DISTRIBUTED OPERATION OF UNCERTAIN DYNAMICAL CYBERPHYSICAL SYSTEMS

A Dissertation

by

RAHUL SINGH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	P.R. Kumar
Committee Members,	Jean-Francois Chamberland-Tremblay
	I-Hong Hou
	Srinivas Shakkottai
	R. Srikant
	Radu Stoleru
Head of Department,	Miroslav Begovic

December 2015

Major Subject: Computer Engineering

ABSTRACT

In this thesis we address challenging issues that are faced in the operation of important cyber-physical systems of great current interest. The two particular systems that we address are communication networks and the smart grid. Both systems feature distributed agents making decisions in dynamic uncertain environments. In communication networks, nodes need to decide which packets to transmit, while in the power grid individual generators and loads need to decide how much to produce or consume in a dynamic uncertain environment. The goal in both systems, which also holds for other cyber-physical systems, is to develop distributed policies that perform efficiently in uncertain dynamically changing environments. This thesis proposes an approach of employing duality theory on dynamic stochastic systems in such a way as to develop such distributed operating policies for cyber-physical systems.

In the first half of the thesis we examine communication networks. Many cyber-physical systems, e.g., sensor networks, mobile ad-hoc networks, or networked control systems, involve transmitting data over multiple-hops of a communication network. These networks can be unreliable, for example due to the unreliability of the wireless medium. However, real-time applications in cyber-physical systems often require that requisite amounts of data be delivered in a timely manner so that it can be utilized for safely controlling physical processes. Data packets may need to be delivered within their deadlines or at regular intervals without large gaps in packet deliveries when carrying sensor readings. How such packets with deadlines can be scheduled over networks is a major challenge for cyber-physical systems.

We develop a framework for routing and scheduling such data packets in a

multi-hop network. This framework employs duality theory in such a way that actions of nodes get decoupled, and results in efficient decentralized policies for routing and scheduling such multi-hop communication networks. A key feature of the scheduling policy derived in this work is that the scheduling decisions regarding packets can be made in a fully distributed fashion. A decision regarding the scheduling of an individual packet depend only on the age and location of the packet, and does not require sharing of the queue lengths at various nodes.

We examine in more detail a network in which multiple clients stream video packets over shared wireless networks. We are able to derive simple policies of threshold type which maximize the combined QoE of the users.

We turn to another important cyber-physical system of great current interest – the emerging smarter grid for electrical power. We address some fundamental problems that arise when attempting to increase the utilization of renewable energy sources. A major challenge is that renewable energy sources are unpredictable in their availability. Utilizing them requires adaptation of demand to their uncertain availability. We address the problem faced by the system operator of coordinating sources of power and loads to balance stochastically time varying supply and demand while maximizing the total utilities of all agents in the system. We develop policies for the system operator that is charged with coordinating such distributed entities through a notion of price. We analyze some models for such systems and employ a combination of duality theory and analysis of stochastic dynamic systems to develop policies that maximize the total utility function of all the agents.

We also address the issue of how the size of energy storage facilities should scale with respect to the stochastic behavior of renewables in order to mitigate the unreliability of renewable energy sources.

TABLE OF CONTENTS

Page

ABSTRACT ii

TABLE OF CONTENTS iv

LIST OF FIGURES vii

1. INTRODUCTION 1

2. DEADLINE CONSTRAINED MULTI-HOP NETWORKS 6

2.1 Introduction 6
2.2 Motivation 6
2.3 Summary 7
2.4 System Model 8
2.5 Previous Works 12
2.6 Characterizing the Rate Region 13
2.6.1 Constrained MDP Formulation 13
2.7 The Dual MDP 15
2.8 Single Packet Transportation Problem 17
2.9 A Decomposition Result 18
2.10 Obtaining the Optimal Policy $\pi(\lambda)$ 20
2.11 Convergence to the Optimal Prices $\lambda^\star$ 21
2.12 Wireless Fading 22
2.13 Jointly Serving Real-Time and Non-Real-Time Data 23
2.14 Constraints on Number of Links and Wireless Interference 23

3. MAXWEIGHT SCHEDULING: ASYMPTOTIC BEHAVIOR OF UNSCALED QUEUE-DIFFERENTIALS IN HEAVY TRAFFIC 27

3.1 Overview 27
3.2 Introduction 27
3.2.1 Basic Notation 31
3.3 System Model 32
3.4 MaxWeight Scheduling Scheme. Heavy Traffic Regime 33
3.4.1 MaxWeight Definition 33
3.4.2 Heavy Traffic Regime 34
3.5 Main Results 34

3.6 Complete Resource Pooling Condition . . . 35
3.7 Background on General-State-Space Discrete-Time Markov Chains . . 36
3.8 Queue Length Process . . . 39
3.9 Steady-State Queue Lengths Deviations from $\boldsymbol{\nu}$ . . . 46
3.10 Limit of the Queue-Differential Process . . . 47
3.11 Generalization to the Case When CRP Condition Does Not Necessarily Hold . . . 53

4. OPTIMIZING QUALITY OF EXPERIENCE OF DYNAMIC VIDEO STREAMING OVER FADING WIRELESS NETWORKS . . . 55

4.1 Overview . . . 55
4.2 Introduction . . . 55
4.3 System Description . . . 56
4.4 Problem Formulation . . . 58
4.5 The Dual MDP . . . 59
4.6 Single Client Problem . . . 60
4.7 Threshold Structure of the Optimal Policy for the Single Client Problem 62
4.8 Solution of Primal MDP . . . 68
4.8.1 Obtaining $\lambda_E^\star$ Iteratively in a Decentralized Fashion . . . 70
4.9 Fading Channels . . . 70

5. INDEX POLICIES FOR OPTIMAL MEAN-VARIANCE TRADE-OFF OF INTER-DELIVERY TIMES IN SINGLE-HOP NETWORKS* . . . 72

5.1 Overview . . . 72
5.2 Introduction . . . 73
5.3 Related Works . . . 74
5.4 System Model . . . 75
5.5 Markov Decision Process Formulation . . . 75
5.6 Whittle Index . . . 77
5.7 The Client Scheduling Problem is Indexable . . . 78
5.8 Bounds on Optimal Reward . . . 91
5.9 Optimality of Index Policy . . . 93
5.10 Simulations . . . 94
5.11 Concluding Remarks . . . 94

6. THE ISO PROBLEM: DECENTRALIZED STOCHASTIC CONTROL VIA BIDDING SCHEMES . . . 97

6.1 Overview . . . 97
6.2 Notation . . . 98
6.3 Introduction . . . 98
6.4 System Model . . . 101

6.5 The ISO Problem . . . 104
6.6 Common and Private Observations . . . 104
6.7 Illustrative Examples . . . 105
6.8 Fundamental Issues . . . 106
6.8.1 Interdependence/ Interconnection of Agents . . . 107
6.8.2 Privacy . . . 107
6.8.3 Decentralized Control with Non-Classical Information Structure . . . 108
6.8.4 Information Sharing/ Signaling . . . 108
6.8.5 Dynamic Market . . . 109
6.8.6 Online Optimization . . . 109
6.8.7 Curse of Dimensionality . . . 109
6.8.8 Big Data: Sufficient Statistics . . . 110
6.9 Problem Statements, Key Questions and Goals . . . 111
6.10 Related Works . . . 113
6.11 Dynamic Programming Approach . . . 116
6.12 A Tree Visualization of System Randomness . . . 117
6.13 Iterative Bidding Schemes . . . 117
6.14 The Deterministic Case . . . 120
6.15 Commonly Observed Noise . . . 122
6.16 Privately Observed Noise . . . 126
6.17 Using Learning Techniques to Eliminate Complexity of $\mathcal{L}(N(t))$ . . . 130
6.18 The Case of Linear Systems . . . 131
6.19 Concluding Remarks . . . 136

7. ON STORAGE AND RENEWABLES: A THEORY OF SIZING AND UNCERTAINTY . . . 138

7.1 Introduction . . . 138
7.2 System Model . . . 140
7.3 Random Walk Model . . . 141
7.4 Reflected Brownian Motion Model . . . 144
7.5 Correlated Uncertainty Between Loads and Renewables . . . 148
7.5.1 Performance Analysis . . . 150
7.6 Conclusions . . . 150

8. CONCLUSION . . . 152

REFERENCES . . . 155

LIST OF FIGURES

FIGURE Page

2.1 Multi-hop sensor network serving a single flow with source node s, and destination node d. Each directed edge corresponds to a link. . . 9

5.1 Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, \theta_2 = 3, R_2 = 1$, p_2 varying from $.1$ to 1. 95

5.2 Reward Optimal Policy vs.Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, p_2 = .6, R_2 = 1$ while θ_2 varies from 1 to 10. 95

5.3 Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 5, R_1 = 5, p_2 = .6, \theta_2 = 5$ while R_2 is varied. 96

6.1 Agents submit bids via Agent $\rightarrow$ ISO, while the ISO sends price-signals for the remaining time horizon through ISO $\rightarrow$ Agent 120

6.2 A Tree based visualization of randomness for a 2 agent system evolving over 2 bid times. The noise values are allowed to be binary and assume the values 0 and 1. 136

6.3 Flowchart depicting the decision flow in Algorithm 1. 137

7.1 Renewable and fossil fuel energy consolidated into a microgrid. . . . 142

1. INTRODUCTION

We address important problems in cyberphysical systems in two areas, the operation of communication networks and the operation of the smart electricity grid. Both systems are examples of cyberphysical systems that are often large and distributed, and important to operate efficiently. A major theme of this thesis is the development of operating policies that perform efficiently, and that can be implemented in a decentralized manner with tractable computation.

The first half of the thesis addresses communication networks. Real-time applications utilizing multi-hop networks, e.g., networked control systems, vehicular networks, video streaming applications, sensor networks etc., require the data packets to be delivered to the application in a timely manner. In many applications the information carried by packets is time-critical and the end-to-end delay constraints need to be respected. In sensor-actuator networks, a regular stream of packets carrying sensory information needs to be delivered to their destinations.

However, transmitting data packets over multi-hop networks poses several challenges. Any policy has to address several important issues while making scheduling decisions.

First, for wireless networks, a policy necessarily has to take into account the unreliable nature of the radio medium. It may make multiple transmission attempts before a packet gets delivered across a link. Since the data transmission rates depend upon transmission power levels, and the network suffers from wireless interference, the choice of transmission power level for a single flow or user affects the data transmission rates of the other users. This introduces dependencies amongst the scheduling choices for different packets, and so nodes in the network

need to coordinate in order to decide the transmit power levels. This requires, generally, a central coordinator which has access to the complete system state including the states of all nodes across the entire network and the states of all the packets at them, and that utilizes them in order to make scheduling decisions.

Furthermore, the wireless nodes have to respect power constraints, which means that the usage of the energy resource should be carefully optimized. One specific consequence is that power allotment across various flows or packets needs to be done on the basis of a packet's age. This key aspect is missing in the multi-hop scheduling policies which make choices based only on queue lengths, and not a packet's age.

Additionally, in the presence of wireless channel fading the knowledge of the prevailing channel conditions can be exploited to schedule data packets opportunistically, and thereby operating the network in an energy efficient manner.

The present state-of-the art multi-hop scheduling policies, e.g., the backpressure policy, require a centralized controller to schedule packet transmissions, which is a major limitation. Additionally, since they are designed to maximize throughput, they may perform poorly with regard to end-to-end delay. Also, they do not consider a packet's age in the system, though some modifications consider a packet's age at a node [55].

In Section 1, we address the above mentioned issues by constructing a multi-hop scheduling policy that is completely decentralized. This eliminates the requirement of a central coordinator in order to make scheduling choices. The policy also supports delay guarantees with respect to delivered packets.

Our key observation is to pose the problem of designing an efficient policy as a constrained Markov decision process (CMDP) that involves finding the optimal decision variables to be applied to an *individual packet*, and not a flow. Thus, the

state of the system is taken to be the age of each packet present at each node. This is in contrast to the technique of letting the system state be described by the vector comprising queue lengths of different flows at various nodes.

Thereafter, we look at the Lagrangian corresponding to the CMDP. The Lagrange multipliers associated with the nodal power constraints have the effect of decoupling the decision variables across different packets. These multipliers can be interpreted as the prices charged by wireless nodes for utilizing their transmission power.

This decoupling allows the computation of the Lagrangian to be carried out in a decentralized fashion. The resulting optimal policy has the form that each node schedules packets transmissions based on the knowledge of their age. An online learning technique can be utilized to learn the unknown network parameters. The policy thus derived ensures that the network transports the maximum number of packets per unit time subject to the constraint that the delivered packets have a delay that is bounded by a threshold which is tunable by the users.

In Sections 2-5, we turn our attention to sensor networks. In Section 2 we consider the problem of delivering a steady stream of sensory measurements, which is important for applications such as networked control or sensor-actuator networks where control loops are closed around sensor measurements. We define a notion of *service smoothness* for a scheduling policy. The service process associated with a policy is smooth if each flow in the network receives data packets in a *non-bursty* manner, or equivalently, there are no large time gaps in between two consecutive packet deliveries. The property of service smoothness is crucial for safety critical cyberphysical systems, since packet starvation would mean that the physical process evolves open loop, possibly making it unstable. We analyze the MaxWeight or Backpressure policy in Section 2, and show that the service process associated with

it is asymptotically smooth in the limit when the network load is close to 1.

Section 3 studies an important specific application - video streaming networks. The goal in such systems is to schedule video packets over a possibly unreliable wireless network so as to maximize the Quality of Experience (QoE) of the end-users. Each user maintains a buffer that contains video packets of different qualities. Video packets are fetched from this buffer, and played for a fixed time duration. An outage is said to occur if the user finds that its buffer is empty, since the video stream is then interrupted. We judge the performance of a policy by several criteria a) the average number of outages, b) how often a new outage occurs, c) average numbers of packets of different quality that are received. We construct decentralized scheduling policies that have an easily implementable threshold structure.

In Section 4 we treat the special case of a single-hop wireless network shared by multiple flows, and examine the optimal trade-off between the two conflicting objectives of maximizing packet throughput and service smoothness. We derive the optimal policy in a closed form expression by solving the dynamic programming optimality equation.

In the second half of the thesis beginning with Section 5, we turn our attention to problems in the emerging smart grid. We address issues that are at the core of efficient operation as we seek to enhance the usage of renewable energy sources such as photovoltaics or wind in place of fossil fuels. The fundamental challenge is that such renewable sources are unreliable, so demand must adapt to supply, called "demand response," rather than the other way around.

The first issue we address is : how can a system operator ensure efficient coordination of multiple generators and loads when both have their own dynamics and possible unreliabilities. Efficient operation of the smart grid entails always

balancing uncertain demand against supply of uncertain power. On the generator side, electric power generation needs to be carried out at the minimum possible cost. On the demand side, consumer needs can be modeled by utility functions of power consumption, and the system operator must maximize the overall utility of all agents in the system. A significant constraint is that the states, models and utilities of the agents cannot be assumed to be known to the system operator. We analyze the problem of coordination through the announcement of dynamic prices for energy. We examine how the system operator can determine clearing prices, and under what conditions optimal coordination is possible through their announcement. An important role is played by the structure of uncertainty. In some cases, there may be no uncertainty at all affecting any of the agents, in others there may be common uncertainty affecting all agents, and in yet others each agent may have private uncertainty. We examine under what conditions the system operator can attain optimality through price-based coordination in each case. We have shown that multiple linear quadratic Gaussian systems can be very efficiently coordinated even when the agents have private uncertainties. It is noteworthy that for this special case of systems of much interest in control systems, there is no need to share the uncertainty "tree", unlike in the general case of privately observed information.

2. DEADLINE CONSTRAINED MULTI-HOP NETWORKS

2.1 Introduction

We consider multi-hop networks serving multiple flows in which packets have to meet hard end-to-end deadline constraints, i.e., if a packet is not delivered to its destination node by its deadline, it is dropped from the network. We address the design of policies for routing and scheduling data packets in the network so as to optimize the network throughput and delay in an energy-efficient manner. The derived policies are highly decentralized in that the decisions regarding a data packet can be based solely on the knowledge of the age of the packet, thus eliminating the need to share the knowledge of the network topology, or queue lengths amongst the nodes. Global coordination is achieved through a notion of "price" for resource usage.

Applications include, but are not limited to, sensor networks, mobile ad-hoc networks [5], video-streaming, and other real-time applications. In sensor-actuator networks or cyber-physical systems, for example, sensors are deployed to sense time-critical processes at the source and send the measurements to a controller at the destination.

2.2 Motivation

Applications such as cyber-physical systems where control-loops are closed over networks and system stability are sensitive to delays. Similarly, the Quality of Service (QoS) requirements for real-time applications such as video streaming, VoIP, real-time surveillance, sensor networks, mobile ad-hoc networks (MANETS), in-vehicular networks etc., entail that the utility of a data packet delivered to its destination depends critically on its *age* [5]. Traditional information theory how-

ever does not consider the age of data [38]. The above objective may also have to be achieved in an energy efficient manner, possibly utilizing rate-adaptation schemes [95]. An important design requirement in multi-hop networks is that data packets need to be scheduled in a decentralized manner due to the absence of a centralized scheduler. Our main contribution is the design of routing and scheduling policies which provide hard-deadline guarantees on packet deliveries at minimum energy expenditure, maximize the network throughput, and additionally are also highly decentralized. In the designed policies, a node can take decisions only on the basis of the age of the packets present with it. This vastly simplifies the network operation when compared to policies in which a node requires the knowledge of queue lengths at its neighboring nodes.

2.3 Summary

We consider multi-hop, multi-flow networks in which a packet is discarded if its age exceeds a certain threshold τ; the results also extend to situations where the specification of the threshold is allowed to vary from packet to packet. Since the wireless channel is unreliable, the outcome of packet transmissions is modeled as a random process. Each node can transmit multiple packets on its out-links and can also transmit and receive packets simultaneously. Nodes can carry out packet transmissions at varying power levels, enabling rate-adaptation techniques. Each node has an average power constraint. The throughput of a flow is defined as the average number of packets meeting the deadline constraint delivered to its destination node per unit time. Our goal is to design decentralized scheduling policies that maximize the total throughput of the network.

Our approach is as follows. We invoke the scalarization principle [49] and pose the problem of maximizing the network throughput subject to nodal power

constraints as a constrained Markov Decision Process (MDP) [4]. We then solve the problem via considering the Lagrangian dual of this MDP. The Lagrange multipliers associated with the power constraints are interpreted as prices for transmitting packets, and the resulting MDP decomposes conveniently into a "unit-packet unit flow" MDP (Section 2.8). It is easily solved, and, importantly, presents a completely decentralized solution, where a node only needs to know the remaining lifetime till deadline, or equivalently the age, of each packet that is present at the node. The introduction of Lagrange multipliers, specifically prices for utilizing power for packet transmissions, gives rise to a tractable and easy to implement policy. These Lagrange multipliers are shown to be computable in a decentralized online fashion.

One can interpret this approach as asymptotically optimal in the same sense as Whittle's indexability [111] approach is asymptotically optimal as the population of bandits increases in proportion [109]. Finally, we also show how our policy is closely connected to the technique of reinforcement learning that is used in Online Machine Learning.

2.4 System Model

We consider networks in which the data-packets have a hard deadline constraint on the time by which they should be delivered to their destination nodes. The communication network of interest is described by a directed graph $G = (V, \mathcal{E})$ as shown in Figure 2.1, where $V = \{1, 2, \ldots, |V|\}$ is the set of nodes that are connected via communication links. A directed edge $i \to j \in \mathcal{E}$ signifies that node i can transmit data packets to node j. For simplicity of exposition, we will neglect contention for the transmission medium, though the results can be extended in appropriate ways as described below. We assume that time is discrete, and evolves over slots numbered $1, 2, \ldots$. One time-slot is the time taken to attempt a packet

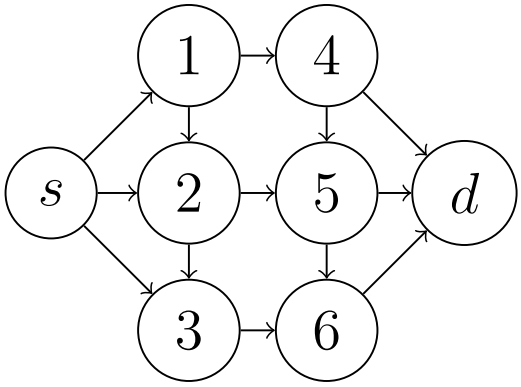

Figure 2.1: Multi-hop sensor network serving a single flow with source node s, and destination node d. Each directed edge corresponds to a link.

transmission over any link in the network. The network is shared by F flows $f_1, f_2, \ldots, f_F$.

The link between any two network nodes is allowed to be random, which enables us to model unreliable wireless channels. If a packet transmission occurs on the link l, then the transmission is successful with probability p_l. We can model the phenomena of wireless fading by allowing the success probability to be a function of time, i.e., probability is $= p_l(t)$. The probability $p_l(t)$ can be assumed to be governed by a finite-state Markov process, whose state is known at the transmitting node. We can also incorporate transmit power control by allowing the success probability $p_l(t, E)$, to depend on the transmission power E.

Each communication node i has an average power constraint P_i. If packet transmissions on link l at time t use $E_l(t)$ units of power, then the nodal power constraints are given by,

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\left\{\sum_{l:l=i\to j} E_l(t)\right\}\right) \leq P_i, \forall i \in \{1, 2, \ldots, V\}. \tag{2.1}$$

(2.2)

We allow a node to transmit and receive packets simultaneously over several outgoing links, employing various techniques such as TDMA, OFDMA, CDMA etc. [53, 70, 93]. The power constraints on the wireless nodes induce constraints on communication rates [38].

Each packet that is generated by the network has a "relative-deadline", or "allowable delay threshold". If the packet is not delivered to its destination within this deadline, it is dropped from the network and will never again be transmitted in any future time-slot. More precisely, if a packet has a relative-deadline of d time-slots, and is generated at the beginning of time-slot τ, then either it is delivered to its destination node by time-slot $\tau + d$, or it is discarded from the network.

We enforce an assumption that, with probability 1, the relative deadline of any packet is bounded by a quantity Δ. Packet arrivals, and their relative deadlines (allowable delay) are governed by a finite-state Markov process. It turns out that the policy does not need to observe this process, or know its statistics. We will make the following assumption: for each packet that is present in the network at any given time t, the time-till-deadline of the packet is known to the node at which it is present.

Our analysis can be extended in a straightforward manner to consider the case when the relative-deadline of a packet is an arbitrary (adapted) stochastic process, and becomes known to the network as soon as the packet is generated. If the relative-deadline can be chosen to be an adapted stochastic process, then some interesting models are possible. For example, suppose the context is video-streaming where there is a frame buffer at the receiver. Then the relative deadline can be taken to equal to the "remaining play time" left in the frame buffer, since we don't want the buffer emptied. In that case Relative Deadline $=-$ (Elapsed time since the last time that Destination Buffer was empty, i.e., the current age of the "busy

epoch") + (Number of packets that arrived at the Destination since then)$\times$ (Time to play one packet). Note that in this case the deadline process depends on the policy being used.

The throughput attained by a flow f under a policy is the average number of packets received per unit time, i.e.,

$$\liminf_{T\to\infty} \mathbb{E}\left(\frac{\sum_{t=1}^{T} d_f(t)}{T}\right), \tag{2.3}$$

where the random variable $d_f(t)$ is equal to 1 if a packet of flow f is delivered to its destination at time t, and 0 otherwise, with the expectation taken under the policy being applied.

A throughput vector $\boldsymbol{\alpha}$ that can be achieved via some scheduling policy will be called an "achievable throughput vector". The set of all achievable throughput vectors constitutes the rate-region, and a scheduling policy that achieves the complete rate-region is said to be throughput-optimal. Thus, under the application of a throughput-optimal policy π, the network can achieve any throughput vector that can be attained by some other scheduling policy.

Note that all the above definitions depend on the manner in which the relative-deadlines of the packets are decided. The rate-region thus depends on the process that decides the relative-deadlines, and thus we might call such networks "deadline-constrained networks".

Vectors will be in bold font, and by $\mathbb{R}_+^N$ we refer to N dimensional vectors which are non-negative component wise.

2.5 Previous Works

Hou et. al [50] have proposed a network model in which multiple flows share a single-hop network, and all the packets across every flow have the same relative deadline. There have been many extensions of this line of work. References [43, 59] consider a similar one-hop network model and characterize the throughput maximizing policy.

Reference [62] considers the challenging problem of scheduling deadline-constrained packets over a multi-hop network, but the proposed policies are not shown to have any provable guarantees on the resulting throughput. To the best of the author's knowledge, [71] is the only work which provides a provable sub-optimal policy for deadline-constrained networks, though it only concerns wired networks. The policies proposed in [71] guarantee only a fraction

$$\frac{1}{\text{length of the longest route in network}}$$

of the maximum possible throughput, i.e., only a small fraction of the capacity region.

Scheduling policies designed for multi-hop networks, e.g., the back-pressure policy [102], are guaranteed to be throughput optimal, i.e., they can stabilize the data queues in the network under any arrival rate vector for which there exists some network policy that can stabilize the network. However they can perform poorly with regard to delay performance [25, 42, 64, 121]. Any optimal scheduling policy needs to take into account how much time each packet has spent in the network, and the channel reliabilities of the links that the packets have to traverse in order to reach the destination node. The backpressure policy schedules packets only on the basis of queue-lengths of nodes. This is one key reason why it can

result in a high end-to-end delay.

2.6 Characterizing the Rate Region

The rate-region of the network (defined in Section 2.4) will be denoted by $\mathbf{\Lambda}$. We note that in order to characterize the set $\mathbf{\Lambda}$, it is sufficient to characterize the set of Pareto-optimal vectors $\alpha \in \mathbf{\Lambda}$, defined as

$$\left\{ \alpha : \alpha \text{is a throughput vector and } \exists \beta \in \mathbb{R}_+^N \text{ such that } \alpha \in \arg\max_{y \in \Lambda} \sum_f \alpha_f \beta_f \right\},$$

since $\mathbf{\Lambda}$ is simply its closed convex hull. The problem of obtaining the set $\mathbf{\Lambda}$ is equivalent to that of finding scheduling policies which maximize a non-negatively weighted sum of throughputs.

2.6.1 Constrained MDP Formulation

The problem of maximizing a non-negatively weighted sum of throughputs subject to rate-constraints can be posed as a Constrained Markov Decision Process (CMDP) [3]. The state of an individual packet present in the network at time t is described by the flow f to which it belongs, and the two tuple (i, s), where i is the node at which it is present, and s is the time-to-go till its deadline. The state of the network at time t is then described by specifying the state of each packet present in the network at time t.

Since the number of packets in the network at any time is bounded, assuming a bounded number of arrivals in any time slot, the system state $X(t)$ takes on finitely many values. A scheduling policy π has to choose, for each time t, at each node, which packets to transmit, from the set of packets available to it. Moreover, it has to choose which link and at what power these packets should be transmitted. The choice made at time t will be denoted $U(t)$.

Since the probability distribution of the system state $X(t+1)$ at time $t+1$ depends only on the system state $X(t)$ at time t and the action $U(t)$ chosen at time t, the problem of maximizing the throughput subject to node-capacity constraints (2.1) is a Constrained Markov Decision Process, where a reward of β_f is received when a packet of flow f is delivered to its destination. Thus a policy maximizing the network throughput solves the following optimization problem:

$$\max_{\pi} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{f}\sum_{t=1}^{T}\beta_f d_f(t)\right), \text{ such that}$$
$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\left\{\sum_{l:l=i\to j} E_l(t)\right\}\right) \leq P_i, \forall i \in \{1,2,\ldots,V\}, \tag{2.4}$$

where $d_f(t)$ is the number of packets of flow f delivered to their destination node at time t. We note that the above CMDP parameterized by the vector $\beta := (\beta_1, \ldots, \beta_F)$ is solved by a Stationary Randomized Policy [3]. Since the state-space of the network, and the number of link-capacity constraints (2.1) is finite, it follows that there is a finite set $\{\pi_1, \pi_2, \ldots, \pi_M\}$ of Stationary Randomized Policies such that for each value of β, there is a policy that belongs to this set and solves the CMDP (2.6) [3]. Let $\gamma_1, \gamma_2, \ldots, \gamma_M$ be the vectors of throughputs associated with the policies $\pi_1, \pi_2, \ldots, \pi_M$. We then have the following characterization of $\mathbf{\Lambda}$.

Lemma 1.

$$\mathbf{\Lambda} = \left\{\alpha : \alpha = \sum_{i=1}^{M}\gamma_i c_i, c_i \geq 0, \sum_i c_i \leq 1\right\},$$

where the c_i, s are scalars.

Note that the number of stationary Markov policies is exponentially large in the following parameter:

Maximum possible number of packets in the network $\times$ Maximum path length of the flows $\times$ Maximum possible relative deadline.

Hence using Lemma 1 to compute $\mathbf{\Lambda}$ is out of the question. Thus we seek to design low complexity decentralized scheduling policies that achieve the region $\mathbf{\Lambda}$.

Since we can restrict ourselves to stationary randomized policies, we can replace the $\limsup$ and $\liminf$ in the definition of CMDP to pose the problem as,

$$\max_{\pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{f}\sum_{t=1}^{T}\beta_f d_f(t)\right), \text{ such that}$$
$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\left\{\sum_{l:l=i\to j} E_l(t)\right\}\right) \leq P_i, \forall i \in \{1,2,\ldots,V\}. \tag{2.5}$$

(2.6)

2.7 The Dual MDP

Letting λ_i be the Lagrange multiplier associated with the power constraint on node i, and denoting $\lambda = \left(\lambda_1, \lambda_2, \ldots, \lambda_{|V|}\right)$, we can write the Lagrangian for the Primal MDP (2.6) as,

$$\mathcal{L}(\pi,\lambda) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{f}\sum_{t}\beta_f d_f(t)\right) - \sum_{i}\lambda_i\left(\mathbb{E}\left(\sum_{t=1}^{T}\left\{\sum_{l:l=i\to j} E_l(t)\right\}\right)\right)$$
$$+ \sum_{i}\lambda_i P_i, \tag{2.7}$$

where the expectation is w.r.t. the policy π that is being used, the random packet transmission outcomes, and the randomness of the process deciding packet arrivals and relative deadlines. Next we note that $E_l(t)$, the total power consumed on link l at time t, is the sum of power spent on transmitting individual packets. Thus if $E_{l,f,n}(t)$ is the amount of power spent on transmitting n-th packet of flow f at time

t on link l, we have,

$$E_l(t) = \sum_{f,n} E_{l,f,n}(t).$$

After some algebraic manipulation, the Lagrangian reduces to,

$$\mathcal{L}(\pi, \lambda) = \lim_{T\to\infty} \frac{1}{T} \sum_{f,n} \mathbb{E}\left(\sum_t \beta_f d_f(t) - \sum_i \lambda_i \left\{\sum_{l:l=i\to j} E_{l,f,n}(t)\right\}\right) + \sum_i \lambda_i P_i. \tag{2.8}$$

We note that for any fixed value of the vector λ, the Lagrangian is a sum of *transportation cost* terms,

$$\mathbb{E}\left(\sum_t \beta_f d_f(t) - \sum_i \lambda_i \left\{\sum_{l:l=i\to j} E_{l,f,n}(t)\right\}\right).$$

(Single Packet Transportation Cost)

This cost involves transporting a single packet from its source node to its destination, and is independent of the actions chosen for other packets in the network. It can be interpreted as incurring a payment of λ_i for using unit power at node i, and accruing a reward of β_f upon delivery of the packet to its destination. Thus, for designing the policy π for maximizing the Lagrangian, we can solve an unconstrained problem of minimizing the Single Packet Transportation Cost.

This yields us the dual function, defined as

$$D(\lambda) = \max_{\pi} \mathcal{L}(\pi, \lambda). \tag{2.9}$$

The Dual function can be obtained in a decentralized fashion, since the introduc-

tion of Lagrange multipliers has decomposed the primal problem into a collection of Single Packet Transportation problems, which are coupled through the node prices λ_i.

The next section analyzes this Single Packet Transportation Cost problem.

2.8 Single Packet Transportation Problem

Consider the single packet transmission cost expression. To make the discussion simple, let us consider the case when wireless fading is absent, i.e., the channel conditions are static. The probability that a packet transmission over link l at time t is successful is given by $p_l(E)$. Below, we omit the subscript f, and relabel the nodes so that the source and destination nodes are labeled as 1 and V respectively. Moreover the time at which the packet is generated is taken to be the time at which network operation begins.

The Single Packet Transportation Problem is described as follows: A single packet is generated at time $t = 0$ at source node $i = 1$, and, if it is not delivered to the destination node V by time $t = d$, then it is discarded from the network. The age of the packet, $X(t)$, evolves as,

$$X(t+1) = X(t) + 1, \text{ if } X(t) < d,$$

with the packet discarded if its age reaches d units. A price of λ_i per unit amount of power has to be paid for transmission over an outgoing link at node i. A reward of β units is paid once the packet gets delivered to the destination node V. Thus,

$$\text{cost } = \begin{cases} \lambda_i \text{ if attempted at link } l = (i,j) \\ 0 \text{ otherwise,} \end{cases}$$

while,

$$\text{reward } = \begin{cases} \beta \text{ if delivered at destination node } V \\ 0 \text{ otherwise.} \end{cases}$$

The single packet transportation problem is to choose the control $U(t)$ so as to minimize the Single Packet Transportation Cost

$$\min \mathbb{E}\left(\sum_t \beta d(t) - \sum_i \lambda_i \left\{\sum_{l:l=i\to j} E_l(t)\right\}\right),$$

(Single Packet Transportation Problem)

where $d(t)$ assumes the value 1 if the packet is delivered to node V at time t, and is 0 otherwise.

It is clear that the state of the packet is described by the two tuple (i, s), where i is the node at which it is present, and s is the time till deadline. Thus we can use Dynamic Programming to solve the problem. Let $V(\cdot)$ denote the value function for the above problem. Then the corresponding Bellman recursion is given by,

$$\begin{aligned} V(i,s) &= \max\{V(i,(s-1)^+), X\}, \text{ where} \\ X &= \max\{\lambda_i + p_{i\to j}(E)V(j,(s-1)^+) + (1-p_{i\to j}(E))\,V(i,(s-1)^+)\}. \end{aligned} \quad (2.10)$$

Solving for the maximizer on the r.h.s. yields the optimal action in the corresponding state.

2.9 A Decomposition Result

We note that the Single Packet Transportation Cost Problem was parametrized by the vector of node prices λ. Let us denote by $\pi_f(\lambda)$ the policy which solves

the Single Packet Transportation Cost Problem corresponding to the case when the packet belongs to flow f, and node prices are set at λ. Also let $\pi(\lambda)$ be the policy that implements the policy $\pi_f(\lambda)$ for each packet belonging to flow f. That is,

$$\mathcal{L}(\pi(\lambda), \lambda) = D(\lambda). \tag{2.11}$$

The constrained optimization problem (2.6) can equivalently be posed as a linear program, in which the variables to be optimized are the *occupation measures* induced by a policy π on the joint state-action space [2–4, 20].

Being a linear program, the duality gap corresponding to (2.6), and its dual problem, defined as,

$$\max_{\lambda} D(\lambda). \tag{2.12}$$

is zero. Let $\lambda^\star$ be the price vector that solves the Dual Problem (2.12). It then follows from (2.11), that the policy $\pi(\lambda^\star)$ solves (2.6) that was posed in its primal form. We thus obtain:

Theorem 1 (Decomposition Result)**.** *The optimal policy for* (2.6) *is fully decentralized: In order for each node i to make a decision regarding a packet present with it at any time t, the node needs to know the flow f that the packet belongs to, and the age of the packet. Once this is known, the node v implements the policy $\pi_f(\lambda^\star)$.*

We may observe the following key features of the analysis that was carried out. In order to solve the Primal Problem (2.6), the network is required to make decisions $U(t)$ based on the knowledge of the network state $X(t)$. The size of the state-space in which the network state $X(t)$ resides is exponential in the quantity:

Number of packets ($\leq M$) $\times$ Deadline threshold bound (Δ) $\times$ Distance between nodes, ($\leq |V|$). Thus an approach based on directly solving the Primal version of (2.6) would have been computationally futile. Moreover, a central coordinator is needed in order to implement the optimal action $U(t)$ as a function of the network state $X(t)$.

These serious limitations have led us to instead consider the Dual Problem (2.12). The introduction of the nodal prices λ_i has the effect of reducing the computational complexity from exponential to linear in the quantity $M \times \Delta \times |V|$. Moreover the obtained solution is highly decentralized.

