ma and R. Bartos, "HTTP live streaming bandwidth management using intelligent segment selection," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pp. 1–5, 2011.

[73] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using openflow-assisted adaptive video streaming," in *Proceedings of the 2013 ACM SIGCOMM workshop on Future human-centric multimedia networking*, pp. 15–20, ACM, 2013.

[74] A. Seetharam, P. Dutta, V. Arya, J. Kurose, M. Chetlur, and S. Kalyanaraman, "On managing quality of experience of multiple video streams in wireless networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 619–631, 2015.

[75] M. Elazab, A. Noureldin, and H. Hassanein, "Integrated cooperative localization for connected vehicles in urban canyons," in *Proc. IEEE Globecom*, 2015.

[76] M. Elazab, A. Noureldin, and H. S. Hassanein, "Integrated cooperative localization for vehicular networks with partial gps access in urban canyons," *Vehicular Communications*, 2016.

[77] M. Elazab, "Integrated cooperative localization in vanets for gps denied environments," Master's thesis, 2015.

[78] A. Mahmoud, A. Noureldin, and H. Hassanein, "Vanets positioning in urban environments : A novel cooperative approach," in *Proc. IEEE VTC (Fall)*, 2015.

[79] A. Mahmoud, A. Noureldin, and H. S. Hassanein, "Distributed vehicle selection for non-range based cooperative positioning in urban environments," in *Proc. IEEE ICC*, pp. 1–6, 2016.

[80] W. Hu and G. Cao, "Energy-aware video streaming on smartphones," in *Proc. IEEE INFOCOM*, pp. 1185–1193, 2015.

[81] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Using crowd-sourced viewing statistics to save energy in wireless video streaming," in *Proc. ACM MobiCom*, pp. 377–388, 2013.

[82] R. Ramamonjison and V. K. Bhargava, "Sum energy-efficiency maximization for cognitive uplink networks with imperfect CSI," in *Proc. IEEE WCNC*, pp. 1012–1017, 2014.

[83] M. Abdel-Rahman and M. Krunz, "Stochastic guard-band-aware channel assignment with bonding and aggregation for DSA networks," *IEEE Trans. Wireless Commun.*, vol. 14, pp. 3888–3898, July 2015.

[84] B. Liu, *Theory and practice of uncertain programming*. Springer, 2002.

[85] P. Kali and S. W. Wallace, *Stochastic programming*. Springer, 1994.

[86] Q. Chen, "Comparing probabilistic and fuzzy set approaches for designing in the presence of uncertainty," 2000.

[87] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[88] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Management science*, vol. 6, no. 1, pp. 73–79, 1959.

[89] A.-C. So and Y. J. Zhang, "Distributionally robust slow adaptive ofdma with soft qos via linear programming," *IEEE J. Select. Areas Commun.*, vol. 31, no. 5, pp. 947–958, 2013.

[90] W.-L. Li, Y. Zhang, A.-C. So, and M. Z. Win, "Slow adaptive ofdma systems through chance constrained programming," *IEEE Trans. Signal Process*, vol. 58, no. 7, pp. 3858–3869, 2010.

[91] M. J. Abdel-Rahman, F. Lan, and M. Krunz, "Spectrum-efficient stochastic channel assignment for opportunistic networks," in *Proc. IEEE GLOBECOM*, pp. 1272–1277, 2013.

[92] S. Parsaeefard, A. R. Sharafat, and M. Rasti, "Robust probabilistic distributed power allocation by chance constraint approach," in *Proc. IEEE PIMRC*, pp. 2162–2167, 2010.

[93] B. L. Miller and H. M. Wagner, "Chance constrained programming with joint constraints," *Operations Research*, vol. 13, no. 6, pp. 930–945, 1965.

[94] G. Classen, D. Coudert, A. M. Koster, and N. Nepomuceno, "Bandwidth assignment for reliable fixed broadband wireless networks," in *Proc. IEEE WoWMoM*, pp. 1–6, 2011.

[95] B. Nunez, P. Adasme, I. Soto, J. Cheng, M. Letournel, and A. Lisser, "A chance constrained approach for uplink wireless ofdma networks," in *Proc. CSNDSP*, pp. 754–757, 2014.

[96] S. S. Venkatesh, *The Theory of Probability: Explorations and Applications*. Cambridge University Press, 2012.

[97] M. Ono and B. C. Williams, "Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint," in *Proc. IEEE CDC*, pp. 3427–3432, 2008.

