Optimal and Suboptimal Policies for Opportunistic Spectrum Access: A Resource Allocation Approach

by

Sahand Haji Ali Ahmad

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2010

Doctoral Committee:

Associate Professor Mingyan Liu, Chair
Professor Romesh Saigal
Professor Demosthenis Teneketzis
Associate Professor Achilleas Anastasopoulos

To myself for my exemplar persistence and hard work which made this possible and also to Mom and dad for their endless encouragement, support and love

ACKNOWLEDGEMENTS

I am much indebted to my thesis advisor, Professor Mingyan Liu. Mingyan has been a great advisor and I consider myself very fortunate to have the opportunity to work with her. For me she was the advisor who made all the difference. Without her valuable help and guidance, the research presented in this thesis would not have been possible in any sorts.

I would also like to express my gratitude toward my thesis committee members. I have had the opportunity of working with Professor Demos Teneketzis as a student. I would like to thank him for all his encouragements and advice throughout the time I worked on this thesis. I learned the fundamentals of stochastic control from him. I would also like to thank Professor Anastasopoulos for his encouragement and help not only with my dissertation but also with my entire graduate studies at university of Michigan. I like to thank Prof.Saigal for his insightful questions and comments on this thesis particularly during the proposal session. There are many other professors in the Electrical Engineering department whom I would like to thank, in particular Professor Momcilovic whom I learned from him in class and also Professor Pradhan whom I passed my qualifying exam under him.

I would like to thank my parents for their unconditional support and love whenever I needed them. they have consistently been supportive and encouraging throughout my graduate studies. Thanks to my brother whom without his early life encouragements I would have never pursued education to the highest level possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER 1

Introduction

Wireless communication technology has proved to be one of the most fundamental modern technologies affecting our daily lives. This is amplified by a wide variety of emerging applications. As we all know wireless spectrum is limited and with growing need and applications of wireless communication systems, interference will be one of the most serious challenges ahead if not the foremost hurdle in achieving widespread applications of wireless systems. This includes wireless sensor networks (WSN) and mobile ad hoc networks (MANETs). There have been several methods proposed for overcoming spectrum limitations; one of the more recent development is the idea of open access, whereby secondary users or unlicensed users are allowed access to spectrum licensed to other, primary users when it is not in active use. This leads to the notion of cognitive radio, a wireless transceiver that is highly agile and spectrum aware and thus can take advantage of instantaneous spectrum availability to dynamically access radio frequencies to perform data transmission. This has been an area of extensive research in recent years.

The overarching theme of this thesis is the investigation of good opportunistic spectrum access and sharing schemes within the above context using a resource allocation framework. The two main analytical approaches taken in this thesis are a

stochastic optimization based approach and a game theoretic approach.

Using the first approach we study an optimal channel sensing and access problem from the point of view of a single (secondary) user. Activities of primary users and possibly other secondary users are modeled as part of a channel of time-varying availability perceived by this single user. With this assumption, the objective is for the user to determine at each instance of time (under a discrete time model) which (subset of) channel(s) to sense – the ones sensed available can subsequently be used for data transmission – so as to maximize its total reward over a finite or infinite horizon. A key assumption here is that the user is limited in its channel sensing capability in that it can only sense a small number of channels at a time compared to the total number of channels available. It must therefore make judicious decisions on which channels to sense over time to maximize its transmission opportunity. This problem is further separated into two cases, where the user is limited to sense one channel at a time, and where the user can sense more than one channel at a time. These constitute the first two technical chapters of this thesis.

Using the second, game theoretic approach we study a spectrum sharing problem in the presence of a group of peer (secondary) users. This is formulated as a generalized form of the classical congestion game, referred to as a network congestion game in this thesis: a group of wireless users share a set of common channels, each tries to find the best channel for itself, and the utility of a channel depends not only on the channel itself but also on how many interfering users are simultaneously using it. This constitutes the third technical chapter of this thesis.

Below we describe each of these two problems in more detail along with our main results and contributions.

1.1 Problem 1: Opportunistic Spectrum Access as a Restless Bandit Problem

This problem (referred to as optimal probing or OP) concerns the opportunistic communication over multiple channels where the state ("good" or "bad") of each channel evolves as independent and identically distributed Markov processes. A user, with limited channel sensing capability, chooses one channel to sense and subsequently access (based on the sensing result) in each time slot. A reward is obtained whenever the user senses and accesses a "good" channel. The objective is to design a channel selection policy that maximizes the expected total (discounted or average) reward accrued over a finite or infinite horizon. This problem can be cast as a Partially Observable Markov Decision Process (POMDP) or a restless multi-armed bandit process, to which optimal solutions are often intractable.

This problem was first introduced by Zhao [1] where they established the optimality of a simple greedy policy – always sensing the channel that has the highest current probability of being available – for the simple case of 2 channels. The approach they used was based on proving the problem by proving optimality of greedy policy almost surely (path wise) for all possible consecutive outcomes of channel states. That approach was not useful for the case of more than 2 channels as myopic policy is no longer path wise optimal (almost surely), so a new approach had to be introduced. The idea that we used was to define a function based on the order of channels played (not necessarily depending on their ranking) and then use a coupling argument to prove that this function achieves its maximum value if channels are played in the order of their ranks. This proved the optimality of the myopic policy in the general case of multiple channels.

Our main results and contributions are as follows: For the finite horizon dis-

counted reward case, we show that the same myopic policy that maximizes the immediate one-step reward is optimal when the channel state transitions are positively correlated over time. When the state transitions are negatively correlated, it is shown that the same policy is optimal when the number of channels is limited to 2 or 3, while presenting a counterexample for the case of 4 channels. The same optimality result is then extended to the case of infinite horizon discounted reward and average reward cases.

We further extended this problem to the case of allowing the users to sense a fixed number (more than one) of channels at a time (referred to as multi-channel optimal probing, or MC-OP). We proved the optimality of the same greedy policy under the same bursty channel condition.

In both the single-play and multiple-play cases, the greedy policy depends only on the ordering of the channel occupancy probabilities. Therefor it is fairly robust to the knowledge of initial probabilities as long as the order of channel occupancies are preserved .

1.2 Problem 2: Opportunistic Spectrum Sharing as a Congestion Game

In this problem we turned our attention to the case where multiple users wish to efficiently share multiple channels. Each user may only use one channel at a time, and each channel may present different value to different users. However, the value of any channel to a user decreases as it is shared by more interfering users. The interference relationship among users is represented by a graph. This is modeled as a network congestion game (NCG), which is a generalization of the classical congestion game (CG). In a classical congestion game [2], multiple users share the same set of resources and a user's payoff for using any resource is a function of the total number

of users sharing it. This game enjoys some very appealing properties, including the existence of a pure strategy Nash equilibrium (NE) and that every improvement path is finite and leads to such a NE (also called the finite improvement property or FIP), which is also a local optimum to a potential function. On the other hand, it does not model well spectrum sharing and spatial reuse in a wireless network, where resources (interpreted as channels) may be *reused* without increasing congestion provided that users are located far away from each other. This motivates to study an extended form of the congestion game where a user's payoff for using a channel is a function of the number of its *interfering* users sharing that channel, rather than the total number of users using the channel. This naturally leads to a network congestion game, whereby users are placed over a network (or a conflict graph).