2.10 Obtaining the Optimal Policy $\pi(\lambda)$

The Bellman recursions (2.10) require the nodes to know the network parameters, i.e., the vector λ, and link reliability functions $p_l(\cdot)$. Let us fix the nodal power prices to be λ for the time being, and try to solve for the policy $\pi(\lambda)$ in a decentralized fashion. If we assume that network parameters are not known by the nodes, then we can use techniques from the field of Online Learning in order to simultaneously learn the network parameters, and adaptively control the network performance via an explore-exploit strategy such as *reinforcement learning*, or learning through delayed rewards [108]. Let a_n be a positive sequence that satisfies $\sum_n a_n = \infty, \sum_n a_n^2 < \infty$. The iterations, indexed by parameter n, are then given by,

$$\begin{aligned} V_{n+1}(i,s) &= 1_n(i,s)\left\{V_{n+1}(i,s)\,(1-a_n) + a_n \max\{V_n(i,(s-1)^+), X\}\right\} \\ &\quad + (1 - 1_n(i,s))\, V_{n+1}(i,s), \text{ where} \\ X &= \max\{\lambda_i + p_{i\to j}(E) V_n(j,(s-1)^+) + (1-p_{i\to j}(E))\, V_n(i,(s-1)^+)\}, \end{aligned} \tag{2.13}$$

which can be viewed as a noisy version of the corresponding Value Iteration algorithm [84], and is thus a stochastic approximation-based scheme [90]. In the above, $1_n(i,s)$ assumes the value 1 if the packet-state at iteration n is (i,s). The iterations converge to the true value function $V(\cdot)$, and thus yield the optimal policy. Note that in performing the above iterations, at iteration step n, a node needs to know the value function $V_n(\cdot)$ of its neighboring nodes. This is not a restriction since the iterations will be performed only at the commencement of network operation. Once the iterations converge, and the optimal policy is obtained, the policy can be implemented in a local fashion.

2.11 Convergence to the Optimal Prices $\lambda^\star$

If we assume that each node i in the network has knowledge of the network parameters, then, it can obtain the optimal policy $\pi(\lambda)$ as a function of the node prices λ. Since the Dual Problem is convex, each node $i \in V$ can use the gradient-descent method in order to solve for the optimal price vector $\lambda^\star$, and implement the optimal policy $\pi(\lambda^\star)$.

Let us now assume that the nodes do not know the network parameters. In the previous section we could use learning-based techniques in order to solve for the optimal policy $\pi(\lambda)$ in a decentralized fashion. Now, in addition to learning optimal policy, we will also "learn" the optimal nodal prices $\lambda^\star$ in a decentralized manner.

One way to achieve this task would be to perform the Value Iterations using reinforcement learning for each price vector λ until convergence, and then update the price λ using gradient descent method. Since the gradient $\frac{\partial \mathcal{L}}{\partial \lambda_i}$ evaluated at the

policy $\pi(\lambda)$ is $= P_i - \bar{P}_i(\pi(\lambda))$, we have,

$$\lambda_{n+1} = \lambda_n(1 - b_n) + b_n(P - \bar{P}(\pi(\lambda))), \tag{2.14}$$

$$\tag{2.15}$$

where $P = (P_1, P_2, \ldots, P_{|V|})$ is the vector consisting of nodal power bounds P_i, and $\bar{P}_i(\pi(\lambda))$ is the average power utilization at node i under the application of policy $\pi(\lambda)$. The iterations converge to the optimal prices $\lambda^\star$.

However this operation is highly inefficient since a lot of time is spent in evaluating the optimal policy $\pi(\lambda)$ for values of prices λ that are "far away" from the optimal prices. An alternative method to achieve the above is via a two time-scale stochastic approximation scheme, which involves Reinforcement Learning and Price Update (2.14) iterations simultaneously by coupling the two. However these two updates occur at two different time-scales, signified by adaptation gains a_n and b_n. The Price Update iterations are to be performed at a much slower time-scale since for a fixed value of λ; the Reinforcement Learning iterations should have nearly converged to yield the optimal $\pi(\lambda)$.

2.12 Wireless Fading

Our model allows us to incorporate wireless fading by letting the link transmission probabilities be a function of time t, as is the energy E, i.e., $p_l(t, E)$. We model the channel conditions as a finite-state Markov process $Y(t)$, with the probabilities $p_l(t, \cdot)$ a function of the channel condition $Y(t)$.

The network state is then described by a) the state of each packet, and b) the channel condition $Y(t)$. The derivation of the optimal policy is carried out along similar lines as before, after having augmented the system state by having

appended the channel state $Y(t)$. We note that now the optimal policy would be of the following form: the decision to be taken by a node i at time t will depend on the state of each packet present in the network, and the channel state $Y(t)$. Thus a central coordinator needs to communicate the prevailing channel conditions to all the nodes.

A simplification is possible if we assume that the process $Y(t)$ is i.i.d., which would eliminate the need for communicating $Y(t)$. Under assumption of block fading [89, 105], the channel state would only need to be communicated periodically.

2.13 Jointly Serving Real-Time and Non-Real-Time Data

In the previous sections we assumed that the utility of a packet that had exceeded its deadline was zero. Instead, we could have allowed its utility to be described by a function $f(t, d)$, where d is the age of the packet when it was delivered to the destination. As an example, if τ is the delay threshold, a soft penalty of $(((d-\tau)^+)^2, \exp^{(d-\tau)^+})$ can be imposed. Such a consideration will enable us to support differential Quality of Experience (QoE) to flows based on whether they serve real-time or non real-time data [96].

2.14 Constraints on Number of Links and Wireless Interference

We have assumed in the preceding sections that the nodes can transmit/receive packets simultaneously on multiple channels. This assumption has the following positive aspects: a) it enables the network to fully exploit the resource sharing techniques available to it, e.g., CDMA, OFDM, TDMA etc. b) it also allows the network to utilize the battery power available to it in a time-slot when it requires it the most, for example when a node has an excess of packets in a state in which it is desirable for them to be transmitted, c) this assumption vastly simplifies the

construction of the optimal policy, which also turns out to be highly decentralized.

It should however be noted that this may not be a realistic assumption, so that we may need to impose constraints of the following form: a) *link usage constraint*, i.e., a limit on the total number of links that can be used by a node simultaneously at any given time t, or, b) *power interference constraint :* one might allow for a more general model in which the amount of wireless link interference depends upon the transmit power level used at the set of links, so that $\vec{R}(t) := (R_1(t), R_2(t), \ldots, R_{|\mathcal{E}|}(t))$, the vector of instantaneous data-rate at time t of the combined set of links, is a function of the instantaneous power transmission level $\vec{P}(t) := (P_1(t), P_2(t), \ldots, P_{|\mathcal{E}|}(t))$, i.e.,

$$\vec{R}(t) = f(\vec{P}(t)).$$

Furthermore, in order to accommodate channel fading into the above model, we can let the function $f(\cdot)$ be determined by the prevailing channel conditions. They could, for example, be assumed to be governed by a Markov process evolving on a finite set of states. This model would allow us to treat the wireless SINR model [6] that is employed widely while analyzing multi hop wireless networks.

In summary, the average power constraint considered in this section gives rise to a very simple model in which the constraints on number of links at the individual nodes' disposal, and/or wireless power interference are "smoothened/relaxed". This smoothening has allowed us to treat the core issue of allocating the resources amongst the data packets in an optimal fashion. The model of average power constraints was amenable to a treatment that showed us how a very simple decentralized policy could transfer data optimally if the network were willing to invest in resource sharing techniques mentioned above (e.g., TDMA, CDMA, OFDM etc.).

Next, we shred some light on treating the *link-usage* and *power interference* constraints.

We would like to mention that the analysis performed in this section can be extended in order to derive policies under the link-usage and interference constraints. However, the resulting policy will not be decentralized, but rather requires the nodes to coordinate and agree on a) which packet to be transmitted at each link (link-usage constraint), or b) the values of the transmission power levels to be used at each link (power interference constraint). The approach to be utilized is based on the recurring theme of handling a constrained stochastic control problem via introducing additional auxiliary variables. Thereafter one could utilize a multiple time-scale learning approach [21, 58] in order to learn the "optimal variables".

Also worth mentioning is the fact that we can recover the optimal policy under link-usage constraints from the policy that was derived in this section under the average power constraints. This approach can be viewed as analogous to the treatment employed for the Multi-Armed Bandit Problem [40, 111], in which one relaxes the hard-constraint (on the total number of arms that can be pulled at any given time) by a corresponding soft-constraint (which removes the hard constraint, but instead constrains the average number of arms that can be pulled per unit time, i.e., a bound on $\frac{\sum_{t=1}^{T} N(t)}{T}$, where $N(t)$ is the number of arms pulled at time t), and recovers an optimal policy for the problem under a hard constraint in the limit as the total number of bandits becomes large, while the fraction of arms that can be pulled is kept constant. The reader can find a detailed discussion of MABP in [40, 111], while [109] shows that the policy recovered for the hard constrained problem, from the solution of the soft constrained problem, is optimal in the limit as the number of bandits becomes large while respecting the proportions of the various types of arms.

Now we show that we can treat the problem of designing an optimal scheduling policy as an MABP. First we identify an individual bandit in the MABP with a data-packet in our network. The decision regarding the choice of the bandit to be pulled at each time is analogous to the problem of choosing which data-packet to be scheduled for transmission. The hard constraint imposed on the number of links that can be used in each time-slot is identified with the hard-constraint on the number of arms that can be pulled in the MABP.

3. MAXWEIGHT SCHEDULING: ASYMPTOTIC BEHAVIOR OF UNSCALED QUEUE-DIFFERENTIALS IN HEAVY TRAFFIC

3.1 Overview

In this section we consider the problem of smoothness of packets delivered in a timely manner over a shared wireless medium to destination nodes. We consider the model of a "generalized switch" serving multiple traffic flows in discrete time. Our interest is in the MaxWeight algorithm [99], and so we suppose that the switch uses that algorithm to make a service decision (scheduling choice) at each time step, which determines the probability distribution of the amount of service that will be provided. We are primarily motivated by the following question: in the heavy traffic regime, when the switch load approaches critical level, will the service processes provided to each flow remain "smooth", i.e., without large gaps in service? Addressing this question reduces to the analysis of the asymptotic behavior of the unscaled queue-differential process in heavy traffic. We prove that the stationary regime of this process converges to that of a positive recurrent Markov chain, whose structure we explicitly describe. This in turn implies asymptotic "smoothness" of the service processes.

3.2 Introduction

Suppose we have a system in which several data traffic flows share a common transmission medium (or channel). Sharing means that in each time slot a scheduler chooses a transmission mode – the subset of flows to serve and corresponding transmission rates; the outcome of each transmission (the number of successfully delivered packets) is random. The scheduler has two key objectives: (a) the time-average (successful) transmission rate of each flow i has to be at least some $\lambda_i > 0$;

(b) the successful transmissions for each flow need to be spread out "smoothly" in time – without large time-gaps between successful transmissions. Such models arise, for example, when the goal is *timely delivery* of information over a shared wireless channel [50].

A very natural way to approach this problem is to treat the model as a queuing system, where services (transmissions) are controlled by a so called MaxWeight scheduler (see [34,99,102]), which serves a set of *virtual queues* (one for each traffic flow), each receiving new work at the rate λ_i. (See e.g., [68].) This automatically achieves objective (a), if its feasible at all; MaxWeight is known to be *throughput optimal* –it stabilizes the queues if that is feasible at all. The MaxWeight stability results, however, do not indicate whether or not the objective (b) is achieved. Specifically, when the system is heavily loaded, i.e. the vector $\boldsymbol{\lambda} = (\lambda_i)$ is within the system *rate region* $\boldsymbol{V}$, but close to its boundary, the steady-state queue lengths under MaxWeight are necessarily large, and it is conceivable that this may result in large time-gaps in service for individual flows. (Note that, if (a) and (b) are the objectives and the queues are virtual, the large queue lengths in themselves are not an issue. As long as (a) and (b) are achieved, minimizing the queue lengths is not important.) Our main results show that this is *not* the case. Namely, in the heavy traffic regime, when $\boldsymbol{\lambda} \to \boldsymbol{\lambda}^\star$, where $\boldsymbol{\lambda}^\star$ is a point on the outer boundary of rate region $\boldsymbol{V}$, the service process remains "smooth", in the sense that its stationary regime converges to that of a positive recurrent Markov chain, whose structure is given explicitly.

To obtain "clean" convergence results, we assume that the amount of new work arriving in the queues in each time slot is random and has a *continuous* distribution. (The amounts of service are random, but discrete.) Under this assumption, the state spaces of the processes that we consider are continuous. On one hand,

this makes the analysis more involved because the notion of positive recurrence is more involved for a continuous state space, as opposed to a countable one. But on the other hand, this makes all stationary distributions absolutely continuous w.r.t. the corresponding Lebesgue measure, making it easier to prove convergence. We emphasize that the assumption of continuous distribution of the arriving work is non-restrictive; if we create virtual queues, artificially, for the purpose of applying MaxWeight algorithm, the structure of the virtual arrival process is within our control.

This problem essentially reduces to analysis of stationary versions of the *queue-differential* process $\boldsymbol{Y}$, which is the projection of the (weighted) queue length process on the subspace $\boldsymbol{\nu}_{\perp}$, orthogonal to the outer *normal cone* $\boldsymbol{\nu}$ to the rate region $\boldsymbol{V}$ at the point $\boldsymbol{\lambda}^{\star}$. As we show, in the heavy-traffic limit, in steady-state, the values of the queue-differential process $\boldsymbol{Y}$ uniquely determine the decisions chosen by MaxWeight scheduler. Note that the process $\boldsymbol{Y}$ is obtained by projection only, without any scaling depending on the system load.

The model that we consider is essentially a "generalized switch" [99]. Some features of our model, namely random service outcome and continuous amounts of arriving work, as well as the objective (b), are motivated by applications such as timely delivery of packets of multiple flows over a shared wireless channel [50]. The model of [50] is a special case of ours; paper [50] introduces a *debt scheme* and proves that it achieves the throughput objective (a); the objective (b) is not considered in [50].

The analysis of MaxWeight stability has a long history, starting from the seminal paper [102], which introduced MaxWeight; heavy traffic analysis of the algorithm originated in [99]. (See, e.g., [34] for an extensive recent review of MaxWeight literature.)

The line of work most closely related to this section, is that in [34, 61, 62]. Paper [34] studies MaxWeight under a heavy traffic regime and under the additional assumption that the normal cone $\boldsymbol{\nu}$ is one-dimensional, i.e., it is a ray. (The latter assumption is usually referred to in the literature as *complete resource pooling* (CRP).) Paper [34] shows, in particular, the stationary distribution tightness of what we call the queue-differential process $\boldsymbol{Y}$ in heavy traffic. Part of our analysis shows the stationary distribution tightness of $\boldsymbol{Y}$ – it is analogous to that in [34] (and we also borrow a lot of notation from [34]). Besides the difference in models, our proof of tightness is more general in that it applies to the non-CRP case – this more general argument is close to that used in [67]. From the tightness of stationary distributions, using the structure of the corresponding continuous state space, we obtain the convergence of the stationary version of (non-Markov) process $\boldsymbol{Y}$ to that of a positive recurrent Markov chain, whose structure we explicitly describe.

Papers [61, 62] consider objective (b) in the heavy traffic regime. They introduce a modification of MaxWeight, called *regular service guarantee* (RSG) scheme, which explicitly tracks the service time-gaps for each flow to dynamically increase the scheduling priority of flows with large current time-gaps. The papers prove that RSG, under certain parameter settings, preserves heavy-traffic queue-length minimization properties of MaxWeight under the CRP condition; at the same time, the papers demonstrate via simulations that RSG improves smoothness (regularity) of the service process. Recall that in this section we focus on the "pure" MaxWeight, without CRP, and formally show the service process smoothness in the heavy traffic limit.

The rest of the section is organized as follows. The formal model is presented in Section 3.3. Section 3.4 describes the MaxWeight algorithm and the heavy traffic asymptotic regime. Our main results, Theorems 2 and 4, are described in Sec-

tion 3.5. (The formal statement of Theorem 4 is in Section 3.10.) The CRP condition is defined in Section 3.6. In Section 3.7 we provide some necessary background and results for general state-space Markov chains. In sections 3.8 – 3.10 we prove our results for the special case when CRP holds. Finally, in Section 3.11 we show how the proofs generalize to the case when CRP does *not* necessarily hold.

3.2.1 Basic Notation

Elements of a Euclidean space $\mathbb{R}^N$ will be viewed as row-vectors, and written in bold font; $\|\boldsymbol{a}\|$ is the usual Euclidean norm of vector $\boldsymbol{a}$. For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{a} \cdot \boldsymbol{b}$ denotes their scalar (dot) product; vector inequalities are understood componentwise; zero vector and the vector of all ones are denoted $\mathbf{0}$ and $\mathbf{1}$, respectively; $\boldsymbol{ab}$ will denote the vector obtained by componentwise multiplication; if all components of $\boldsymbol{b}$ are non-zero, $\frac{\boldsymbol{a}}{\boldsymbol{b}}$ will denote the vector obtained by componentwise division; statement "$\boldsymbol{a}$ is a positive vector" means $\boldsymbol{a} > \mathbf{0}$. The closed ball of radius r centered at $\boldsymbol{x}$ is $B_r(\boldsymbol{x})$. The positive orthant of $\mathbb{R}^N$ is denoted $\mathbb{R}^N_+ = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x} \geq \mathbf{0}\}$.

For numbers a and b, we denote $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$, $a^+ = a \vee 0$. For vectors $\boldsymbol{a} \leq \boldsymbol{b}$, we denote by $[\boldsymbol{a}, \boldsymbol{b}]$ the rectangle $\times_{i=1}^N [a_i, b_i]$ in $\mathbb{R}^N$.

We always consider Borel σ-algebra $\mathcal{B}(\mathbb{R}^N)$ (resp. $\mathcal{B}(\mathbb{R}^N_+)$) on $\mathbb{R}^N$ (resp. $\mathcal{B}(\mathbb{R}^N_+)$), when the latter is viewed as measurable space. Lebesgue measure on $\mathbb{R}^N$ is denoted by $\mathcal{L}$. When we consider a linear subspace of $\mathbb{R}^N$, we endow it with the Euclidean metric and the corresponding Borel σ-algebra and Lebesgue measure.

For a random process $\boldsymbol{W}(t),\ t = 0, 1, 2, \ldots$, we often use notation $\boldsymbol{W}(\cdot)$ or simply $\boldsymbol{W}$.

3.3 System Model

We consider a system of N flows served by a "switch", which evolves in discrete time $t = 0, 1, \ldots$. At the beginning of each time-slot, the scheduler has to choose from a finite number K of "service-decisions". If the service decision $k \in \{1, \ldots, K\}$ is chosen, then independently of the past history the flows get an amount of service, given by a random non-negative vector. Furthermore, we assume that (if decision k is chosen), there is a finite number $\mathcal{O}_k$ of possible service-vector outcomes, i.e. with probability $p^{k,j}, j = 1, \ldots, \mathcal{O}_k$, it is given by a non-negative vector $\boldsymbol{v}^{k,j} = (v_1^{k,j}, \ldots, v_N^{k,j})$. The expected service vector for decision k is denoted $\boldsymbol{\mu}^k = (\mu_1^k, \ldots, \mu_N^k) = \sum_{j=1}^{\mathcal{O}_k} \boldsymbol{v}^{k,j} p^{k,j}$. We assume that vectors $\boldsymbol{\mu}^k$ are non-zero and different from each other; and that for each i there exists k such that $\mu_i^k > 0$. We will use notations

$$S_i^{max} = \max v_i^{k,j} \text{ over all } k \text{ and } j; \quad \boldsymbol{S}^{max} = (S_1^{max}, \ldots, S_N^{max}).$$

We denote by $\boldsymbol{S}(t) = (S_1(t), \ldots, S_N(t))$ the (random) realization of the service vector at time t, and call $\boldsymbol{S}(\cdot)$ the service process.

After the service at time t is completed, a random amount of work arrives into the queues, and it is given by a non-negative vector $\boldsymbol{A}(t) = (A_1(t), \ldots, A_N(t))$. The values of $\boldsymbol{A}(t)$ are i.i.d. across times t, and $\boldsymbol{A}(\cdot)$ is called the arrival process. The mean arrival rates of this process are given by vector

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N) = E\boldsymbol{A}(t).$$

We will now make assumptions on the distribution of $\boldsymbol{A}(t)$. The distribution is absolutely continuous w.r.t. Lebesgue measure, it is concentrated on the rectangle

$[\mathbf{0}, \boldsymbol{A}^{\max}]$ for some constant vector $\boldsymbol{A}^{\max} > \boldsymbol{S}^{\max}$; moreover, on this rectangle the distribution density $f(\boldsymbol{x})$ is both upper and lower bounded by positive constants, i.e. $0 < \delta_* \leq f(\boldsymbol{x}) \leq \delta^*$.

If $\boldsymbol{Q}(t) \triangleq (Q_1(t), \ldots, Q_N(t))$ is the vector of queue lengths at time t, then for each $i = 1, \ldots, N$

$$\begin{aligned} Q_i(t+1) &= (Q_i(t) - S_i(t))^+ + A_i(t), \\ &= Q_i(t) + A_i(t) - S_i(t) + U_i(t), \end{aligned} \tag{3.1}$$

where $U_i(t) = (S_i(t) - Q_i(t))^+$ is the amount of service "wasted" by flow i at time t.

3.4 MaxWeight Scheduling Scheme. Heavy Traffic Regime

3.4.1 MaxWeight Definition

Let a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N) > \mathbf{0}$ be fixed. MaxWeight scheduling algorithm chooses, at each time t, a service decision

$$k \in \arg\max_l \left((\boldsymbol{\gamma} \boldsymbol{Q}(t)) \cdot \boldsymbol{\mu}^l \right); \tag{3.2}$$

with ties broken according to any well defined rule.

Under MaxWeight, the queue length process $\boldsymbol{Q}(\cdot)$ is a discrete time Markov chain with (continuous) state space $\mathbb{R}_+^N$. System stability is understood as positive Harris recurrence of this Markov chain.

Denote the system *rate region* by

$$\boldsymbol{V} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}_+^N : \boldsymbol{x} \leq \sum_k \psi_k \boldsymbol{\mu}^k \text{ for some } \psi_k \geq 0, \sum_k \psi_k = 1 \right\} \tag{3.3}$$

It is well known (see [34, 99, 102]) that, in general, under

MaxWeight the system is stable as long as the vector of mean arrival rates $\boldsymbol{\lambda}$ is such that $\boldsymbol{\lambda} < \boldsymbol{x} \in \boldsymbol{V}$. (Scheduling rules having this property are sometimes called "throughput-optimal".) This is true for our model as well as will be shown in Section 3.8. (Establishing this fact is not difficult, but it does not directly follow from previous work, because we have continuous state space.)

3.4.2 Heavy Traffic Regime

We will consider a sequence of systems, indexed by $n \to \infty$, operating under MaxWeight scheduling. (Variables pertaining to n-th system will be supplied superscript (n).) The switch parameters will remain unchanged, but the distribution of $\boldsymbol{A}^{(n)}(t)$ changes with n: namely, for each n it has density $f^{(n)}$ which satisfies all conditions specified in Section 3.3, and $f^{(n)}$ uniformly converges to some density f^*. Note that, automatically, the limiting density $f^\star$ (as well as each $f^{(n)}$) satisfies bounds $0 < \delta_* \leq f^\star(\boldsymbol{x}) \leq \delta^*$ in the rectangle $[\boldsymbol{0}, \boldsymbol{A}^{max}]$, and is zero elsewhere. The arrival process $\boldsymbol{A}^\star(\cdot)$, such that the distribution of $\boldsymbol{A}^\star(t)$ has density $f^\star$, has the arrival rate vector $\boldsymbol{\lambda}^\star$. Correspondingly, $\boldsymbol{\lambda}^{(n)} \to \boldsymbol{\lambda}^\star$.

We assume that $\boldsymbol{\lambda}^\star > \boldsymbol{0}$ is a maximal element of rate region $\boldsymbol{V}$, i.e. $\boldsymbol{x} \geq \boldsymbol{\lambda}^\star$ and $\boldsymbol{x} \in \boldsymbol{V}$ only when $\boldsymbol{x} = \boldsymbol{\lambda}^\star$. Thus, $\boldsymbol{\lambda}^\star$ lies on the outer boundary of $\boldsymbol{V}$. We further assume that for each n, $\boldsymbol{\lambda}^{(n)}$ lies in the *interior* of $\boldsymbol{V}$; therefore, the system is stable for each n (under the MaxWeight algorithm).

The (limiting) system, with arrival process $\boldsymbol{A}^\star(\cdot)$ is called critically loaded.

3.5 Main Results

Consider the sequence of systems described in Section 3.4, in the heavy traffic regime. Under any throughput-optimal scheduling algorithm, for each n, the steady-state average amount of service provided to each flow i is greater or equal to its arrival rate λ_i. (It may, and typically will, be greater if the wasted service is

taken into account.)

We now define the notion of *asymptotic smoothness* of the steady-state service process. Informally, it means the property that as the system load approaches critical, the steady state service processes are such that for each flow the probability of a T-long gap (without any service at all) uniformly vanishes, as $T \to \infty$.

For each n, consider the cumulative service process $\boldsymbol{G}^{(n)}(\cdot)$ in steady state. Namely,

$$\boldsymbol{G}^{(n)}(t) \triangleq \sum_{\tau=1}^{t} \boldsymbol{S}^{(n)}(\tau), \quad t = 1, 2, \dots$$

Definition. *We call the service process asymptotically smooth, if*

$$\max_i \lim_{T\to\infty} \left(\limsup_{n\to\infty} P\left(G_i^{(n)}(T) = 0 \right) \right) = 0. \tag{3.4}$$

Our key result (Theorem 4 in Section 3.10) shows that a "queue-differential" process, which determines scheduling decisions in the system under MaxWeight in heavy traffic, is such that its stationary version converges to that of stationary positive Harris recurrent Markov chain, whose structure we describe explicitly. This result, in particular, will imply the following

Theorem 2. *Consider the sequence of systems described in Section 3.4, in the heavy traffic regime. Under MaxWeight scheduling, the service process is asymptotically smooth.*

The proof is given in Section 3.10.

3.6 Complete Resource Pooling Condition

To improve exposition, we first give detailed proofs of our main results for the special case, when the following *complete resource pooling* (CRP) condition holds.

(In Section 3.11 we will show how the proof generalizes to the case without the CRP condition.) Assume that vector $\boldsymbol{\lambda}^\star$ is such that there is a unique (up to scaling) outer normal vector $\boldsymbol{\nu} > \mathbf{0}$ to $\boldsymbol{V}$ at point $\boldsymbol{\lambda}^\star$; we choose $\boldsymbol{\nu}$ so that $\|\boldsymbol{\nu}\| = 1$. Denote by

$$\boldsymbol{V}^\star \triangleq \arg\max_{\boldsymbol{x}\in\boldsymbol{V}} \boldsymbol{\nu}\cdot\boldsymbol{x} \tag{3.5}$$

the outer face of $\boldsymbol{V}$ where $\boldsymbol{\lambda}^\star$ lies. Given our assumptions on $\boldsymbol{\lambda}^\star$, it lies in the relative interior of $\boldsymbol{V}^\star$.

By $\boldsymbol{\nu}_\perp$ we denote the subspace of $\mathbb{R}^N$ orthogonal to $\boldsymbol{\nu}$. For any vector $\boldsymbol{a}$, we denote by $\boldsymbol{a}_\star \triangleq (\boldsymbol{a}\cdot\boldsymbol{\nu})\,\boldsymbol{\nu}$ its orthogonal projection on the (one-dimensional) subspace spanned by $\boldsymbol{\nu}$, and by $\boldsymbol{a}_\perp \triangleq \boldsymbol{a} - \boldsymbol{a}_\star$ its orthogonal projection on the $(N-1)$-dimensional subspace $\boldsymbol{\nu}_\perp$.

The following observations and notations will be useful. There is a $\delta > 0$ such that the entire set

$$B^\delta_{\boldsymbol{\lambda}^\star} \triangleq \{\boldsymbol{y}\in\boldsymbol{V}^\star : \|\boldsymbol{y}-\boldsymbol{\lambda}^\star\| \le \delta\}, \tag{3.6}$$

also lies in the relative interior of $\boldsymbol{V}^\star$.

3.7 Background on General-State-Space Discrete-Time Markov Chains

We will briefly discuss some notions and results from [75] and [44] on the stability of discrete time Markov Chains (MC), which will be used in later sections. Throughout this section we will assume that the Markov Chain $\boldsymbol{\Phi} = \{\Phi(0), \Phi(1), \ldots\}$ is evolving on a locally compact separable metric space $\boldsymbol{X}$ whose Borel σ-algebra will be denoted by $\mathcal{B}$. P_η and E_η are used to denote the probabilities and expectations conditional on Φ_0 having distribution η, while $P_{\boldsymbol{x}}$ and $E_{\boldsymbol{x}}$ are used when η is

concentrated at $\boldsymbol{x}$. The transition function of $\mathbf{\Phi}$ is denoted by $P(\boldsymbol{x}, A), \boldsymbol{x} \in \boldsymbol{X}, A \in \mathcal{B}$. The iterates $P^t,\ t = 0, 1, 2, \ldots$, are defined inductively by

$$P^0 \triangleq I, P^t \triangleq PP^{t-1}, t \geq 1,$$

where I is the identity transition function.

Definition. *(i) ϕ-irreducibility: A Markov Chain $\mathbf{\Phi} = \{\Phi(0), \Phi(1), \ldots\}$ is called ϕ irreducible if there exists a finite measure ϕ such that $\sum_{k=1}^{\infty} P^k(\boldsymbol{x}, A) > 0$ for all $\boldsymbol{x} \in \boldsymbol{X}$ whenever $\phi(A) > 0$. Measure ϕ is called an irreducibility measure.*

(ii) Harris Recurrence: If $\mathbf{\Phi}$ is ϕ-irreducible and $P_x(\Phi(t) \in A \quad \text{i.o.}) \equiv 1$ whenever $\phi(A) > 0$, then $\mathbf{\Phi}$ is called Harris recurrent. [Abbreviation 'i.o.' means 'infinitely often'.]

(iii) Invariant Measure: A σ-finite measure π on $\mathcal{B}$ with the property

$$\pi\{A\} = \pi P\{A\} \triangleq \int \pi(d\boldsymbol{x}) P(\boldsymbol{x}, A), \forall A \in \mathcal{B},$$

is called an invariant measure.

(iv) Positive Harris Recurrence: If $\mathbf{\Phi}$ is Harris Recurrent with a finite invariant measure π, then it is called positive Harris Recurrent.

(v) Boundedness in Probability: If for any $\epsilon > 0$ and any $\boldsymbol{x} \in \boldsymbol{X}$, there exists a compact set D such that

$$\liminf_{t \to \infty} P_{\boldsymbol{x}}(\Phi(t) \in D) \geq 1 - \epsilon, \tag{3.7}$$

then the Markov process $\mathbf{\Phi}$ is called bounded in probability.

(vi) Small Sets: A set C is called small if for all $\boldsymbol{x} \in C$ and some integer $l \geq 1$, we

have

$$P^l(\boldsymbol{x}, \cdot) \geq \nu(\cdot), \tag{3.8}$$

where $\nu(\cdot)$ *is a sub-probability measure, i.e.* $\nu(\boldsymbol{X}) \leq 1$.

(vii) For a probability distribution $\boldsymbol{a} = (a_1, a_2 \ldots)$ *on* $\{1, 2, \ldots\}$*, the Markov transition function* $K_{\boldsymbol{a}}$ *is defined as*

$$K_{\boldsymbol{a}} \triangleq \sum_{i=1}^{\infty} a_i P^i.$$

(viii) Petite Sets: A set $A \in \mathcal{B}$ *and a sub-probability measure* ψ *on* $\mathcal{B}(\boldsymbol{X})$ *are called petite if for some probability distribution* $\boldsymbol{a}$ *on* $\{1, 2, \ldots\}$ *we have*

$$K_{\boldsymbol{a}}(\boldsymbol{x}, \cdot) \geq \psi(\cdot), \forall \boldsymbol{x} \in A.$$

(ix) Non-evanescence: A Markov chain $\boldsymbol{\Phi}$ *is called non-evanescent if* $P_{\boldsymbol{x}}\{\boldsymbol{\Phi} \to \infty\} = 0$ *for each* $\boldsymbol{x} \in \boldsymbol{X}$*. [Event* $\{\boldsymbol{\Phi} \to \infty\}$ *consists of the outcomes such that the sequence* $\Phi(t)$ *visits any compact set at most a finite number of times.]*

The following proposition states some results from [75].

Proposition 1. *(i) If a set* A *is small and for some probability distribution* $\boldsymbol{a}$ *on* $\{1, 2, \ldots\}$ *and a set* $B \in \mathcal{B}$*, we have*

$$\inf_{\boldsymbol{x} \in B} K_{\boldsymbol{a}}(\boldsymbol{x}, A) > 0, \tag{3.9}$$

then B *is petite.*

(ii) Suppose that every compact subset of $\boldsymbol{X}$ *is petite. Then* $\boldsymbol{\Phi}$ *is positive Harris recurrent if and only if it is bounded in probability.*

(iii) Suppose that every compact subset of $\boldsymbol{X}$ is petite. Then $\boldsymbol{\Phi}$ is Harris recurrent if and only if it is non-evanescent.

The following result is form from [44]. It is stated in a form convenient for its application in this section.

Proposition 2. *Let $L(\boldsymbol{x})$ be a non-negative (Lyapunov) function such that the Markov process $\boldsymbol{\Phi}$ satisfies the following two conditions, for some positive constants κ, δ, D:*
(a) $E\left[L(\Phi(t+1)) - L(\Phi(t))|\Phi(t) = \boldsymbol{x}\right] < -\delta$*, for any state $\boldsymbol{x}$ such that* $L(\boldsymbol{x}) \geq \kappa > 0$.
(b) $|L(\Phi(t+1)) - L(\Phi(t))| < D$.
Then there exist constants $\eta > 0$ and $0 < \rho < 1$ such that

$$P(L(\Phi(t)) \geq u \mid L(\Phi(0)) = b) \leq$$

$$\rho^t \exp(\eta(b-u)) + \frac{1-\rho^t}{1-\rho} D \exp(\eta(\kappa - u)),\ u \geq 0. \tag{3.10}$$

3.8 Queue Length Process

Recall that $\boldsymbol{Q}^{(n)}(\cdot)$ is the queue length process for the n-th system under MaxWeight. In this section we prove that for all n, the process $\boldsymbol{Q}^{(n)}(\cdot)$ is positive Harris recurrent. The proof uses a Lyapunov drift argument which is fairly standard (in fact, there is more than one way to prove stability of $\boldsymbol{Q}^{(n)}(\cdot)$), except, since our state space is continuous, as a first step we will show that all compact sets are petite.

Some simple preliminary observations are given in the following lemma.

Lemma 2. *(i) The points $\boldsymbol{x} \in \mathbb{R}_+^N$, such that*
$k \in \arg\max_\ell (\boldsymbol{\gamma x}) \cdot \boldsymbol{\mu}^l$ is non-unique, form a set of zero Lebesgue measure. Moreover, if $\boldsymbol{x} > \mathbf{0}$ is such that $k \in \arg\max_\ell (\boldsymbol{\gamma x}) \cdot \boldsymbol{\mu}^l$ is unique, then for a sufficiently small

$\epsilon > 0$ *the decision* k *is also the unique element of* $\arg\max_\ell(\boldsymbol{\gamma y}) \cdot \boldsymbol{\mu}^l$ *for all* $\boldsymbol{y} \in B_\epsilon(\boldsymbol{x})$. *(ii) The one-step transition function* $P^{(n)}(\boldsymbol{x}, \cdot)$ *of the process* $\boldsymbol{Q}^{(n)}(\cdot)$ *is such that, uniformly in* n *and* $\boldsymbol{x} \in \mathbb{R}^N_+$*, the distribution* $P^{(n)}(\boldsymbol{x}, \cdot)$ *is absolutely continuous with the density upper bounded by* δ^* *and, in the rectangle* $[\mathbf{0}, \boldsymbol{A}^{max} - \boldsymbol{S}^{max}]$*, lower bounded by* δ_*.