[98] F. Oldewurtel, C. N. Jones, and M. Morari, "A tractable approximation of chance constrained stochastic mpc based on affine disturbance feedback," in *Proc. of IEEE Conf. on Decision and Control*, pp. 4731–4736, 2008.

[99] U. A. Ozturk, M. Mazumdar, and B. A. Norman, "A solution to the stochastic unit commitment problem using chance constrained programming," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1589–1598, 2004.

[100] Q. Wang, Y. Guan, and J. Wang, "A chance-constrained two-stage stochastic program for unit commitment with uncertain wind power output," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 206–215, 2012.

[101] M. Ono and B. C. Williams, "Decentralized chance-constrained finite-horizon optimal control for multi-agent systems," in *Proc. IEEE CDC*, pp. 138–145, 2010.

[102] Q.-T. Thieu and H.-Y. Hsieh, "Use of chance-constrained programming for solving the opportunistic spectrum sharing problem under rayleigh fading," in *Proc. IEEE IWCMC*, pp. 1792–1797, 2013.

[103] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *Proc. European Wireless*, pp. 1–6, 2014.

[104] W. Xu, A. Tajer, X. Wang, and S. Alshomrani, "Power allocation in miso interference channels with stochastic csit," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1716–1727, 2014.

[105] J. E. Mitchell, "Polynomial interior point cutting plane methods," *Optimization Methods and Software*, vol. 18, no. 5, pp. 507–534, 2003.

[106] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear algebra and its applications*, vol. 284, no. 1, pp. 193–228, 1998.

[107] A. Noureldin, T. B. Karamat, and J. Georgy, *Fundamentals of Inertial Navigation, Satellite-Based Positioning and Their Integration*. Springer, 2013.

[108] K. Huber and S. Haykin, "Improved bayesian mimo channel tracking for wireless communications: incorporating a dynamical model," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2458–2466, 2006.

[109] L. Mihaylova, D. Angelova, S. Honary, D. R. Bull, C. N. Canagarajah, and B. Ristic, "Mobility tracking in cellular networks using particle filtering," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3589–3599, 2007.

[110] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Cellular coverage optimization: A radio environment map for minimization of drive tests," in *Cognitive Communication and Cooperative HetNet Coexistence*, pp. 211–236, 2014.

[111] N. Kolehmainen, J. Puttonen, P. Kela, T. Ristaniemi, T. Henttonen, and M. Moisio, "Channel quality indication reporting schemes for utran long term evolution downlink," in *Proc. IEEE VTC (Spring)*, pp. 2522–2526, 2008.

[112] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); physical layer procedures," Technical Specification TS 36.213 v12.5.0, 3GPP, 2015.

[113] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects," Technical Report TR 36.814 V9.0.0, 3GPP, 2010.

[114] Y. Li, M. Reisslein, and C. Chakrabarti, "Energy-efficient video transmission over a wireless link," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1229–1244, 2009.

[115] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, and H. Holktamp, "Flexible power modeling of lte base stations," in *Proc. IEEE WCNC*, pp. 2858–2862, 2012.

[116] G. Auer, V. Giannini, I. Gódor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, C. Desset, and O. Blume, "Cellular energy efficiency evaluation framework," in *Proc. IEEE VTC (Spring)*, pp. 1–6, 2011.

[117] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE J. Select. Areas Commun.*, vol. 32, no. 2, pp. 297–306, 2014.

[118] A. ParandehGheibi, M. Médard, A. Ozdaglar, and S. Shakkottai, "Avoiding interruptions—a qoe reliability function for streaming media applications," *IEEE J. Select. Areas Commun.*, vol. 29, no. 5, pp. 1064–1074, 2011.

[119] Y. Xu, E. Altman, R. El-Azouzi, M. Haddad, S. Elayoubi, and T. Jimenez, "Analysis of buffer starvation with application to objective QoE optimization of streaming services," *Multimedia, IEEE Transactions on*, vol. 16, no. 3, pp. 813–827, 2014.

[120] N. Bui, F. Michelinakis, and J. Widmer, "Mobile network resource optimization under imperfect prediction," in *Proc. IEEE WoWMoM*, pp. 1–6, 2015.

[121] N. Y. Soltani, S.-J. Kim, and G. B. Giannakis, "Chance-constrained optimization of ofdma cognitive radio uplinks," *IEEE Trans. Commun.*, vol. 12, no. 3, pp. 1098–1107, 2013.

[122] 3GPP, "E-UTRA; base station (BS) radio transmission and reception (release 10)," Technical Specification TS 36.104 V10.2.0, Dec. 2011.