Our main results and contributions are as follows. We study fundamental properties of the above network congestion game; in particular, we seek to answer under what conditions on the underlying network this game possesses the FIP or a NE. In case that number of channels is exactly two, we proved the problem has the finite improvement property (FIP) property; as a result, a Nash equilibrium always exists for any underlying graph, that any greedy updating strategy by the users will lead to such an equilibrium. We then proved by a counter-example that when there exist three or more channels, the problem does not have the finite improvement property so the game will not be a potential game. We also proved that for certain special types of graph, in the form of a tree or a loop, an NE exists. Finally, we showed that when the channels are equally perceived by any user (need not be the same for all users), this game has an exact potential function so that both the NE and the FIP property exist.

1.3 Organization of Dissertation

The remainder of this thesis is organized as follows. In Chapter 2 we study the first problem, opportunistic spectrum access in the case of sensing one channel at a time. The more general case, multiple channel sensing is studied in Chapter 3. Then in Chapter 4 we study the network congestion game for opportunistic spectrum sharing. We conclude the thesis in Chapter 5.

CHAPTER 2

Opportunistic Spectrum Access as a Restless Bandit Problem

2.1 Introduction

We consider a communication system in which a sender has access to multiple channels, but is limited to sensing and transmitting only on one at a given time. We explore how a smart sender should exploit past observations and the knowledge of the stochastic state evolution of these channels to maximize its transmission rate by switching opportunistically across channels.

We model this problem in the following manner. As shown in Figure 1, there are n channels, each of which evolves as an independent, identically-distributed, two-state discrete-time Markov chain. The two states for each channel — "good" (or state 1) and "bad" (or state 0) — indicate the desirability of transmitting over that channel at a given time slot. The state transition probabilities are given by p_{ij}, $i, j = 0, 1$. In each time slot the sender picks one of the channels to sense based on its prior observations, and obtains some fixed reward if it is in the good state. The basic objective of the sender is to maximize the reward that it can gain over a given finite time horizon. This problem can be described as a partially observable Markov decision process (POMDP) [3] since the states of the underlying Markov chains are not fully observed. It can also be cast as a special case of the class of

restless multi-armed bandit problems [4]; more discussion on this is given in Section 2.7.

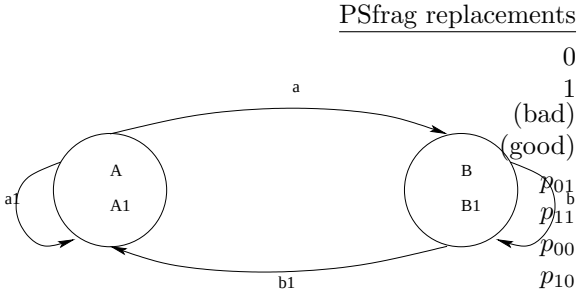

Figure 2.1: The Markov channel model.

This formulation is broadly applicable to several domains. It arises naturally in opportunistic spectrum access (OSA) [5, 6], where the sender is a secondary user, and the channel states describe the occupancy by primary users. In the OSA problem, the secondary sender may send on a given channel only when there is no primary user occupying it. It pertains to communication over parallel fading channels as well, if a two-state Markovian fading model is employed. Another interesting application of this formulation is in the domain of communication security, where it can be used to develop bounds on the performance of resource-constrained jamming. A jammer that has access to only one channel at a time could also use the same stochastic dynamic decision making process to maximize the number of times that it can successfully jam communications that occur on these channels. In this application, the "good" state for the jammer is precisely when the channel is being utilized by other senders (in contrast with the OSA problem).

In this chapter we examine the optimality of a simple myopic policy for the opportunistic access problem outlined above. Specifically, we show that the myopic policy is optimal for arbitrary n when $p_{11} \geq p_{01}$. We also show that it is optimal for $n = 3$ when $p_{11} < p_{01}$, while presenting a finite horizon counter example showing that it is in general not optimal for $n \geq 4$. We also generalize these results to related

formulations involving discounted and average rewards over an infinite horizon.

These results extend and complement those reported in prior work [7]. Specifically, it has been shown in [7] that for all n the myopic policy has an elegant and robust structure that obviates the need to know the channel state transition probabilities and reduces channel selection to a simple round robin procedure. Based on this structure, the optimality of the myopic policy for $n = 2$ was established and the performance of the myopic policy, in particular, the scaling property with respect to n, analyzed in [7]. It was conjectured in [7] that the myopic policy is optimal for any n. This conjecture was partially addressed in a preliminary conference version [8], where the optimality was established under certain restrictive conditions on the channel parameters and the discount factor. In the present chapter, we significantly relax these conditions and formerly prove this conjecture under the condition $p_{11} \geq p_{01}$. We also provide a counter example for $p_{11} < p_{01}$.

We would like to emphasize that compared to earlier work [7, 8], the approach used in this chapter relies on a coupling argument, which is the key to extending the optimality result to the arbitrary n case. Earlier techniques were largely based on exploiting the convex analytic properties of the value function, and were shown to have difficulty in overcoming the $n = 2$ barrier without further conditions on the discount factor or transition probabilities. This observation is somewhat reminiscent of the results reported in [9], where a coupling argument was also used to solve an n-queue problem while earlier versions [10] using value function properties were limited to a 2-queue case. We invite the interested reader to refer to [11], an important manuscript on monotonicity in MDPs which explores the power as well as the limitation of working with analytic properties of value functions and dynamic programming operators as we had done in our earlier work. In particular, [11, Section

9.5] explores the difficulty of using such techniques for multi-dimensional problems where the number of queues is more than $n = 2$; [11, Chapter 12] contrasts this proof technique with the stochastic coupling arguments, which our present work uses.

The remainder of this chapter is organized as follows. We formulate the problem in Section 2.2 and illustrate the myopic policy in Section 2.3. In Section 2.4, we prove that the myopic policy is optimal in the case of $p_{11} \geq p_{01}$, and show in Section 2.5 that it is in general not optimal when this condition does not hold. Section 2.6 extends the results from finite horizon to infinite horizon. We discuss our work within the context of the class of restless bandit problems as well as some related work in this area in Section 2.7. Section 2.8 concludes the chapter.

2.2 Problem Formulation

We consider the scenario where a user is trying to access the wireless spectrum to maximize its throughput or data rate. The spectrum consists of n independent and statistically identical channels. The state of a channel is given by a two-state discrete time Markov chain shown in Figure 2.1.

The system operates in discrete time steps indexed by t, $t = 1, 2, \cdots, T$, where T is the time horizon of interest. At time t^-, the channels (i.e., the Markov chains representing them) go through state transitions, and at time t the user makes the channel sensing and access decision. Specifically, at time t the user selects one of the n channels to sense, say channel i. If the channel is sensed to be in the "good" state (state 1), the user transmits and collects one unit of reward. Otherwise the user does not transmit (or transmits at a lower rate), collects no reward, and waits until $t + 1$ to make another choice. This process repeats sequentially until the time horizon expires.