Proof. Statement (i) easily follows from the finiteness of the set of decisions k. Statement (ii) easily follows from the assumptions on the arrival process distribution and the fact that $\boldsymbol{S}^{max} < \boldsymbol{A}^{max}$. □

Lemma 3. *For any* $\boldsymbol{x} > \mathbf{0}$*, there exists* $\epsilon > 0$ *such that the set* $B_\epsilon(\boldsymbol{x})$ *is small for the process* $\boldsymbol{Q}^{(n)}(\cdot)$.

Proof. Consider rectangle
$H = [\boldsymbol{x} + (1/3)(\boldsymbol{A}^{max} - \boldsymbol{S}^{max}), \boldsymbol{x} + (2/3)(\boldsymbol{A}^{max} - \boldsymbol{S}^{max})]$. Choose $\epsilon > 0$ small enough, so that $\epsilon < (1/3)\min_i(A_i^{max} - S_i^{max})$ and $\epsilon < \min_i x_i$. Then, $B_\epsilon(\boldsymbol{x})$ lies in the interior of $\mathbb{R}^N_+$ and every point in $B_\epsilon(\boldsymbol{x})$ is strictly smaller than any point in H. Lemma 2(ii) implies that for any $\boldsymbol{y} \in B_\epsilon(\boldsymbol{x})$, the distribution $P^{(n)}(\boldsymbol{y}, \cdot)$ has a density lower bounded by δ_* in H. □

Lemma 4. *For the Markov process* $\boldsymbol{Q}^{(n)}(\cdot)$*, any compact set is petite.*

Proof. Consider a compact set $G \subset \mathbb{R}^N_+$; of course, G is bounded. Fix arbitrary $\boldsymbol{x} > 0$ and pick $\epsilon > 0$ small enough, so that $B_\epsilon(\boldsymbol{x})$ is small and lies in the interior of $\mathbb{R}^N_+$. Pick small $\delta > 0$ such that any point in $\{\|y\| \le \delta\}$ is strictly less than any point in $B_\epsilon(\boldsymbol{x})$.

It is easy to verify that there exists an integer $\tau > 0$ such that the following holds uniformly in $\boldsymbol{Q}^{(n)}(0) \in G$:

$$P\{\|\boldsymbol{Q}^{(n)}(\tau)\| \le \delta\} \ge \alpha \text{ for some } \alpha > 0. \tag{3.11}$$

Indeed, suppose first that for all $t = 0, 1, \ldots$, $\boldsymbol{A}^{(n)}(t) = \mathbf{0}$. (This is a probability zero event, of course, but let's consider it anyway.) Then, for any $\delta_3 > 0$ there exist $\delta_1, \delta_2 > 0$, such that the following holds: with probability at least some $\delta_1 > 0$, the norm $\|\boldsymbol{Q}^{(n)}(t)\|$ decreases at least by some $\delta_2 > 0$, at each time t when $\|\boldsymbol{Q}^{(n)}(t)\| \geq \delta_3 > 0$. This implies that for some $\tau > 0$ and $\delta_4 > 0$, $\boldsymbol{Q}^{(n)}(0) \in G$ implies $P\{\|\boldsymbol{Q}^{(n)}(\tau)\| \leq \delta_3\} \geq \delta_4$. Now, using this and the fact that with a positive probability $\boldsymbol{A}^{(n)}(t)$ can be "very close to $\mathbf{0}$," we can easily establish property (3.11). (We omit rather trivial details.)

Next, it is easy to show that there exists an integer $\tau_1 > 0$ such that the following holds uniformly in $\|\boldsymbol{Q}^{(n)}(0)\| \leq \delta$:

$$P\{\|\boldsymbol{Q}^{(n)}(\tau_1)\| \in B_\epsilon(\boldsymbol{x})\} \geq \alpha_1 \text{ for some } \alpha_1 > 0. \tag{3.12}$$

Here we use Lemma 2(ii), which shows that at each time step the distribution of the increments of $\boldsymbol{Q}^{(n)}(\cdot)$ has a density lower bounded by δ_* in $[\mathbf{0}, \boldsymbol{A}^{max} - \boldsymbol{S}^{max}]$.

From (3.11) and (3.12) we see that uniformly in $\boldsymbol{Q}^{(n)}(0) \in G$, $P\{\|\boldsymbol{Q}^{(n)}(\tau + \tau_1)\| \in B_\epsilon(\boldsymbol{x})\} \geq \alpha\alpha_1$. Application of Theorem 1(i) shows that G is petite (and, moreover, that it is small). □

To prove stability, we will apply Proposition 1 which requires the following

Lemma 5. *Consider the scalar projection $\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(\cdot)\|$, $t = 0, 1, \ldots$ of the the Markov process $\boldsymbol{Q}^{(n)}$ starting with a fixed initial state $\boldsymbol{Q}^{(n)}(0)$, such that $\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(0)\| = b$. Then, uniformly on all large n we have,*

$$P(\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\| \geq u) \leq \rho^t \exp(\eta(b-u)) + \frac{1-\rho^t}{1-\rho} D \exp(\eta(\kappa - u)),\ u \geq 0, \tag{3.13}$$

for some constants $\eta, \kappa, D > 0$ and $1 > \rho > 0$ which depend on n. Consequently, the process $\boldsymbol{Q}^{(n)}(\cdot)$ is bounded in probability.

Proof. We will use notation $L(\boldsymbol{x}) = \|\sqrt{\boldsymbol{\gamma}}\boldsymbol{x}\|$. Then $L(\boldsymbol{Q}^{(n)}(0)) = b$. Clearly, $|L(\boldsymbol{Q}^{(n)}(t+1)) - L(\boldsymbol{Q}^{(n)}(t))|$ is uniformly bounded by a constant, given our assumptions on the arrival and service processes. We will show that the drift (average increment) of $L(\boldsymbol{Q}^{(n)}(t+1))) - L(\boldsymbol{Q}^{(n)}(t)))$ is upper bounded by some $-\tilde{\delta} < 0$ when $\|L(\boldsymbol{Q}^{(n)}(t))\| \geq \kappa$ for some $\kappa > 0$.

Consider a fixed $\boldsymbol{Q}^{(n)}(t)$ and denote $\Delta L = E[L(\boldsymbol{Q}^{(n)}(t+1)) - L(\boldsymbol{Q}^{(n)}(t))]$. Clearly,

$$\begin{aligned}\Delta L &= E\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t+1)\| - \|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\| \\ &\leq \frac{1}{2\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|}\left(E\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t+1)\|^2 - \|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|^2\right), \end{aligned} \tag{3.14}$$

where the inequality follows from the concavity of the function $\sqrt{x}$. Substitute the value of $\boldsymbol{Q}^{(n)}(t+1)$ from equation (3.1), concentrate on the numerator of the above expression to obtain,

$$\begin{aligned}&E\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t+1)\|^2 - \|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|^2 \\ &= E\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t) + \sqrt{\boldsymbol{\gamma}}\left(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t)\right)\|^2 \\ &\qquad - \|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|^2 \\ &= E\left[\|\sqrt{\boldsymbol{\gamma}}(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t))\|^2\right. \\ &\qquad \left.+2\left(\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\sqrt{\boldsymbol{\gamma}}\left(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t)\right)\right)\right] \\ &= E\left[\|\sqrt{\boldsymbol{\gamma}}(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t))\|^2\right. \\ &\qquad \left.+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t))\right] \\ &= E\left[\|\sqrt{\boldsymbol{\gamma}}(\boldsymbol{A}^{(n)}(t) - \boldsymbol{S}^{(n)}(t) + \boldsymbol{U}^{(n)}(t))\|^2\right. \\ &\left.+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\boldsymbol{U}^{(n)}(t)\right.\end{aligned}$$

$$+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\boldsymbol{A}^{(n)}(t)-\boldsymbol{S}^{(n)}(t)\right)\Big]$$

$$\leq b_1+b_2+2E\left[\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\boldsymbol{A}^{(n)}(t)-\boldsymbol{S}^{(n)}(t)\right)|\boldsymbol{Q}^{(n)}(t)\right], \tag{3.15}$$

where b_1 is a uniform bound on $\|\sqrt{\boldsymbol{\gamma}}(\boldsymbol{A}^{(n)}(t)-\boldsymbol{S}^{(n)}(t)+\boldsymbol{U}^{(n)}(t))\|^2$, and b_2 is a uniform bound on $\|2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\boldsymbol{U}^{(n)}(t)\|$ which follows from the property that $U_i(t)>0$ only when $Q_i(t)$ is sufficiently small.

To simplify exposition and avoid introducing additional notation, let us assume that $\boldsymbol{\lambda}^{(n)}-\boldsymbol{\lambda}^\star=-\epsilon\boldsymbol{\nu}$ for some $\epsilon>0$. (If not, then instead of $\boldsymbol{\lambda}^\star$ in this proof we can use $\boldsymbol{\lambda}^{\star\star}$, which the orthogonal projection of $\boldsymbol{\lambda}^{(n)}$ on $\boldsymbol{V}^\star$.) Combining (3.14) and (3.15), we obtain

$$\begin{aligned}
2\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|\Delta L &\leq b_1+b_2+2E\left[\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\boldsymbol{A}^{(n)}(t)-\boldsymbol{S}^{(n)}(t)\right)\right]\\
&= b_1+b_2+2E\left[\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\boldsymbol{A}^{(n)}(t)-\boldsymbol{\lambda}^\star+\boldsymbol{\lambda}^\star-\boldsymbol{S}^{(n)}(t)\right)\right]\\
&= b_1+b_2-2\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\star\|+2E\left[\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)\cdot\left(\boldsymbol{\lambda}^\star-\boldsymbol{S}^{(n)}(t)\right)\right]\\
&\leq b_1+b_2-2\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\star\|-\delta\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\perp\|,
\end{aligned} \tag{3.16}$$

where the last inequality follows from the definition of Max Weight (see (3.2)) and the set $B^\delta_{\boldsymbol{\lambda}^\star}$ (see (3.6)). If $\|\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\|\geq x$, then at least one of $\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\star\|$ or $\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\perp\|$ is greater than or equal to $x/\sqrt{2}$. After some algebraic manipulations we obtain ($\gamma_{min}=\min_i\gamma_i$),

$$\begin{aligned}
&\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|>x \implies \|\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\|>\sqrt{\gamma_{\min}}x\\
\implies &\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\star\|\vee\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\perp\|\geq\frac{\sqrt{\gamma_{\min}}x}{\sqrt{2}}\\
\implies &\delta\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\star\|+\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)\right)_\perp\|\geq(\epsilon\wedge\delta)\frac{\sqrt{\gamma_{\min}}x}{\sqrt{2}}.
\end{aligned}$$

Substituting the above in inequality (3.16) we see that the drift is upper bounded by

$$-(\epsilon \wedge \delta)\frac{\sqrt{\gamma_{\min}}x}{2\sqrt{2}} + \frac{b_1 + b_2}{\|\sqrt{\boldsymbol{\gamma}}\boldsymbol{Q}^{(n)}(t)\|}.$$

This quantity is uniformly bounded by a negative constant for sufficiently large x. Application of Proposition 2 completes the proof. □

Now the positive recurrence of $\boldsymbol{Q}^{(n)}(\cdot)$ follows from Proposition 1. In fact, we will prove the following stronger statement.

Theorem 3. *For each $n = 1, 2, \ldots$, the Markov process $\boldsymbol{Q}^{(n)}(\cdot)$ is positive Harris recurrent and hence has a unique invariant probability distribution, which will be denoted $\chi^{(n)}$. Moreover, if $\boldsymbol{Q}^{(n)}(\infty)$ is the (random) process state in stationary regime (i.e. it has distribution $\chi^{(n)}$),*

$$E[\|\boldsymbol{Q}^{(n)}(\infty)\|^r] < \infty, \forall r > 0.$$

Proof. By Lemma 4 any compact set is petite. Since $\boldsymbol{Q}^{(n)}(\cdot)$ is also bounded in probability (Lemma 5), by Proposition 1 $\boldsymbol{Q}^{(n)}(\cdot)$ is positive Harris recurrent.

For a function $f(\cdot)$ and fixed $b > 0$, denote $T_b f(\cdot) = f(\cdot) \wedge b$. Consider the process starting from an arbitrary fixed initial state $\boldsymbol{Q}^{(n)}(0)$. Since the process is positive Harris recurrent, we can apply the ergodic theorem to obtain (note that $T_b\|\cdot\|$ is a bounded continuous function):

$$E\left(T_b\|\boldsymbol{Q}^{(n)}(\infty)\|^r\right) = \lim_{m\to\infty}\frac{1}{m}\sum_{t=0}^{m} E\left[T_b\|\boldsymbol{Q}^{(n)}(t)\|^r\right]. \tag{3.17}$$

On the other hand,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{t=0}^{m}E\left[T_b\|\boldsymbol{Q}^{(n)}(t)\|^r\right]\le\lim_{m\to\infty}\frac{1}{m}\sum_{t=0}^{m}E\left[\|\boldsymbol{Q}^{(n)}(t)\|^r\right]$$
$$<C, \tag{3.18}$$

for some constant $C>0$, where the second inequality follows from (3.13). Combining (3.17) and (3.18), we have

$$E\left(T_b\|\boldsymbol{Q}^{(n)}(\infty)\|^r\right)\le C,\quad\forall b>0, \tag{3.19}$$

and therefore, by monotone convergence theorem,

$$E\left(\|\boldsymbol{Q}^{(n)}(\infty)\|^r\right)=\lim_{b\to\infty}E\left(T_b\|\boldsymbol{Q}^{(n)}(\infty)\|^r\right)\le C.$$

□

Lemma 6. *Uniformly on all (large) n and the distributions of $\boldsymbol{Q}^{(n)}(0)$, the distribution of $\boldsymbol{Q}^{(n)}(1)$ is absolutely continuous w.r.t. Lebesgue measure, with the density upper bounded by δ^*.*

We omit the proof, which is straightforward, given our assumptions on the distribution of $\boldsymbol{A}^{(n)}(t)$.

Lemma 7. *As $n\to\infty$, $\|\boldsymbol{Q}^{(n)}(\infty)\|\to\infty$ in probability.*

Proof. The proof is by contradiction. Suppose, for some fixed $C>0$ the compact set $D=\{\boldsymbol{x}\in\mathbb{R}^N:\|\boldsymbol{x}\|\le C\}$ is such that

$$\limsup_{n\to\infty}\chi^{(n)}(D)=\beta>0. \tag{3.20}$$

Suppose $\boldsymbol{Q}^{(n)}(t) \in D$. Then, using the same argument as in the proof of Lemma 4, it is easy to see that for any $\epsilon > 0$ there exists time $\tau \geq 1$, such that $P\{\|\boldsymbol{Q}^{(n)}(t+\tau)\| \leq \epsilon\} \geq \beta_1 > 0$. This in turn implies that, with probability at least some $\beta_2 > 0$, for at least one flow i the amount of wasted service $U_i^{(n)}(t+\tau) \geq \epsilon_2 > 0$. This implies that, for at least one i,

$$\limsup_{n\to\infty} E[U_i^{(n)}(\infty)] \geq \beta_1\beta_2\epsilon_2 > 0.$$

This, however, contradicts the fact that the process is stable for all large n. □

3.9 Steady-State Queue Lengths Deviations from $\boldsymbol{\nu}$

Let us consider the process $\boldsymbol{Y}^{(n)}(\cdot)$, defined as

$$\boldsymbol{Y}^{(n)}(t) := (\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t))_{\perp}.$$

Lemma 8. *The steady-state expected norm $E\|\boldsymbol{Y}^{(n)}(\infty)\|$ is uniformly bounded in n.*

Proof. As we did in the proof of Lemma 5, to simplify exposition, assume that $\boldsymbol{\lambda}^{(n)} - \boldsymbol{\lambda}^{\star} = -\epsilon\boldsymbol{\nu}$. (If not, in this proof we would consider the projection $\boldsymbol{\lambda}^{\star\star}$ of $\boldsymbol{\lambda}^{(n)}$ on $\boldsymbol{V}^{\star}$, instead of $\boldsymbol{\lambda}^{\star}$. Consider Lyapunov function $L(\boldsymbol{Q}) = \sum_{i=1}^{N}\gamma_i Q_i^2$. By Theorem 3, $EL(\boldsymbol{Q}^{(n)}(\infty)) < \infty$. The conditional drift of $L(\boldsymbol{Q})$ in one time step is given by (let $\boldsymbol{Q}^{(n)}(t) = \boldsymbol{Q}^{(n)}$, $\boldsymbol{A}^{(n)}(t) = \boldsymbol{A}^{(n)}$, and so on, to simplify notation)

$$\begin{aligned}
&E\left[L(\boldsymbol{Q}^{(n)}(t+1)) - L(\boldsymbol{Q}^{(n)}(t))|\boldsymbol{Q}^{(n)}\right] \\
&= E\left[\sum_{i=1}^{N}\gamma_i\left(Q_i^{(n)} + A_i^{(n)} - S_i^{(n)} + U_i^{(n)}\right)^2|\boldsymbol{Q}^{(n)}\right] - \sum_{i=1}^{N}\gamma_i\left(Q_i^{(n)}\right)^2 \\
&= E\left[\sum_{i=1}^{N}\gamma_i\left(A_i^{(n)} - S_i^{(n)} + U_i^{(n)}\right)\left(2Q_i^{(n)} + A_i^{(n)} - S_i^{(n)} + U_i^{(n)}\right)|\boldsymbol{Q}^{(n)}\right]
\end{aligned}$$

$$
\begin{aligned}
&= E\left[\sum_{i=1}^{N}\gamma_i\left(A_i^{(n)}-S_i^{(n)}+U_i^{(n)}\right)^2+2\gamma_i Q_i^{(n)}\left(A_i^{(n)}-S_i^{(n)}+U_i^{(n)}\right)|\boldsymbol{Q}^{(n)}\right]\\
&\le b_1+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)\cdot\left(\boldsymbol{\lambda}^{(n)}-E\left(\boldsymbol{S}^{(n)}|\boldsymbol{Q}^{(n)}\right)\right)\\
&= b_1+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)\cdot\left(\boldsymbol{\lambda}^{(n)}-\boldsymbol{\lambda}^{\star}+\boldsymbol{\lambda}^{\star}-E\left(\boldsymbol{S}^{(n)}|\boldsymbol{Q}^{(n)}\right)\right)\\
&= b_1-2\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)_{\star}\|+2\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)\cdot\left(\boldsymbol{\lambda}^{\star}-E\left(\boldsymbol{S}^{(n)}|\boldsymbol{Q}^{(n)}\right)\right)\\
&\le b_1-2\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)_{\star}\|+2\min_{\boldsymbol{y}\in B_{\boldsymbol{\nu}}^{\delta}}\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)\cdot\left(\boldsymbol{\lambda}^{\star}-\boldsymbol{y}\right)\\
&\le b_1-2\epsilon\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}\right)_{\star}\|-2\delta\|\left(\boldsymbol{\gamma}\boldsymbol{Q}\right)_{\perp}\|, \qquad (3.21)
\end{aligned}
$$

where b_1 depends only on $\boldsymbol{\gamma}, \boldsymbol{A}^{max}, \boldsymbol{S}^{max}$, and the last inequality follows from the definition of MaxWeight and $B_{\boldsymbol{\lambda}^{\star}}^{\delta}$. Now consider the process $\boldsymbol{Q}^{(n)}(\cdot)$ in stationary regime, and take the expectation of both parts of (3.21). We obtain,

$$
2\delta E\left[\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(\infty)\right)_{\perp}\|\right]+2\epsilon E\left[\|\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(\infty)\right)_{\star}\|\right]\le b_1. \qquad (3.22)
$$

Recalling that $\left(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(\infty)\right)_{\perp}=\boldsymbol{Y}^{(n)}(\infty)$, we see that
$E\|\boldsymbol{Y}^{(n)}(\infty)\|$ is uniformly bounded. □

3.10 Limit of the Queue-Differential Process

We now define a Markov chain $\boldsymbol{Y}^{\star}(\cdot)$, which, in the sense that will be made precise later, is a limit of the (non-Markov) process $\boldsymbol{Y}^{(n)}(\cdot)$ as $n\to\infty$.

Define $\boldsymbol{Y}^{(n)}(t)$ as the orthogonal projection of $\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t)$ on the subspace $\boldsymbol{\nu}_{\perp}$. We call $\boldsymbol{Y}^{(n)}(\cdot)$ a *queue-differential* process. (Obviously, *under the CRP condition*, the queue-differential process is equal to the "queue deviation" process $\boldsymbol{Y}^{(n)}(\cdot)=(\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t))_{\perp}$ in Section 3.9. When CRP does not hold, the "deviation" and "differential" processes are defined differently. This will be discussed in Section 3.11.) Denote by $\boldsymbol{Y}^{(n)}(\infty)$ the corresponding projection of the steady-state $\boldsymbol{Q}^{(n)}(\infty)$, and

by $\Gamma^{(n)}$ its distribution.

Markov chain $\boldsymbol{Y}^\star(\cdot)$ is defined formally as follows. (We will show below that, in fact, the distribution $\Gamma^{(n)}$ converges to the stationary distribution $\Gamma^\star$ of $\boldsymbol{Y}^\star(\cdot)$.) The state space of $\boldsymbol{Y}^\star(\cdot)$ is $\boldsymbol{\nu}_\perp$. Assume that at time t the "scheduler" chooses decision

$$k \in \arg\max_{l:\boldsymbol{\mu}^l \in \boldsymbol{V}^\star} (\boldsymbol{Y}^\star(t)) \cdot \boldsymbol{\mu}^l, \tag{3.23}$$

which determines the corresponding random amount of service $\boldsymbol{S}(t)$, provided to the "queues" given by vector $\boldsymbol{Q}^\star(t) = \boldsymbol{Y}^\star(t)/\boldsymbol{\gamma}$. After that the (random) amount $\boldsymbol{A}^\star(t)$ of new "work" arrives and is added to the "queues." Finally, the new queue lengths vector $\boldsymbol{Q}^\star(t) - \boldsymbol{S}(t) + \boldsymbol{A}^\star(t)$ is transformed into $\boldsymbol{Y}^\star(t+1)$ via componentwise multiplication by $\boldsymbol{\gamma}$ and orthogonal projection on $\boldsymbol{\nu}_\perp$. (Note that both $\boldsymbol{Q}^\star(t)$ and $\boldsymbol{Y}^\star(t)$ may have components of any sign. Also, there is no "wasted service" here.) In summary, the one step evolution is described by

$$\boldsymbol{Y}^\star(t+1) = \boldsymbol{Y}^\star(t) + \left(\boldsymbol{\gamma}\boldsymbol{A}^\star(t) - \boldsymbol{\gamma}\boldsymbol{S}(t)\right)_\perp. \tag{3.24}$$

Informally, one can interpret the process $\boldsymbol{Y}^\star(\cdot)$ as the queue-differential process $\boldsymbol{Y}^{(n)}(\cdot)$, when n is very large and the queue length vector $\boldsymbol{Q}^{(n)}$ is both large and has a small angle with $\boldsymbol{\nu}$. Under these conditions, *the only service decisions k that can be chosen are such that $\boldsymbol{\mu}^k \in \boldsymbol{V}^\star$, and the choice is uniquely determined by $\boldsymbol{Y}^{(n)}(\cdot)$.*

Let $\tilde{P}(\boldsymbol{x}, \cdot)$ denote the one-step transition function for the Markov process $\boldsymbol{Y}^\star(\cdot)$. If $\boldsymbol{x} \in \boldsymbol{\nu}_\perp$, then let $\tilde{B}_\epsilon(\boldsymbol{x}) := \{\boldsymbol{y} \in \boldsymbol{\nu}_\perp : ||\boldsymbol{y} - \boldsymbol{x}|| \leq \epsilon\}$. The following fact is analogous to Lemma 2.

Lemma 9. *(i) The points $\boldsymbol{y} \in \boldsymbol{\nu}_\perp$, such that*

$$k \in \arg\max_{l:\boldsymbol{\mu}^l \in \boldsymbol{V}^\star} \boldsymbol{y} \cdot \boldsymbol{\mu}^l \tag{3.25}$$

is non-unique, form a set of zero Lebesgue measure. Moreover, if $\boldsymbol{y}$ is such that the corresponding decision k is unique, then for a sufficiently small $\epsilon > 0$ the decision k is also the unique element of

$$\arg\max_{l:\boldsymbol{\mu}^l \in \boldsymbol{V}^\star} \boldsymbol{z} \cdot \boldsymbol{\mu}^l$$

for all $\boldsymbol{z} \in \tilde{B}_\epsilon(\boldsymbol{y})$.

(ii) There exist small $\epsilon > 0$ and constant $c_* > 0$, $c^* > 0$ such that $\tilde{P}(\boldsymbol{x}, \cdot)$ is absolutely continuous and, moreover, uniformly in $\boldsymbol{x} \in \boldsymbol{\nu}_\perp$, the density of $\tilde{P}(\boldsymbol{x}, \cdot)$ is lower bounded by c_* on set $\tilde{B}_\epsilon(\boldsymbol{x})$ and is upper bounded by c^* everywhere.

Proof. Statement (i) is obvious. Statement (ii) follows from our assumptions on the distribution of $\boldsymbol{A}^\star(t)$, the fact that $\boldsymbol{A}^{max} > \boldsymbol{S}^{max}$, and the one-step evolution rule (3.24). We omit details. □

Lemma 10. *For the Markov chain $\boldsymbol{Y}^\star(\cdot)$, every compact set is petite.*

The proof easily follows from Lemma 9, by using the argument analogous to that in the proof of Lemma 4. We omit details.

Next, we establish some properties of a stationary distribution $\Gamma^\star$ of the Markov process $\boldsymbol{Y}^\star(\cdot)$, *assuming a stationary distribution exists*. This will help us later prove that the stationary distribution in fact exists and is unique.

Lemma 11. *If $\Gamma^\star$ is a stationary distribution of $\boldsymbol{Y}^\star(\cdot)$, then $\Gamma^\star$ is equivalent to the Lebesgue measure $\tilde{\mathcal{L}}$, i.e. $\Gamma^\star \ll \tilde{\mathcal{L}}$ and $\tilde{\mathcal{L}} \ll \Gamma^\star$.*

Proof. $\Gamma^\star \ll \tilde{\mathcal{L}}$: This follows from Lemma 9.

$\tilde{\mathcal{L}} \ll \Gamma^\star$: It suffices to show that $\Gamma^\star(\tilde{B}_r(\boldsymbol{z})) > 0$ for any $\boldsymbol{z} \in \boldsymbol{\nu}_\perp$ and $r > 0$. Consider

the process $\boldsymbol{Y}^\star(\cdot)$ with the distribution of $\boldsymbol{Y}^\star(0)$ equal to $\Gamma^\star$. (Then the process is of course stationary.) Fix any $0 < \beta < 1$ and choose a compact set $D \subset \boldsymbol{\nu}_\perp$ such that $\Gamma^\star(D) \geq \beta$. Using Lemma 9 we can easily show that there exists time $\tau > 0$ and a constant $\Delta > 0$, such that, uniformly in $\boldsymbol{Y}^\star(0) = \boldsymbol{x} \in D$,

$$P\{\boldsymbol{Y}^\star(\tau) \in \tilde{B}_r(\boldsymbol{z}) \mid \boldsymbol{Y}^\star(0) = \boldsymbol{x}\} \geq \Delta,$$

and therefore

$$\Gamma^\star(\tilde{B}_r(\boldsymbol{z})) \geq \beta\Delta > 0.$$

□

Lemma 12. *Suppose $\Gamma^\star$ is a stationary distribution of $\boldsymbol{Y}^\star(\cdot)$. Then $\tilde{P}_{\boldsymbol{x}}(\boldsymbol{Y}^\star \to \infty) = 0,\ \Gamma^\star - a.s.$, and hence $\tilde{P}_{\boldsymbol{x}}(\boldsymbol{Y}^\star(t) \to \infty) = 0,\ \tilde{\mathcal{L}} - a.s.$.*

Proof. The proof is by contradiction. Let $\boldsymbol{Y}^\star(0)$ have the stationary distribution $\Gamma^\star$, and assume that $\exists \quad \epsilon > 0, \epsilon_1 > 0$ such that

$$\Gamma^\star(\{\boldsymbol{x} : \tilde{P}_{\boldsymbol{x}}(\boldsymbol{Y}^\star \to \infty) \geq \epsilon_1\}) \geq \epsilon.$$

This would imply that $\limsup_{t\to\infty} P(\boldsymbol{Y}^\star(t) \in D) \leq 1 - \epsilon\epsilon_1$ for every compact set $D \subset \boldsymbol{\nu}_\perp$. This is impossible, because the distribution of $\boldsymbol{Y}^\star(t)$ is equal to $\Gamma^\star$ for all t. □

Lemma 13. *If process $\boldsymbol{Y}^\star(\cdot)$ has a stationary distribution, it is non-evanescent.*

Proof. Consider process $\boldsymbol{Y}^\star(\cdot)$ with fixed initial state $\boldsymbol{Y}^\star(0) = \boldsymbol{x}$. Consider one-step transition. The distribution of $\boldsymbol{Y}^\star(1)$ is absolutely continuous with respect to $\tilde{\mathcal{L}}$. Thus, by Lemma 12, with probability 1, $\boldsymbol{z} = \boldsymbol{Y}^\star(1)$ is such that $\tilde{P}_{\boldsymbol{z}}(\boldsymbol{Y}^\star \to \infty) = 0$. Then, $\tilde{P}_{\boldsymbol{x}}(\boldsymbol{Y}^\star \to \infty) = 0$. □

Lemma 14. *Suppose* $\Gamma^\star$ *is a stationary distribution of* $\boldsymbol{Y}^\star(\cdot)$*. Then, the Markov chain is positive Harris recurrent, and therefore* $\Gamma^\star$ *is its unique stationary distribution.*

Proof. Since every compact set is petite (Lemma 10) and the process is non-evanescent (Lemma 13), it is Harris recurrent by Proposition 1. But since it has a finite invariant measure $\Gamma^\star$, $\boldsymbol{Y}^\star(\cdot)$ is positive Harris recurrent. □

We now show the existence of a stationary distribution of $\boldsymbol{Y}^{(\star)}(\cdot)$.

Lemma 15. *Every weak limit point* $\Gamma^{(\star)}$ *of the sequence of distributions* $\Gamma^{(n)}$ *is a stationary distribution of the process*
$\boldsymbol{Y}^{(\star)}(\cdot)$.

Proof. Let $\Gamma^\star$ be a weak limit of $\Gamma^{(n)}$ along a subsequence on n. We can make the following observations.
(a) Observe that uniformly on all (large) n and the distributions of $\boldsymbol{Q}^{(n)}(0)$, the distribution of $\boldsymbol{Y}^{(n)}(1)$ is absolutely continuous w.r.t. Lebesgue measure, with the upper bounded density. (This easily follows from Lemma 6 and the fact that $\|\boldsymbol{Q}^{(n)}(1) - \boldsymbol{Q}^{(n)}(0)\|$ is uniformly bounded.) Then, we see that $\Gamma^\star$ is absolutely continuous with bounded density.
(b) Consider any point $\boldsymbol{y} \in \boldsymbol{\nu}_\perp$ such that the decision k in (3.25) is unique and a small $\epsilon > 0$ such that this decision k is also unique for all $\boldsymbol{z} \in \tilde{B}_\epsilon(\boldsymbol{y})$. (See Lemma 9(i).) Then, there exists a sufficiently large $C > 0$ such that, uniformly in n, conditions $\|\boldsymbol{Q}^{(n)}(t)\| \geq C$ and $\boldsymbol{Y}^{(n)}(t) \in \tilde{B}_\epsilon(\boldsymbol{y})$ imply that the same decision k will be unique at time t for the process $\boldsymbol{Q}^{(n)}(\cdot)$.

Using these two observations, Lemma 7, and the fact that the distribution of $\boldsymbol{A}^{(n)}(t)$ converges to that of $\boldsymbol{A}^\star(t)$, we can choose a further subsequence of n along which the following property holds. The stationary versions of processes $\boldsymbol{Q}^{(n)}(\cdot)$

and the process $\boldsymbol{Y}^\star(\cdot)$ with distribution of $\boldsymbol{Y}^\star(0)$ equal to $\Gamma^\star$, can be constructed on one common probability space, so that with probability 1:

(c) for all large n, the same decision k is chosen at time 0 in the processes $\boldsymbol{Q}^{(n)}(\cdot)$ and $\boldsymbol{Y}^\star(\cdot)$;

(d) $\boldsymbol{Y}^{(n)}(0) \to \boldsymbol{Y}^\star(0)$ and $\boldsymbol{Y}^{(n)}(1) \to \boldsymbol{Y}^\star(1)$.

This, in turn, implies that for any bounded continuous function g we have,

$$E\left[g(\boldsymbol{Y}^\star(0))\right] = \lim_{n\to\infty} E\left[g(\boldsymbol{Y}^{(n)}(0))\right],$$
$$E\left[g(\boldsymbol{Y}^\star(1))\right] = \lim_{n\to\infty} E\left[g(\boldsymbol{Y}^{(n)}(1))\right].$$

But, $E\left[g(\boldsymbol{Y}^{(n)}(0))\right] = E\left[g(\boldsymbol{Y}^{(n)}(1))\right]$ for all n. Therefore, $E\left[g(\boldsymbol{Y}^\star(0))\right] = E\left[g(\boldsymbol{Y}^\star(1))\right]$. This proves stationarity of $\Gamma^\star$. □

Theorem 4. *The Markov process $\boldsymbol{Y}^\star(\cdot)$ is positive Harris recurrent. The sequence $\Gamma^{(n)}$ [i.e., the distributions of $\boldsymbol{Y}^{(n)}(\infty)$] weakly converges to the unique stationary distribution $\Gamma^\star$ of $\boldsymbol{Y}^\star(\cdot)$.*

Proof. This follows from Lemma 15 and Lemma 14. □

We are finally in position to give a

of Theorem 2. By Theorem 4, the process $\boldsymbol{Y}^\star(\cdot)$ is positive Harris recurrent. Moreover, we know that it is such that every compact set is petite. We can pick any compact set D such that $\Gamma^\star(D) > 0$, and using Nummelin splitting view the process $\boldsymbol{Y}^\star(\cdot)$ as having an atom state, with finite average return time to this atom. We see that the cumulative "service process" $\boldsymbol{G}^\star(\cdot)$ corresponding to $\boldsymbol{Y}^\star(\cdot)$ in steady-state is such that

$$\max_i \lim_{T\to\infty} P(G_i^\star(T) = 0) = 0.$$

Finally, the argument used in the proof of Lemma 15 shows that the stationary versions of processes $\boldsymbol{Y}^\star(\cdot)$ and $\boldsymbol{Q}^{(n)}(\cdot)$ for all (large) n can be constructed on a common probability space in a way such that, w.p.1, for any $T > 0$

$$\boldsymbol{G}^{(n)}(T) \to \boldsymbol{G}^\star(T).$$

This implies (3.4). □

3.11 Generalization to the Case When CRP Condition Does Not Necessarily Hold

If CRP condition does not necessarily hold, let $\boldsymbol{\nu}$ denote the *normal cone* to $\boldsymbol{V}$ at point $\boldsymbol{\lambda}^\star$; it has dimension $d \geq 1$. (In the CRP case, $d = 1$ and $\boldsymbol{\nu}$ is a ray.) Fix any positive vector $\boldsymbol{\nu}'$ which lies in the relative interior of $\boldsymbol{\nu}$. Then, $\boldsymbol{V}^\star$ is defined more generally as

$$\boldsymbol{V}^\star = \arg\max_{\boldsymbol{x} \in \boldsymbol{V}} \boldsymbol{\nu}' \cdot \boldsymbol{x};$$

it is a $(N-d)$-dimensional face of V. By $\boldsymbol{\nu}_\perp$ we denote the $(N-d)$-dimensional *subspace* orthogonal to $\boldsymbol{\nu}$.

We will denote by $\boldsymbol{x}_\star$ the projection of a vector $\boldsymbol{x}$ on the normal cone $\boldsymbol{\nu}$; that is, $\boldsymbol{x}_\star$ is the closest to $\boldsymbol{x}$ point of $\boldsymbol{\nu}$. Then let $\boldsymbol{x}_\perp = \boldsymbol{x} - \boldsymbol{x}_\star$, and let $\boldsymbol{x}_{\perp,sp}$ be the orthogonal projection of $\boldsymbol{x}$ on the *subspace* $\boldsymbol{\nu}_\perp$. Note the difference between the definitions of $\boldsymbol{x}_\perp$ and $\boldsymbol{x}_{\perp,sp}$. (In the CRP case, $\boldsymbol{x}_\perp \equiv \boldsymbol{x}_{\perp,sp}$. In the non-CRP case they are in general different.) We always have $\|\boldsymbol{x}_{\perp,sp}\| \leq \|\boldsymbol{x}_\perp\|$. Note that, *if $\boldsymbol{x}_\star$ lies in the relative interior of $\boldsymbol{\nu}$, then $\boldsymbol{x}_\star = \boldsymbol{x}_{\perp,sp}$*.