[123] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[124] Gurobi, "Gurobi Optimization." http://www.gurobi.com/. Accessed Mar. 29th, 2017.

[125] G. Piro, N. Baldo, and M. Miozzo, "An LTE module for the ns-3 network simulator," in *Proc. Int. ICST Conf. on Simulation Tools and Techniques*, pp. 415–422, 2011.

[126] H. Abou-zeid, H. S. Hassanein, and R. Atawia, "Towards mobility-aware predictive radio access: modeling; simulation; and evaluation in lte networks," in *Proc. ACM MSWiM*, pp. 109–116, 2014.

[127] Gurobi, "Gurobi Optimization." http://www.gurobi.com/products/features-benefits/. Accessed Mar. 29th, 2015.

[128] MATLAB, "MATLAB Optimization." http://www.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html/. Accessed Mar. 29th, 2017.

[129] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA based radio access*. John Wiley & Sons, 2009.

[130] A. Ben-Tal and A. Nemirovski, "Selected topics in robust convex optimization," *Mathematical Programming*, vol. 112, no. 1, pp. 125–158, 2008.

[131] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2013.

[132] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi, *Complexity and approximation: Combinatorial optimization problems and their approximability properties*. Springer Science & Business Media, 2012.

[133] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proc. IEEE ISM*, pp. 494–499, 2011.

[134] L. G. M. Ballesteros, S. Ickin, M. Fiedler, J. Markendahl, K. Tollmar, and K. Wac, "Energy saving approaches for video streaming on smartphone based on QoE modeling," in *Proc. IEEE CCNC*, pp. 103–106, IEEE, 2016.

[135] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 1–12, Apr. 2006.

[136] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Long-term fairness in multi-cell networks using rate predictions," in *Proc. IEEE GCC Conf. and Exhibition (GCC)*, pp. 131–135, 2013.

# Appendix

## Appendix A

The objective function and all constraints in Eq. 5.12 are linear except the QoS one. The convexity of this first constraint will be checked using the Hessian matrix, which should be positive semidefinite [123]. Let the QoS constraint for the first user ($i = 0$) at time $t = 1$ be denoted as $f(x_{0,0}, x_{0,1}, \zeta_{0,1})$. In the standard form, the constraint is represented as follows

$$f(x_{0,0}, x_{0,1}, \zeta_{0,1}) = -\sum_{t'=0}^{1} \bar{r}_{0,t'} x_{0,t'} - Q^{-1}_{1-\zeta_{0,1}} \sqrt{\sum_{t'=0}^{1} x^2_{0,t'} \sigma^2_{0,t'}} \tag{1}$$

For the ease of representation, let $f(x_{0,0}, x_{0,1}, \zeta_{0,1})$, $x_{0,0}$, $x_{0,1}$ and $\zeta_{0,1}$ be denoted as $\mathcal{F}$, $x_0$, $x_1$ and $\zeta$ respectively. The Hessian matrix $H$ can then be defined as follows

$$H = \nabla^2 \mathcal{F} = \begin{bmatrix} \frac{\partial^2 \mathcal{F}}{\partial x_0^2} & \frac{\partial^2 \mathcal{F}}{\partial x_0 \partial x_1} & \frac{\partial^2 \mathcal{F}}{\partial x_0 \partial \zeta} \\ \frac{\partial^2 \mathcal{F}}{\partial x_1 \partial x_0} & \frac{\partial^2 \mathcal{F}}{\partial x_1^2} & \frac{\partial^2 \mathcal{F}}{\partial x_1 \partial \zeta} \\ \frac{\partial^2 \mathcal{F}}{\partial \zeta \partial x_0} & \frac{\partial^2 \mathcal{F}}{\partial \zeta \partial x_1} & \frac{\partial^2 \mathcal{F}}{\partial \zeta^2} \end{bmatrix} \tag{2}$$

$$\frac{\partial^2 \mathcal{F}}{\partial x_0^2} = -Q_{1-\zeta}^{-1}\sigma_{0,0}^2 \frac{x_1^2\sigma_{0,1}^2}{\left(\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2}\right)^3} \tag{3}$$

$$\frac{\partial^2 \mathcal{F}}{\partial x_1^2} = -Q_{1-\zeta}^{-1}\sigma_{0,1}^2 \frac{x_0^2\sigma_{0,0}^2}{\left(\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2}\right)^3} \tag{4}$$

$$\frac{\partial^2 \mathcal{F}}{\partial \zeta^2} = -\frac{\partial^2 Q_{1-\zeta}^{-1}}{\partial \zeta^2}\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2} \tag{5}$$