As mentioned earlier, this abstraction is primarily motivated by the following multi-channel access scenario where a secondary user seeks spectrum opportunity in between a primary user's activities. Specifically, time is divided into frames and at the beginning of each frame there is a designated time slot for the primary user to reserve that frame and for secondary users to perform channel sensing. If a primary user intends to use a frame it will simply remain active in a channel (or multiple channels) during that sensing time slot (i.e., reservation is by default for a primary user in use of the channel), in which case a secondary user will find the channel(s) busy and not attempt to use it for the duration of that frame. If the primary user is inactive during this sensing time slot, then the remainder of the frame is open to secondary users. Such a structure provides the necessary protection for the primary user as channel sensing (in particular active channel sensing that involves communication between a pair of users) conducted at arbitrary times can cause undesirable interference.

Within such a structure, a secondary user has a limited amount of time and capability to perform channel sensing, and may only be able to sense one or a subset of the channels before the sensing time slot ends. And if all these channels are unavailable then it will have to wait till the next sensing time slot. In this chapter we will limit our attention to the special case where the secondary user only has the resources to sense one channel within this slot. Conceptually our formulation is easily extended to the case where the secondary user can sense multiple channels at a time within this structure, although the corresponding results differ, see e.g., [12].

Note that in this formulation we do not explicitly model the cost of channel sensing; it is implicit in the fact that the user is limited in how many channels it can sense at a time. Alternative formulations have been studied where sensing costs are

explicitly taken into consideration in a user's sensing and access decision, see e.g., [13] and [14].

In this formulation we have assumed that sensing errors are negligible. Techniques used in this chapter may be applicable in proving the optimality of the myopic policy under imperfect sensing and for a general number of channels. The reason behind this is that our proof exploits the simple structure of the myopic policy, which remains when sensing is subject to errors as shown in [15].

Note that the system is not fully observable to the user, i.e., the user does not know the exact state of the system when making the sensing decision. Specifically, channels go through state transition at time t^- (or anytime between $(t-1,t)$), thus when the user makes the channel sensing decision at time t, it does not have the true state of the system at time t, which we denote by $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_n(t)] \in \{0,1\}^n$. Furthermore, even after its action (at time t^+) it only gets to observe the true state of one channel, which goes through another transition at or before time $(t+1)^-$. The user's action space at time t is given by the finite set $\{1, 2, \cdots, n\}$, and we will use $a(t) = i$ to denote that the user selects channel i to sense at time t. For clarity, we will denote the outcome/observation of channel sensing at time t following the action $a(t)$ by $h_{a(t)}(t)$, which is essentially the true state $s_{a(t)}(t)$ of channel $a(t)$ at time t since we assume channel sensing to be error-free.

It can be shown (see e.g., [3, 16, 17]) that a sufficient statistic of such a system for optimal decision making, or the *information state* of the system [16, 17], is given by the conditional probabilities of the state each channel is in given all past actions and observations. Since each channel can be in one of two states, we denote this information state or belief vector by $\bar{\omega}(t) = [\omega_1(t), \cdots, \omega_n(t)] \in [0,1]^n$, where $\omega_i(t)$ is the conditional probability that channel i is in state 1 at time t given all past states,

actions and observations [1]. Throughout the chapter $\omega_i(t)$ will be referred to as the information state of channel i at time t, or simply the channel probability of i at time t.

Due to the Markovian nature of the channel model, the future information state is only a function of the current information state and the current action; i.e., it is independent of past history given the current information state and action. It follows that the information state of the system evolves as follows. Given that the state at time t is $\bar{\omega}(t)$ and action $a(t) = i$ is taken, $\omega_i(t+1)$ can take on two values: (1) p_{11} if the observation is that channel i is in a "good" state ($h_i(t) = 1$); this occurs with probability $P\{h_i(t) = 1|\bar{\omega}(t)\} = \omega_i(t)$; (2) p_{01} if the observation is that channel i is in a "bad" state ($h_i(t) = 0$); this occurs with probability $P\{h_i(t) = 0|\bar{\omega}(t)\} = 1-\omega_i$. For any other channel $j \neq i$, the corresponding $\omega_j(t+1)$ can only take on one value (i.e., with probability 1): $\omega_j(t+1) = \tau(\omega_j(t))$ where the operator $\tau : [0,1] \to [0,1]$ is defined as

$$\tau(\omega) := \omega p_{11} + (1-\omega)p_{01},\ 0 \leq \omega \leq 1. \tag{2.1}$$

These transition probabilities are summarized in the following equation for $t = 1, 2, \cdots, T-1$:

$$\begin{aligned} &\{\omega_i(t+1)|\bar{\omega}(t), a(t)\} \\ &= \begin{cases} p_{11} & \text{with prob. } \omega_i(t) \text{ if } a(t) = i \\ p_{01} & \text{with prob. } 1-\omega_i(t) \text{ if } a(t) = i \\ \tau(\omega_i(t)) & \text{with prob. 1 if } a(t) \neq i \end{cases}, \ i = 1, 2, \cdots, n, \end{aligned} \tag{2.2}$$

Also note that $\bar{\omega}(1) \in [0,1]^n$ denotes the initial condition (information state in the form of conditional probabilities) of the system, which may be interpreted as the the

[1] Note that this is a standard way of turning a POMDP problem into a classic MDP (Markov decision process) problem by means of information state, the main implication being that the state space is now uncountable.

user's initial belief about how likely each channel is in the good state before sensing starts at time $t = 1$. For the purpose of the optimization problems formulated below, this initial condition is considered given, which can be any probability vector [2].

It is important to note that although in general a POMDP problem has an uncountable state space (information states are probability distributions), in our problem the state space is countable for any given initial condition $\bar{\omega}(1)$. This is because as shown above, the information state of any channel with an initial probability of ω can only take on the values $\{\omega, \tau^k(\omega), p_{01}, \tau^k(\omega), p_{11}, \tau^k(\omega)\}$, where $k = 1, 2, \cdots$ and $\tau^k(\omega) := \tau(\tau^{k-1}(\omega))$, which is a countable set.