In this notation, the entire development in Sections 3.8 and 3.9 is carried out essentially as is, with very minor adjustments.

The development in Section 3.10 is carried out with small adjustments, which are as follows. The queue differential process is defined as $\boldsymbol{Y}^{(n)}(t) = (\boldsymbol{\gamma}\boldsymbol{Q}^{(n)}(t))_{\perp,sp}$.

Correspondingly, the one step evolution of $\boldsymbol{Y}^\star(\cdot)$ is defined by (3.23) and

$$\boldsymbol{Y}^\star(t+1) = \boldsymbol{Y}^\star(t) + (\boldsymbol{\gamma}\boldsymbol{A}^\star(t) - \boldsymbol{\gamma}\boldsymbol{S}(t))_{\perp,sp}\,.$$

Therefore, the state space for both $\boldsymbol{Y}^{(n)}(\cdot)$ and $\boldsymbol{Y}^\star(\cdot)$ is $\boldsymbol{\nu}_\perp$.

The proof of the key Lemma 15 requires, in addition to Lemma 7, the following Lemma 16. Let $h(\boldsymbol{x})$ denote the distance from $\boldsymbol{x}_\star$ to the relative boundary of the cone $\boldsymbol{\nu}$. (To be precise, $h(\boldsymbol{x})$ is defined as the distance from $\boldsymbol{x}_\star$ to the set {relative boundary on the cone $\boldsymbol{\nu}$} $\setminus$ {boundary of the positive orthant $\mathbb{R}^N_+$}.)

Lemma 16. *As $n \to \infty$, $h(\boldsymbol{Q}^{(n)}(\infty) \to \infty$ in probability.*

This lemma is easily proved, because the contrary, along with Lemmas 15 and Lemma 8, would imply that the frequency of choosing scheduling decisions outside $\boldsymbol{V}^\star$ would not vanish, as $n \to \infty$; that would contradict stability when n is large.

Then, in the proof of Lemma 15, in the statement (b), the condition $\|\boldsymbol{Q}^{(n)}(t)\| \geq C$ is replaced by $h(\boldsymbol{Q}^{(n)}(t)) \geq C$; also, Lemma 16 is used along with Lemma 7.

The statement of Theorem 4 and the proof of Theorem 2 remain unchanged.

4. OPTIMIZING QUALITY OF EXPERIENCE OF DYNAMIC VIDEO STREAMING OVER FADING WIRELESS NETWORKS

4.1 Overview

We address the problem of video streaming packets from an Access Point (AP) to multiple clients over a shared wireless channel with fading. In such systems, each client maintains a buffer of packets from which to play the video, and an outage occurs in the streaming whenever the buffer is empty. Clients can switch to a lower-quality of video packet, or request packet transmission at a higher energy level, in order to minimize the number of outages plus the number of outage periods and the number of low-quality video packets streamed, while there is an average power constraint on the AP. We pose the problem of choosing the video quality and transmission power as a Constrained Markov Decision Process (CMDP). We show that the problem involving N clients decomposes into N MDPs, each involving only a single client, and furthermore that the optimal policy has a threshold structure, in which the decision to choose the video-quality and power-level of transmission depends solely on the buffer-level.

4.2 Introduction

Scheduling packets for video streaming over a shared wireless downlink is of increasing attention [1]. Predominantly, this problem has been addressed with the goal of minimizing the average number of outages, i.e., time-slots during which a client has no packet to play [63,83], [118], [28,29,48,104,119]. The primary objective in these works is to minimize the average number of outages suffered during the streaming, i.e., time-slots during which a client has no packet to play. However the models considered in these works do not incorporate the communication con-

straints imposed by the network over which the streaming occurs. Typically clients streaming video files will share a common wireless channel, which again typically has a constraint on the average power. The access point (AP) has to choose the power level at which to transmit individual packets to each client so as to maximize the total Quality of Experience (QoE) experienced by the clients. The system also has an additional degree of freedom in that the AP can transmit lower quality packets on occasion, leading to a softer loss of video quality than an abrupt outage. Another important aspect is that the quality of video streaming experienced by a client depends not only on the number of outages, but also on the number of "outage-periods", i.e., number of interruption periods as well. Thus an outage lasting 10 time-slots is not the same as 10 outages each lasting 1 time-slot. The QoE experienced by a client thus has to take into account several metrics: the average number of outages, the average number of outage-periods, and the quality of video-packets streamed. In this paper we address this overall problem. While we focus here on the single "last-hop" case for ease of exposition and brevity, our results can be generalized to multi-hop networks as well.

4.3 System Description

Consider a system where a wireless channel is shared by N clients for the purpose of streaming video packets. It is assumed that the system evolves over discrete time-slots, and one time-slot is taken by the access point (AP) for attempting one packet transmission.

Client n maintains a buffer of size B_n packets and plays a packet for a duration of T_n time-slots. Once it has finished playing a video-packet, it looks for the next packet in the buffer. In case the buffer is empty, there is an "outage", meaning that the video streaming is interrupted, and the client has to wait for a packet to be

delivered to its buffer before it can resume the video streaming.

The wireless channels connecting the clients to the AP are assumed to be random. For ease of exposition, we will derive the results for the case when the channel conditions are fixed. These results carry over to the case of fading channels in a straight-forward manner. Later, in Section 4.9, we will outline the results for the case of fading channels.

There are Q_n different video-qualities $\{1, 2, \ldots, Q_n\}$ of packets that can be transmitted for client n, with class 1 video quality providing the best viewing experience. Similarly there are $\{\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_n\}$ different power levels at which the packets for client n can be transmitted. We let $\hat{E}_1 = 0$, i.e. a user may choose to not request packet in a time-slot. The probability that the packet for client n is successfully delivered upon a transmission attempt, $P_n(q, E)$, depends on the amount of power E used in the packet transmission and the quality of video packet q that was attempted. We also incorporate an average power constraint on the AP.

The basic problem considered is that of scheduling the AP's packet transmissions to clients so as to maximize the combined Quality of Experience (QoE) of the clients. The QoE of a single client depends on multiple factors

1. The average number of outages.

2. How "often" the video gets interrupted, i.e., the number of outage-periods, or the number of time-slots in which the transition from "non-outage" to outage occurs.

3. The number of packets of different quality types that are streamed.

4.4 Problem Formulation

We denote by $O_n(s)$ the random variable that assumes the value 1 if the n-th client faces an outage at time s, and 0 otherwise, and by $E_n(s)$ the transmission power utilized by the n-th client at time-slot s. Also, let $I_n(q,s)$ be the random variable that takes the value 1 if a packet of quality q is delivered to client n in time-slot s.

The Constrained Markov Decision Process (CMDP) of interest is then to choose the quality of video packets and transmission power for each client, in order to

$$\text{Minimize} \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_n\sum_s \Bigg(O_n(s) + \sum_{q=1}^{Q_n} \lambda_{q,n} I_n(q,s) + \lambda_{O,n} |O_n(s)\left(O_n(s-1)-1\right)| \Bigg)$$

$$\text{subject to ,} \tag{4.1}$$

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_n\sum_s E_n(s) \le \bar{E}. \qquad \text{(Primal MDP)}$$

Note that the term $|O_n(s)\left(O_n(s-1)-1\right)|$ assumes the value 1 if time-slot s is the beginning of an outage-period for client n, and is 0 otherwise. It thereby measures the number of outage periods incurred. The parameters $\{\lambda_{q,n}\}_{q=1}^{Q_n}, \lambda_{O,n} \quad n = 1,2,\ldots,N$ are employed for tuning the QoS to account for the relative importance placed on each of the objectives. We note that for $i > j$, $\lambda_{i,n} > \lambda_{j,n}$ for all n, since we assumed that the video quality of a packet is less if the packet belongs to a higher valued class.

Thus the above problem is a CMDP in which the system state at time t is described by the N dimensional vector $L(t) := (l_1(t), l_2(t), \ldots, l_N(t))$, where $l_n(t)$ is the amount of play time remaining in the buffer of client n at time t.

The central difficulty which arises is that the cardinality of the state-space of the system increases exponentially with the number of clients N, and thus the problem is computationally infeasible as formulated above.

We show in this paper that the problem of serving N clients can be decomposed into N separate problems each involving only a single client. Thus the computational complexity of the problem grows linearly in the number of clients. Moreover, we show that the optimal policy is easily implementable since it has a simple threshold structure.

4.5 The Dual MDP

The Lagrangian associated with a policy π for the system (4.1) is given by,

$$\begin{aligned}\mathcal{L}(\pi,\lambda_E) = \limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_n\sum_s\Bigg(O_n(s) + \sum_{q=1}^{Q_n}\lambda_{q,n}I_n(q,s) \\ + \lambda_{O,n}|O_n(s)\left(O_n(s-1)-1\right)|\Bigg) \\ + \lambda_E\left(\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_n\sum_s E_n(s) - \bar{E}\right),\end{aligned} \tag{4.2}$$

where λ_E is the Lagrangian multiplier associated with the average power constraint. The associated Lagrange dual is,

$$D(\lambda_E) = \min_{\pi}\mathcal{L}(\pi,\lambda_E). \tag{4.3}$$

Next we present a useful bound on the dual, the proof of which follows from the super-additivity of $\limsup$ and sub-additivity of $\liminf$ operations.

Lemma 17.

$$D(\lambda_E) \geq \min_{\pi} \sum_{n} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}\Bigg(O_n(s) + \lambda_E E_n(s) + \lambda_{O,n}|O_n(s)\left(O_n(s-1)-1\right)| + \sum_{q=1}^{Q_n}\lambda_{q,n}I_n(q,s)\Bigg) - \lambda_E\bar{E}. \tag{4.4}$$

4.6 Single Client Problem

We consider minimizing the bound obtained in Lemma 17. Observing the bound, we find that we have decomposed the original problem (4.1) into N single-client problems, i.e., the expression in the r.h.s. of (4.4) is the sum of the costs of N clients, in which the cost of a single client depends only on the action chosen for it in each time-slot.

The problem for the single client is described as follows. We omit the subscript n in the following discussion. The channel connecting the client to the AP is random. The client maintains a buffer of capacity B time-slots of play-time video (this assumption is equivalent to the assumption of maintaining a buffer of B packets since a packet is played for T time-slots), and in each time-slot, the AP has to choose two quantities, which together comprise the control action chosen for the client:

- The video quality $q \in \{1, 2, \ldots, Q\}$.
- The power $E \in \{\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_n\}$ at which to carry out packet transmission.

The state of the client is thus described by $l(t)$, the play-time duration of the packets present in the buffer at time t.

If the client is scheduled a packet transmission of quality q at an power E at time t, and the remaining playtime at time t, $l(t)$, is less than or equal to $B-T+1$, then the system state at time $t+1$ is $(l(t)-1)^+ + T$ with a probability $P(q,E)$, while it is $(l(t)-1)^+$ with a probability $P(q,E)$. However if the value of remaining playtime $l(t)$ is strictly greater than $B-T+1$, then the system state at time $t+1$ is $l(t)-1$ with a probability 1.

We let

$$\mathcal{S}(x) := \begin{cases} (x-1)^+ + T, \text{ if } x \leq B-T+1, \\ x-1, \text{ if } B-T+1 < x \leq B, \end{cases} \tag{4.5}$$

$$\mathcal{F}(x) := (x-1)^+, \tag{4.6}$$

be the transitions associated with the remaining play-times associated for a successful and failed packet transmission respectively. The control action at time t will be denoted $\boldsymbol{u}(t) := (q(t), E(t))$, where $q(t), E(t)$ are the video quality and transmission power level chosen at time t.

The transmissions at power level E incur a cost of $\lambda_E \times E$. There is a penalty of 1 units upon an outage at time t. A penalty of amount λ_q units is imposed if a packet of quality q is delivered to it, while a penalty of λ_O units is imposed at time t in case there was no outage at time-slot $t-1$, and an outage occurs in time-slot t, i.e. if a new outage-period begins at time t.

Since the probability distribution of the system state at time $t+1$ is completely determined by the system state at time t, and the action (q,E) chosen at time t, i.e., requested video quality and power level at which transmission occurs, the single client problem is a Markov Decision Process (MDP) involving only a finite number of actions and states, and thus is solved by a stationary Markov policy [85].

Denote by π_n a policy for the client n. The single client problem is to solve,

$$\min_{\pi} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}\Bigg(O(s)+\lambda_E E(s) \\ +\lambda_O|O(s)\left(O(s-1)-1\right)|+\sum_{q=1}^{Q}\lambda_q I(q,s)\Bigg). \tag{4.7}$$

Denote by $\pi_n^{\star}(\lambda_E)$, the optimal policy which solves the single client problem. We also let

$$V_n(\lambda_E) = \min_{\pi} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_{s=1}^{t}\Bigg(O(s)+\lambda_E E(s) \\ +\lambda_O|O(s)\left(O(s-1)-1\right)|+\sum_{q=1}^{Q}\lambda_q I(q,s)\Bigg), \tag{4.8}$$

be the optimal cost, and $V_n(\lambda_E,\pi)$ be the cost associated with a policy π.

4.7 Threshold Structure of the Optimal Policy for the Single Client Problem

We will suppress the subscript n in the following discussion, and begin with a discussion of the $\beta \in (0,1)$ discounted infinite horizon cost problem for the single client. Let

$$V_\beta(x) = \min_{\pi} \liminf_{t\to\infty} \mathbb{E}\Bigg[\sum_{t=0}^{\infty}\beta^t\left(O(t)+\lambda_E E(t) \\ +\lambda_O|O(t)\left(O(t-1)-1\right)|+\sum_{q=1}^{Q}\lambda_q I(q,s)\right)\Bigg] \tag{4.9}$$

be the minimum β-discounted infinite horizon cost for the system starting in state x at time 0, where x can assume values in the set $\{0,1,\ldots,B\}$. The function $V_\beta^s(x)$ is similarly defined to be the minimum β-discounted cost incurred in s time-slots

for the system starting in state x, i.e.,

$$V_\beta^s(x) = \min_{\pi^s} \mathbb{E}_x \left[\sum_{t=0}^{s} \beta^t \left(O(t) + \lambda_E E(t) + \lambda_O |O(t) \left(O(t-1) - 1 \right)| + \sum_{q=1}^{Q} \lambda_q I(q,s) \right) \right],$$

where π^s is a policy for the s horizon β-discounted problem. The quantities $V_\beta(x), V_\beta^s(x)$ should not be confused with the quantities $V_n(\lambda_E)$ defined in the previous section. We have,

$$\begin{aligned} V_\beta^s(x) &= \min_{(q,E)} 1(x=0) + \lambda_E E + P(q,E) \left[\lambda_q + \beta V_\beta^{s-1}(\mathcal{S}(x)) \right] \\ &\quad + (1 - P(q,E)) \left[1(x=1)\lambda_O + \beta V_\beta^{s-1}(\mathcal{F}(x)) \right] \\ &= 1(x=0) + 1(x=1)\lambda_O + \left[\beta V_\beta^{s-1}(\mathcal{F}(x)) \right] \\ &\quad + \min_{\boldsymbol{u}} \{ C(\boldsymbol{u}) - P(\boldsymbol{u}) D_s^\beta(x) \}, \end{aligned} \tag{4.10}$$

where

$$C(\boldsymbol{u}) := \lambda_E E + P(q,E)\lambda_q, \tag{4.11}$$

is the one-step cost associated with the action $\boldsymbol{u} = (q, E)$, and for $s = 1, 2, \ldots$,

$$D_s^\beta(x) := 1(x=1)\lambda_O + \beta \left\{ V_\beta^{s-1}(\mathcal{F}(x)) - V_\beta^{s-1}(\mathcal{S}(x)) \right\}. \tag{4.12}$$

We assume that a lower video quality packet, or a higher power packet transmission leads to an increase in the success of packet transmission $P(q, E)$, i.e., an increase in cost is associated with a higher transmission success probability.

Definition. ***Threshold policy****: We say a policy is of threshold-type if it satisfies the following for each stage* s*:*

- *Fix any $E \in \{\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_n\}$. If the policy chooses the action (q, E) in state x, then it does not choose the actions $\{(\hat{q}, E) : \hat{q} < q\}$ for any state.* $1 \leq y \leq x$

- *Fix any $q \in \{Q_1, Q_2, \ldots, Q_n\}$. If the policy chooses the action (q, E) in state x, then it does not choose the actions $\{(q, \tilde{E}) : \tilde{E} < E\}$ for any state* $1 \leq y \leq x$.

If $x, y \in \{1, 2, \ldots, B\}$ are such that $x > y$, and let $\boldsymbol{u}_x, \boldsymbol{u}_y$ be the actions chosen by a threshold policy π in states x and y. Then it is also easily verified that $P(\boldsymbol{u}_x) < P(\boldsymbol{u}_y)$.

Next we present a useful lemma that is easily proved. In the following, $(\boldsymbol{u}, \pi)$ is the policy that follows the action $\boldsymbol{u}$ in the first slot, and then follows policy π, while $V_\beta^{s,\pi}(x)$ is the cost achieved under the policy π in s time-slots for the system starting in state x.

Lemma 18. *Let $\boldsymbol{u}_1, \boldsymbol{u}_2$ be two actions where $P(\boldsymbol{u}_2) > P(\boldsymbol{u}_1)$, or equivalently, $P(\boldsymbol{u}_2) > P(\boldsymbol{u}_1)$. Then,*

$$
\begin{aligned}
&V_\beta^{s,(\boldsymbol{u}_2,\pi^\star)}(\mathcal{F}(x)) - V_\beta^{s,(\boldsymbol{u}_1,\pi^\star)}(\mathcal{S}(x)) = P(\boldsymbol{u}_1)\left\{\beta V_\beta^{s-1}(\mathcal{S}(\mathcal{F}(x))) - V_\beta^{s-1}(\mathcal{S}(\mathcal{S}(x)))\right\} \\
&+ (1 - P(\boldsymbol{u}_2))\left\{1(\mathcal{F}(x) = 1)\lambda_O + \beta V_\beta^{s-1}(\mathcal{F}(\mathcal{F}(x))) - V_\beta^{s-1}(\mathcal{F}(\mathcal{S}(x)))\right\} \\
&+ C(\boldsymbol{u}_2) - C(\boldsymbol{u}_1) \\
&= P(\boldsymbol{u}_1)\left\{\beta V_\beta^{s-1}(\mathcal{F}(\mathcal{S}(x))) - V_\beta^{s-1}(\mathcal{S}(\mathcal{S}(x)))\right\} \\
&+ (1 - P(\boldsymbol{u}_2))\left\{1(\mathcal{F}(x) = 1)\lambda_O + \beta V_\beta^{s-1}(\mathcal{F}(\mathcal{F}(x)))\right. \\
&\left.- V_\beta^{s-1}(\mathcal{S}(\mathcal{F}(x)))\right\} + C(\boldsymbol{u}_2) - C(\boldsymbol{u}_1).
\end{aligned}
$$

Lemma 19. *For $s = 1, 2, \ldots$, the functions $D_s^\beta(x)$ are decreasing in x for $x \in \{1, 2, \ldots, B - T + 1\}$.*

Proof. Within this proof, let $\pi_s^\star$ be the optimal policy for the β-discounted s time-

slots problem, and let $(\boldsymbol{u}, \pi^{\star}_{s-1})$ be the policy for s time-slots which takes the action $\boldsymbol{u}$ at the first time-slot, and then follows the policy $\pi^{\star}_{s-1}$. In order to prove the claim, we will use induction on s, the number of time-slots.

Let us assume that the statement is true for the functions $D^{\beta}_{z}(x)$, for all $z \leq s$. In particular the function,

$$1(x=1)\lambda_O + \beta \left\{ V^{s-1}_{\beta}(\mathcal{F}(x)) - V^{s-1}_{\beta}(\mathcal{S}(x)) \right\}, \tag{4.13}$$

is decreasing for $x \in \{1, 2, \ldots, B-T+1\}$.

First we will prove the decreasing property for $x \in \{2, 3, \ldots, B-T+1\}$. Now the assumption (4.13) made above, and (4.10), together imply that $\pi^{\star}_{s}$ is of threshold-type.

Fix an $x \in \{1, 2, \ldots, B-T\}$ and denote by $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4$, the optimal actions at stage s for the states $\mathcal{S}(x), \mathcal{F}(x), \mathcal{S}(x+1), \mathcal{F}(x+1)$ respectively. Note that the threshold nature of $\pi^{\star}_{s}$ implies that,

$$\begin{aligned} &P(\boldsymbol{u}_1) < P(\boldsymbol{u}_2), P(\boldsymbol{u}_3) < P(\boldsymbol{u}_4) \text{ and }, \\ &P(\boldsymbol{u}_3) < P(\boldsymbol{u}_1), P(\boldsymbol{u}_4) < P(\boldsymbol{u}_2). \end{aligned}$$

This is true because as the value of state decreases in the interval $\{1, 2, \ldots, B\}$, a threshold policy switches to an action that has a higher transmission success probability. So it follows from Lemma 18 that

$$\begin{aligned} &V^{s}_{\beta}(\mathcal{F}(x+1)) - V^{s}_{\beta}(\mathcal{S}(x+1)) \\ &\leq V^{s,(\boldsymbol{u}_2, \pi^{\star}_{s-1})}_{\beta}(\mathcal{F}(x+1)) - V^{s}_{\beta}(\mathcal{S}(x+1)) \\ &= C(\boldsymbol{u}_2) - C(\boldsymbol{u}_3) \end{aligned}$$

$$
\begin{aligned}
&+ P_c(\boldsymbol{u}_3) \times \beta \left[V_\beta^{s-1}(\mathcal{F}(\mathcal{S}(x+1))) - V_\beta^{s-1}(\mathcal{S}(\mathcal{S}(x+1)))\right] \\
&+ (1 - P_c(\boldsymbol{u}_2)) \times \\
&\big\{1(\mathcal{F}(x+1) = 1) + \beta V_\beta^{s-1}(\mathcal{F}(\mathcal{F}(x+1))) \\
&\qquad\qquad - V_\beta^{s-1}(\mathcal{S}(\mathcal{F}(x+1)))\big\} \\
&\leq C(\boldsymbol{u}_2) - C(\boldsymbol{u}_3) \\
&+ P_c(\boldsymbol{u}_3) \times \beta \left[V_\beta^{s-1}(\mathcal{S}(\mathcal{F}(x))) - V_\beta^{s-1}(\mathcal{S}(\mathcal{S}(x)))\right] \\
&+ (1 - P_c(\boldsymbol{u}_2)) \times \\
&\left[1(\mathcal{F}(x) = 1) + \beta V_\beta^{s-1}(\mathcal{F}(\mathcal{F}(x))) - V_\beta^{s-1}(\mathcal{S}(\mathcal{F}(x)))\right] \\
&\leq V_\beta^s(\mathcal{F}(x)) - V_\beta^s(\mathcal{S}(x)),
\end{aligned}
$$

where the first inequality follows since a sub-optimal action in the state $\mathcal{F}(x+1)$ increases the cost-to-go for s time-slots, the second inequality is a consequence of the assumption that the functions $V_\beta^{s-1}(\mathcal{F}(x)) - V_\beta^{s-1}(\mathcal{S}(x))$ are decreasing in x, while the last inequality follows from the fact that a sub-optimal action in the state $\mathcal{S}(x)$ will increase the cost-to-go for s time-slots. Thus we have proved the decreasing property of $D_{s+1}^\beta(\cdot)$ for $x \in \{2, 3, \ldots, B - T + 1\}$, and it remains to show that $D_{s+1}^\beta(1) > D_{s+1}^\beta(2)$.

Once again let $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4$ be the optimal actions at stage s for the states $T, 0, T+1, 1$ respectively. Using the same argument as above (i.e., assuming that the actions taken in stage s at states $T, T+1$ are the same, and the actions taken in the states $0, 1$ are the same), it follows that

$$
\begin{aligned}
&D_{s+1}(1) - D_{s+1}(2) \geq \\
&\qquad (1 + \lambda_O - \beta\lambda_O) - \left(V_\beta^s(T) - V_\beta^s(T+1)\right).
\end{aligned}
$$

However then $V_\beta^s(T) - V_\beta^s(T+1) \leq 1 + \lambda_O - \beta\lambda_O$ (for s stages, apply the same actions for the system starting in state T, as that for a system starting in state $T+1$, and note that the two systems couple at a stage $t-1$, when the latter system hits the state 1 at any stage t. The hitting stage is of course random). This gives us,

$$D_{s+1}(1) - D_{s+1}(2) \geq 0,$$

and thus we conclude that the function $D_{s+1}(x)$ is decreasing for $x \in \{1, 2, \ldots, B\}$. In order to complete the proof, we notice that for $s = 1$, we have,

$$D_1^\beta(x) = 1(x = 1)\lambda_O,$$

and thus the assertion of Lemma is true for $s = 1$. This completes the proof. □ □

Theorem 5. *Consider the single client problem discussed in Section 4.6. There is a threshold policy that is Blackwell optimal [18], i.e., it is optimal for all values of* $\beta \in (\hat{\beta}, 1)$ *for some* $\hat{\beta} \in (0, 1)$*, and is also optimal for the Average cost problem. Thus* $\pi_n^\star(\lambda_E)$ *is of threshold-type and can be obtained in time* $O(B^{E \times Q})$ *via comparing the costs of all threshold-type policies.*

Proof. Fix a q and let $E_i, E_j, i > j$ be two power levels. Without loss of generality, let $\boldsymbol{u}_1 = (q, E_i), \boldsymbol{u}_2 = (q, E_j)$. Clearly $C(\boldsymbol{u}_1) > C(\boldsymbol{u}_2)$ (4.11). In the Bellman equation (4.10), consider the term depending on $\boldsymbol{u}$, i.e. the term $C(\boldsymbol{u}) - P(\boldsymbol{u})D_s^\beta(x)$. For $x, y \in \{1, 2, \ldots, B - T + 1\}$, $x > y$, we have,

$$\begin{aligned}
&C(\boldsymbol{u}_1) - P(\boldsymbol{u}_1)D_s^\beta(x) - \big(C(\boldsymbol{u}_2) - P(\boldsymbol{u}_2)D_s^\beta(x)\big) \\
&- \{C(\boldsymbol{u}_1) - P(\boldsymbol{u}_1)D_s^\beta(y) - \big(C(\boldsymbol{u}_2) - P(\boldsymbol{u}_2)D_s^\beta(y)\big)\} \\
&= \left(P(\boldsymbol{u}_1) - P(\boldsymbol{u}_2)\right)\big(D_s^\beta(y) - D_s^\beta(x)\big)
\end{aligned}$$

$$\geq 0,$$

where the last inequality follows from Lemma 19. Thus it follows that if action $\boldsymbol{u}_1$ is preferred over action $\boldsymbol{u}_2$ for any state x, then $\boldsymbol{u}_1$ will also be preferred over action $\boldsymbol{u}_2$ for any state $y < x$, $y \in \{1, 2, \ldots, B-T+1\}$. Finally note that it follows from the Bellman equation (4.10) and (4.5), that the optimal action for states $x > B-T+1$ is to let $E = 0$ (since any packet that is received will be lost due to buffer overflow). The proof for variations in power levels is similar. Thus it follows from the definition of a threshold policy, that the optimal policy is of threshold type.

Finally note that the statement regarding Blackwell optimality follows from the result in the above paragraph, and because the state-space is finite. □ □

We note that the computational complexity of obtaining the optimal threshold policy, $O(B^{E\times Q})$, is polynomial in B, the buffer size. However the computational complexity of policy iteration is $O(2^B)$, and thus using policy iteration is infeasible for large buffer sizes, while the search for the optimal threshold policy is still feasible. Thus Theorem 5 offers computational advantages also.

4.8 Solution of Primal MDP

We now present the solution of the Primal Problem.

Lemma 20. $D(\lambda_E) = \sum_n V_n(\lambda_E) - \lambda_E \bar{E}$.

Proof. Let $\pi^\star(\lambda_E) := \otimes \pi_n^\star(\lambda_E)$ be the policy obtained by following the policy $\pi_n^\star(\lambda_E)$ for each client n. Then from the definition of dual function, Lagrangian (4.2), cost associated with a policy π (4.8) and Lemma 17, we have

$$\mathcal{L}(\pi, \lambda_E) \geq D(\lambda_E) \geq \sum_n V_n(\lambda_E, \pi) - \lambda_E \times \bar{E}. \tag{4.14}$$

However since the policy $\pi^\star(\lambda_E)$ is stationary, (all the $\liminf$ and $\limsup$ become $\lim$ in the definition of its Lagrangian, and associated rewards in the single-client problem change to $\lim$), we have that

$$\mathcal{L}(\pi^\star(\lambda_E), \lambda_E) = \sum_n V_n(\lambda_E) - \lambda_E \times \bar{E},$$

which, along with (4.14) gives us $D(\lambda_E) = \sum_n V_n(\lambda_E) - \lambda_E \bar{E}$. □ □

Theorem 6. *Consider the Primal MDP* (4.1) *and its associated dual problem defined in* (4.3). *There exists a price $\lambda_E^\star$ such that $(\pi^\star(\lambda_E^\star), \lambda_E^\star)$ is an optimal primal-dual pair and thus the policy $\pi^\star(\lambda_E^\star)$ solves the Primal MDP.*

Proof. We observe that there is a one-to-one correspondence between any stationary randomized policy, and the measure it induces on the state-action space, and thus the Primal MDP can be posed as a linear program [2, 20]. Thus it follow from Slater's condition [15], that for the Primal MDP, strong duality holds if there exists a policy π that satisfies the constraints $\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\sum_n \sum_s E_n(s) < \bar{E}$. However the policy which never schedules any packets incurs a net power expenditure of 0, and thus the Slater's condition is true for the Primal MDP if $\bar{E} > 0$. The claim of the Theorem then follows from Lemma 19. □ □

We note that the policy $\pi^\star(\lambda_E^\star)$ is a decentralized policy. That is, the decision to choose the video-quality and power-level at each time t for client n, i.e., $(q_n(t), E_n(t))$ can be taken by client n itself, and doesn't require the AP to coordinate the clients. Thus a client n need not know the state values of other clients, $l_m(t)$ for $m \neq n$, nor does the AP need to know the values of $l_n(t)$. Thus the policy is easy to implement.

4.8.1 Obtaining $\lambda_E^\star$ Iteratively in a Decentralized Fashion

We note that in order to implement the optimal policy $\pi^\star(\lambda_E^\star)$ as in Theorem 8, we need to find the optimal value of the price $\lambda_E^\star$. We will iterate on the price λ_E using the sub-gradient method [80], and since the problem is concave, the price will converge to the optimal value $\lambda_E^\star$. Moreover the iterations involving price-updates are decentralized, i.e., the clients need only the knowledge of the current price λ_E for the iteration.

Now since $D(\lambda_E) = \mathcal{L}(\pi^\star(\lambda_E), \lambda_E)$, we have,

$$\frac{\partial D}{\partial \hat{\lambda}_v} = \bar{E} - \mathbb{E}_{\pi^\star(\lambda_E)} \sum_n \tau(n, \pi^\star(\lambda_E)), \tag{4.15}$$

where $\mathbb{E}_{\pi^\star(\lambda_E)} \sum_n \tau(n, \pi^\star(\lambda_E))$ is the expected cost incurred on the power over all the users. This is the total "congestion" at the AP. The iteration for λ_E is,

$$\lambda_E^{k+1} = \lambda_E^k - \alpha_k g_k,$$

where d_k is the sub-gradient evaluated in (4.15).

4.9 Fading Channels

The results in the previous sections can be extended in a straight forward manner to the case of fading channels. Let the channel conditions for client n be described by a Markov process evolving on finitely many states $\{1, 2, \ldots, C_n\}$ having a transition matrix Π_n. The state of client n is described by the vector $\boldsymbol{x}_n(t) := (l_n(t), c_n(t))$, where $l_n(t)$ is the play-time duration of the packets present in the buffer at time t, and $c_n(t)$ is the channel condition at time t. If the client n is scheduled a packet transmission of quality q at an power E at time t, then the

system state at time $t+1$ is $(\mathcal{S}(l(t)), \tilde{c})$ with a probability $P_{n,c_n(t)}(q, E)\Pi(c_n(t), \tilde{c})$, while it is $(\mathcal{F}(l(t)), \tilde{c})$ with a probability $P_{n,c_n(t)}(q, E)\Pi(c_n(t), \tilde{c})$.

However now the cost associated to an action $\boldsymbol{u}$ will also depend on the channel condition, i.e.,

$$C_c(\boldsymbol{u}) := \lambda_E E + P_c(l, E)\lambda_q, \tag{4.16}$$

and a threshold policy will have a threshold structure for each value of channel condition (as defined in Section 4.6).

5. INDEX POLICIES FOR OPTIMAL MEAN-VARIANCE TRADE-OFF OF INTER-DELIVERY TIMES IN SINGLE-HOP NETWORKS*

5.1 Overview

A problem of much current practical interest is the replacement of the wiring infrastructure connecting approximately 200 sensor and actuator nodes in automobiles by an access point. This is motivated by the considerable savings in automobile weight, simplification of manufacturability, and future upgradability.

A key issue is how to schedule the nodes on the shared access point so as to provide regular packet delivery. In this and other similar applications, the mean of the inter-delivery times of packets, i.e., throughput, is not sufficient to guarantee service-regularity. The time-averaged variance of the inter-delivery times of packets is also an important metric.

So motivated, we consider a wireless network where an Access Point schedules real-time generated packets to nodes over a fading wireless channel. We are interested in designing simple policies which achieve optimal mean-variance tradeoff in interdelivery times of packets by minimizing the sum of time-averaged means and variances over all clients. Our goal is to explore the full range of the Pareto frontier of all weighted linear combinations of mean and variance so that one can fully exploit the design possibilities.

We transform this problem into a Markov decision process and show that the problem of choosing which node's packet to transmit in each slot can be formulated

---

*Reprinted with permission from "Index policies for optimal mean-variance trade-off of inter-delivery times in real-time sensor networks" by Rahul Singh, Xueying Guo and P.R. Kumar, INFOCOM 2015, Copyright 2015, IEEE.

as a bandit problem. We establish that this problem is indexable and explicitly derive the Whittle indices. The resulting Index policy is optimal in certain cases. We also provide upper and lower bounds on the cost for any policy. Extensive simulations show that Index policies perform better than previously proposed policies.

5.2 Introduction

Traditionally, throughput and delay have been used as performance metrics to judge quality of service (QoS) [24, 35, 46, 77, 100, 117, 120]. The steady-state variance of inter-delivery times of packets is considered as a measure of service regularity in [62]. Motivated by cyber-physical systems applications serving sensors, we address the problem of achieving an optimal "mean-variance trade-off" in the inter-delivery times of packets of N clients sharing K channels.

We consider an access point with K channels shared by N clients. The clients desire a high throughput with high service regularity. We can associate a reward function $\frac{\theta_i}{\bar{D}_i} - var(D_i)$ with client i, where θ_i is the parameter that client i uses to tune its trade-off between its throughput $\frac{1}{\bar{D}_i}$ (where $\bar{D}_i$ is the mean inter-delivery time between packets of client i) and the service regularity $var(D_i)$, the variance of the inter-delivery times for client i. By varying θ_i one can explore the full range of design freedom along the Pareto frontier of all mean-variance tradeoffs. In summary, the net function which captures the trade-off is,

$$\sum_{i=1}^{N} R_i \left(\frac{\theta_i}{\bar{D}_i} - var(D_i) \right),$$

where $R_i > 0$ is the weight attached to client i, and θ_i is a tunable parameter permitting full exploration of the Pareto frontier.

Our contributions can be summarized as follows. We show how one may obtain

tractable decoupled solutions for the problem of scheduling the clients by addressing it as a Restless Multi-Armed Bandit Problem [111]. In particular we obtain the Whittle indices in a closed form, which yields a very elegant solution based merely on comparing the indices of the clients. We also derive upper bounds on the achievable performance of any policy. Simulation results show that the performance of the obtained Index policy is very close to optimal.

5.3 Related Works

The steady-state variance of the inter-delivery times of packets of clients as a measure of service regularity has been considered in [62]. References [62] and [61] consider the scenario where multiple queues are sharing a server and deal with the problem of stabilizing the queues while ensuring an optimal delay and service regularity. [88,97] perform an analysis of the pathwise starvations in service for the case of a single-hop multi-user wireless network.