$$\frac{\partial^2 \mathcal{F}}{\partial x_0 \partial x_1} = \frac{\partial^2 \mathcal{F}}{\partial x_1 \partial x_0} = Q_{1-\zeta}^{-1}\sigma_{0,0}^2\sigma_{0,1}^2 \frac{x_0 x_{0,}\sigma_{0,0}^2\sigma_{0,1}^2}{\left(\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2}\right)^3} \tag{6}$$

$$\frac{\partial^2 \mathcal{F}}{\partial x_0 \partial \zeta} = \frac{\partial^2 \mathcal{F}}{\partial \zeta \partial x_0} = -\frac{\partial Q_{1-\zeta}^{-1}}{\partial \zeta} \frac{x_0\sigma_{0,0}^2}{\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2}} \tag{7}$$

$$\frac{\partial^2 \mathcal{F}}{\partial x_1 \partial \zeta} = \frac{\partial^2 \mathcal{F}}{\partial \zeta \partial x_1} = -\frac{\partial Q_{1-\zeta}^{-1}}{\partial \zeta} \frac{x_1\sigma_{0,1}^2}{\sqrt{\sum_{t'=0}^{1} x_{t'}^2\sigma_{0,t'}^2}} \tag{8}$$

The function (F) is convex if the Hessian matrix is positive semidefinite. In particular, all the principle minors should be positive or zero.

- The value of satisfaction degree of individual chance constraint (i.e., $\zeta$) should be less than 0.5 to satisfy the constraint (summation of $\zeta$) for $\beta > 0.5$. Accordingly, the inverse of Q function $Q_{1-\zeta}^{-1}$ is less than 0. Thus, all the first order principle minors are positive.

- The first second order principle minor (constructed by deleting the third row and column) is always positive for all the values of $x_0$ and $x_1$, $\sigma_{0,0}$ and $\sigma_{0,0}$. However, this is not the case for the other second order principle minors whose positiveness depend on the actual values of $x_0$ and $x_1$, $\sigma_{0,0}$ and $\sigma_{0,0}$. For illustration, the value

of a second order principle (constructed by deleting the second row and column) is calculated as follows

$$\Delta_5 = \Bigg( - Q^{-1}_{1-\zeta} \sigma^2_{0,0} \frac{x_1^2 \sigma^2_{0,1}}{\big(\sqrt{\sum_{t'=0}^{1} x^2_{t'} \sigma^2_{0,t'}}\big)^3} \Bigg) \tag{9}$$
$$\Bigg( - \frac{\partial^2 Q^{-1}_{1-\zeta}}{\partial \zeta^2} \sqrt{\sum_{t'=0}^{1} x^2_{t'} \sigma^2_{0,t'}} \Bigg) - \Bigg( - \frac{\partial Q^{-1}_{1-\zeta}}{\partial \zeta} \frac{x_0 \sigma^2_{0,0}}{\sqrt{\sum_{t'=0}^{1} x^2_{0,t'} \sigma^2_{0,t'}}} \Bigg)^2$$

It can be observed that $\Delta_5$ is only positive for specific values of allocation decisions and the variance. For instance, by assuming the variance $\sigma_0$ is greater than the variance $\sigma_1$, the second term will be greater than the first term, and thus $\Delta_5 < 0$. Accordingly, the Hessian matrix is neither positive nor negative semidefinite and hence the problem is non-convex.

## Appendix B

All the equations in Eq. 5.14 are linear and thus convex except the second constraint whose convexity is checked as follows

$$\begin{aligned} F(y, \beta) &= \sum_{\forall t \in \mathcal{T}} Q(y_{i,t}) - 1 + \beta \\ \nabla F(y, \beta) &= Q'(y_{i,t}) = -\frac{1}{\sqrt{2\pi}} e^{\frac{-y^2_{i,t}}{2}} \\ \nabla^2 F(y, \beta) &= Q''(y_{i,t}) = \frac{1}{\sqrt{2\pi}} y_{i,t} e^{\frac{-y^2_{i,t}}{2}}. \end{aligned} \tag{10}$$

Since we assume $\beta \geq 0.5$ for practical QoS levels, the constraint holds iff $\sum_{\forall t \in \mathcal{T}} Q(y_{i,t}) \leq 0.5$. This implies that $Q(y_{i,t}) \leq 0.5$ which occurs when $y_{i,t} \geq 0$. The Hessian matrix is a diagonal matrix of positive entries that represents its eigenvalues. Accordingly, the Hessian matrix is positive semidefinite and this proves the convexity of function.