For compactness of presentation we will further use the operator τ to denote the above probability distribution of the information state (the entire vector):

$$\bar{\omega}(t+1) = \tau(\bar{\omega}(t), a(t)), \tag{2.3}$$

by noting that the operation given in (2.2) is applied to $\bar{\omega}(t)$ element-by-element. We will also use the following to denote the information state given observation outcome:

$$\tau(\bar{\omega}(t), a(t)|h_{a(t)}(t) = 1) = \tag{2.4}$$

$$(\tau(\omega_1(t)), \cdots, \tau(\omega_{a(t)-1}(t)), p_{11}, \tau(\omega_{a(t)+1}(t)), \cdots, \tau(\omega_n(t))) \tag{2.5}$$

$$\tau(\bar{\omega}(t), a(t)|h_{a(t)}(t) = 0) \tag{2.6}$$

$$= (\tau(\omega_1(t)), \cdots, \tau(\omega_{a(t)-1}(t)), p_{01}, \tau(\omega_{a(t)+1}(t)), \cdots, \tau(\omega_n(t))) \tag{2.7}$$

The objective of the user is to maximize its total (discounted or average) expected reward over a finite (or infinite) horizon. Let $J_T^\pi(\bar{\omega})$, $J_\beta^\pi(\bar{\omega})$, and $J_\infty^\pi(\bar{\omega})$ denote, respectively, these cost criteria (namely, finite horizon, infinite horizon with discount, and infinite horizon average reward) under policy π starting in state $\bar{\omega} = [\omega_1, \cdots, \omega_n]$.

[2] That is, the optimal solutions are functions of the initial condition. A reasonable choice, if the user has no special information other than the transition probabilities of these channels, is to simply use the steady-state probabilities of channels being in state "1" as an initial condition (i.e., setting $\omega_i(1) = \frac{p_{10}}{p_{01}+p_{10}}$).

The associated optimization problems ((P1)-(P3)) are formally defined as follows.

$$\text{(P1): } \max_{\pi} J_T^{\pi}(\bar{\omega}) = \max_{\pi} E^{\pi}[\sum_{t=1}^{T} \beta^{t-1} R_{\pi_t}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

$$\text{(P2): } \max_{\pi} J_{\beta}^{\pi}(\bar{\omega}) = \max_{\pi} E^{\pi}[\sum_{t=1}^{\infty} \beta^{t-1} R_{\pi_t}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

$$\text{(P3): } \max_{\pi} J_{\infty}^{\pi}(\bar{\omega}) = \max_{\pi} \lim_{T\to\infty} \frac{1}{T} E^{\pi}[\sum_{t=1}^{T} R_{\pi_t}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

where β ($0 \leq \beta \leq 1$ for (P1) and $0 \leq \beta < 1$ for (P2)) is the discount factor, and $R_{\pi_t}(\bar{\omega}(t))$ is the reward collected under state $\bar{\omega}(t)$ when channel $a(t) = \pi_t(\bar{\omega}(t))$ is selected and $h_{a(t)}(t)$ is observed. This reward is given by $R_{\pi_t}(\bar{\omega}(t)) = 1$ with probability $\omega_{a(t)}(t)$ (when $h_{a(t)}(t) = 1$), and 0 otherwise.

The maximization in (P1) is over the class of deterministic Markov policies.[3]. An admissible policy π, given by the vector $\pi = [\pi_1, \pi_2, \cdots, \pi_T]$, is thus such that π_t specifies a mapping from the current information state $\bar{\omega}(t)$ to a channel selection action $a(t) = \pi_t(\bar{\omega}(t)) \in \{1, 2, \cdots, n\}$. This is done without loss of optimality due to the Markovian nature of the underlying system, and due to known results on POMDPs. Note that the class of Markov policies in terms of information state are also known as seperated policies (see [17]). Due to finiteness of (unobservable) state spaces and action space in problem (P1), it is known that an optimal policy (over all random and deterministic, history-dependent and history-independent policies) may be found within the class of separated (i.e. deterministic Markov) policies (see e.g., [17, Theorem 7.1, Chapter 6]), thus justifying the maximization and the admissible policy space.

In Section 2.6 we establish the existence of a stationary separated policy π^*, under which the supremum of the expected discounted reward as well as the supremum of

[3]A Markov policy is a policy that derives its action only depending on the current (information) state, rather than the entire history of states, see e.g., [17].

expected average cost are achieved, hence justifying our use of maximization in (P2) and (P3). Furthermore, it is shown that under this policy the limit in (P3) exists and is greater than the limsup of the average performance of any other policy (in general history-dependent and randomized). This is a strong notion of optimality; the interpretation is that the most "pessimistic" average performance under policy π^* ($\liminf \frac{1}{T} J_T^{\pi^*}(\cdot) = \lim \frac{1}{T} J_T^{\pi^*}(\cdot)$) is greater than the most "optimistic" performance under any other policy π ($\limsup \frac{1}{T} J_T^{\pi}(\cdot)$). In much of the literature on MDP, this is referred to as the *strong optimality* for an expected average cost (reward) problem; for a discussion on this, see [18, Page 344].

2.3 Optimal Policy and the Myopic Policy

2.3.1 Dynamic Programming Representations

Problems (P1)-(P3) defined in the previous section may be solved using their respective dynamic programming (DP) representations. Specifically, for problem (P1), we have the following recursive equations:

$$\begin{aligned} V_T(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega})] \\ V_t(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega}) + \beta V_{t+1}(\tau(\bar{\omega}, a))] \\ &= \max_{a=1,\cdots,n} \left(\omega_a + \beta\omega_a V_{t+1}\left(\tau\left(\bar{\omega}, a|1\right)\right) + \beta(1-\omega_a) V_{t+1}\left(\tau\left(\bar{\omega}, a|0\right)\right)\right), \end{aligned} \tag{2.8}$$

for $t = 1, 2, \cdots, T-1$, where $V_t(\bar{\omega})$ is known as the value function, or the maximum expected future reward that can be accrued starting from time t when the information state is $\bar{\omega}$. In particular, we have $V_1(\bar{\omega}) = \max_\pi J_T^\pi(\bar{\omega})$, and an optimal deterministic Markov policy exists such that $a = \pi_t^*(\bar{\omega})$ achieves the maximum in (2.8) (see e.g., [18] (Chapter 4)). Note that since $\mathcal{T}$ is a conditional probability distribution (given in (2.3)), $V_{t+1}(\mathcal{T}(\bar{\omega}, a))$ is taken to be the expectation over this distribution when its argument is $\mathcal{T}$, with a slight abuse of notation, as expressed in (2.8).

Similar dynamic programming representations hold for (P2) and (P3) as given below. For problem (P2) there exists a unique function $V_\beta(\cdot)$ satisfying the following fixed point equation:

$$\begin{aligned} V_\beta(\bar{\omega}) &= \max_{a=1,\cdots,n} E[R_a(\bar{\omega}) + \beta V_\beta(\tau(\bar{\omega}, a))] \\ &= \max_{a=1,\cdots,n} \left(\omega_a + \beta\omega_a V_\beta\left(\tau\left(\bar{\omega}, a|1\right)\right) + \beta(1-\omega_a)V_\beta\left(\tau\left(\bar{\omega}, a|0\right)\right)\right) . \end{aligned} \tag{2.9}$$

We have that $V_\beta(\bar{\omega}) = \max_\pi J^\pi_\beta(\bar{\omega})$, and that a stationary separated policy π^* is optimal if and only if $a = \pi^*(\bar{\omega})$ achieves the maximum in (2.9) [19, Theorem 7.1].