A detailed introduction to Restless Multi-Armed Bandit Problems (RMBP) can be found in [40]. RMBP and its relaxation were first introduced in [111]. The RMBP model has been used earlier in works such as [66], which considered the problem of choosing an appropriate channel for up and downlink transmissions in multichannel access. Reference [8] is another notable work which uses the RMBP model and derives index policies for optimizing convex holding costs in a multiclass queue.

We also note that optimality of Index policies has been established in certain cases as the population of arms goes to infinity [109] and extensive simulations have shown that Index policies have "good" performance even in the finite population regime [8], [57]. References [32,45,54,74] consider minimization of variance as an objective in Markov Decision Process.

5.4 System Model

We consider the situation where time has been discretized into slots, and the duration of a slot corresponds to the time taken to attempt a packet transmission. Each client is assumed to have one packet at the beginning of each slot. In each slot, a scheduler chooses K out of the N clients, and attempts to deliver their packets. Channel unreliability is modeled by supposing that if client i is served in slot t, then the packet is delivered with probability p_i, independent of the past attempts. Moreover the service times are independent across clients. The scheduler has to choose the K clients transmitted in each slot so as to maximize the reward function,

$$\sum_{i=1}^{N} R_i \left(\frac{\theta_i}{\bar{D}_i} - var(D_i) \right), \tag{5.1}$$

where $\bar{D}_i$ and $var(D_i)$ are the mean and variance of the inter-delivery times of packets for client i in the steady state distribution.

5.5 Markov Decision Process Formulation

The system state at time t is given by the vector $\mathbf{s}(t) := (s_1(t), \ldots, s_N(t))$, with $s_i(t)$ denoting the time slots elapsed between the latest delivery of a packet of client i, and t. Because time is discretized, the state vector $\mathbf{s}(t)$ is updated only at the beginning of slot t, and remains unchanged within the slot. The state thus

evolves as,

$$s_i(t+1) = \begin{cases} s_i(t)+1 \text{ if no packet of client } i \text{ is} \\ \text{delivered in slot } t, \\ 0 \text{ if a packet of client } i \text{ is delivered in} \\ \text{slot } t. \end{cases}$$

The Access Point (AP) takes a decision at the beginning of the slot t to grant channel access to K clients by choosing a control $\mathbf{u}(t) \in \{0,1\}^K, \sum_i^N u_i(t) = K$, where $u_i(t) = 1$ implies that client i will be granted channel access in slot t. The decision can be based on the entire past history of the system up to time t.

The "reward earned" at time t when the system is in state $\mathbf{s}$ is given by

$$\sum_{i=1}^{N} R_i \left(\theta_i 1\left(s_i = 0\right) - s_i\right),$$

and thus is solely a function of the system state $\mathbf{s}$. With this set-up, the process $\mathbf{s}(t)$ becomes a controlled Markov process.

For a positive discount factor $\beta < 1$, the β-discounted optimization problem is to design control policy $\mathbf{u}(t)$ so as to maximize the expected infinite horizon discounted reward,

$$\liminf_{T\to\infty} \mathbb{E} \sum_{t=0}^{T} \beta^t \left(\sum_{i=1}^{N} R_i \left(\theta_i 1\left(s_i = 0\right) - s_i\right) \right). \tag{5.2}$$

Similarly the average reward problem is to maximize the expected infinite horizon

time-average reward,

$$\liminf_{T\to\infty} \mathbb{E}\frac{1}{T}\sum_{t=0}^{T}\left(\sum_{i=1}^{N} R_i\left(\theta_i 1\left(s_i=0\right)-s_i\right)\right). \tag{5.3}$$

It is easily verified that the above reward function reduces to,

$$\sum_{i=1}^{N} R_i\left(\frac{\theta_i}{\mathbb{E}\left(D_i\right)}-\mathbb{E}\left(\frac{D_i\left(D_i+1\right)}{2}\right)\right), \tag{5.4}$$

and thus differs slightly from the original reward function (5.1).

5.6 Whittle Index

We will pose the MDP of the previous section as a Restless Multiarmed Bandit Problem (RMBP). First we briefly describe the RMBP. A detailed discussion can be found in [40, 111].

Consider a bandit which has N arms modeled as Markov processes. At each time a player can choose to play any $K < N$ arms and collect a reward from each arm, where the reward is a function of the current state of the arm that is played. The time evolution of each arm depends on whether it was chosen to play or not; thus the bandits (arms) are "restless" and evolve even if they are not played. The player has to choose the K arms to play at each time, so as to maximize the expected reward.

A "Whittle" policy, or "Index-based" policy, for the RMBP, calibrates each of the N arms by deriving N positive functions (called "index functions") $W_i(\cdot), i = 1,\ldots,N$, which are defined for each possible value that the state of arm i can assume. At time t the policy simply chooses to play the K arms having the K largest values of $W_i(s_i(t))$. After a re-labeling so that $W_1(s_1(t)) \geq W_2(s_2(t)) \geq W_N(s_N(t))$,

the choices at time t are

$$u_i(t) = \begin{cases} 1 \text{ for } i = 1, 2, \ldots, K, \\ 0 \text{ otherwise.} \end{cases}$$

The derivation of the functions $W_i(\cdot)$ follows the following procedure. Each arm is considered in isolation from the rest of the arms, and the reward function is now modified so that the player receives, in addition to the original reward of the arm, a "subsidy" each time that he chooses not to play the arm (chooses "passive action"), and the goal once again is to maximize the average reward. After having solved this problem, let us denote by $\Pi(w)$ the set of states that an optimal policy chooses to not play arm (stay passive). Then the arm is said to be *indexable* if for any two values of subsidies w_1, w_2, we have $w_1 > w_2 \implies \Pi(w_2) \subseteq \Pi(w_1)$, and the original MDP is said to be indexable if all the N arms are indexable. In case the MDP is indexable, the *Whittle Index* as a function of the state value s is defined as the smallest value of subsidy that makes an optimal policy choose the passive action when the client is in state n, i.e.,

$$W(n) = \inf\{w : n \in \Pi(w)\}. \tag{5.5}$$

Thus, the Whittle index measures, in a sense, the "value" of an arm as a function of the present state, and the Whittle or Index policy chooses those K arms which have the highest value amongst the N arms.

5.7 The Client Scheduling Problem is Indexable

We will consider the β-discounted MDP, show that it is indexable and derive the corresponding Whittle index. The results for the average reward MDP will be

obtained by letting $\beta \to 1$. We begin with a brief description of the single-arm β discounted reward problem.

Consider the following single client β discounted bandit problem parametrized by w and β. The subscripts are suppressed for convenience since the discussion below applies to each of the N clients. Thus $s(t), p$ are used in place of $s_i(t), p_i$.

There is a single client, whose state at time t, $s(t)$, is the time-elapsed-since-last-packet-delivery. At each time-slot, we can choose from the following two control actions: either attempt the transmission of a packet for it (active), or stay idle (passive). The reward earned at time t is $= -Rs(t) + w + R\theta 1\{s(t) = 0\}$ if the client chooses the passive action of not transmitting, while a reward of $-Rs(t) + R\theta 1\{s(t) = 0\}$ is earned if client chooses the active action of transmitting. If the action at time t is active, then $s(t+1)$, the state at time $t+1$, becomes 0 with probability p, and $s(t)+1$ with probability $1-p$. If the action at time t is passive, then $s(t+1) = s(t) + 1$. The costs are additive over time after discounting by a factor β^t. A policy whether to be active or remain passive at time t when the system state at time t is $s(t) = s$.

We will prove that there is an optimal policy which is of **threshold type**, i.e. there is a threshold "elapsed time since last delivery" T (which depends on β, w, p), such that the policy which keeps the client passive in slot t if $s(t) < T$, and active if $s(t) \geq T$, is optimal.

By $c_i(T)$ we will denote the β-discounted reward earned by a policy when the system starts with an initial state value of i at time 0, and the policy with threshold at T is used. Let τ_i be the first time that state i is hit, i.e. $\tau_i = \min\{t \geq 1 : s(t) = i\}$. By "reward earned in the cycle $i \to j \to 0 \to i$" we will mean the reward earned by the system starting in state i in the time slots $0, \ldots, \tau_{i-1}$, while operating under the policy with threshold at j. Expressions involving reward-functions belonging to a

single value of threshold are at times not mentioned as a function of threshold. X_p is a random variable that is geometrically distributed with parameter p. Also, we define $X := \mathbb{E}\beta^{X_p}$ and $Y := \mathbb{E}X_p\beta^{X_p}$.

Lemma 21. *Consider the single client β discounted MDP.*

1. *$c_i(i+1) - c_i(i)$ is a linear increasing function of the subsidy w for all $i \geq 0$ It is strictly negative when $w = 0$.*

2. *For each $n \geq 0$, there exists a unique value of the subsidy, denoted $W(n)$, such that $c_n(n+1) = c_n(n)$.*

3. *$W(n) \geq W(n-1)$; thus $W(n)$ form an increasing sequence.*

4. *For all values of thresholds T, if $j > i \geq T$, then $c_i(T) > c_j(T)$.*

Proof. For $T \geq 0$, the infinite horizon discounted reward earned starting in state i and following a policy with threshold $T + i$ is,

$$\begin{aligned} c_i(i+T) = w\sum_{j=0}^{T-1}\beta^j - \sum_{j=0}^{T-1} R\left(i+j\right)\beta^j + R\beta^T\left[\mathbb{E}\left[-\sum_{j=0}^{X_p-1}\left(i+T+j\right)\beta^j\right]\right] \\ + \beta^T\left(\mathbb{E}\beta^{X_p}\right)\left[R\theta + \sum_{j=0}^{i}\left(w - Rj\right)\beta^j\right] + \beta^{T+i}\left(\mathbb{E}\beta^{X_p}\right)c_i(i+T). \end{aligned}$$

Thus $c_i(i+T)$ depends on w as,

$$\begin{aligned} &\left[w\sum_{j=0}^{T-1}\beta^j + w\beta^T\left(\mathbb{E}\beta^{X_p}\right)\sum_{j=0}^{i-1}\beta^j\right] \Big/ \left[1 - \beta^{T+i}\left(\mathbb{E}\beta^{X_p}\right)\right] \\ =& w\left[\frac{1-\beta^T}{1-\beta} + \beta^T\frac{p\beta}{p\beta+1-\beta}\cdot\frac{1-\beta^i}{1-\beta}\right] \Big/ \left(1 - \beta^{T+i}\frac{p\beta}{p\beta+1-\beta}\right) \\ =& \frac{w\left[1-\beta+p\beta-\beta^T(1-\beta+p\beta^{i+1})\right]}{(1-\beta)\left(1-\beta+p\beta-\beta^{T+i+1}p\right)}. \end{aligned} \tag{5.6}$$

Thus $c_i(i+1) - c_i(i)$ depends on w as, $\frac{w(1-\beta)(1-\beta+p\beta)}{(1-\beta+p\beta-p\beta^{i+1})(1-\beta+p\beta-p\beta^{i+2})}$, which is linear and increasing in w.

Now we consider the case when $w = 0$. If C_1 is the cost of cycle $i \to i \to 0 \to i$, then it follows via a simple coupling argument that the cost of cycle $i \to i+1 \to 0 \to i+1$, denoted C_2, is given by,

$$C_2 = -Ri + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j,$$

and thus to prove the second result of the first statement, we only have to show that

$$\frac{C_1}{1-\beta^i X} - \frac{-Ri + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p-1} \beta^j}{1 - \beta^i \beta X} > 0.$$

This is equivalent to showing that,

$$C_1 > -Ri \cdot \frac{1-\beta^i X}{1-\beta} - R\beta \left(\mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j \right) \cdot \frac{1-\beta^i X}{1-\beta}$$

We observe that $-Ri \cdot \frac{1-\beta^i X}{1-\beta} - R\beta \left(\mathbb{E} \sum_{j=0}^{X_p-1} \beta^j \right) \cdot \frac{1-\beta^i X}{1-\beta}$ is the reward earned over the cycle $i \to i \to 0 \to i$ if one were to modify the original cost function and instead charge a penalty of $-Ri$ for value of states $s(t) \leq i$ and a penalty of $-Rs(t)$ if $s(t) > i$. However since the original reward function is $= -Rs(t) + R\theta \mathbf{1}\{s(t) = 0\}$ (note that $w = 0$), a simple coupling argument shows that the reward earned is lower with the modified function. This completes the proof of first statement.

Note that from the first statement it follows that $c_n(n+1) - c_n(n)$ is a linear increasing function of w which is less than 0 at $w = 0$. Hence there exists a value of w such that the function $c_n(n+1) - c_n(n)$ vanishes, and moreover vanishes at

a unique point since the slope of this function is strictly positive. This value of w, where the function $c_n(n+1) - c_n(n)$ vanishes, is $W(n)$.

Let C_1, C_2 be the costs of cycles $n \to n \to 0 \to n$ and $n \to n+1 \to 0 \to n$. It is seen that,

$$c_n(n) = \frac{C_1}{1-\beta^n X}, \quad c_n(n+1) = \frac{C_2}{1-\beta^n \beta X}. \tag{5.7}$$

Using a coupling argument we obtain,

$$C_2 = (W(n) - Rn) + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j. \tag{5.8}$$

Combining (5.7),(5.8) and the fact that for $w = W(n)$ we have $c_n(n) = c_n(n+1)$,

$$\frac{C_1}{1-\beta^n X} = \frac{(W(n) - Rn) + \beta C_1 - R\beta \mathbb{E} \sum_{j=0}^{X_p-1} \beta^j}{1-\beta^{n+1} X}, \text{or ,}$$

$$C_1 (1-\beta) = \left(W(n) - Rn - R\beta \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j \right) (1 - \beta^n X). \tag{5.9}$$

Now let us check if under the value of subsidy set to $W(n)$, we have $c_{n-1}(n) > c_{n-1}(n-1)$. If this is the case, then from the first statement of this lemma, we will deduce that $W(n-1) < W(n)$. Now, $c_{n-1}(n) > c_{n-1}(n-1)$ is equivalent to showing

$$\frac{W(n) - R(n-1) + \beta C_1 - \beta^n X \left(W(n) - R(n-1)\right)}{1-\beta^n X} > \frac{C_1 + R\mathbb{E} \sum_{j=0}^{X_p-1} \beta^j - \beta^{n-1} X \left(W(n) - R(n-1)\right)}{1-\beta^{n-1} X}.$$

After some algebraic manipulations and using (5.9) it can be shown that proving the above inequality is equivalent to proving $X > 0$, which indeed is true. This completes the proof of third statement.

For the fourth statement, using a coupling argument, we obtain, $c_j(T) = c_i(T) - R(j-i)\sum_{j=0}^{X_p-1}\beta^j$, and hence $c_j(T) < c_i(T)$. □

Lemma 22. *Let the subsidy be $w = W(n)$. Then for the single client β discounted MDP,*

1. $c_i(n) = c_i(n+1), \forall i \geq 0$.

2. $c_{i-1}(n) \geq c_i(n), \forall i \geq 1$.

Proof. Firstly recall that for subsidy $= W(n)$, $c_n(n) - c_n(n+1) = 0$. Thus for $i = 0, 1, \ldots, n-1$,

$$c_i(n) - c_i(n+1) = \beta^{n-i}\left(c_n(n) - c_n(n+1)\right) = 0. \tag{5.10}$$

For $i \geq n+1$,

$$c_i(n+1) - c_i(n) = \beta X\left(c_0(n+1) - c_0(n)\right) = 0,$$

where the last equality follows from (5.10). This proves the first statement.

To prove the second result, consider the following cases:

- For $i > n$, Lemma 21 implies that the inequality is true.

- For $2 \leq i \leq n$, denote d_i as the cost incurred in the cycle $n \to 0 \to i-1$. Then both $c_i(n)$, and $c_i(n+1)$ can be derived in terms of d_i. When subsidy is equal to $W(n)$, we have $c_i(n) = c_i(n+1)$, i.e.,

$$-\beta^{n-i}(1-\beta)d_i = W(n)\left(\beta^n X - \beta^{n-i}\right) + R\beta^{n-i}n - \beta^n X i \tag{5.11}$$

$$+ R\frac{\beta^{n-i}}{1-\beta}\left(\beta(1-X) - \beta^{i+1}X + \beta^{n+1}X^2\right), \qquad (5.12)$$

where the first equality follows from statement 1. Similarly, $c_{i-1}(n) - c_i(n) \geq 0$ is equivalent to

$$\sum_{j=0}^{n-i-1}\left(W(n) - Ri - Rj\right)\beta^j + \beta^{n-i}d_i \geq \sum_{j=0}^{n-i}\left(W(n) - Ri + R - Rj\right)\beta^j + \beta^{n-i+1}d_i$$
$$- \beta^n X\left(W(n) - Ri + R\right),$$

i.e.,

$$-\beta^{n-i}(1-\beta)d_i + R\frac{1-\beta^{n-i}}{1-\beta} + \left(W(n) - nR + R\right)\beta^{n-i} - \beta^n X\left(W(n) - Ri + R\right) \geq 0,$$

or,

$$\frac{\left(1-\beta^n X\right)\left(\beta^i - \beta^{n+1}X\right)}{\beta^i\left(1-\beta\right)} \geq 0,$$

where the second-last equivalence follows from (5.11). We note that the last inequality holds trivially for all $\beta \in (0,1)$ and hence the statement 2 holds for $i = 2, \ldots, n$.

- $i = 1$. We compare the cost incurred by the system starting in state 0 over the cycle $0 \to n \to 0$ (say C_0) with the cost incurred over the cycle $j \to n \to 0 \to j$ when starting in state j (denoted C_j) via coupling the processes associated with the two systems constructed on the same probability space. Clearly $C_0 > C_j$. Thus $c_0(T) > c_j(T)$ for any value of threshold T.

□

Lemma 23. *The function $w+p\beta\left(c_i(T)-c_0(T)\right)$ (which depends on w,i,T) is linear, increasing in w. Also,*

$$W(n)+p\beta\left(c_{n+1}(n)-c_0(n)\right)=0 \text{ for } n=0,1,\ldots. \tag{5.13}$$

Proof. We consider the following cases:

i) For $i \leq T$, it follows from (5.6) that the function $w+p\beta\left(c_i(T)-c_0(T)\right)$ depends on w as

$$\frac{1-\beta-p\beta+p\beta^{T-i+1}}{1-\beta+p\beta-p\beta^{T+1}}w. \tag{5.14}$$

We have $1-\beta+p\beta-p\beta^{T+1}>0, \forall \beta<1$. Also, $1-\beta-p\beta+p\beta^{T-i+1} \geq 1-2\beta+\beta^{T-i+1}>0$ since the function

$$1-2\beta+\beta^k \geq 0, \forall k>1, \beta \in (0,1).$$

Thus, in the expression (5.14) the coefficient of w is positive.

ii) For $i \geq T+1$, we have,

$$c_i(T)=\mathbb{E}\sum_{j=0}^{X_p-1}\beta^j\left(-i-j\right)+Xc_0(T).$$

The dependence of $c_0(T)$ on w can be obtained from (5.6). Combining, $w+p\beta\left(c_i(T)-c_0(T)\right)$ depends on w as,

$$\frac{1-\beta}{1-\beta+p\beta-p\beta^{T+1}}w,$$

which has a positive slope with respect to w.

This completes the proof of first statement. Note that for $w = W(n)$, we have

$$c_n(n+1) = c_n(n).$$

This implies

$$\begin{aligned} -Rn + W(n) + \beta c_{n+1}(n+1) &= -Rn \\ &\quad + \beta\left(pc_0(n) + (1-p)c_{n+1}(n)\right) \text{ i.e.} \\ W(n) + \beta c_{n+1} &= \beta\left(p(c_0 + (1-p)c_{n+1}\right) \text{ and so} \\ W(n) + p\beta\left(c_{n+1} - c_0\right) &= 0. \end{aligned}$$

Above, in the second implication, we have used the first statement of Lemma 22 to remove the dependence of $c_i(\cdot)$ on the threshold values. □

Theorem 7. *For the β-discounted MDP with subsidy $w \in [W(n), W(n+1))$, the policy with threshold at n is optimal. Thus the MDP is indexable and $W(n)$ is the Whittle index when the state is n.*

Proof. Fix a $w \in [W(n), W(n+1))$. If the policy is indeed optimal, then the Dynamic Programming optimality equation would be satisfied. Hence we only need to verify the inequality

$$\begin{aligned} -Ri + w + \beta c_{i+1} &\geq -Ri + \beta\left[(1-p)\, c_{i+1} + pc_0\right], \\ &\quad \text{for } i = 0, 1, \ldots, n, \\ \text{or, equivalently,} \quad w + \beta p\left(c_{i+1} - c_0\right) &\geq 0, \end{aligned} \tag{5.15}$$

with strict inequality holding if $w \in (W(n), W(n+1))$, and equality holding for

$i = n, w = W(n)$. Similarly for $i = n+1, n+2, \ldots$ we have to verify the inequality

$$w + \beta p \left(c_{i+1} - c_0\right) \leq 0. \tag{5.16}$$

We will first prove (5.15). We use superscripts to distinguish between costs c_i calculated under different values of subsidy. We have,

$$\begin{aligned}
&w + \beta p \left(c_{i+1}^{w} - c_0^{w}\right) \geq W(n) + \beta p \left(c_{i+1}^{W(n)} - c_0^{W(n)}\right) \\
&= p\beta \left(c_0^{W(n)} - c_{n+1}^{W(n)}\right) + p\beta \left(c_{i+1}^{W(n)} - c_0^{W(n)}\right) \\
&= p\beta \left(c_{i+1}^{W(n)} - c_{n+1}^{W(n)}\right) \\
&\geq 0,
\end{aligned}$$

where the first inequality and equality follow from Lemma 23, and the last inequality follows from Lemma 22.

To prove (5.16) we have,

$$\begin{aligned}
w + \beta p \left(c_{i+1}^{w} - c_0^{w}\right) &\leq W(n+1) + \beta p \left(c_{i+1}^{W(n+1)} - c_0^{W(n+1)}\right) \\
&= p\beta \left(c_0^{W(n+1)} - c_{n+2}^{W(n+1)}\right) \\
&\quad + p\beta \left(c_{i+1}^{W(n+1)} - c_0^{W(n+1)}\right) \\
&= p\beta \left(c_{i+1}^{W(n+1)} - c_{n+2}^{W(n+1)}\right) \\
&\leq 0,
\end{aligned}$$

where first two steps follow from Lemma 23, and the last inequality follows from Lemma 22. This completes the optimality of the policy with threshold at $W(n)$.

Following 5.5, the Whittle index for the state n is thus given by

$$\inf\{w : n \in \Pi(w)\} = \inf\{w : w \geq W(n)\} = W(n),$$

where the first equality follows from the first statement of Theorem. □

We now proceed to explicitly derive the values of the indices $W(n)$.

Theorem 8.

$$\begin{aligned}
W(n) &= \frac{p\beta(f_1 - f_2 - f_3 + f_4)}{f_5}, \textit{ where }, \\
f_1 &= \frac{1-\beta^n}{(1-\beta)^2} \cdot ((1-X)\left[n(1-\beta)+\beta\right] - Y(1-\beta)), \\
f_2 &= \frac{\beta(1-\beta^n) - \beta^n n(1-\beta)}{(1-\beta)^2} \cdot (1-X), \\
f_3 &= \frac{1-X}{1-\beta}(1-\beta^n X), \\
f_4 &= \theta\,(1-X), \\
f_5 &= 1 - \beta^n X - p\beta\left(\frac{1-\beta^n}{1-\beta}\right)(1-X) \\
&= \frac{1-\beta}{1-\beta+p\beta}.
\end{aligned}$$

Proof. From (5.13) we have,

$$\begin{aligned}
W(n) &= p\beta(c_0 - c_{n+1}) \\
&= p\beta(c_0 - c_n - \mathbb{E}\sum_{j=0}^{X_p}\beta^j). \qquad (5.17)
\end{aligned}$$

Now,

$$c_0 - c_n = \frac{C_0 - C_n}{1 - \beta^n \mathbb{E}\beta^{X_p}}, \qquad (5.18)$$

where C_0, C_n are the costs over the cycles $0 \to n \to 0$ and $n \to n \to 0 \to n$. We can compute $C_0 - C_n$ as,

$$C_0 - C_n = \left(\mathbb{E} \sum_{j=0}^{X_p - 1} (n+j)\beta^j \right) (1 - \beta^n) \tag{5.19}$$

$$+ \left(\sum_{j=0}^{n-1} (W(n) - j)\beta^j \right) (1 - \mathbb{E}\beta^{X_p}) + \theta \left(1 - \beta^{X_p}\right). \tag{5.20}$$

Combining (5.17,5.18,5.19) and setting $\Delta = \mathbb{E} \sum_{j=0}^{X_p - 1} (n+j)\beta^j$, we have,

$$W(n) = p\beta \left(\frac{\Delta(1-\beta^n) + \left(\sum_{j=0}^{n-1} (W(n) - j)\beta^j \right) \left(1 - \mathbb{E}\beta^{X_p}\right)}{1 - \beta^n \mathbb{E}\beta^{X_p}} \right.$$
$$\left. - \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j + \frac{\theta \left(1 - \mathbb{E}\beta^{X_p}\right)}{1 - \beta^n \mathbb{E}\beta^{X_p}} \right),$$

or,

$$W(n) \left[1 - p\beta \cdot \frac{\left(\sum_{j=0}^{n-1} \beta^j \right) \left(1 - \mathbb{E}\beta^{X_p}\right)}{1 - \beta^n \mathbb{E}\beta^{X_p}} \right] =$$
$$p\beta \left(\frac{\Delta(1-\beta^n) + \left(\sum_{j=0}^{n-1} -j\beta^j \right) \left(1 - \mathbb{E}\beta^{X_p}\right)}{1 - \beta^n \mathbb{E}\beta^{X_p}} \right.$$
$$\left. - \mathbb{E} \sum_{j=0}^{X_p - 1} \beta^j + \frac{\theta \left(1 - \mathbb{E}\beta^{X_p}\right)}{1 - \beta^n \mathbb{E}\beta^{X_p}} \right),$$

which simplifies to,

$$W(n) \cdot f_5 = p\beta(f_1 - f_2 - f_3 + f_4). \tag{5.21}$$

□

Theorem 9. *The Whittle indices for the average cost MDP are given by,*

$$W^{Avg}(n) = \lim_{\beta \to 1} W^{\beta}(n) = nRp \cdot \left(\frac{n}{2} + \frac{1-p}{1+p} + \frac{1}{2} \right) + Rp\theta. \tag{5.22}$$

Proof. The expression (5.22) is easily derived from (5.21). It remains to show that the quantities $W^{\text{Avg}}(n)$ are indeed Whittle indices for the average-cost problem. Fix the subsidy to be w, and without loss of generality let $w \in \left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1) \right)$. Below we use superscripts to exhibit the dependence of the cost on β. Now,

$$\begin{aligned}
&c_0^{\beta}(n) = \\
&\quad \frac{1}{1-\beta^n X} \cdot \left(w \frac{1-\beta^n}{1-\beta} + \frac{\beta(1-\beta^n) - n\beta^{n+1}(1-\beta)}{(1-\beta)^2} - \right. \\
&\quad \left. \beta^n \sum_{j=1}^{X_p - 1} (n+j)\beta^j + \frac{R\theta}{1-\beta^n X} \right), \text{ and so} \\
&\lim_{\beta \uparrow 1} (1-\beta) c_0^{\beta}(n) = \\
&\lim_{\beta \uparrow 1} \left(w \frac{1-\beta^n}{1-\beta^n X} + \frac{\beta(1-\beta^n) - n\beta^{n+1}(1-\beta)}{(1-\beta)(1-\beta^n X)} \right. \\
&\left. -(1-\beta)\beta^n \sum_{j=1}^{X_p - 1} (n+j)\beta^j + \frac{R\theta\,(1-\beta)}{1-\beta^n X} \right) \\
&= w \frac{np}{np+1} + \frac{Rp(n^2+n)}{2(np+1)} + \frac{Rp\theta}{np+1} \qquad (5.23) \\
&< \infty.
\end{aligned}$$

Since for each m, $W^{\beta}(m) \to W^{\text{Avg}}(m)$, it follows from Theorem 8 that there exists a $\beta^{\star}(w)$ such that the policy with the threshold at n is optimal for the single client β-discounted MDP for all $\beta \in (\beta^{\star}(w), 1)$. However since $\lim_{\beta \uparrow 1}(1-\beta) c_0^{\beta}(n)$ exists, the policy with threshold at n is also optimal for the average cost problem. However since w can assume any value in the interval $\left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1) \right)$, the policy with threshold at n is optimal for the average cost MDP for each value

of subsidy $w \in \left(W^{\text{Avg}}(n), W^{\text{Avg}}(n+1)\right)$. Thus,

$$\inf\{w : \text{optimal policy chooses active at } n\} \leq W^{\text{Avg}}(n). \tag{5.24}$$

Similarly, picking subsidy $w < W^{\text{Avg}}(n)$ shows that the active action is not optimal for any value of subsidy $w < W^{\text{Avg}}(n)$. Hence,

$$\inf\{w : \text{optimal policy chooses active at } n\} = W^{\text{Avg}}(n), \tag{5.25}$$

and we obtain that $W^{\text{Avg}}(n)$ are indeed the Whittle indices for the average cost problem. □

We note that the expression (5.23) is the average reward earned under the subsidy w and threshold at n. We will denote this quantity as $C^{Avg}(W, n)$.

5.8 Bounds on Optimal Reward

Lemma 24. *For the average cost MDP, the reward obtained under any policy is upper-bounded by the value of the following optimization problem:*

$$\begin{aligned} &\max \sum_{i=1}^{N} R_i \left[\bar{D}_i^2 + \theta_i \frac{1}{\bar{D}_i}\right] \\ &\textit{such that } \sum_{i=1}^{N} \frac{1}{\bar{D}_i p_i} \leq 1, \bar{D}_i \geq 0, i = 1, \ldots, N. \end{aligned} \tag{5.26}$$

Proof. The random reward earned in time steps $1, 2, \ldots, t$ is given by,

$$C(t) := \sum_{i=1}^{N} \frac{R_i}{t} \left[-\sum_{l=1}^{N_i(t)} D_i(l)^2 + \theta_i N_i(t) \right],$$

where $N_i(t)$ is the number of packets of client i delivered by time t and $D_i(l)$ is

the interdelivery time of l-th packet of client i. Let us assume that the average interdelivery-time for client i under a policy is equal to $\bar{D}_i$. Thus,

$$\begin{aligned}
\liminf_{t\to\infty} \mathbb{E}C(t) &\leq \limsup_{t\to\infty} \mathbb{E}C(t) \\
&\leq \mathbb{E}\limsup_{t\to\infty} C(t) \\
&= \mathbb{E}\limsup_{t\to\infty} \sum_{i=1}^{N} R_i \left[\frac{\sum_{l=1}^{N_i(t)} D_i(l)^2}{t} + \frac{\theta_i N_i(t)}{t} \right] \\
&\leq \sum_{i=1}^{N} R_i \left[\bar{D}_i^2 + \theta_i \frac{1}{\bar{D}_i} \right],
\end{aligned}$$

where the second inequality follows from Fatou's lemma and the last is Jensen's inequality. Thus solving the optimization problem (5.26) gives a lower bound on the performance of any policy. We note that the constraint $\sum_{i=1}^{N} \frac{1}{\bar{D}_i p_i} \leq 1, \bar{D}_i \geq 0$ is simply the capacity of the wireless channel.

□

Next we consider the Lagrangian relaxation of the RMBP [110]. For this, we relax the constraint of choosing K arms at each time, to the constraint that one plays K arms on average, i.e., $\lim_{t\to\infty} \frac{\text{Total numbers of arms played by time } t}{t} = K$. Clearly the maximum possible reward in the relaxed problem is greater than or equal to the reward earned by any policy for the original RMBP. Also since the Index policy is the optimal solution to this relaxed problem ([111]), its value function serves as an upper-bound for the value function of the RMBP.

Lemma 25. *Let $C^{Avg,i}$ be the average reward earned by the policy maximizing the single-client average reward under the subsidy W (5.23). Then the reward for the*

average cost MDP obtained by any policy is less than or equal to,

$$\begin{aligned}
&\inf_{W>0} \sum_{i=1}^{N} C^{Avg,i}(W) - W(N-K) \\
&= \inf_{W>0} \left(\sum_{i=1}^{N} W \frac{n_i p_i}{n_i p_i + 1} + \frac{R_i p_i (n_i^2 + n_i)}{2(n_i p_i + 1)} \right. \\
&\left. + \frac{R_i p_i \theta_i}{n_i p_i + 1} - W\left(N - K\right) \right), \\
&= \inf_{W>0} \left[W \left(\sum_{i=1}^{N} \frac{n_i p_i}{n_i p_i + 1} + K - N \right) + \frac{R_i p_i (n_i^2 + n_i)}{2(n_i p_i + 1)} \right. \\
&\quad \left. + \frac{R_i p_i \theta_i}{n_i p_i + 1} \right],
\end{aligned}$$

where n_i is such that $W \in (W(n_i), W(n_i + 1))$.

5.9 Optimality of Index Policy

Now we consider several special cases of interest.

Theorem 10. *Consider the average cost problem for the case where all the clients are identical, i.e., $R_i \equiv 1$ and $p_i \equiv p$ for all the clients. The index policy is optimal in this case.*

Proof. Firstly we note that in this symmetric case, the Index policy serves the client with the largest value of the state, i.e. the policy is, "largest time-since-last-service-first". We will prove the result only for the case of two clients, each having channel reliability p. The case where there are multiple such clients follows in a straightforward manner.

Consider the time-horizon at t. If (s_1, s_2) is the initial value of the state vector, and $R_t(\mathbf{s})$ is the maximum reward that can be earned when there are t time-slots

to go, then the Dynamic Programming optimality equation becomes,

$$R_t\left[(s_1,s_2)\right] = -\left(s_1+s_2\right) + (1-p)R_{t-1}\left[(s_1+1,s_2+1)\right] + p\max\{R_{t-1}\left[(0,s_2+1)\right], R_{t-1}\left[(s_1+1,0)\right]\},$$

where the optimal action corresponds to the one maximizing the expression on the right hand side. Let us assume without loss of generality that $s_1 < s_2$. Then $R_{t-1}\left[(0,s_2+1)\right] \leq R_{t-1}\left[(s_1+1,0)\right]$, which implies that the optimal action is to serve client 2. □

5.10 Simulations

We have carried out simulations to compare the performance of the optimal policy which was obtained via the Policy Iteration tool-box in Matlab vs. the Index policy which was obtained in Theorem 9. We present three plots in Figures 5.1-5.3. In all the cases considered 2 clients share a single channel. To obtain Figure 5.1, we fix client 1's parameter as $p_1 = .8, \theta_1 = 3, R_1 = 1$, while for client 2 we fix $\theta_2 = 3, R_2 = 1$ and vary p_2 from 0 to 1. For Figure 5.2, we fix Client 1 parameters to be $p_1 = .8, \theta_1 = 3, R_1 = 1$ while for Client 2 we fix $p_2 = .6, R_2 = 1$ and vary the value of θ_2 from 1 to 10. To obtain Figure 5.3, we fix Client 1's parameters as $p_1 = .8, \theta_1 = 5, R_1 = 5$, and for Client 2 we fix the parameters $p_2 = .6, \theta_2 = 5$ while varying the value of R_2.

We observe that Index policy gives near-optimal performance in all the cases.

5.11 Concluding Remarks

We have proposed an analytical framework for exploring the full range of mean vs. variance tradeoffs in inter-delivery times in wireless sensor networks, i.e. Throughput vs. Service Regularity trade-off. The problem can be formulated as

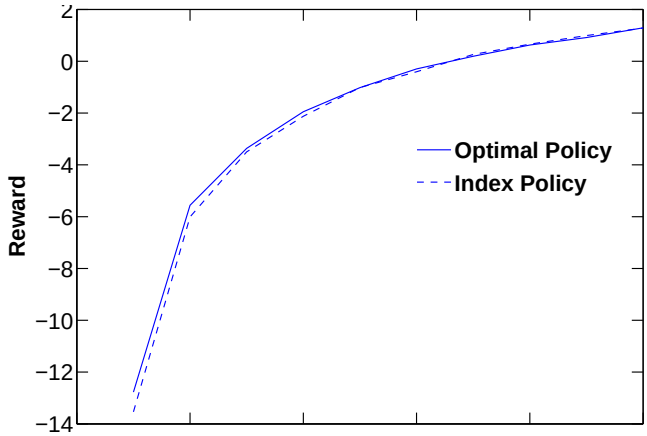

Figure 5.1: Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, \theta_2 = 3, R_2 = 1$, p_2 varying from $.1$ to 1.

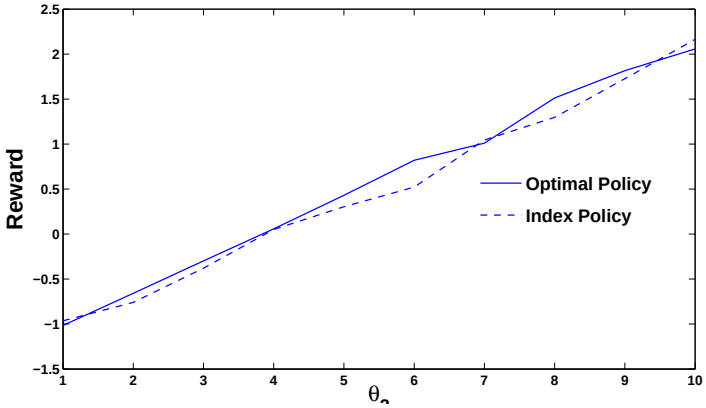

Figure 5.2: Reward Optimal Policy vs.Index Policy for $p_1 = .8, \theta_1 = 3, R_1 = 1, p_2 = .6, R_2 = 1$ while θ_2 varies from 1 to 10.

Restless Multiarmed Bandit Problem and indices can be obtained in closed form. Simulations indicate near-optimal performance of the resulting Index policy.

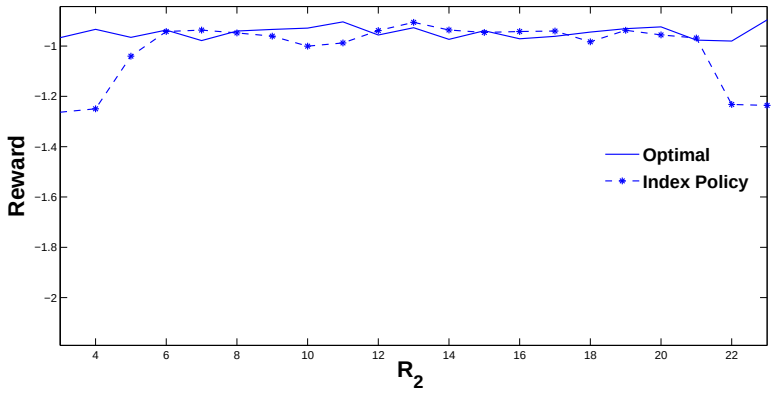

Figure 5.3: Reward Optimal Policy vs. Index Policy for $p_1 = .8, \theta_1 = 5, R_1 = 5, p_2 = .6, \theta_2 = 5$ while R_2 is varied.

6. THE ISO PROBLEM: DECENTRALIZED STOCHASTIC CONTROL VIA BIDDING SCHEMES

6.1 Overview

We will consider a smart-grid connecting various agents, modeled as stochastic dynamical systems, who may be electricity consumers/producers. At each discrete time instant, which may represent a 15 minute interval, they will be drawing/supplying some quantity of electrical energy into the grid. We are given the task of maximizing the total utility of this system subject to the constraint that energy generated at each time equals the energy consumed. On the demand side, the optimal solution specifies an optimal demand response, with, say, consumers shifting their demand to the "energy-rich" time of the day, while maintaining some desirable level of overall service. On the generation side, there may be a mix of power from renewable energy sources as well as fossil fuels. The former, such as solar or wind power, may themselves be stochastic, and only amenable to curtailment but not enhancement, while the latter are more controllable sources though with restrictions on ramping rates and the like. This model also allows modeling of energy storage services who may wish to store energy when it is cheap and supply it when it is expensive. The model can also incorporate "prosumers" who may produce or consume energy depending on environmental conditions and load states. Given the stochastic behavior of the loads, the optimal solution specifies how the power is to be generated in the most efficient manner to balance demand.

This task of mediating between generation and demand has to be accomplished without the need for the agents to communicate amongst themselves about their system states; in fact they should not even need to reveal their individual sys-

tem's dynamics or model or utility or cost functions This mediation task is to be accomplished by an agency called a system operator, which basically obtains the electricity bids by the agents, and eventually declares the market clearing price. In response to this price, the agents submit new bids (see Figure 6.1). We show that a simple iterative procedure yields the optimal solution to the above Independent System Operator (ISO) problem. Thereby we solve a decentralized stochastic control problem with price mediation, but without agents sharing any information even about their individual system models, states or utilities.

6.2 Notation

Throughout, random variables will be denoted by capitals and their realizations in small. Equalities between random variables are to be understood in an *almost sure* sense.

6.3 Introduction

We consider the problem faced by the electricity grid operator, called the Independent System Operator (ISO). In the context where the ISO knows or estimates the net demand of the loads, it is faced with the task of allocating the required power among different generators so that the total cost of production is minimized, and the power flow can be delivered over the network [26, 37, 115][1]. The former problem can be solved via the generators bidding their marginal cost curves, as in a Walrasian auction, and the ISO performing the optimization to obtain and declare the market clearing price. The optimization simply amounts to minimizing a cost function over a simplex [16], and in the convex case the local minimum is indeed the global minimum. This is an exemplary model in which the ISO is able to determine the optimal solutions without the generators revealing their systems.

[1]There are additional aspects such as security against contingencies, etc., that we neglect here.

However the above deterministic static model with a fixed demand is insufficient for the oncoming era when we want to maximize the integration of renewable energy sources into the energy system. Renewable energy sources such as wind and photo voltaic are dynamic and vary unpredictably with time. Thus, modeling generation of renewable power requires a dynamic stochastic system, not a deterministic static system.

Dynamic models can also be used to model features such as ramping constraints that are important for modeling fossil fuel generators that may also supply a portion of the power mix. On the load side, demand response is a strategy of importance in integrating renewables. When trying to employ renewable energy we need to make the level of demand compatible with the availability of renewable energy, in contrast to the traditional scenario where demand is inflexible and supply needs to match whatever demand is. Thus loads are controllable and also need to be modeled. Loads generally have dynamic constraints since some loads such as air conditioners can be deferred for a while but not indefinitely. So they also need to be modeled as dynamic systems. Further, since environmental variables such as temperature are involved, future loads may be uncertain. Also, since economic incentives may be used to shape demand, and human beings may be in the loop, their response may also be uncertain. Hence loads generally also will need to modeled as stochastic dynamic systems.

Such dynamic models can also model storage devices where the state is the amount of energy stored. They can also be used to model prosumers, such as homes with solar panels, which may switch at uncertain times from being consumers to generators. Therefore we model all the agents involved, whether generators or loads or storage devices, as stochastic dynamical systems.

Our goal in operating this system is to maximize total utility, or equivalently

minimize total systemwide cost. There are however several constraints on information sharing that need to be respected in arriving at a solution. An important constraint is that the individual agents, whether loads or generators, may be averse to sharing system states with each other. More fundamentally, they may not even be willing to share their individual system models with each other. Similarly for their individual utility functions. There are several reasons for this, ranging from the competitive nature of commercial enterprises, to protecting privacy of states in the case of consumers.

The overall systemwide optimality of such a system is sought to be achieved by an Independent System Operator (ISO), which plays the role of the mediator. This mediator needs to both determine the optimal demand response over time as well as allocate it over time among the lowest cost generators, all in the face of stochastic uncertainty, and to do so at minimum systemwide cost to all agents. The ISO would like to achieve this through economic mechanisms that do not entail revealing system models or states. In particular the ISO would ideally like to simply determine prices and leave each agent to its own selfish utility maximization as in general equilibrium theory [10].

The fundamental question examined in this section is whether and how this optimality can be attained given stochastic dynamical system models for the agents, and what form the mediation process or tatonnement [79]. Our contribution is to show that there are iterative interaction processes under which the ISO can indeed perform this task. We address the complexity of this task under several scenarios. The complexity is very high in the general case. However, in the case where the agents can be modeled as linear Gaussian stochastic systems and the cost functions are quadratic, we show that a much simpler scheme yields the systemwide global optimum.

6.4 System Model

Consider a smart-grid consisting of M agents, each of which may act as a producer, consumer or even as both, e.g., a "prosumer" such as a home with a solar panel, or a storage device that can absorb power when in charging mode or supply it when discharging. Each such agent is modeled as a stochastic dynamical system. The following are the key ingredients of our system:

1. *Randomness* is modeled through a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The "state of the world" ω lies in the set Ω, and captures "random" phenomena such as unpredicted weather (example the wind-speed), or unexpected events that occur while producing power in power-plants (example coal shortage, or a damaged wind-turbine) etc. The state of the world ω affects the agent i through the random processes $N_i(t)$ and $N_c(t), t = 0, 1, 2, \ldots, T-1$. (Throughout, all functions are assumed to be measurable with respect to $\mathcal{F}$). In the sequel we will regard $N_c(t)$ as a "common" uncertainty that affects all agents, while $N_i(t)$ is a "private" uncertainty specific to agent i. The precise probabilistic assumptions are described in detail in the sequel.

2. *Agents* are modeled as stochastic dynamical systems. As mentioned earlier, each agent may correspond to a producer, consumer, prosumer, or storage. Associated with each agent i is its state at time t, denoted $X_i(t)$, that takes values in some set, and evolves as,

$$X_i(t+1) = f_i^t(X_i(t), U_i(t), N_i(t), N_c(t)), \tag{6.1}$$

where $U_i(t)$ is the amount of electricity supplied (negative if consumed) to the grid by agent i at time t. Note that $U_i(t) \in \mathbb{R}$, and the evolution of the

smart-grid occurs over discrete time-slots. Each such time-slot may represent the 15-minute bidding times of the real-time market implemented by the ISO. The function f_i^t captures the system dynamics corresponding to agent i.

3. *Observations* are available to an agent i at time t. They are modeled as random variables whose realizations are available to the agent i at time t. As will be discussed later, we will partition the observations into a set of common observations, that are observed by all the agents, and a set of private observations that are available exclusively to agent i. A detailed discussion of observation structures is provided in Section 6.8.

4. One-step *Cost function* of an agent i, denoted $c_i(\cdot)$ (or its negative, a one-step utility function $-c_i(\cdot)$), which is a function of the state of agent i, and denotes the cost incurred by the agent i as a function of its state and possibly action in a period. As an example, for the producers, this cost could be composed of several factors such as labor, coal, etc.. For the consumers, this could represent the cost incurred due to the high temperature of house/business facility, or the cost incurred due to a delay in performing a task resulting from non-purchase of electricity.

5. *System Operating Cost* is the expected value of the sum of the finite horizon total costs incurred over the time duration $\{1, 2, \ldots, T\}$ by all the agents, i.e., the quantity,

$$\mathbb{E}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} c_i\left(X_i(t), U_i(t)\right)\right). \tag{6.2}$$

The time horizon T can, for example, be chosen to be 96 which corresponds to one day, with the time slots t corresponding to be the "bidding times"

which have a separation of 15 minutes. Since the grid consists of consumers and producers, the cost (6.2) is the total electricity generation cost minus the utility provided to the consumers.

6. *Power Flow Equations* are a set of algebraic equations that have to be satisfied by the electrical variables, voltage and current magnitudes and phase angles, over the grid at each time t, imposing, for example, some constraints on the quantities $U_i(t), t = 1, 2, \ldots, T-1$. Such equations are derived from the underlying physical phenomena, specifically Kirchoff's laws, together with some constraints on the power transmission lines (such as line capacity) etc. A basic constraint, and one that we will centrally focus on, is that the total generation must equal to total consumption at each time t, leading to the constraint $\sum_{i=1}^{N} U_i(t) = 0$ at each time t.

7. *Independent System Operator (ISO)* is an agency that accepts electricity purchase/sale bids that are submitted by the agents for each time slot $t = 1, 2, \ldots, T-1$. In our model of the ISO, we allow for the agents to iterate on the bids before the market clearing price is declared. Once the iterations have converged, the ISO declares the market clearing prices, and the agents purchase/sell the agreed electrical energies at the prices declared by the ISO.

8. *Bidding Schemes* A typical bidding scheme discussed in the section will involve agents submitting their bids to the ISO, and the ISO declaring market clearing prices. The bid function of agent i corresponding to time t will declare, as a function of its past information, the amount of electricity that agent i will be willing to purchase/generate.

   Depending on the assumptions made upon the system model, we will pro-

pose multiple types of bidding schemes. Below we describe one such bidding scheme, with details of other specific schemes provided in the sequel.

After collecting the bids, the ISO updates the market prices based on the bids it receives. An iteration of *price updates* followed by *bid updates*, continues till the market prices and the bids converge. This entire process can be repeated at each discrete time instant (which could be every 15 mins) in real-time.

6.5 The ISO Problem

With the above set-up in place, the ISO problem is to ensure a systemwide optimization, i.e., minimize the total cost (6.2). The physical laws governing the individual power-plants, wind-farms, and loads, etc., have to be respected as well. Another important constraint is maintaining energy balance in each period. These aspects give rise to constraints, and we arrive at the following constrained stochastic dynamic control problem, which we call the ISO Problem,

$$\begin{aligned}
&\min \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{i=1}^{M} c_i\left(X_i(t), U_i(t)\right)\right\} \\
&\text{such that } \sum_i U_i(t) = 0, t = 1, 2, \ldots, T-1, \\
&\text{and } X_i(t+1) = f_i^t(X_i(t), U_i(t), N_i(t), N_c(t)), \text{ for} \\
&i = 1, 2, \ldots, N, \text{ and } t = 1, 2, \ldots, T-1. \qquad \text{(ISO Problem)}
\end{aligned}$$

The expectation above is taken with respect to the combined uncertainty or "noise" process $N(t) := (N_1(t), N_2(t), \ldots, N_M(t), N_c(t))$.

6.6 Common and Private Observations

The randomness ω manifests itself in the collection of primitive random variables $\{N_i(t)\}_{i=1}^{M}, N_c(t) \quad t = 0, 1, \ldots, T-1$. The process $N_c(t)$ will be assumed to be

observed by all agents, and is thus a common observation, while the process $N_i(t)$ is observed only by agent i. The ISO Problem has to be solved under these observation constraints. As we will see, the decomposition of the observations clarifies the task of constructing the mediation schemes to be followed by the agents and the ISO.

6.7 Illustrative Examples

In this section we provide examples to illustrate how the set up of Sections 6.4 and 6.5 can be utilized to model some of the problems faced by the ISO.

Consider a smart-grid comprised of three agents:

1. Agent $\mathcal{A}_1$ is a coal-plant electricity producer, whose state is described by the speed of the turbine $X_1(t)$. The costs that it incurs at time t can be classified into three types:

   - Ramping cost, which is equal to the square of its ramp-rate at time t, i.e. $(X_1(t) - X_1(t-1))^2$.
   - Coal cost, which is given by the market price of coal $N_1(t)$ times the amount of coal used, i.e. $X_1(t)$.

   The total cost incurred by the producer at time t is simply $X_1(t)N_1(t) + (X_1(t) - X_1(t-1))^2$.

2. The second agent $\mathcal{A}_2$ is a consumer, who wants to maintain the temperature of his house/facility $X_2(t)$ close to some prescribed temperature, say 0 units. Denoting his temperature by $X_2(t)$, it evolves as,

$$X_2(t+1) = X_2(t) + N_2(t) + U_2(t),$$

where $N_2(t)$ is the heat supplied to the facility from sources other than electricity. Suppose that the "discomfort" cost at time t due to a too high/too low temperature is given by $X_2(t)^2$.

3. The third agent $\mathcal{A}_3$ is a wind-farm operator, who owns a wind-farm and a storage facility in which it stores the excess wind energy to be sold at a later time. Thus if $X_3(t)$ is the amount of energy in storage at time t,

$$X_3(t+1) = X_3(t) + \alpha(N_3(t) - U_3(t)),$$

where $N_3(t)$ is the amount of wind-energy that it receives at time t, $U_3(t)$ is the amount of electricity, and $0 < \alpha < 1$ is the efficiency of the storage facility. The cost incurred by it is some function of the state of the turbine. For example if the state of turbine is "broken", then he incurs some maintenance cost, etc. We will denote this cost function by $c(\cdot)$.

Combining, we see that the ISO is given the task of optimizing the cost[2],

$$\mathbb{E}\left\{\sum_t X_1(t)N_1(t) + (X_1(t) - X_1(t-1))^2 + X_2(t)^2 + c(X_3(t))\right\}.$$

6.8 Fundamental Issues

The ISO Problem poses challenges with regard to multiple issues. It is a multi-agent problem subject to constraints. Examples of constraints are power flow equations or privacy constraints. The objectives of the agents are not all aligned and may have conflicts amongst themselves.

[2]From a technical point of view, one can condition this on the past observations, as is standard in stochastic control, to eliminate the dependence on $N_1(t)$.

6.8.1 *Interdependence/ Interconnection of Agents*

First and foremost, the ISO Problem cannot be solved by considering each of the agents separately in isolation from the other agents. This is because the operating cost and the power-flow constraint are a function of the *combined* actions chosen by the agents. For example power balance requires $\sum_i U_i(t) = 0$ for each t. [3]

6.8.2 *Privacy*

The agents may not want to disclose their state values $X_i(t)$. In fact, they may not even want to disclose their *system dynamics*, i.e., the functions f_i^t or the laws of the noise processes $N_i(t)$. In the example presented in Section 6.7, the system dynamics of agent $\mathcal{A}_1$ may depend upon trade secrets of the power-plant, and it may want to keep it secret in order to maintain a competitive edge over other firms. Similarly agent $\mathcal{A}_2$ risks losing its privacy if it reveals the value of its room-temperature $X_2(t)$; for example if the temperature is high, then it may reveal that the occupant may not be in the house. Even if privacy were not an issue, sharing the complete system observation amongst the agents requires huge overhead in terms of communication costs, processing times and constant updates, etc, and may be impossible in practice.

In summary, the agents would like the common observations and knowledge of each others' systems to be as little as possible. Nevertheless the ISO is required in our formulation to minimize the expected value of the sum over all agents of their total cost over a time horizon.

[3] This condition applies even if there are storage units, by taking their power input/output into account in the balance.

6.8.3 *Decentralized Control with Non-Classical Information Structure*

Consider a stochastic dynamical system in which multiple agents (controllers) have access to different sets of observations, and act at multiple times so as to minimize a cost that depends on the system state at each time t. This problem is the core of decentralized stochastic control [113, 114] with non-classical information structures, and is in general a difficult problem. Based on its observations, each agent has an aposteriori belief about the system state $X(t)$, and its control action $U(t)$ may depend on this belief. The key difficulty stems from the fact that since the observation sets are different, the agents have differing beliefs about the state of the system $X(t)$.

In the ISO Problem, clearly, if at time t, each agent i communicates the value of its state $X_i(t)$ to the aggregator, and the aggregator has complete knowledge of each agents' system dynamics (functions f_i^t and laws of processes $N_i(t)$), then the problem reduces to a case of centralized control. However this will generally not be the case because of the privacy constraints imposed in Section 6.8.2.

Thus the ISO Problem lies in the domain of decentralized control [12, 30, 81, 106, 113, 122].

6.8.4 *Information Sharing/ Signaling*

One approach in decentralized stochastic control takes the following two steps [12, 94,106]. First, the information available to different controllers is structured/classified as common and private information [12]. After this, the controllers try to communicate some of their private observations to other agents via some "channel". This channel can be a physical channel, for example a noisy communication channel. Or, in case a communication channel is not present, then, since the evolution of

the dynamical system is affected by the control that is applied by each of the controllers (agents), agent i can use its dynamical system itself as a channel to *signal* its private observation to other agent(s). The agents would then have to design appropriate encoding-decoding schemes for signaling in order to ensure the design of optimal control.

The bidding schemes proposed by us in this section signal the private observation of agents using "market prices" as signals.

6.8.5 Dynamic Market

The nature of the electricity grid is inherently dynamic. Thus the states of agents are continually changing (for example, due to the state of the power generation plant, wind speed, failure of a unit, or temperature of a consumer's building, etc.). Any solution to the ISO Problem necessarily has to accommodate these variations. The operating schemes proposed in this section are adaptive to such dynamics in the system. In fact, in our solution, the agents do not even need to know how many or what other agents are present in the network.

6.8.6 Online Optimization

Any solution to the ISO Problem Problem has to be in real-time, keeping in mind the dynamic nature of the grid (Section 6.8.5) and the fact that the real-time markets operate in time-slots having gaps of 15 min duration. This imposes a constraint due to the computational resources available, and it is important to obtain a solution which is computationally feasible.

6.8.7 Curse of Dimensionality

While the ISO Problem can be viewed as a constrained Markov Decision Process (MDP) [4], the current state of our knowledge does not allow us to handle general

MDPs with different observation patterns and different cost functions. Thus the results encountered in the field of MDP, such as dynamic programming, are not applicable.

Even if we assume that there is a centralized controller (the ISO) that observes the states of agents, the complexity of solving the MDP using Dynamic Programming is proportional to the cardinality of the associated state-space. It suffers from the curse of dimensionality [13]. In our case, the size of the state-space increases exponentially with the number of agents M. Thus a blind application of the results from MDP theory would not lead us too far since the ISO Problem would quickly become intractable as the number of agents is increased. This calls for developing new techniques for solving the stochastic optimization problem.

6.8.8 Big Data: Sufficient Statistics

The complexity of ISO Problem Problem scales with the time horizon T, and the number of agents. Beginning from the time that the grid operation begins, i.e. $t = 0$, each agent continually collects observations over time. If we denote by $\mathcal{I}_i(t)$ the observations collected by the agent i until time t, then the set $\mathcal{I}_i(t)$ increases as time passes. Since the optimal action of agent i at time t is a function of the observations $\mathcal{I}_i(t)$, it has to keep a record of entire past observations that it has received.

However, we would want to know whether it is possible that agents can discard some of these observations, or lossily compress them, while still retaining the ability to make optimal decisions? In other words, is there a function which maps/compresses the observation $\mathcal{I}_i(t)$, such that an optimal control law is a function of the compressed observation? If the answer to the above question is "yes", then we have essentially constructed a sufficient statistic for the ISO Problem [19,

101].

It is a well known fact that for the case of centralized control, the knowledge of the present state of the system suffices as a sufficient statistic, and thus the centralized controller need not remember the values of past system states, or the inputs it applied to the system in the past [101]. However this is not true for the case of decentralized control [12, 114]. In fact, in decentralized control, the structural results imply that the sufficient statistics reside in infinite-dimensional spaces, namely the beliefs of agents about the system states of other agent [12, 116], and also depend upon the policy being implemented by all the agents.

Is it possible to obtain data-reduction of the same scale as in the case of centralized controller? That is, is it possible to construct a policy where it suffices for the agents to only keep track of the values of their own system state, and yet take optimal decisions? We will show that the answer is in the affirmative, and that it is indeed possible to do so under a variety of observation structures. This enables the agents to discard a huge amount of data that is not required for the purpose of control.

6.9 Problem Statements, Key Questions and Goals

Having laid out the key issues in the previous section, we now proceed to formulate the problems that will be solved in the next few sections.

As has been pointed out earlier in Section 6.8.2, if each agent reveals its private observations to other agents or to the ISO, then the problem can be reduced to classical "centralized stochastic control", the solution to which can be obtained in principle via Dynamic Programming. However as discussed earlier, in the power system context, sharing all information involves too much communication and revelation, and thus infeasible, and even if that is somehow accomplished, computing

the optimal centralized solution is computationally infeasible. Nevertheless the goal is to drive the system to such an optimal operation.

Our approach is to optimally coordinate the M dynamic systems through announced "prices". In the power system context, the Independent System Operator (ISO) is indeed the agent specifically assigned to do this. The question therefore is: Can M independent systems be driven to an overall optimal operation through an intermediary such as the ISO making price announcements? As we will show later, the ISO can achieve this solution amongst the agents under appropriate assumptions, by declaring market-prices of the electricity.

The key questions of interest are the following:

1. Is it possible to achieve the exact optimal performance as attained by centralized control despite the fact that the agents do not share their observations, i.e., the system is decentralized, and moreover do not even share the dynamics of their systems or their utility/cost functions?

2. If the answer to the above is "yes", then what kind of schemes achieve optimal centralized performance while still allowing each agent complete confidentiality about its dynamic system model and state?

3. What are the "sufficient statistics"? Is there a scheme where each agent simply keeps track of the value of its present state?

4. How computationally expensive is it?

5. Is the scheme real-time implementable?

We will analyze all these questions, and obtain positive results under various models. We will show that there exist simple "iterative bidding schemes" (IBS) which

yield the same performance as that of the optimal centralized controller under some models to the above formulated ISO Problem.

6.10 Related Works

We note that no similar results appear to be known to the authors for the general decentralized stochastic control problem. Team problems have been extensively studied, for example in [73, 106, 122], but the formulations are very restrictive in the sense that each agent needs to know the system dynamics of the other agents. Even when the models are known, there are still considerable difficulties in decentralized stochastic control. When agents do not share observations, severe complexity can set in, even in an otherwise linear quadratic Gaussian problem, as pointed out by Witsenhausen in his counterexample of a two stage problem [113]. The role of observation, signaling [122], and the trade-off between communication and control are evident from Witsenhausen's counterexample [113]. Reference [30] considers decentralized stochastic control under the restrictive assumption that the interaction between agents is "weak". There are some recent structural results [12] and results regarding sufficient statistics [116] under these restrictive assumptions, moreover the proposed solutions suffer from the curse of dimensionality. Reference [69,81] contains some heuristic approaches. Reference [91] applies progressive hedging to deal with the uncertainties on the production side, though the solution is centralized, and doesn't provide any theoretical guarantees.

As we will show below, the ISO Problem formulated here provides an excellent example of decentralized control systems with non-classical observation patterns in which signaling can successfully result in globally optimum performance. The agents need not signal not only their observations or state values, but in fact even

their individual system dynamics and their individual cost-functions. Each of the algorithms constructs concrete signaling schemes which encode-decode the information required in order to recover the same performance as that of centralized control. From the economics side, this work is an extension of general equilibrium theory [11]. To the author's knowledge there does not appear to be any similar result for coordinating multiple LQG systems or the efficiency of the simplified signaling.

Looked at from the power system end, there have been many efforts since the deregulation of the electricity sector on a market-based framework to clear the system. Ilic et al. [51] proposed a two-layered approach that internalizes individual constraints of market participants while allowing the ISO to manage the spatial complexity. The approximated MPC algorithm is shown to perform well in many realistic applications.

In order to analyze the strategic interactions between the ISO and market participants, game theoretical approaches have been proposed. Zhu et al. [123] use a Stackelberg game framework for economic dispatch with demand response. The approach uses a two person game with the ISO as leader and agents aggregated into second player. The agents change their demand based on price signal so as to maximize their payoff function. The Economic Dispatch (ED) problem considered is a single time interval conventional dispatch without transmission line constraints. Bu and Yu [23] models the interactions between electricity retailers and customers as a Stackelberg game. This work considers the case of a monopoly retailer where observations about customers' utility and consumption pattern are available. Jia and Tong [56] uses a Stackelberg formulation to study the energy consumption scheduling problem for customers who are subjected to a time-varying price which is determined one day ahead of time. The trade-off

between consumer surplus and retailer profit under different pricing schemes is investigated.

Song et al. [98] applies a Markov decision process (MDP) model to the bidding problem for generators participating in electricity market. Gajjar et al. [36] extends this approach and uses actor-critic learning. Gao et al. [39] present a method for obtaining the bidding strategy of market participants using parametric linear programming. However, it assumes that market participants have complete observations of system conditions and competitor strategies.

Wang et al. [107] formulates the trading of energy by storage units as a non-cooperative game. Under certain assumptions on the strategy space and utility functions, a Nash equilibrium is shown to exist. An iterative algorithm is used to reach equilibrium, following which a double auction is conducted. Mohsenian-Rad et al. [78] proposes a distributed algorithm to obtain the optimal energy consumption schedule for each agent. The problem of determining the agent energy consumption schedule for the whole day is formulated as a deterministic linear program. Two problems are considered with two different objectives of: minimizing the energy cost, and minimizing the peak to average ratio of demand.

One of the major challenges in the above approaches is how to elicit optimal demand response without revealing the inherent dynamic nature of the loads to the ISO. In this thesis, we model the agents as stochastic dynamical systems and generate the optimal demand response in a decentralized and adaptive manner, thus maximizing the sum total of the utilities of the agents, which in turn facilitates maximum renewable penetration.

6.11 Dynamic Programming Approach

In this section we suppose that the ISO has access to the private observations of the agents, and describe a simple algorithm to solve the resulting ISO Problem. This assumption is impractical, but is intended to serve as a prelude to later sections. It helps in illustrating the key challenges discussed in earlier sections.

Let us assume that the evolution of each agent is described by a Controlled Markov Decision Process (MDP). Specifically in the problem formulation ISO Problem, we let the noise processes $N_i(t)$ be i.i.d. across times and agents. We assume that the functions $f_i^t(\cdot)$ and the laws of the noise processes $N_i(t)$ are known to the ISO. Moreover the ISO has knowledge of the state of each agent i, i.e. $X_i(t)$. Under the above assumptions, the ISO can solve the Bellman recursions to obtain the optimal control policy through value iteration of the following form,

$$\begin{aligned} V_t(x) &= \min_{u:\sum_i u_i=0} \left(\sum_i c_i(x_i) + \mathbb{E}V_{t-1}(f(x,u,N(t))) \right) \\ u_t(x) &= \arg \min_{u:\sum_i u_i=0} \left(\sum_i c_i(x_i) + \mathbb{E}V_{t-1}(f(x,u,N(t))) \right), \end{aligned} \tag{6.3}$$

where x represents the combined system state. Since at each time t, the ISO has access to the realization of the system state at time t, i.e. $X(t) := (X_1(t), X_2(t), \ldots, X_M(t))$ it can implement the optimal inputs U_i which have been obtained by solving the recursions (6.3).

We note that a similar algorithm can be implemented if we instead assume that each agent knows the functions $f_i^t(\cdot)$ laws of the noise processes $N_i(t)$ for all i, and the combined system state at each time t, i.e. $X(t)$.

Remark. *The proposed Algorithm is such that the agents agree on the choice of the optimal control policy before the system starts at $t = 0$. It achieves co-ordination*

amongst the agents by carefully designing the system so as to mimic a centralized controller. Note that under a centralized implementation where the ISO has complete knowledge of the system, the proposed solution clearly does not solve the issues of privacy. Another major concern is that the algorithm obviously suffers from the curse of dimensionality, and hence may be impractical to implement in real-time markets.

6.12 A Tree Visualization of System Randomness

A tree visualization of the system randomness will be insightful in the discussions to follow. Recall (6.1), the combined system comprising of the M agents evolves according to,

$$X(t+1) = f^t(X(t), U(t), N(t)). \tag{6.4}$$

Let us assume for the time being that the noise process $N(t)$ is allowed to assume finitely many values at each time. We then construct an uncertainty tree of depth T, in which the root node corresponds to initial system state, and a path from the root to a leaf node corresponds to a unique realization of the noise sequence $(N(0), N(1), \ldots, N(T-1))$, Figure 6.2.

6.13 Iterative Bidding Schemes

The key contribution of this work is to propose solutions to the ISO Problem in the form of Iterative Bidding Schemes (IBS), as in Walrasian tatonnement [10]. Here we explain what is meant by such an IBS. Such schemes intertwine two simple processes, which we call *Bid Update* and *Price Update*. We begin by defining the key elements of the IBS, the bid function and the price function. These will be combined to form the Bid Update and Price Update processes, which will then combine to yield the IBS in a bottom-up manner.

Bid Function: A bid sequence by agent i specifies to the ISO how much electricity that agent will purchase (negative if supplying) in every time period from that time till the final time. At time t it is a sequence of the form $(u_i(t), u_i(t+1), \ldots, u_i(T))$. Let us, for the time being, assume that the noise process $N(t)$ is observed by all the agents. Then, a bid function (in short just "bid") of an agent i is a function which specifies to the ISO, at any time t, as a function of the past history of observed noise $N(s), s < t$, how much electricity it will purchase at each instant in the future. In order to conceptualize a bid function, let us look at the uncertainty tree shown in the Figure 6.2. The bid function of each agent then, simply specifies, for each node in the tree, the amount of electricity that agent i is willing to purchase when the system passes through that node if it ever does so. We also note that the function $U_i(t)$ is adapted to the filtration $\mathcal{F}_t$, and hence is non-anticipative w.r.t. the noise process. The bid function of agent i will be denoted U_i in short.

A *price function* is a function announced by the ISO, which specifies for each time t, as a function of the past history of observed noise $N(s), s < t$, the price $\lambda(t)$ at which electricity will be sold in the market. In the tree example of Figure 6.2, this corresponds to the market clearing price corresponding to each node of the tree. The price function $\{\lambda(\omega, t)\}$ is also an $\mathcal{F}_t$-adapted stochastic process, which will be denoted by λ.

Bid Update: Let us suppose, for the time being, that the ISO has declared a price function λ via some mechanism. In the *Bid Update*, each agent i changes its bid in response to the price function λ. In order to derive its new bid, it solves the following problem, dubbed Agent i's Problem,

$$\min \mathbb{E}\left\{\sum_t c_i(X_i(t), U_i(t)) + \lambda(t)U_i(t)\right\}$$

Agent i's Problem.

We notice that the bid function $U_i(t)$ obtained after solving the above problem would minimize the agent i's total net utility , defined as the utility $-c_i(X_i(t))$ it derives from its state being $X_i(t)$, minus the amount $\lambda(t)U_i(t)$ it pays for the electricity at the price $\lambda(t)$.

Price Update is the mechanism via which the ISO updates the price function in response to the agents having submitted their updated bids via the Bid Update mechanism. Since in our context here the sole purpose of the ISO is to make sure that the net demand equals net supply, we will consider a simple rule by which it raises prices if demand exceeds supply and reduces otherwise. Suppose the previous price was λ^k and the bid was U^k. Then, for each possible sample path ω, the Price Update is,

$$\lambda^{k+1}(t) = \lambda^k(t)\,(1-\alpha_k) + \alpha_k \left(\sum_i U_i^k(t) \right)$$

Price Update.

where $\alpha_k > 0$ is an "adaptation gain". (One choice is $\alpha_k = 1/k$, which satisfies the twin conditions $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < +\infty$, which is a common convergence condition in stochastic approximation [21]). Figure 6.1 summarizes the IBS technique.

It will be the object of the following section, to show that an iteration of Bid Update-Price Update can solve the ISO Problem under some observation structures. We begin with the simplest of all cases, the deterministic case.

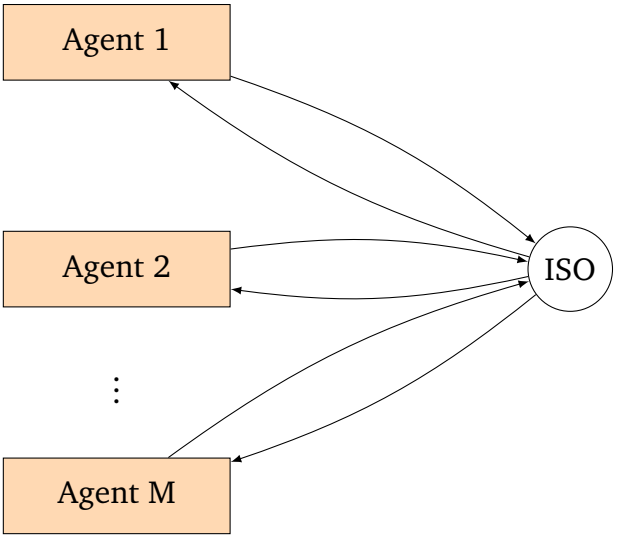

Figure 6.1: Agents submit bids via Agent $\rightarrow$ ISO, while the ISO sends price-signals for the remaining time horizon through ISO $\rightarrow$ Agent

6.14 The Deterministic Case

The simple example considered in Section 6.11 illustrates the key difficulty involved in a decentralized stochastic control, namely that of achieving co ordination amongst the agents. The centralized algorithm above circumvents this difficulty by having the agents explicitly communicate their system states $X_i(t)$ to the ISO, or, equivalently, to every other agent. This however does not meet our constraints on what can be communicated or revealed by agents.