For problem (P3), we will show that there exist a bounded function $h_\infty(\cdot)$ and a constant scalar J satisfying the following equation:

$$\begin{aligned} J + h_\infty(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega}) + h_\infty(\tau(\bar{\omega}, a))] \\ &= \max_{a=1,\cdots,n} \left(\omega_a + \omega_a h_\infty\left(\tau\left(\bar{\omega}, a|1\right)\right) + (1-\omega_a)h_\infty\left(\tau\left(\bar{\omega}, a|0\right)\right)\right). \end{aligned} \tag{2.10}$$

The boundedness of h_∞ and the immediate reward implies that $J = \max_\pi J^\pi_\infty(\bar{\omega})$, and that a stationary separated policy π^* is optimal in the context of (P3) if $a = \pi^*(\bar{\omega})$ achieves the maximum in (2.10) [19, Theorems 6.1-6.3].

Solving (P1)-(P3) using the above recursive equations is in general computationally heavy. Therefore, instead of directly using the DP equations, the focus of this chapter is on examining the optimality properties of a simple, greedy algorithm. We define this algorithm next and show its simplicity in structure and implementation.

2.3.2 The Myopic Policy

A myopic or greedy policy ignores the impact of the current action on the future reward, focusing solely on maximizing the expected immediate reward. Myopic policies are thus stationary. For (P1), the myopic policy under state $\bar{\omega} = [\omega_1, \omega_2, \cdots, \omega_n]$

is given by

$$a^*(\bar{\omega}) = \arg \max_{a=1,\cdots,n} E[R_a(\bar{\omega})] = \arg \max_{a=1,\cdots,n} \omega_a. \tag{2.11}$$

In general, obtaining the myopic action in each time slot requires the successive update of the information state as given in (2.2), which explicitly relies on the knowledge of the transition probabilities $\{p_{ij}\}$ as well as the initial condition $\bar{\omega}(1)$. Interestingly, it has been shown in [7] that the implementation of the myopic policy requires only the knowledge of the initial condition and the order of p_{11} and p_{01}, but not the precise values of these transition probabilities. To make the present chapter self-contained, below we briefly describe how this policy works; more details may be found in [7].

Specifically, when $p_{11} \geq p_{01}$ the conditional probability updating function $\tau(\omega)$ is a monotonically increasing function, i.e., $\tau(\omega_1) \geq \tau(\omega_2)$ for $\omega_1 \geq \omega_2$. Therefore the ordering of information states among channels is preserved when they are not observed. If a channel has been observed to be in state "1" (respectively "0"), its probability at the next step becomes $p_{11} \geq \tau(\omega)$ (respectively $p_{01} \leq \tau(\omega)$) for any $\omega \in [0, 1]$. In other words, a channel observed to be in state "1" (respectively "0") will have the highest (respectively lowest) possible information state among all channels.

These observations lead to the following implementation of the myopic policy. We take the initial information state $\bar{\omega}(1)$, order the channels according to their probabilities $\omega_i(1)$, and probe the highest one (top of the ordered list) with ties broken randomly. In subsequent steps we stay in the same channel if the channel was sensed to be in state "1" (good) in the previous slot; otherwise, this channel is moved to the bottom of the ordered list, and we probe the channel currently at the top of the list. This in effect creates a round robin style of probing, where the channels are cycled through in a fixed order. This circular structure is exploited in

Section 2.4 to prove the optimality of the myopic policy in the case of $p_{11} \geq p_{01}$.

When $p_{11} < p_{01}$, we have an analogous but opposite situation. The conditional probability updating function $\tau(\omega)$ is now a monotonically decreasing function, i.e., $\tau(\omega_1) \leq \tau(\omega_2)$ for $\omega_1 \geq \omega_2$. Therefore the ordering of information states among channels is reversed at each time step when they are not observed. If a channel has been observed to be in state "1" (respectively "0"), its probability at the next step becomes $p_{11} \leq \tau(\omega)$ (respectively $p_{01} \geq \tau(\omega)$) for any $\omega \in [0, 1]$. In other words, a channel observed to be in state "1" (respectively "0") will have the lowest (respectively highest) possible information state among all channels.

As in the previous case, these similar observations lead to the following implementation. We take the initial information state $\bar{\omega}(1)$, order the channels according to their probabilities $\omega_i(1)$, and probe the highest one (top of the ordered list) with ties broken randomly. In each subsequent step, if the channel sensed in the previous step was in state "0" (bad), we keep this channel at the top of the list but completely reverse the order of the remaining list, and we probe this channel. If the channel sensed in the previous step was in state "1" (good), then we completely reverse the order of the entire list (including dropping this channel to the bottom of the list), and probe the channel currently at the top of the list. This alternating circular structure is exploited in Section 2.5 to examine the optimality of the myopic policy in the case of $p_{11} < p_{01}$.

2.4 Optimality of the Myopic Policy in the Case of $p_{11} \geq p_{01}$

In this section we show that the myopic policy, with a simple and robust structure, is optimal when $p_{11} \geq p_{01}$. We will first show this for the finite horizon discounted cost case, and then extend the result to the infinite horizon case under both discounted

and average cost criteria in Section 2.6.

The main assumption is formally stated as follows.

Assumption 2.1. *The transition probabilities p_{01} and p_{11} are such that*

$$p_{11} - p_{01} \geq 0. \tag{2.12}$$

The main theorem of this section is as follows.

Theorem 2.2. *Consider Problem (P1). Define $V_t(\bar{\omega}; a) := E[R_a(\bar{\omega}) + \beta V_{t+1}(\mathcal{T}(\bar{\omega}, a))]$, i.e., the value of the value function given in Eqn (2.8) when action a is taken at time t followed by an optimal policy. Under Assumption 2.1, the myopic policy is optimal, i.e. for $\forall t, 1 \leq t < T$, and $\forall \bar{\omega} = [\omega_1, \cdots, \omega_n] \in [0,1]^n$,*

$$V_t(\bar{\omega}; a = j) - V_t(\bar{\omega}; a = i) \geq 0, \tag{2.13}$$

if $\omega_j \geq \omega_i$, for $i = 1, \cdots, n$.

The proof of this theorem is based on backward induction on t: given the optimality of the myopic policy at times $t+1, t+2, \cdots, T$, we want to show that it is also optimal at time t. This relies on a number of lemmas introduced below. The first lemma introduces a notation that allows us to express the expected future reward under the myopic policy.

Lemma 2.3. *There exist T n-variable functions, denoted by $W_t()$, $t = 1, 2, \cdots, T$, each of which is a polynomial of order 1*[4] *and can be represented recursively in the following form:*

(2.14)

$$W_t(\bar{\omega}) = \omega_n + \omega_n \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-1}), p_{11}) + (1-\omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-1})),$$

where $\bar{\omega} = [\omega_1, \omega_2, \cdots, \omega_n]$ and $W_T(\bar{\omega}) = \omega_n$.

[4]Each function W_t is affine in each variable, when all other variables are held constant.

Proof. The proof is easily obtained using backward induction on t given the above recursive equation and noting that $W_T()$ is one such polynomial and the mapping $\tau()$ is a linear operation. □

Corollary 1. When $\bar{\omega}$ represents the ordered list of information states $[\omega_1, \omega_2, \cdots, \omega_n]$ with $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_n$, then $W_t(\bar{\omega})$ is the expected total reward obtained by the myopic policy from time t on.

Proof. This result follows directly from the description of the policy given in Section 2.3.2.

When $\bar{\omega}$ is the ordered list of information states, the recursive expression in (2.14) gives the expected reward of the following policy: probe the n-th channel; for the next step, if the current sensing outcome is "1", then continue to probe the n-th channel; if the current sensing outcome is "0", then drop this channel to the bottom of the list (it becomes the first channel) while moving the i-th channel to the $(i+1)$-th position for all $i = 1, \cdots, n-1$; repeat this process. (This is essentially the same description as given in Section 2.3 for the case of $p_{11} \geq p_{01}$.) To see that this is the myopic policy, note that under the above policy, at any time the list of channel probabilities are increasingly ordered. This is because for any $0 \leq \omega \leq 1$, we have $p_{01} \leq \tau(\omega) \leq p_{11}$ when $p_{11} \geq p_{01}$. Furthermore, under the assumption $p_{11} \geq p_{01}$, $\tau(\omega)$ is a monotonically increasing function. Therefore under this policy, when starting out with increasingly ordered information states, this ordered is maintained in each subsequent time step. As expressed in (2.14), at each step it's always the n-th channel that is probed. Since the n-th channel has the largest probability of being available, this is the myopic policy. □

Proposition 2.4. *The fact that W_t is a polynomial of order 1 and affine in each of*

its elements implies that

$$W_t(\omega_1, \cdots, \omega_{n-2}, y, x) - W_t(\omega_1, \cdots, \omega_{n-2}, x, y)$$

$$= (x-y)[W_t(\omega_1, \cdots, \omega_{n-2}, 0, 1) - W_t(\omega_1, \cdots, \omega_{n-2}, 1, 0)] \ . \tag{2.15}$$

Similar results hold when we change the positions of x and y.

To see this, consider $W_t(\omega_1, \cdots, \omega_{n-2}, x, y)$ and $W_t(\omega_1, \cdots, \omega_{n-2}, y, x)$, as functions of x and y, each having an x term, a y term, an xy term and a constant term. Since we are just swapping the positions of x and y in these two functions, the constant term remains the same, and so does the xy term. Thus the only difference is the x term and the y term, as given in the above equation. This linearity result will be used later in our proofs.

The next lemma establishes a necessary and sufficient condition for the optimality of the myopic policy.

Lemma 2.5. *Consider Problem (P1) and Assumption 1. Given the optimality of the myopic policy at times $t+1, t+2, \cdots, T$, the optimality at time t is equivalent to:*

$$W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq W_t(\omega_1, \ldots, \omega_n), \tag{2.16}$$

$$\text{for all } \omega_1 \leq \cdots \leq \omega_i \leq \cdots \leq \omega_n.$$

Proof. Since the myopic policy is optimal from $t+1$ on, it is sufficient to show that probing ω_n followed by myopic probing is better than probing any other channel followed by myopic probing. The former is precisely given by the RHS of the above equation; the latter by the LHS, thus completing the proof. □

Having established that $W_t(\bar{\omega})$ is the total expected reward of the myopic policy for an increasingly-ordered vector $\bar{\omega} = [\omega_1, \cdots, \omega_n]$, we next proceed to show that

we do not decrease this total expected reward $W_t(\bar{\omega})$ by switching the order of two neighboring elements ω_i and ω_{i+1} if $\omega_i \geq \omega_{i+1}$. This is done in two separate cases, when $i+1 < n$ (given in Lemma 2.7) and when $i+1 = n$ (given in Lemma 2.8), respectively. The first case is quite straightforward, while proving the second cased turned out to be significantly more difficult. Our proof of the second case (Lemma 2.8) relies on a separate lemma (Lemma 2.6) that establishes a bound between the greedy use of two identical vectors but with a different starting position. The proof of Lemma 2.6 is based on a coupling argument and is quite instructive. Below we present and prove Lemmas 2.6, 2.7 and 2.8.

Lemma 2.6. *For* $0 < \omega_1 \leq \omega_2 \leq \ldots \leq \omega_n < 1$ *, we have the following inequality for all* $t = 1, 2, \cdots, T$*:*

$$1 + W_t(\omega_2, \ldots, \omega_n, \omega_1) \geq W_t(\omega_1, \ldots, \omega_n). \tag{2.17}$$

Proof. This lemma is the key to our main result and its proof is by using coupling argument along any sample path. It is however also lengthy, and for this reason has been relegated to the Appendix. □

Lemma 2.7. *For all* j*,* $1 \leq j \leq n-3$*, and all* $x \geq y$*, we have*

$$W_t(\omega_1, \ldots, \omega_j, x, y, \ldots, \omega_n) \leq W_t(\omega_1, \ldots, \omega_j, y, x, \ldots, \omega_n) \tag{2.18}$$

Proof. We prove this by induction over t. The claim is obviously true for $t = T$, since both sides will be equal to ω_n, thereby establishing the induction basis. Now

suppose the claim is true for all $t+1, \cdots, T-1$. We have

$$\begin{aligned} & W_t(\omega_1, \cdots, \omega_{j-1}, x, y, \cdots, \omega_n) \\ = \; & \omega_n(1 + \beta W_{t+1}(\tau(\omega_1), \cdots, \tau(x), \tau(y), \cdots, \tau(\omega_{n-1}), p_{11})) \\ + \; & (1-\omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \cdots, \tau(x), \tau(y), \cdots, \tau(\omega_{n-1})) \\ \leq \; & \omega_n(1 + \beta W_{t+1}(\tau(\omega_1), \cdots, \tau(y), \tau(x), \cdots, \tau(\omega_{n-1}), p_{11})) \\ + \; & (1-\omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \cdots, \tau(y), \tau(x), \cdots, \tau(\omega_{n-1})) \\ = \; & W_t(\omega_1, \cdots, \omega_{j-1}, y, x, \cdots, \omega_n) \end{aligned} \tag{2.19}$$

where the inequality is due to the induction hypothesis, and noting that $\tau()$ is a monotone increasing mapping in the case of $p_{11} \geq p_{01}$. □

Lemma 2.8. *For all $x \geq y$, we have*

$$W_t(\omega_1, \ldots, \omega_j, \ldots, \omega_{n-2}, x, y) \leq W_t(\omega_1, \ldots, \omega_j, \ldots, \omega_{n-2}, y, x). \tag{2.20}$$

Proof. This lemma is proved inductively. The claim is obviously true for $t = T$. Assume it also holds for times $t+1, \cdots, T-1$. We have by the definition of $W_t()$ and due to its linearity property:

$$\begin{aligned} & W_t(\omega_1, \ldots, \omega_{n-2}, y, x) - W_t(\omega_1, \ldots, \omega_{n-2}, x, y) \\ = \; & (x-y)(W_t(\omega_1, \ldots, \omega_{n-2}, 0, 1) - W_t(\omega_1, \ldots, \omega_{n-2}, 1, 0)) \\ = \; & (x-y)\left(1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) - \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11})\right). \end{aligned}$$

But from the induction hypothesis we know that

$$W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) \geq W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}, p_{01}) \tag{2.21}$$

This means that

$$\begin{aligned} & 1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) - \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}) \\ \geq \; & 1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}, p_{01}) - \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}) \geq 0\ , \end{aligned}$$

where the last inequality is due to Lemma 2.6 (note that in that lemma we proved $1 + A \geq B$, which obviously implies $1 + \beta A \geq \beta B$ for $0 \leq \beta \leq 1$ that is used above). This, together with the condition $x \geq y$, completes the proof. □

We are now ready to prove the main theorem.