In this section we show that it is possible under certain convexity assumptions, for the agents to not communicate their state values, but still attain the same performance as centralized control. We establish this result here for the case of deterministic systems. Sections 6.15 and 6.16 show that the same idea carries over to the stochastic setting; however the procedure requires additional operations associated with encoding the randomness $N(t)$.

The deterministic version of ISO Problem can be stated as follows:

$$\min \sum_{t=1}^{T} \sum_{i=1}^{N} c_i\left(x_i(t), u_i(t)\right)$$

$$\text{such that } \sum_i u_i(t) = 0, \text{ for } t = 1, 2, \ldots, T-1,$$

$$\text{and } x_i(t+1) = f_i^t(x_i(t), u_i(t)), \text{ for}$$

$$i = 1, 2, \ldots, N, \text{ and } t = 1, 2, \ldots, T-1. \qquad \text{Deterministic ISO Problem}$$

The intermediate variables $x_i(t)$ can be expressed in terms of the inputs $u_i := (u_i(1), u_i(2), \ldots, u_i(T-1))$ and thus the cost term $\sum_{t=1}^{T} \sum_{i=1}^{N} c_i\left(x_i(t), u_i(t)\right)$ can also be expressed solely as a function of the inputs $u_i, i = 1, 2 \ldots, M$. Convexity plays a major role, as first identified by Arrow [10].

Assumption 1 (Convexity Assumption)**.** *For $i = 1, 2, \ldots, M$, the function $\sum_{t=1}^{T} c_i\left(x_i(t), u_i(t)\right)$ is convex in the input vector $(u_i(1), u_i(2), \ldots, u_i(T-1))$.*

We will now derive a decentralized solution to the ISO Problem under Assumption 1, and show that under the resulting solution, the system achieves the same performance as that of optimal centralized control.

Employing the definition of each $x_i(t)$ as $f_i^t\left(x_i(t-1), u_i(t-1)\right)$, each $x_i(t)$ can be written as a function of $(u_i(0), u_i(1), \ldots, u_i(t-1))$, since $x_i(0)$ is regarded as fixed. The associated Lagrangian and dual function are given by,

$$\mathcal{L}\left(u, \lambda\right) := \sum_{i=1}^{M} \left\{ \sum_t c_i(x_i(t), u_i(t)) + \lambda(t) u_i(t) \right\},$$

$$D(\lambda) := \min_u \mathcal{L}\left(u, \lambda\right),$$

where $u := (u_1, u_2, \ldots, u_M)$, and $\lambda := (\lambda(1), \lambda(2), \ldots, \lambda(T-1))$. Note that the

Lagrangian is the sum of the costs $\sum_t \{c_i(x_i(t), u_i(t)) + \lambda(t)u_i(t)\}$ incurred by each individual agent. Hence, given the Lagrange multipliers λ, the inputs u_i minimizing the Lagrangian can be calculated in a decentralized fashion, with each agent i solving its own problem,

$$\min \sum_t c_i(x_i(t), u_i(t)) + \lambda(t)u_i(t), \tag{6.5}$$

$$\text{subject to } x_i(t+1) = f_i^t(x_i(t), u_i(t)). \tag{6.6}$$

Agent i's Problem

Each agent i then communicates this optimal $u_i(t)$ to the ISO by submitting its bid. This would enable the computation of the dual function at each value of $\lambda(t)$.

Note that the sub-gradient with respect to λ of the Dual function $D(\lambda)$ is $\sum_i u_i^k(t)$. Since the dual problem of finding the optimizing prices $\lambda(t)$ in order to maximize $D(\lambda(t))$ is convex, it can be solved via the sub-gradient method [**?**, 16, 22, 86].

$$\lambda^{k+1}(t) = \lambda^k(t)\left(1 - \alpha^k\right) + \alpha^k \left(\sum_i u_i^k(t)\right), t \geq 0, \tag{6.7}$$

where k is the index which keeps track of the iteration number. The iterations end when the price vector $\lambda(t)$ converges to the optimal value $\lambda^\star(t)$. The resulting solution is optimal for the ISO Problem due to the convexity assumptions.

6.15 Commonly Observed Noise

We now turn attention to the stochastic case. In this section we will consider the case where the noise affecting all agents is the same, i.e., $N_i \equiv N_c$, and is observed by all agents. The agents also know the laws $\mathcal{L}(N_c(\cdot))$.

The solution (6.3) proposed in Section 6.11 using the Dynamic Programming

approach suffers from severe drawback that the value of the state and the system dynamics of each agent are assumed to be known to the ISO.

However under the convexity assumption, the ISO Problem has a low complexity decentralized solution. As in Section 6.11, it is assumed that the agents evolve as controlled MDPs,

$$X_i(t+1) = f_i^t(X_i(t), U_i(t), N_c(t)),$$

where the noise process $N_c(t)$ is observed by all the agents.

The knowledge of the system dynamics $f_i^t(\cdot)$ and the processes $X_i(t)$ is kept private, and is known only to the agent i. We make the following assumption on the cost function, which is the stochastic counterpart of Assumption 1.

Assumption 2. *The function*

$$\sum_t c_i(X_i(t), U_i(t)), i = 1, 2, \ldots, M, \tag{6.8}$$

is convex in the vector $\{U_i(t)\}_{t=0}^{T-1}$ *for fixed* $\{N_c(t)\}_{t=0}^{T-1}$.

We next present an iterative algorithm composed of Bid and Price updates. The bid submitted by each agent i is a random process that maps the space $\Omega \times \{1, 2, \ldots, T-1\}$ to $\mathbb{R}$. This is akin to Arrow's [10] approach of treating each product available at a certain time and place as a separate product. Furthermore, the bid process is adapted to the filtration $\mathcal{F}_t$. In words, at each time t, it specifies to the ISO, as a function of the past noise $N(s), s < t$, the amount of electricity that the agent is willing to purchase at time t.

Theorem 11. *Algorithm 1 solves the ISO Problem when the cost functions* $c_i(\cdot), i = 1, 2, \ldots, M$ *satisfy the Assumption 2.*

Algorithm 1

Assumption: The law of the combined noise process $\mathcal{L}(N(t))$ is common knowledge of all agents. The noise process $N_c(t)$ affecting the agents is observed by all.
$k = 0$

repeat
Each agent i solves the problem

$$\min \mathbb{E}\left\{\sum_t c_i(X_i(t), U_i(t)) + \lambda^k(t) U_i(t)\right\}, \qquad \text{(Agent i's Problem)}$$

for the optimal $\{U_i^k(t), 0 \leq t \leq T-1\}$, and submits to ISO, i.e. Bid Update.
ISO declares new prices via the 2.14, i.e.

$$\lambda^{k+1}(t) = \lambda^k(t)\left(1-\alpha^k\right) + \alpha^k\left(\sum_i U_i(t)\right). \qquad \text{(Price Update)}$$

$k \to k+1$
until $U_i^k(t)$ converge to $U_i^\star(t)$
Each agent implements $U_i^\star(t)$ =0

Figure 6.3 lays out the decision flow involved while implementing the algorithm.

Remark. *Comparing the algorithm proposed above with that proposed in the Section 6.11, we note that the present algorithm mitigates the curse of dimensionality since the dual function at each value of price process $\lambda(t)$ can be computed by agents individually. Thus the computational complexity of the proposed scheme is linear in the number of agents (M). Of course, for each agent, the complexity does grow with the number of its own states.*

Privately Observed Noise Communicated to ISO *A subtle point to note is that solving the Agent i's Problem does not require the agents to know the noise process $N_c(t)$. It suffices for them to know the law $\mathcal{L}(N_c(\cdot))$.*

Hence, instead of assuming that the noise sequence $N_c(t)$ is commonly observed, we could have equivalently assumed that each agent i was affected by private noise $N_i(t)$, that was observed only by it. The private observations could then have been communicated to the ISO. The noises $N_1(t), N_2(t), \ldots, N_M(t)$ need not be independent. The ISO would then know the combined noise process $N(t) := (N_1(t), N_2(t), \ldots, N_M(t))$, and can implement the optimal $U(t)$.

This result can be further extended as follows. The ISO does not really need to know the true value of the combined noise process $N(t)$. It only needs to know the "label" or "index" of the noise values for the purpose of communicating to the agents the prices for each such label.

The agents can hide the actual value of the noise by mapping their noise process $N_i(t)$ to some other process $\hat{N}(t)$. For example, in the uncertainty tree discussed in Section 6.12, the agents could re label the noise values 0 to 1, and value 1 to 2. This technique thus enables the agents to maintain privacy to some extent.

Algorithm 2

Assumption: The law of the combined noise process $\mathcal{L}(N(t))$ is common knowledge of all agents and ISO.
for bidding times $s = 0$ to $T-1$ **do**
$k = 0$

repeat
Each agent i solves the problem

$$\min \mathbb{E}\left\{\sum_{t\geq s} c_i(X_i(t), U_i(t)) + \lambda^k(t)U_i(t)\right\}, \qquad \text{(Agent i's Problem)}$$

for the optimal $\{U_i^k(t), 0 \leq t \leq T-1\}$, and submits it to ISO.
ISO declares new prices via the 2.14, i.e.

$$\lambda^{k+1}(t) = \lambda^k(t)\left(1-\alpha^k\right) + \alpha^k\left(\sum_i U_i^k(t)\right), t \geq s.$$

$k \to k+1$
until $U_i^k(t)$ converge to $U_i^\star(t), \quad t \geq s$
ISO implements $U_i^\star(s)$
end for=0

Next, we investigate the problem when we remove the assumption that the system noise is commonly observed by all.

Assumption: Suppose that the agents are affected only by privately observed noises, i.e., they evolve as,

$$X_i(t+1) = f_i^t(X_i(t), U_i(t), N_i(t)), \text{ where} \qquad (6.9)$$
$$i = 1, 2, \ldots, M, \text{ and } t = 0, 1, \ldots, T-1,$$

and the noise $N_i(t)$ is observed only by the agent i. The noises $\{N_1(t), N_2(t), \ldots, N_M(t)\}$ may be dependent random variables. There can also be dependence across time.

Even though the agents do not observe the private noise of other agents, they are assumed to know the laws of the combined noise process, $\mathcal{L}(N(\cdot))$. In the context of the uncertainty tree of Section 6.12, the agents know the topology of the tree, and the transition probabilities along the edges.

In order to construct an algorithmic solution for the private noise case, we revisit Algorithm 1 where it makes use of the assumption that the process $N_c(t)$ is commonly observed. We will construct algorithm for the present case from 1. Each agent i could perform its Bid Updates (by solving Agent i's Problem) based only on a) the prices λ that had been declared by the ISO, and b) the $\mathcal{L}(N(t))$. Similarly the operations that went in performing the Price Updates involved the bids that had been submitted by agents. Thus the optimal actions could be calculated in a decentralized fashion without the agents knowing the noise sequences. In the context of the tree, the agents need to know the labels of the nodes of the tree and the transition probabilities. If a transition from one node to another is caused by many different random events transpiring at different agents' system, they do not need to know what transpired at each agent's system. The ISO needs to know even less. It only needs to know the labels of the nodes, but does not need to know the probabilities of the transitions from node to node. However after the Price and Bid Update iterations in Algorithm 1 converge they yield the optimal action $U^\star(t)$ at each time t as a function of the past values of the combined noise process $N(s), s < t$ or in the tree context, the optimal action $U^\star$ at each node in the tree. Once calculated, we assumed that the process $N(t)$ ($N_c(t)$) was observed by all the agents only to ensure that all the agents agreed on the choice of the action at time t, i.e., $U^\star(t)$. In summary, Algorithm 1 required the agents to observe the combined

noise sequence $N(t)$ only in order to implement the optimal action, not in order to calculate it.

Upon closer inspection of Algorithm 1, we find that even though the optimal actions $U^\star(t)$ for bid-times $t \geq 1$ are expressed as functions of $N(s), s = 1, 2, \ldots, t-1$, the action to be taken at time $t = 0$, i.e., $U^\star(0)$ is not a function of the noise process. Or, stated differently, the $U_i(0)$ s are measurable w.r.t. the sigma-algebra $\mathcal{F}_0 = \{\Omega, \emptyset\}$. Hence, in order to implement the converged quantities $U_i^\star(0)$ for the first time-slot, the agents do not need to know the private observations such as noise or system states of other agents.

A similar process can therefore be used at every time t. The agents need only to share the topology of the *remaining* uncertainty tree from the current node, i.e., the laws $\mathcal{L}(N(t))$ for the *remaining* times $t = 1, 2, \ldots, T-1$. Then the bid-price update iteration can take place just as though that were the initial time. Thus we see that a repeated application of the Algorithm 1 at each time t, followed by sharing the laws of the noise process for the remaining bid-times would enable the agents to implement the optimal actions at each time t. This yields us Algorithm 2 detailed below.

Theorem 12. *Algorithm 2 solves the ISO Problem when the cost functions satisfy Assumption 1, with each agent i having access only to its private noise $N_i(t)$, while the law of the combined noise process, i.e. $\mathcal{L}(N(t))$ is known publicly.*

Proof. Let us first consider a version of the Commonly Observed Noise Problem in which the noise process $N(t)$ assumes only finitely many values.

Let us suppose that $x(0)$ is fixed, without loss of generality. Let p_v denote the probability of node v in the uncertainty tree. The depth of the node in the tree indicates time, as can be seen from Fig 6.2. Every Markov policy specifies an action

$U(v) := (U_0(v), U_1(v), \ldots, U_M(v))$ satisfying $\sum_i U_i(v) = 0$ for every node v in the tree. This is easily seen by recursion starting at the root which corresponds to the initial time and state of the system, and noting that each node then also indicates the state of the system at that time. On the other hand, a "tree policy" that specifies a $U(v) := (U_0(v), U_1(v), \ldots, U_M(v))$ satisfying $\sum_i U_i(v) = 0$ for every node v in the tree may be slightly more general than a Markov policy since two nodes in the tree at the same depth may correspond to the same state $X(t)$ but a tree policy may prescribe different actions for them. We will consider this slightly more general class of tree policies, which also contains an optimal policy since we know that the smaller class of Markov policies contains an optimal policy for a finite horizon MDP.

For every such policy, for every node v, there is a unique sequence of actions $U^v := \{U(0), U(1), \ldots, U(t)\}$ that was taken in the preceding t steps, where t denotes the depth of the node v. Note that the state $X(t)$ at time t corresponding to the node v is determined by (v, u^v). The centralized optimization problem can then be written as the following optimization problem,

$$\begin{aligned} &\min \sum_{i=1}^{M} \sum_{v} p_v c_i\left(v, U^v\right) \\ &\text{such that } \sum_{i} U_i(v) = 0, \forall v. \end{aligned}$$

Note that $c_i(v, u)$ is convex in u. Hence this is a convex programming problem with no duality gap. Associating Lagrange multiplier $\lambda(v)$ with the constraint $\sum_i U_i(v) = 0$, and letting $\lambda := \{\lambda(v)\}$, we obtain,

$$\mathcal{L}\left(U, \lambda\right) := \sum_{i=1}^{M} \sum_{v} p_v \left\{ \sum_{v} c_i(v, u^v) + \lambda(v) U_i(v) \right\}.$$

We will call the process $\lambda(v)$ as the "price process".

Each agent submits a bid for each possible partial realization v of the noise process, while the ISO specifies a price at each v. Now the proof parallels the proof in the deterministic case. □

6.17 Using Learning Techniques to Eliminate Complexity of $\mathcal{L}(N(t))$

The approaches discussed in the previous sections relied on the assumption that the knowledge of $\mathcal{L}(N(\cdot))$ was global. However this is not a practical assumption because of privacy concerns of agents. Even if privacy were not an issue, the set of possible noise sequences grows exponentially with the number of agents and the length of time horizon T, which makes the sharing of huge amounts of information impractical. However, as seen in Sections 6.15 and 6.16, the knowledge of $\mathcal{L}(N(t))$ was required by the Agents in order to solve Agent i's Problem, which formed a crucial component of the Bid Update step.

It is of interest to determine whether it is indeed possible to solve the ISO Problem without assumption of the knowledge of $\mathcal{L}(N(t))$? For the general case in which the agents are modeled by an MDP, we can "learn" what we need as we go along, rather than needing to know a-priori the exponentially large uncertainty tree. That is, we can simply learn the cumulative impact of what we need to know. This is similar to the assumption of "rational expectations" in Arrows model of uncertainty [10], whereby agents can make inferences about the system from private observations as well as by observing prices. This can be achieved via the technique of Stochastic Approximation or other learning techniques [21, 58, 90].

The key idea involved in eliminating the need for knowledge of $\mathcal{N}(t)$ is similar to the Q-learning technique employed in machine learning. The previous sections used Iterative Bidding Techniques in order to converge to the optimal prices $\lambda^\star$.

The agents knew how to respond to price changes because they could calculate the optimal bids and update them. Since now the bid update is not possible on account of insufficient knowledge of $\mathcal{N}(t)$, the agents can try to *learn* the optimal bids as a function of the price λ. The combined system comprised of M agents and the ISO would have to "learn" the optimal price function, and the optimal bid as a function of prices. This can be achieved via the two time-scale learning algorithms proposed in [21]. More specifically, the Bid Updates would now involve reinforcement learning [21, 108]. Price Updates

6.18 The Case of Linear Systems

This section treats the special case of the ISO Problem when the M agents of interest have linear Gaussian dynamics. The noises of all agents are independent. Each agent i has a quadratic cost criterion, i.e., the functions $c_i(\cdot)$ are quadratic in x_i, u_i, with weighting matrices $Q_i \geq 0$ and $R_i > 0$. Let us call this the Distributed Constrained LQG (DCLQG) Problem.

$$\begin{aligned}
&\min \mathbb{E}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} X_i^{\intercal}(t)Q_iX_i(t) + U_i^{\intercal}(t)R_iU_i(t)\right)\\
&\text{subject to } X_i(t+1) = A_iX_i(t) + B_iU_i(t) + B_iN_i(t),\\
&t = 0, \ldots, T-1,\\
&\text{and } \sum_i U_i(t) = 0, t = 0, \ldots, T-1. \qquad \text{DCLQG}
\end{aligned}$$

(The case of time-varying systems is analogous to time-invariant systems, and omitted for brevity). We will assume that the system dynamics given by (A_i, B_i), the cost functions given by (Q_i, R_i), and the observation structure are all private. None of the agents have knowledge of each others' system parameters, and the state process X_i is observed only by the agent i.

We will derive an Iterative Bidding Scheme which is much simpler than the algorithm proposed in Section 6.16 in the following critical aspect. The bid function submitted at time t specifying the quantity of electricity that agent i is willing to purchase at times $t, t+1, \ldots, T-1$ does not depend on the outcomes of noise sequence $N(s), s > t$. It is simply a vector $(u_i(t), u_i(t+1), \ldots, u_i(T-1))$ comprising of $T-t+1$ entries. This is a drastic reduction in complexity of the bidding scheme. At each time t, the following iteration takes place: Each agent bids a vector of future purchases in response to prices announced by the ISO for future power, and the ISO updates the prices in return, until convergence.

The key to showing the optimality of such a simple bidding scheme lies in utilizing the certainty equivalence property of LQG systems [82].

Definition (Certainty Equivalence)**.** *A stochastic control problem is said to possess the property of certainty equivalence if the optimal policy for the stochastic control problem coincides with the optimal policy for the corresponding deterministic control problem in which the noise is absent.*

Theorem 13. *The following bidding scheme achieves optimality for the ISO Problem with LQG agents. At time t, in response to the k-th iterate of the price sequence $(\lambda^k(t), \lambda^k(t+1), \ldots, \lambda^k(T))$, agent i announces the optimal open loop sequence $(u_i^k(t), u_i^k(t+1), \ldots, u_i^k(T))$ for the deterministic LQ problem:*

$$\min \sum_{s \geq t} x_i^\intercal(t) Q_i x_i(t) + (u^k)_i^\intercal(t) R_i u_i(t) + \lambda^k(t) u_i(t)$$

$$\text{s.t. } x_i(s+1) = A_i x_i(s) + B_i u_i^k(s) \text{ for } s = t, t+1, ..., T.$$

In response, the ISO adjusts the prices according to: $\lambda^{k+1}(s) = \lambda^k(s)\left(1-\alpha^k\right) + \alpha^k\left(\sum_i u_i^k(s)\right), s \geq t$. This process is iterated till it converges to $(u_i^\star(t), \ldots, u_i^\star(T-1))$

Algorithm 3

for bidding times $s = 0$ to $T - 1$ **do**

$k = 0$

Initialize $\lambda^k(t), t \geq s$ to some arbitrary value.

repeat

Each agent i solves the problem

$$\min \sum_{t \geq s}^{T} x^{\intercal}(t) Q_i x(t) + u_i^{\intercal}(t) R_i u_i(t) + \lambda^k(t) u_i(t) \text{subject to} \tag{6.10}$$

$$x_i(t+1) = A_i x_i(t) + B_i u_i(t), t = 0, 1, \ldots, T-1. \quad \text{(Agent i's Problem)}$$

and submits the optimal value, denoted $u^k(t)$ to the ISO.

ISO updates the prices via the 2.14,

$$\lambda^{k+1}(t) = \lambda^k(t) \left(1 - \alpha^k\right) + \alpha^k \left(\sum_i u_i^k(t) \right), t \geq s.$$

Increment k by 1

until $u_i^k(t)$ converge to $u_i^\star(t)$

implement $u^\star(s)$

end for=0

and $(\lambda^\star(t), \ldots, \lambda^\star(T-1))$. *At time* t*, the price is set at* $\lambda^\star(t)$ *and agent* i *applies the input* $u_i^\star(t)$. *This entire procedure is repeated at time* $t+1$.

Proof. Let

$$x := (x_1, x_2, \ldots, x_M), u := (u_1, u_2, \ldots, u_M),$$
$$A := diag(A_1, A_2, \ldots, A_M), B := diag(B_1, B_2, \ldots, B_M),$$
$$Q = diag(Q_1, Q_2, \ldots, Q_M), R = diag(R_1, R_2, \ldots, R_M),$$

and consider the following deterministic LQR problem with no noise, corresponding to the noisy LQG problem,

$$\begin{aligned} &\min \sum_{t=1}^{T} x^\intercal(t) Q x(t) + u^\intercal(t) R u(t) \\ &\text{subject to } x(t+1) = Ax(t) + Bu(t), \\ &\sum_{i=1}^{M} u_i(t) = 0 \text{ for } t = 0, \ldots, T-1. \end{aligned} \tag{6.11}$$

Since the state is affine in u, the cost is convex in u. Hence this centralized problem can be solved by the Bid-Price iteration between the agents and the ISO. In particular, at each time 0, the end result of the scheme is the optimal action $u(0)$. This is arrived at by the ISO announcing a sequence of prices for all future times and the agents bidding their consumptions/generation sequences at all future times.

Now note that due to the energy balance at each time, agent M is forced to choose $u_M(t) = -\sum_i^{M-1} u_i(t)$ for all t. Hence one can substitute this value for $u_M(t)$ and obtain a standard LQ problem where there is no separate energy balance constraint. For this reduced but standard deterministic linear quadratic problem, the optimal solution is given by linear feedback $u(0) = \Gamma(0)x(0)$, where $\Gamma(\cdot)$ is the

optimal feedback gain.

Now consider the corresponding reduced stochastic LQG problem where there is white Gaussian noise in the state equations (6.11). By Certainty Equivalence [82], the same feedback law as in the deterministic reduced LQ problem is also optimal. In particular, in state $x(0)$ at time 0, $u(0) = \Gamma(0)x(0)$ continues to be optimal. Now, in our proposed bidding scheme for the LQG problem, each agent bids on the basis of a private deterministic system for itself. Hence it leads to the same Bid-Price iteration result at time 0. Hence it arrives at the same $u(0)$, which however is also optimal for the stochastic LQG problem.

Thus we see that the Bid-Price iteration scheme determines the optimal actions for the agents at time 0. Now our scheme for the LQG problem repeats such a Bid-Price scheme iteration at each time t. Each $x(t)$ can be regarded as an initial state for the system started at time t, and the same argument as above shows that the actions $u(t)$ that it results in for the agents at time 1 are also optimal. □

We note the following important features of the proposed algorithm. The critical feature that there is an iteration of bids at each time t is important. Also important is that at each stage it is the future sequence of prices that is iterated. If the bids are not iterated to convergence the resulting prices and actions will be sub-optimal.

It is important to note that the alternative of announcing a "bid curve" of price vs. generation for a single time t, a la a Walrasian auction, does not work in the dynamic case. The reason is that the current optimal generation depends on future prices, so iteration of price at only one time is not sufficient to ensure optimal decision when agents are dynamic systems.

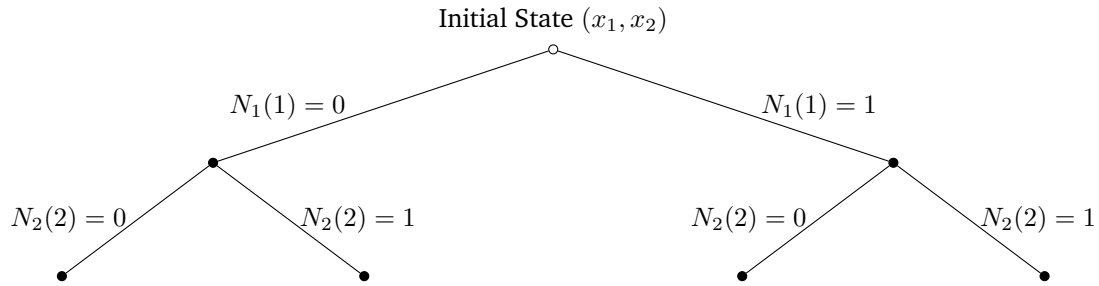

Figure 6.2: A Tree based visualization of randomness for a 2 agent system evolving over 2 bid times. The noise values are allowed to be binary and assume the values 0 and 1.

6.19 Concluding Remarks

We have posed the ISO Problem of maximizing the total utility/minimizing the total operating costs of the electricity grid, while obtaining minimal information from each agent, as a decentralized stochastic control problem. We have shown that the Distributed Constrained LQG problem DCLQG admits a simple and decentralized solution utilizing iterative bidding schemes, which attains the same performance as that of an optimal centralized control policy. Under the proposed policy, the sufficient statistics are vastly simplified, and each agent i needs to keep track of its present state $X_i(t)$. This is in contrast with the general case of decentralized stochastic control, in which the agents need to keep track of the entire history in order to implement an optimal policy, which is also generally intractable to compute. So, not only is our Algorithm decentralized, and easy to implement, but it also leads to a large reduction in the amount of data to be communicated. We further note that our Algorithm is privacy preserving.

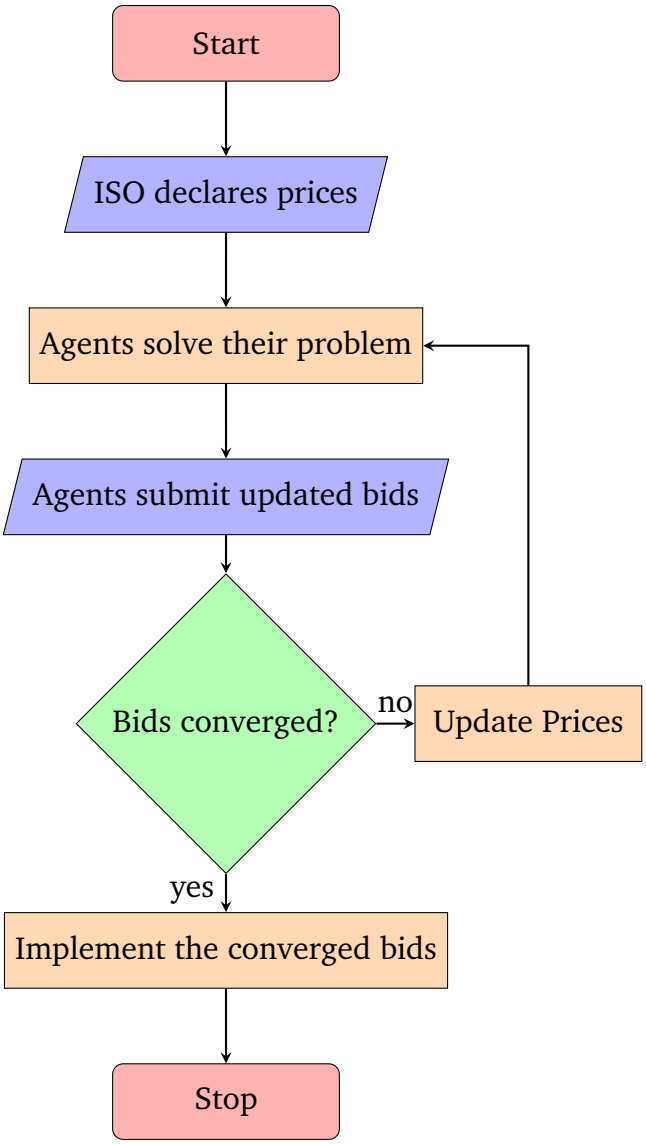

Figure 6.3: Flowchart depicting the decision flow in Algorithm 1.

7. ON STORAGE AND RENEWABLES: A THEORY OF SIZING AND UNCERTAINTY

In this section we study the fundamental problem of sizing energy storage given an uncertainty level of variable resources in a microgrid. A queuing-theoretic model is introduced, which provides unique insights into the coupling between energy storage size and uncertainty level of the net load. The proposed model lends itself to three levels of details: a random walk model with single uncertainty from one net load, a reflected Brownian motion model with more uncertain resources, and a model with a collection of Markov-type power producers and consumers. It is shown that the fundamental requirement of energy storage sizing can be approximately derived from the aforementioned three queuing theory models. Numerical examples suggest that this approach can be applied to microgrid planning and operation in assessing the optimal size of energy storage , as well as the potential curtailment of renewable energy.

7.1 Introduction

This section is motivated by the increasing penetration of variable resources around the world. A fundamental question arises with increasing deployment of variable resources: What is the amount of energy storage needed for high penetration of renewable power? We propose a theoretical framework that provides rigorous yet simple tools to address this question.

A historical comparison is in order. In the early 1900s, when telephone exchanges were being built across the country, there were several questions related to how large the exchanges had to be in terms of the number of lines, in order to ensure that the percentage of blocked calls was below a specified value, while

carrying a specified volume of calls with certain holding times. In response to this challenge, queuing theory was invented by Erlang [33], [31].

A similar problem arises today in the context of the integration of clean energy and energy storage resources. We use the word "storage" in a rather broad context. It includes large batteries, large buildings with controllable thermal energy, as well as a large number of coordinated electric vehicles *[41, 65, 87, 103]*.

Storage can be used to mitigate the unreliability of such stochastic time variation of energy sources, but, depending on the demand, there will necessarily be a need for balancing power, likely from conventional generation (typically fossil fuel-based). In a future grid whose objective is to lower the carbon emission, it is important to characterize not only the mean power drawn from the fossil fuel, but also its variation, e.g., peak-to-mean ratio.

Thus, one would like to determine how the magnitude of the storage interacts with the stochastic temporal unreliability of renewables and their spectral content, in terms of determining what nature of demands can be supported, and what the resulting peak as well as mean need is for augmenting energy sources. This is the goal of the section.

To elucidate how one may address these interrelated issues, we pose the problem in a simple yet fundamental mathematical model. The scope of this section is limited. We attempt to only show what analytical tools and theoretical techniques can be brought to bear to address the nexus of these issues. Future extensions will, one hopes, obtain more useful results employing realistic models of the phenomena involved.

The rest of the section is organized as follows. We provide a description of the system model in Section 7.2, which is followed by a derivation of the system performance in Section 7.3 for the simple case when the energy delivery in a micro-

grid follows a "random-walk" model. Section 7.4 provides an exposition of the Brownian motion model, a model which is justified when one is dealing with a large number of power producers and consumers in the electricity market. Section 7.5 considers the case when we are dealing with more complex dynamics occurring at the level of producers and consumers. Section 7.6 presents concluding remarks.

7.2 System Model

We consider the abstract model, shown in Figure 7.1, of a micro-grid which supplies its customers from a portfolio of renewable power, fossil fuel generation, and a storage device. The storage unit has an energy capacity of B. This is similar to a scenario of a micro-grid operator trying to schedule all the resources to a community of customers in the isolated operating mode.

Once the storage energy hits the level 0, the conventional generators are utilized to meet any power demand that exceeds renewable power supply since the operator has no more stored renewable energy to supply. Thus, at all times, the aggregate excess demand of the consumers is fulfilled either from renewable and storage, or from the conventional generation.

At the other extreme, whenever the storage reaches its maximum value of B, the supplier is forced to curtail any excess renewable energy being produced ("overflow of renewable energy"), thus leading to spill of some renewable energy.

With the above set-up in place, we will be interested in answering the following questions: What is the average amount of fossil power consumed? How does it change with the size of the storage? Does the amount of fossil power required also depend on the statistics of the renewable energy source ([9, 27]? How does it depend on the "spectral content" of the time variation of the renewable energy source (since high frequency variations are better buffered by the source rather

than low frequency variations)? What parameters of the renewal energy supply process quantify this dependence? Similar questions are of interest concerning the quantity of renewable curtailment due to "overflows." As a finer measure of performance, one might also be interested in second-order moments of the amount of fossil energy utilized, since that is related to issues such as peak-to-average generation ratio, and the amount of renewable energy wasted.

Fundamentally it is the difference between two stochastic processes (renewable power and load) that is relevant to answer the above questions [1] Our goal in this section is to show how these two stochastic processes interact with the storage capacity B to determine the answers to the above questions. To illustrate the methodology, we carry forward this analysis to obtain somewhat explicit back-of-the-envelope type answers for certain simple stochastic process models. Doubtless, it is important in practice to determine answers for models of higher fidelity, and this can be, and we hope will be, pursued along the lines indicated in this section, though at the expense of greater mathematical complexity.

In the next three sections, we provide illustratory answers to three models of stochastic time variation.

7.3 Random Walk Model

We begin with the simplest stochastic process, a random walk, to model the renewable sources and loads. Let us denote the energy level in the storage at time t by $V(t)$, where the time parameter t assumes the discrete values $0, 1, \ldots$. For simplicity, we suppose that the energy-levels of the storage are discretized, so that $V(t) \in \{0, 1, 2, \ldots, B\}$. An important quantity, in fact the only relevant quantity,

[1] In this section we assume there is no line constraint. This is not a very restrictive assumption since when the total load in the grid is not very high, then the analysis of the system can be reduced to the case of a single node, which is free of any line constraint.

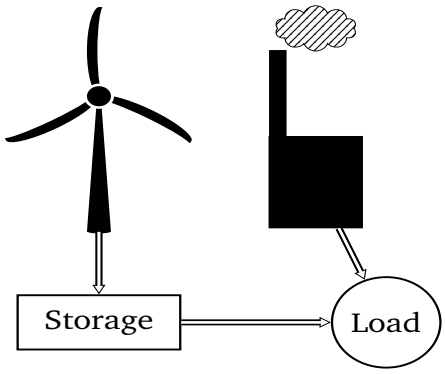

Figure 7.1: Renewable and fossil fuel energy consolidated into a microgrid.

is the *net-put*, which is the difference between the renewable power supply at a certain time minus the demand at that time. If the "net-put" to the storage at time t is $X(t)$, then the value of storage at time $t+1$ is given by

$$V(t+1) = (V(t) + X(t))^+ \wedge B,$$

where $a \wedge b :=$Min(a,b), and $a^+ :=$Max$(a,0)$. (Since we are in discrete time, the "energy" net-put in one time unit can also be called "power"). We note that the assumption that the system evolves over discrete times is not restrictive, since we can always sample a continuous-time system at an arbitrarily high frequency and perform an analysis of the resulting discretized system.

For simplicity, we begin by supposing that $X(0), X(1), \ldots$ are independent and identically distributed random variables, with a known distribution. Under these assumptions, $V(t)$ is a Markov process evolving on a finite state space $\{0, 1, \ldots, B\}$, and, under a mild irreducibility condition that we assume, has a unique stationary distribution. Let us denote by $V(\infty)$ the random variable having the stationary

distribution of the process $V(t)$.