Proof of Theorem 2.2: The basic approach is by induction on t. The optimality of the myopic policy at time $t = T$ is obvious. So the induction basis is established. Now assume that the myopic policy is optimal for all times $t + 1, t + 2, \cdots, T - 1$, and we will show that it is also optimal at time t. By Lemma 2.5 this is equivalent to establishing the following

$$W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq W_t(\omega_1, \ldots, \omega_n). \tag{2.22}$$

But we know from Lemmas 2.7 and 2.8 that,

$$\begin{aligned} & W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_i, \omega_n) \\ \leq \; & W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_i, \omega_{n-1}, \omega_n) \leq \ldots \leq W_t(\omega_1, \ldots, \omega_n) \; , \end{aligned}$$

where the first inequality is the result of Lemma 2.8, while the remaining inequalities are repeated application of Lemma 2.7, completing the proof.

We would like to emphasize that from a technical point of view, Lemma 2.6 is the key to the whole proof: it leads to Lemma 2.8, which in turn leads to Theorem 2.2. While Lemma 2.8 was easy to conceptualize as a sufficient condition to prove the main theorem, Lemma 2.6 was much more elusive to construct and prove. This, indeed, marks the main difference between the proof techniques used here vs. that used in our earlier work [8]: Lemma 2.6 relies on a coupling argument instead of the convex analytic properties of the value function.

2.5 The Case of $p_{11} < p_{01}$

In the previous section we showed that a myopic policy is optimal if $p_{11} \geq p_{01}$. In this section we examine what happens when $p_{11} < p_{01}$, which corresponds to the case when the Markovian channel state process exhibits a negative auto-correlation over a unit time. This is perhaps a case of less practical interest and relevance. However, as we shall see this case presents a greater degree of technical complexity and richness than the previous case. Specifically, we first show that when the number of channels is three ($n = 3$) or when the discount factor $\beta \leq \frac{1}{2}$, the myopic policy remains optimal even for the case of $p_{11} < p_{01}$ (the proof for two channels in this case was given earlier in [7]). We thus conclude that the myopic policy is optimal for $n \leq 3$ or $\beta \leq 1/2$ regardless of the transition probabilities. We then present a counter example showing that the the myopic policy is not optimal in general when $n \geq 4$ and $\beta > 1/2$. In particular, our counter example is for a finite horizon with $n = 4$ and $\beta = 1$.

2.5.1 $n = 3$ or $\beta \leq \frac{1}{2}$

We start by developing some results parallel to those presented in the previous section for the case of $p_{11} \geq p_{01}$.

Lemma 2.9. *There exist T n-variable polynomial functions of order 1, denoted by $Z_t(), t = 1, 2, \cdots, T$, i.e., each function is linear in all the elements, and can be represented recursively in the following form:*

$$\begin{aligned} Z_t(\bar{\omega}) := \omega_n(1 + \beta Z_{t+1}(p_{11}, \tau(\omega_{n-1}), \ldots, \tau(\omega_1))) \\ +(1 - \omega_n)\beta Z_{t+1}(\tau(\omega_{n-1}), \ldots, \tau(\omega_1), p_{01}). \end{aligned} \tag{2.23}$$

where $Z_T(\bar{\omega}) = \omega_n$.

Corollary 2. $Z_t(\bar{\omega})$ given in (2.23) represents the expected total reward of the myopic policy when $\bar{\omega}$ is ordered in increasing order of ω_i.

Similar to Corollary 1, the above result follows directly from the policy description given in Section 2.3.2.

It follows that the function Z_t also has the same linearity property presented earlier, i.e.

$$\begin{aligned} & Z_t(\omega_1, \cdots, \omega_{n-2}, y, x) - Z_t(\omega_1, \cdots, \omega_{n-2}, x, y) \\ = \; & (x-y)(Z_t(\omega_1, \cdots, \omega_{n-2}, 0, 1) - Z_t(\omega_1, \cdots, \omega_{n-2}, 1, 0)) \; . \end{aligned} \tag{2.24}$$

Similar results hold when we change the positions of x and y.

In the next lemma and theorem we prove that the myopic policy is still optimal when $p_{11} < p_{01}$ if $n = 3$ or $\beta \leq 1/2$. In particular, Lemma 2.10 below is the analogy of Lemmas 2.7 and 2.8 combined.

Lemma 2.10. *At time t ($t = 1, 2, \cdots, T$), for all $j \leq n-2$, we have the following inequality for $\forall 1 \geq x \geq y \geq 0$ if either $n = 3$ or $\beta \leq 1/2$:*

$$Z_t(\omega_1, \ldots, \omega_j, y, x, \omega_{j+3}, \ldots, \omega_n) \;\; \geq \;\; Z_t(\omega_1, \ldots, \omega_j, x, y, \omega_{j+3}, \ldots, \omega_n). \tag{2.25}$$

Proof. We prove this by induction on t. The claim is obviously true for $t = T$. Now suppose it's true for $t+1, \cdots, T-1$. Due to the linearity property of Z_t,

$Z_t(\omega_1, \ldots, \omega_j, y, x, \omega_{j+3}, \ldots, \omega_n) - Z_t(\omega_1, \ldots, \omega_j, x, y, \omega_{j+3}, \ldots, \omega_n)$

$= (x-y)\left(Z_t(\omega_1, \ldots, \omega_j, 0, 1, \omega_{j+3}, \ldots, \omega_n) - Z_t(\omega_1, \ldots, \omega_j, 1, 0, \omega_{j+3}, \ldots, \omega_n)\right)$

Thus it suffices to show that

$Z_t(\omega_1, \ldots, \omega_j, 0, 1, \omega_{j+3}, \ldots, \omega_n) \geq Z_t(\omega_1, \ldots, \omega_j, 1, 0, \omega_{j+3}, \ldots, \omega_n)$.