Next, we define two stochastic processes that are relevant to the system performance. Let $L(t)$ be the *loss* in the renewable energy at time t, i.e.,

$$L(t) = (V(t-1) + X(t-1))^+ - V(t).$$

As its name implies, this is the renewable energy that is indeed curtailed because the storage is full. Also, denote by $F(t)$ the fossil energy expended to meet the demands,

$$F(t) := V(t) - (V(t-1) + X(t-1)) \wedge B.$$

Clearly,

$$V(t) = V(t-1) + X(t-1) + F(t) - L(t). \tag{7.1}$$

The average energy wasted due to overflows is then simply,

$$\bar{L} := \mathrm{E}\,(V(\infty) + X)^+ - \left[(V(\infty) + X)^+ \wedge B\right]$$

where X is a random variable having the same distribution as $X(0), X(1), \ldots$. Using (7.1), under a mild aperiodicity assumption, in steady state the expected value of $V(t)$ is the same as that of $V(t+1)$, and so $\bar{L} = \bar{F} + \bar{X}$.

Similarly, the second moments of the steady state loss and fossil energy can be calculated once the stationary distribution of the process $V(t)$ is known. The following result from [92] is useful:

Theorem 14. *Let* $S(n) := \sum_{i=0}^{n} X(i)$, *and* $\tau\,[u, v) := \inf\{n \geq 0 : S(n) \notin [u, v)\}$.

Then $\mathbb{P}\left(V(\infty) \geq x\right) = \mathbb{P}\left(S\left(\tau\left(x-B, x\right)\right) \geq x\right)$.

To illustrate how we can determine the quantities of significant interest, let us consider a "simple random walk", where X assumes value 1 with a probability $p < \frac{1}{2}$ and -1 with probability $q = 1 - p > \frac{1}{2}$; a similar analysis can be carried out for more general models. The steady state distribution is

$$\mathbb{P}\left(V(\infty) = x\right) = \frac{1-\rho}{1-\rho^{B+1}}\rho^x, \text{ for } x = 0, 1, \ldots, B,$$

where $\rho := \frac{p}{q}$. Thus the renewable-energy lost is given by,

$$\bar{L} = \frac{1-\rho}{1-\rho^{B+1}}\rho^B p \tag{7.2}$$

and its standard deviation is given by, $\sigma_L := \frac{1-\rho}{1-\rho^{B+1}}\rho^B p$.

The formulas for loss and variability provide valuable insight. It follows from (7.2) that for a fixed value of ρ, the wastage suffered by the energy supplier decreases roughly exponentially as the size of the storage is increased. Since the storage capacity B comes at a cost, either fixed or operating, such an insight as that provided by (7.2) might be useful for windfarm operators faced with the issue of choosing the right location to place windfarms and the right size of storage. The location of the windfarm fixes the quantity ρ, and the storage-size corresponds to B. The ability to do such back-of-the-envelope calculations is potentially important for providing an intuitive but quantitative understanding of sizing for storage.

7.4 Reflected Brownian Motion Model

We next consider an alternative continuous-time model, where the "cumulative net-put process" (i.e., the net-put process of the previous section integrated over time) is a Brownian motion. One justification for such a model is as a limit of a

sequence of markets operating with many renewable power producers, in which the n-th market has a number n of energy suppliers and consumers such that the mismatch between the total demand of the consumers and the energy supplied in a time-epoch is (after an appropriate scaling) of the order $\sqrt{n}$ with a high probability. As the number of individual renewable resources, n increases to ∞, it can be shown that the energy traded in the market converges weakly to Brownian motion ([17]). The advantage of using such a Brownian approximation is that it allows us to stochastic calculus to obtain simple and elegant expressions for the relevant quantities. Such an approach has been successfully used in stochastic flow systems, and the analysis here follows the approach of modeling queuing systems by reflected Brownian motion [14, 47, 112].

As in the previous section, there is a single storage unit, which is "fed" by a Brownian motion. The system evolves in continuous time, and the *cumulative net-put process* $X(t)$, obtained by subtracting the total energy demand of the consumers till time t from the net renewable energy produced until time t, is a Brownian motion with a drift μ, and a variance σ.

Due to the storage capacity of B, the process denoting the storage level at time t, $V(t) \in [0, B]$, is a Brownian motion constrained between the barriers at levels 0 and B. We assume that the system starts at time $t = 0$ with an initial storage level $V(0) = x$.

If $L(t)$ and $F(t)$ are the cumulative renewable energy wasted due to overflows, and the amount of fossil fuel energy used till time t, respectively, then,

1. $V(t) = X(t) + F(t) - L(t)$, $V(t) \in [0, B]$ for all $t \geq 0$.

2. F can increase only when V is 0, and L can increase only when $V = B$, i.e., the fossil fuel sources are turned on only to meet the excess demand when

the energy level in the storage hits the level 0. Likewise, excess renewable energy is wasted only when the energy level in the storage is B.

Thus the average loss and the average fossil energy used are ([47]),

$$\bar{L} = \lim_{t\to\infty} \frac{L(t)}{t}, \bar{F} = \lim_{t\to\infty} \frac{F(t)}{t}.$$

The associated variances are given by,

$$\sigma_L^2 = \lim_{t\to\infty} \frac{Var(L(t))}{t}, \sigma_F^2 = \lim_{t\to\infty} \frac{Var(F(t))}{t}.$$

To obtain the above quantities of interest, we decompose the paths of the process $V(t)$ into i.i.d. cycles and use the resulting regenerative structure. Define the following,

$$T_0 := \inf\{t \geq 0 : V(t) = 0\}.$$

$$V_{n+1}^\star(t) := V(T_n + t), L_{n+1}^\star(t) := L(T_n + t) - L(T_n),$$
$$U_{n+1}^\star(t) := U(T_n + t) - U(T_n), \text{ where}$$
$$T_{n+1} := \text{ smallest } t > T_n \text{ such that}$$
$$V(t) = 0 \text{ and } Z(s) = b \text{ for some } s \in (T_n, t)\,.$$

We see that the regeneration times are $T_1, T_2, \ldots$, and letting $\tau := T_1 - T_0$, we have,

$$\bar{L} = \frac{\mathbb{E}_0\left[L(\tau)\right]}{\mathbb{E}_0(\tau)}, \quad \bar{F} = \frac{\mathbb{E}_0\left[F(\tau)\right]}{\mathbb{E}_0\left(\tau\right)}.$$

Let us denote by $\pi(\cdot)$ the stationary distribution of $V(t)$ (existence of which can be shown using renewal theory). Also assume that $V(0) = 0$. Then if f is any real-valued twice continuously differentiable function, we have, $f(V(t)) = f(V(0)) + \sigma \int_0^t f'(V)\mathrm{d}X + \int_0^t \Gamma f(V)\mathrm{d}s + f'(0)F(t) - f'(B)L(t)$. Let $t = \tau$ in the above, and note that $f(V(\tau)) = f(V(0)) = f(0)$. Taking the expectations of both sides, noting that $\mathbb{E}\int_0^t f'(V)\mathrm{d}X = 0$, and performing some algebraic manipulations, we have,

$$\int_0^B \Gamma f(z)\pi(\mathrm{d}z) + f'(0)\bar{F} - f'(B)\bar{L} = 0. \tag{7.3}$$

By proper choice of the functions f in equation (7.3), such as $f(v) = v$ or $f(v) = v^2$, one can calculate relevant quantities of interest, which leads us to the following theorem.

Theorem 15. *If $\mu = 0$, then $\bar{L} = \bar{F} = \frac{\sigma^2}{2B}$ and π is the uniform distribution on $[0, B]$. Otherwise, for $\mu < 0$, where the renewable energy is not sufficient to fully meet the loads, let $\theta = \frac{2\mu}{\sigma^2}$. Then,*

$$\bar{L} = \frac{\mu}{1 - \exp^{-\theta B}}, \bar{F} = \frac{\mu}{\exp^{\theta B} - 1}. \tag{7.4}$$

$$SCV(L) = \begin{cases} \frac{2B}{3} \text{ if } \mu = 0, \\ \frac{2(1-\exp(2\theta B)) + 4\theta B \exp(\theta B)}{-\theta(1-\exp(\theta B))^2}, \end{cases}$$

where $SCV(L)$ is the squared coefficient of variation of loss L, i.e., its standard deviation divided by its mean.

The relation (7.4) is the limit of the relation (7.2), justifying the random-walk model discussed in Section 7.3 as a possibly reasonable assumption for a market

having a large number of players. Moreover, as earlier, we note that if $\mu < 0$ (i.e., on average the renewable supply is lesser than the demand), then the renewable energy wastage decreases exponentially with the size of the storage B.

7.5 Correlated Uncertainty Between Loads and Renewables

The assumptions in the previous sections assume that the energy source and consumers behave in independent and identically distributed manner over time. This is not exactly the practical case. Inspired by stochastic fluid analysis, in this section we extend the analysis to allow for more detailed models of individual generators and loads ([7, 52, 60, 72, 76]). Suppose there are m sources and n consumers which are "coupled" through the storage. That is, the energy produced by the sources is used to fill the storage, and the consumers withdraw from it. This approach is potentially applicable for commercial and industrial loads.

To illustrate the approach, suppose that each source can be in one of two states, active or passive, and the time taken to transition from one state to the other is exponentially distributed. When a source is active, it produces energy at a constant rate c_1, while no energy is produced when it is passive. The rates of transition from active to passive and vice-versa are f_1, r_1 respectively. Similarly each consumer transitions from active to passive at rate f_2, and vice versa at a rate r_2, and consumes energy at a constant rate c_2 units while in the active state. Such a system can be described by a Markov process with state (V, i, j), where $V(t)$ is the storage level at time t, and market state $(i(t), j(t))$ where i and j are the numbers of active sources and consumers respectively, with $0 \leq i \leq n_1$ and $0 \leq j \leq n_2$; . The processes describing the number of active sources/consumers are birth-death

processes. Letting

$$\begin{aligned} p_1(t;i) &:= \mathrm{P}(i \text{ sources are active at time } t), \\ p_2(t;j) &:= \mathrm{P}(j \text{ consumers are active at time } t), \text{ and} \\ \mathbf{p}_i(t) &:= \left(p_2(t;0), p_2(t;1), \ldots, p_2(t;n_i)\right), \end{aligned}$$

we have, $\frac{d}{dt}\mathbf{p}_i(t) = \mathbf{p}_i(t)\mathbf{M}_i$, where, $\mathbf{M}_i$ is the corresponding transition rate matrix. Let $p(t;i,j) = p_1(t;i)p_2(t;j)$,

$$\mathbf{p}(t) := \left(\mathbf{p}(t;0,0), \mathbf{p}(t;0,1), \cdots, \mathbf{p}(t;n_1,n_2)\right),$$

(under lexicographic ordering),

$$\begin{aligned} &P(t,x;i,j) := \\ &P(\text{ storage level } \leq x, \text{ market state } = (i,j) \text{ at time } t), \end{aligned}$$

and $\mathbf{P}(t,x)$ the lexicographic arrangement of $\{P(t,x;i,j)\}$. Then,

$$\frac{d}{dt}\mathbf{p}(t) = \mathbf{p}(t)\mathbf{M}, \text{ where } \mathbf{M} := \mathbf{M}_1 \otimes \mathbf{I}(n) + \mathbf{I}(m) \otimes \mathbf{M}_2,$$

and $\mathbf{I}(k)$ is the $k+1$ dimensional identity matrix and $\otimes$ is the Kronecker product, and

$$\frac{\partial}{\partial t}\mathbf{P} + \frac{\partial}{\partial x}\mathbf{PD} = \mathbf{PM}, t \geq 0, 0 < x < B, \tag{7.5}$$

where the "drift matrix" D is given by, $\mathbf{D} := c_1\mathbf{E}(n_1)\otimes\mathbf{I}(n_2) - c_2\mathbf{I}(n_1)\otimes\mathbf{E}(n_2)$, with $\mathbf{E}(n_2) = \text{diag}(0,1,\cdots,n_2)$. Letting π be the continuous steady state solution of the equa-

tion (7.5), we obtain, $\frac{d}{dx}\boldsymbol{\pi}(x)\mathbf{D} = \boldsymbol{\pi}(x)\mathbf{M}, \quad 0 \leq x \leq B$. That is, for any $x \in (0, B)$, $P($ storage content $\leq x$ and market state is$(i, j)) = \pi(x; i, j)$. Spectral expansion yields, $\boldsymbol{\pi}(x) = \sum_l a_l \exp(z_l x)\boldsymbol{\phi}(l)$, where $\{z_l, \boldsymbol{\phi}(l)\}$ are solutions of the eigenvalue problem, $z\boldsymbol{\phi}\boldsymbol{D} = \boldsymbol{\phi}\boldsymbol{M}$, with a_l to be determined by the boundary conditions.

7.5.1 Performance Analysis

Let $w_1(i), w_2(j), 0 \leq i \leq n_1, 0 \leq j \leq n_2$ be the stationary probabilities that i sources and j consumers are active, and let $\mathbf{w}_k$ for $k = 1, 2$ be the vectors comprising these probabilities. The $w_k(i)$'s can be easily solved to obtain the stationary distribution of the market process: $\mathbf{w} := \mathbf{w}_1 \otimes \mathbf{w}_2$. The total average *energy produced* is then simply $c_1 \sum_{i=0}^{m} i w_1(i)$, while the average *demand* is $c_2 \sum_{j=0}^{n} j w_2(j) = \frac{n c_2 r_2}{f_2 + r_2}$.

Since the storage levels will occasionally hit the boundaries at 0 and B, it is clear that the energy utilized will be lesser than both of the above quantities. Thus the renewable energy lost due to the limitation on the size of the storage is simply the sum $\sum P($storage is full and market state is $(i, j))\,(c_1 i - c_2 j)$ over all the states in which the loss-rate $(c_1 i - c_2 j)$ is positive.

7.6 Conclusions

This section addresses the problem of energy storage sizing in a microgrid setting with high penetration of intermittent resources such as wind and solar. By considering the energy storage as a service provider, we propose a queuing theoretical approach to study the fundamental coupling between energy storage sizing and uncertainty levels from the net load. Three models with different levels of details provide a suite of tools for microgrid operators to determine an optimal size of energy storage given a level of renewable and load uncertainties. Numerical examples based on realistic wind and load data suggest that the proposed approach could be a new avenue of research for optimal sizing of energy storage

in renewable-rich power systems.

This is only a first step toward systematically understanding the fundamental role of uncertainty on sizing of energy storage. There are many fruitful directions for future research. One is to analyze the impact of line constraints on the optimal location of energy storage. Also it would be worthwhile to assess the effectiveness of this framework in a larger-scale realistic system.

8. CONCLUSION

We conclude by summarizing our key results, and explaining their significance in the larger scheme of things. Then, we point out some ready extensions of these results. This is followed by identifying outstanding challenges which will need breakthroughs.

We have studied two important classes of cyberphysical systems involving the operation of distributed uncertain dynamic systems. Within the first class of systems, communication networks, we have developed a framework for scheduling data transmissions over multi-hop wireless networks under a hard delay bound. We have obtained the policies that are highly decentralized, and involve low computational complexity. Our approach enables us to design policies that take "packet level" decisions, rather than "queue length level" decisions. This has allowed us to optimize network performance with regards to providing hard delay guarantees, while operating in a decentralized mode. While in this thesis we have dealt only with the case of power constraints on individual wireless nodes, it is possible to extend the results to more general cases in which there is wireless interference between the multiple channels. However, in that case, the convergence speed of the algorithms will be slower since the state space is larger.

We have also addressed the problem of smoothness of packet delivery, important when they are carrying sensor measurements in sensor-actuator networks or networked control. We have shown that the MaxWeight scheduler provides asymptotically smooth service. However, it is a centralized scheduler and requires the nodes to share their queue lengths continually. In general it is not known whether there is a decentralized scheduler that can guarantee asymptotically smooth service

process for the case of multi-hop networks.

For the problem of video streaming, which is growing at an increasing rate, it is important to design policies that enhance Quality of Experience. We have developed ..... Mention your major results ...An open problem is that presently we do not have a theory to account for specific demands such as the order in which packets arrive at the destination or creating multiple copies of a video packet of varying video quality.

In the second application area, smart grid, focused on in this thesis, we have examined the problem faced by the Independent System Operator (ISO). This is a problem involving decentralized stochastic control of dynamical system through price coordination by a system operator. Past works in the decentralized stochastic control literature have assumed that the system dynamics of the combined system are known to each agent , which is not a realistic or tenable assumption in large systems such as the power grid where in addition to the sheer magnitude of information sharing that would entail, the agents may even be averse to violating their privacy or competitive advantage by disclosing such information. In contrast, in economics, general equilibrium theory has been developed without such assumptions. We have carefully analyzed the role played by uncertainty whether it is present, and is it common or private, in allowing optimal coordination. We have shown that the specific case of multiple linear quadratic Gaussian systems is very amenable to optimal price-based coordination without enormous complexity of information sharing to handle the uncertainties involved. It is worthy of noting that we can still attain optimal centralized performance without the uncertainty "tree" being known globally, in comparison with the generally privately observed noise case.

One future problem of interest centers is how to coordinate when the band-

width for communication is constrained. It is of interest to determine how system performance is affected by such communication constraints.

In the Privately Observed Noise case, Algorithm 2 required the topology of the uncertainty tree of the combined system comprising the M agents to be globally known. This assumption allowed the agents to bid in a non-anticipative manner on each possible sample path ω, which gave rise to a highly adaptive Iterative Bidding scheme, which was shown to be optimal. However, knowledge of the uncertainty tree's topology might be impractical since its size grows exponentially in the number of users. It is highly desirable to remove this assumption. It would, however, then be appropriate to compare the performance of the scheme with the performance within some restricted class of policies rather than with the optimal centralized policy.

REFERENCES

[1] Cisco visual networking index (vni). http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

[2] Alan S. Manne. Linear programming and sequential decisions. *Management Science*, 6:259 –267, 1960.

[3] Eitan Altman. Constrained Markov Decision Processes with Total Cost Criteria: Occupation Measures and Primal LP. *Mathematical Methods of Operations Research*, 43(1):45–72, 1996.

[4] Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, March 1999.

[5] J. Andrews, S. Shakkottai, R. Heath, N. Jindal, M. Haenggi, R. Berry, Dongning Guo, M. Neely, S. Weber, S. Jafar, and A. Yener. Rethinking Information Theory for Mobile Ad Hoc Networks. *Communications Magazine, IEEE*, 46(12):94–101, December 2008.

[6] J.G. Andrews, R.K. Ganti, M. Haenggi, N. Jindal, and S. Weber. A primer on spatial modeling and analysis in wireless networks. *IEEE Communications Magazine*, 48(11):156–163, November 2010.

[7] Anick, D. and Mitra, D. and Sondhi, M. M. Stochastic Theory of a Data-Handling System with Multiple Sources. *Bell System Technical Journal*, 61(8):1871–1894, 1982.

[8] P. S. Ansell, K. D. Glazebrook, J. Nio-Mora, and M. O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathe-*

matical Methods of Operations Research, 57(1):21–39, 2003.

[9] Jay Apt. The spectrum of power from wind turbines. *Journal of Power Sources*, 169(2):369–374, 2007.

[10] Kenneth J Arrow. An extension of the basic theorems of classical welfare economics. *Proceedings of the Second Berkeley Symposium on mathematical Statistics and Probability*, pages 507–532, 1951.

[11] Kenneth J Arrow. General Economic Equilibrium: Purpose, Analytic Techniques, Collective Choice. *American Economic Review*, 64(3):253–72, June 1974.

[12] Aditya Mahajan Ashutosh Nayyar and Demosthenis Teneketzis. Chapter 4: The Common-Information Approach to Decentralized Stochastic Control. In *Information and Control in Networks*, pages 123–156. Springer-Verlag, 2014.

[13] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[14] Berger, Arthur W. and Whitt, Ward. The Brownian Approximation for Rate-Control Throttles and the G/G/1/C Queue. *Discrete Event Dynamic Systems*, 2(1):7–60, 1992.

[15] D. P. Bertsekas. *Nonlinear Programming*. Athena scientific. Athena Scientific, 1999.

[16] Dimitri P. Bertsekas, Asuman E. Ozdaglar, and Angelia Nedic. *Convex analysis and optimization*. Athena scientific optimization and computation series. Athena Scientific, Belmont (Mass.), 2003.

[17] Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., 1999.

[18] David Blackwell. Discrete Dynamic Programming. *Annals of Mathematical Statistics*, 33:719–726, 1962.

[19] David Blackwell and M.A. Girshick. *Theory of games and statistical decisions*. John Wiley and Sons, New York, 1954. Republished by Dover in 1979. MR:0070134. Zbl:0056.36303.

[20] Vivek S. Borkar. Control of Markov Chains with Long-Run Average Cost Criterion. In W. Fleming and P.L. Lions, editors, *The IMA Volumes in Mathematics and Its Applications*, pages 57–77. Springer, 1988.

[21] Vivek S. Borkar. *Stochastic Approximation : A Dynamical Systems Viewpoint*. Cambridge University Press New Delhi, Cambridge, 2008.

[22] Stephen Boyd. `https://web.stanford.edu/class/ee392o/subgrad_method.pdf`.

[23] Shengrong Bu and F. R. Yu. A game-theoretical scheme in the smart grid with demand-side management: Towards a smart cyber-physical power infrastructure. *IEEE Transactions on Emerging Topics in Computing*, 1(1):22–32, June 2013.

[24] L. Bui, R. Srikant, and A Stolyar. Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing. In *INFOCOM 2009, IEEE*, pages 2936–2940, April 2009.

[25] L.X. Bui, R. Srikant, and A. Stolyar. A novel architecture for reduction of delay and queueing structure complexity in the back-pressure algorithm.

IEEE/ACM Transactions on Networking, 19(6):1597–1609, Dec 2011.

[26] H.-P. Chao, S.S. Oren, A. Papalexopoulos, D.J. Sobajic, and R. Wilson. Interface between engineering and market operations in restructured electricity systems. *Proceedings of the IEEE*, 93(11):1984–1997, Nov 2005.

[27] Aimee E. Curtright and Jay Apt. The character of power output from utility-scale photovoltaic systems. *Progress in Photovoltaics: Research and Applications*, 16(3):241–247, 2008.

[28] Luca De Cicco, Saverio Mascolo, and Vittorio Palmisano. Feedback control for adaptive live video streaming. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, MMSys '11, pages 145–156, New York, NY, USA, 2011. ACM.

[29] J. De Vriendt, D. De Vleeschauwer, and D. Robinson. Model for estimating qoe of video delivered using http adaptive streaming. In *IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), 2013*, pages 1288–1293, May 2013.

[30] Demosthenis Teneketzis. *Perturbation Methods in Decentralized Stochastic Control*. PhD thesis, Massachusetts Institute of Technology, November 1976.

[31] Arne Jensen E. Brockmeyer, HL Halstrm and Agner Krarup Erlang. The life and works of AK Erlang. 1948.

[32] Eitan Altman and Adam Shwartz. Markov decision problems and state-action frequencies. *SIAM J. CONTROL AND OPTIMIZATION*, 29(4):786–809, 1991.

[33] AK Erlang. Probability and Telephone Calls. *Nyt Tidsskrift Mat. Ser. B*, 20:33–39, 1909.

[34] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst. Theory Appl.*, 72(3-4):311–359, Dec 2012.

[35] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst. Theory Appl.*, 72(3-4):311–359, December 2012.

[36] G. R. Gajjar, S. A. Khaparde, P. Nagaraju, and S. A. Soman. Application of actor-critic learning algorithm for optimal bidding problem of a genco. *IEEE Transactions on Power Systems*, 18(1):11–18, February 2003.

[37] F.D. Galiana, F. Bouffard, J.M. Arroyo, and J.F. Restrepo. Scheduling and pricing of coupled energy and primary, secondary, and tertiary reserves. *Proceedings of the IEEE*, 93(11):1970–1983, Nov 2005.

[38] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

[39] Feng Gao, Gerald B Sheble, Kory W Hedman, and Chien-Ning Yu. Optimal bidding strategy for gencos based on parametric linear programming considering incomplete information. *International Journal of Electrical Power & Energy Systems*, 66:272–279, mar 2015.

[40] J.C. Gittins K. Glazebrook and R. Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[41] Christophe Guille and George Gross. A conceptual framework for the vehicle-to-grid (v2g) implementation. *Energy Policy*, 37(11):4379 – 4390, 2009.

[42] G.R. Gupta and N. Shroff. Delay analysis for multi-hop wireless networks. In *Proc. IEEE INFOCOM 2009*, pages 2356–2364, April 2009.

[43] K. Jagannathan H. Ahmed and S. Bhashyam. Fair scheduling with deadline guarantees in single-hop networks. In *Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014*, pages 1–7, Jan 2014.

[44] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14(3):502–525, June 1982.

[45] Hajime Kawai. A variance minimization problem for a Markov decision process. *European Journal of Operational Research*, 31(1):140–145, 1987.

[46] Haozhi Xiong, Ruogu Li, A. Eryilmaz and E. Ekici. Delay-aware cross-layer design for network utility maximization in multi-hop networks. *Selected Areas in Communications, IEEE Journal on*, 29(5):951–959, May 2011.

[47] Harrison, J. Michael. *Brownian motion and stochastic flow systems*. Wiley series in probability and mathematical statistics. Wiley, New York, 1985.

[48] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen. Initial delay vs. interruptions: Between the devil and the deep blue sea. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 1–6, July 2012.

[49] C.L. Hwang and A.S.M. Masud. *Multiple Objective Decision Making Methods and Applications*. Springer-Verlag New York, Inc., 1979.

[50] I-Hong Hou and V.S. Borkar and P.R. Kumar. A Theory of QoS for Wireless. In *IEEE INFOCOM 2009*, pages 486–494, April 2009.

[51] M. D. Ilić, Jhi-Young Joo, Le Xie, M. Prica, and Rotering N. A decision-making framework and simulator for sustainable electric energy systems. *IEEE Transactions on Sustainable Energy*, 2(1):37–49, January 2011.

[52] Isi Mitrani and Ram Chakka. Spectral expansion solution for a class of markov models: application and comparison with the matrix-geometric method. *Performance Evaluation*, 23(3):241 – 260, 1995.

[53] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti. Achieving single channel, full duplex wireless communication. In *ACM MobiCom*, 2010.

[54] J.A. Filar, L.C.M. Kallenberg and H.M. Lee. Variance-penalized markov decision processes. *Math. Oper. Res.*, 14(1):147–161, March 1989.

[55] Bo Ji, Changhee Joo, and N.B. Shroff. Delay-based back-pressure scheduling in multi-hop wireless networks. In *IEEE INFOCOM, 2011 Proceedings*, pages 2579–2587, April 2011.

[56] Liyan Jia and Lang Tong. Day ahead dynamic pricing for demand response in dynamic environments. In *IEEE 52nd Annual Conference on Decision and Control (CDC)*, pages 5608–5613, December 2013.

[57] K.D. Glazebrook, D. Ruiz-Hernandez and C. Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672, 2006.

[58] H. J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer Verlag, New York, 1997.

[59] Kyu Seob Kim, Chih-Ping Li and E. Modiano. Scheduling multicast traffic with deadlines in wireless networks. In *Proc. IEEE INFOCOM, 2014*, pages 2193–2201, April 2014.

[60] L. C. G. Rogers. Fluid Models in Queueing Theory and Wiener-Hopf Factorization of Markov Chains. *The Annals of Applied Probability*, 4(2):390–413, 1994.

[61] Bin Li, Ruogu Li, and Atilla Eryilmaz. Heavy-traffic-optimal scheduling with regular service guarantees in wireless networks. In *Proceedings of the Fourteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '13, pages 79–88. ACM, 2013.

[62] Ruogu Li, A. Eryilmaz, and Bin Li. Throughput-optimal wireless scheduling with regulated inter-service times. In *INFOCOM, 2013 Proceedings IEEE*, pages 2616–2624, April 2013.

[63] Guanfeng Liang and Guanfeng Liang. Effect of delay and buffering on jitter-free streaming over random vbr channels. *Multimedia, IEEE Transactions on*, 10(6):1128–1141, Oct 2008.

[64] Jian Ni R. Srikant Libin Jiang, Mathieu Leconte and Jean Walrand. Fast Mixing of Parallel Glauber Dynamics and Low-Delay CSMA Scheduling. *IEEE Transactions on Information Theory*, 58(10):6541–6555, Dec 2012.

[65] David Lindley. Smart grids: The energy storage problem. *Nature 463*, pages 18–20, 2010.

[66] Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, Nov 2010.

[67] A. L. Stolyar M. Andrews, K. Jung. Stability of the Max-Weight Routing and Scheduling Protocol in Dynamic Networks and at Critical Loads. *STOC*, pages 145–154, June 2007.

[68] K. Ramanan A. L. Stolyar M. Andrews, K. Kumaran and R. Vijayakumar. Providing quality of service over a shared wireless link. *IEEE Communications Magazine*, 39(2):150–154, 2001.

[69] M. De Lara, P. Carpentier, J.P. Chancelier and V. Leclere. Optimization methods for the smart grid. *Report Commissioned by Conseil Franais de l'Energie*, Oct 2014.

[70] M. Duarte and A. Sabharwal. Full-duplex wireless communications using off-the-shelf radios: Feasibility and first results. In *Proceedings of ASILOMAR*, pages 1558–1562, 2010.

[71] Z. Mao, C.E. Koksal, and N.B. Shroff. Optimal online scheduling with arbitrary hard deadlines in multihop communication networks. *Networking, IEEE/ACM Transactions on*, PP(99):1–1, 2014.

[72] Marcel F. Neuts. *Matrix-geometric solutions in stochastic models - an algorithmic approach*. Dover Publications, 1994.

[73] Jacob Marschak and Roy Radner. *Economic Theory of Teams*. Yale University Press, 1972.

[74] Masami Kurano. Markov decision processes with a minimum-variance criterion. *Journal of Mathematical Analysis and Applications*, 123(2):572–583, 1987.

[75] S. P. Meyn and R.L. Tweedie. Stability of Markovian Processes I: Criteria for Discrete-Time chains. *Advances in Applied Probability*, 24(3):542–574, 1992.

[76] Mitra, Debasis. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20(3):pp. 646–

676, 1988.

[77] M.J. Neely, E. Modiano and Chih-ping Li. Fairness and optimal stochastic control for heterogeneous networks. *IEEE/ACM Transactions on Networking*, 16(2):396–409, April 2008.

[78] A. H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Transactions on Smart Grid*, 1(3):320–331, Dec 2010.

[79] Michio Morishima. *Walras' Economics : A Pure Theory of Capital and Money*. Cambridge University Press, 1977.

[80] Krzysztof C. Kiwiel N.Z. Shor and Andrzej Ruszcaynski. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag New York, Inc., New York, NY, USA, 1985.

[81] P. Carpentier, J.P. Chancelier, G. Cohen and M. De Lara. *Stochastic Multi-Stage Optimization. At the Crossroads between Discrete Time Stochastic Control and Stochastic Programming*. Springer International Publishing, 2015.

[82] P. R. Kumar and P. Varaiya. *Stochastic systems: Estimation, identification and adaptive control*. Prentice Hall Inc., Englewood Cliffs, 1986.

[83] Ali ParandehGheibi, Muriel Mdard, Asuman E. Ozdaglar, and Srinivas Shakkottai. Avoiding Interruptions - A QoE Reliability Function for Streaming Media Applications. *IEEE Journal on Selected Areas in Communications*, 29(5):1064–1074, 2011.

[84] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition,

1994.

[85] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[86] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[87] F. Rahimi and A. Ipakchi. Demand response as a market resource under the smart grid paradigm. *IEEE Transactions on Smart Grid*, 1(1):82–88, June 2010.

[88] Rahul Singh, I-Hong Hou and P.R. Kumar. Pathwise performance of debt based policies for wireless networks with hard delay constraints. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 7838–7843, Dec 2013.

[89] T.S. Rappaport. *Wireless Communications: Principles and practice*. Prentice Hall, New York, 2002.

[90] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, Sept. 1951.

[91] S.M. Ryan, R.J.-B. Wets, D.L. Woodruff, C. Silva-Monroy, and J.-P. Watson. Toward scalable, parallel progressive hedging for stochastic unit commitment. In *2013 IEEE Power and Energy Society General Meeting (PES)*, pages 1–5, July 2013.

[92] S. Asmussen. *Applied Probability and Queues*. Wiley, 1987.

[93] S. Zhang and V. K. N. Lau. Resource allocation for OFDMA system with orthogonal relay using rateless code. *IEEE Transactions on Wireless Communications,*, 7, Nov. 2008.

[94] N.R. Sandell, P. Varaiya, M. Athans, and M.G. Safonov. Survey of decentralized control methods for large scale systems. *Automatic Control, IEEE Transactions on*, 23(2):108–128, Apr 1978.

[95] S. Shakkottai, T.S. Rappaport, and P.C. Karlsson. Cross-layer design for wireless networks. *IEEE Communications Magazine*, 41(10):74–80, Oct 2003.

[96] Sanjay Shakkottai and Alexander L. Stolyar. Scheduling algorithms for a mixture of real-time and non-real-time data in hdr. In Nelson L.S. da Fonseca Jorge Moreira de Souza and Edmundo A. de Souza e Silva, editors, *Teletraffic Engineering in the Internet Era. Proceedings of the International Teletraffic Congress - ITC-I7*, volume 4 of *Teletraffic Science and Engineering*, pages 793 – 804. Elsevier, 2001.

[97] Rahul Singh, I-Hong Hou, and P.R. Kumar. Fluctuation analysis of debt based policies for wireless networks with hard delay constraints. In *IEEE INFOCOM, 2014 Proceedings*, pages 2400–2408, April 2014.

[98] H. Song, Chen-Ching Liu, J. Lawarree, and R. W. Dahlgren. Optimal electricity supply bidding by markov decision process. *IEEE Transactions on Power Systems*, 15(2):618–624, May 2000.

[99] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 02 2004.

[100] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 2004.

[101] Charlotte Striebel. Sufficient statistics in the optimum control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12(3):576 – 592, 1965.

[102] L. Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, Dec 1992.

[103] A.A. Thatte and Le Xie. Towards a unified operational value index of energy storage in smart grid environment. *IEEE Transactions on Smart Grid*, 3(3):1418–1426, Sept 2012.

[104] Guibin Tian and Yong Liu. Towards agile and smooth video adaptation in dynamic http streaming. In *Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '12, pages 109–120, 2012.

[105] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, New York, NY, USA, 2005.

[106] Jan H. van Schuppen and Tiziano Villa. *Coordination Control of Distributed Systems*. Springer Publishing Company, Incorporated, 2014.

[107] Yunpeng Wang, W. Saad, Zhu Han, H. V. Poor, and T. Basar. A game-theoretic approach to energy trading in the smart grid. *IEEE Transactions on Smart Grid*, 5(3):1439–1450, May 2014.

[108] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

[109] Richard R. Weber and Gideon Weiss. On an Index Policy for Restless Bandits. *Journal of Applied Probability*, 27(3):pp. 637–648, 1990.

[110] P. Whittle. Multi-armed bandits and the Gittins index. *J. R. Statist. Soc. B*, 42:143–149, 1980.

[111] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

[112] Williams, R. J. Asymptotic variance parameters for the boundary local times of reflected brownian motion on a compact interval. *Journal of Applied Probability*, 29(4):pp. 996–1002, 1992.

[113] H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

[114] H.S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, Nov 1971.

[115] F.F. Wu, K. Moslehi, and A. Bose. Power system control centers: Past, present, and future. *Proceedings of the IEEE*, 93(11):1890–1908, Nov 2005.

[116] Jeffrey Wu. *Sufficient statistics for team decision problems*. PhD thesis, Stanford University, November 2013.

[117] Xiaojun Lin and Ness B. Shroff. Joint rate control and scheduling in multihop wireless networks. In *in Proceedings of IEEE Conference on Decision and Control*, pages 1484–1489, 2004.

[118] Yuedong 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*, 16(3):813–827, April 2014.

[119] Yuedong Xu, S.E. Elayoubi, E. Altman, and R. El-Azouzi. Impact of flow-level dynamics on qoe of video streaming in wireless networks. In *INFOCOM, 2013 Proceedings IEEE*, pages 2715–2723, April 2013.

[120] Xueying Guo, Sheng Zhou, Zhisheng Niu and P.R. Kumar. Optimal wake-up mechanism for single base station with sleep mode. In *International Teletraffic Congress (ITC)*, Sept 2013.

[121] Lei Ying, Sanjay Shakkottai, Aneesh Reddy, and Shihuan Liu. On combining shortest-path and back-pressure routing over multihop wireless networks. *IEEE/ACM Trans. Networking*, 19(3):841–854, June 2011.

[122] Serdar Yuksel and Tamer Basar. *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Systems & Control: Foundations & Applications. Springer, New York, NY, 2013.

[123] Quanyan Zhu, P. Sauer, and T. Basar. Value of demand response in the smart grid. In *IEEE Power and Energy Conference at Illinois (PECI)*, pages 76–82, February 2013.