We treat the case when $j < n-2$ and $j = n-2$ separately. Indeed, without loss of generality, let $j = n-3$ (the proof follows exactly for all $j \leq n-3$ with more

lengthy notations). At time t we have

$$\begin{aligned}
&Z_t(\omega_1,\ldots,\omega_{n-3},0,1,\omega_n) - Z_t(\omega_1,\ldots,\omega_{n-3},1,0,\omega_n)\\
&= \omega\beta(Z_{t+1}(p_{11},p_{11},p_{01},\tau(\omega_{n-3}),\ldots,\tau(\omega_1)) - Z_{t+1}(p_{11},p_{01},p_{11},\tau(\omega_{n-3}),\ldots,\tau(\omega_1)))\\
&+ (1-\omega)\beta(Z_{t+1}(p_{11},p_{01},\tau(\omega_{n-3}),\ldots,\tau(\omega_1),p_{01}) - Z_{t+1}(p_{01},p_{11},\tau(\omega_{n-3}),\ldots,\tau(\omega_1),p_{01}))\\
&\geq 0
\end{aligned}$$

where the last inequality is due to the induction hypothesis.

Now we will consider the case when $j = n-2$.

$$\begin{aligned}
Z_t(\omega_1,\ldots,\omega_{n-2},0,1) - Z_t(\omega_1,\ldots,\omega_{n-2},1,0) &= 1 + \beta Z_{t+1}(p_{11},p_{01},\tau(\omega_{n-2}),\ldots,\tau(\omega_1)) -\\
&\beta Z_{t+1}(p_{11},\tau(\omega_{n-2}),\ldots,\tau(\omega_1),p_{01}).
\end{aligned}$$

Next we show that if $\beta \leq 1/2$ or $n = 3$ the right hand side of above equation is non-negative.

If $\beta \leq 1/2$, then

$$\begin{aligned}
&1 + \beta Z_{t+1}(p_{11},p_{01},\tau(\omega_{n-2}),\ldots,\tau(\omega_1)) - \beta Z_{t+1}(p_{11},\tau(\omega_{n-2}),\ldots,\tau(\omega_1),p_{01})\\
&\geq 1 - \frac{\beta}{1-\beta} \geq 0.
\end{aligned}$$

If $n = 3$, then

$$\begin{aligned}
&1 + \beta Z_{t+1}(p_{11},p_{01},\tau(\omega_1)) - \beta Z_{t+1}(p_{11},\tau(\omega_1),p_{01})\\
&= 1 + \beta(\tau(\omega_1) - p_{01})(Z_{t+1}(p_{11},0,1) - Z_{t+1}(p_{11},1,0))\\
&\geq 1 - \beta(Z_{t+1}(p_{11},0,1) - Z_{t+1}(p_{11},1,0))\\
&\geq 0
\end{aligned}$$

where the first inequality is due to the fact that $-1 \leq \tau(\omega_1) - p_{01} \leq 0$ and the last inequality is given by the induction hypothesis. □

Theorem 2.11. *Consider Problem (P1). Assume that $p_{11} < p_{01}$. The myopic policy is optimal for the case of $n = 3$ and the case of $\beta \leq 1/2$ with arbitrary n. More precisely, for these two cases, $\forall t$, $1 \leq t \leq T$, we have*

$$V_t(\bar{\omega}; a = j) - V_t(\bar{\omega}; a = i) \geq 0, \tag{2.26}$$

if $\omega_j \geq \omega_i$ for $i = 1, \cdots, n$.

Proof. We prove by induction on t. The optimality of the myopic policy at time $t = T$ is obvious. Now assume that the myopic policy is optimal for all times $t+1, t+2, \cdots, T-1$, and we want to show that it is also optimal at time t. Suppose at time t the channel probabilities are such that $\omega_n \geq \omega_i$ for $i = 1, \cdots, n-1$. The myopic policy is optimal at time t if and only if probing ω_n followed by myopic probing is better than probing any other channel followed by myopic probing. Mathematically, this means

$$Z_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq Z_t(\omega_1, \ldots, \omega_n), \qquad \text{for all } \omega_1 \leq \omega_i \leq \omega_n.$$

But this is a direct consequence of Lemma 2.10, completing the proof. □

2.5.2 A 4-channel Counter Example

The following example shows that the myopic policy is not, in general, optimal for $n \geq 4$ when $p_{11} < p_{01}$.

Example 2.1. Consider an example with the following parameters: $p_{01} = 0.9, p_{11} = 0.1, \beta = 1$, and $\bar{\omega} = [.97, .97, .98, .99]$. Now compare the following two policies at time $T - 3$: play myopically (I), or play the .98 channel first, followed by the myopic policy (II). Computation reveals that

$$V^I_{T-3}(.97, .97, .98, .99) = 2.401863 < V^{II}_{T-3}(.97, .97, .98, .99) = 2.402968$$

which shows that the myopic policy is not optimal in this case.

It remains an interesting question as to whether such counter examples exist in the case when the initial condition is such that all channel are in the good state with the stationary probability.

2.6 Infinite Horizon

Now we consider extensions of results in Sections 2.4 and 2.5 to (P2) and (P3), i.e., to show that the myopic policy is also optimal for (P2) and (P3) under the same conditions. Intuitively, this holds due to the fact that the stationary optimal policy of the finite horizon problem is independent of the horizon as well as the discount factor. Theorems 2.12 and 2.13 below concretely establish this.

We point out that the proofs of Theorems 2.12 and 2.13 do not rely on any additional assumptions other than the optimality of the myopic policy for (P1). Indeed, if the optimality of the myopic policy for (P1) can be established under weaker conditions, Theorems 3 and 4 can be readily invoked to establish its optimality under the same weaker condition for (P2) and (P3), respectively.

Theorem 2.12. *If myopic policy is optimal for (P1), it is also optimal for (P2) for* $0 \leq \beta < 1$. *Furthermore, its value function is the limiting value function of (P1) as the time horizon goes to infinity, i.e., we have* $\max_\pi J_\beta^\pi(\bar{\omega}) = \lim_{T\to\infty} \max_\pi J_T^\pi(\bar{\omega})$.

Proof. We first use the bounded convergence theorem (BCT) to establish the fact that under any deterministic stationary Markov policy π, we have $J_\beta^\pi(\bar{\omega}) = \lim_{T\to\infty} J_T^\pi(\bar{\omega})$. We prove this by noting that

$$\begin{aligned} J_\beta^\pi(\bar{\omega}) &= E^\pi[\lim_{T\to\infty} \sum_{t=1}^{T} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}] \\ &= \lim_{T\to\infty} E^\pi[\sum_{t=1}^{T} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}] = \lim_{T\to\infty} J_T^\pi(\bar{\omega}) \end{aligned}$$

where the second equality is due to BCT for $\sum_{t=1}^{T}\beta^{t-1}R_{\pi(t)}(\bar{\omega}(t)) \leq \frac{1}{1-\beta}$. This proves the second part of the theorem by noting that due to the finiteness of the action space, we can interchange maximization and limit.

Let π^* denote the myopic policy. We now establish the optimality of π^* for (P2). From Theorem 1, we know:

$$J_T^{\pi^*}(\bar{\omega}) \;=\; \max_{a=i}\left\{\omega_i + \beta\omega_i J_{T-1}^{\pi^*}\left(\tau\left(\bar{\omega},i|1\right)\right)\beta\left(1-\omega_i\right)J_{T-1}^{\pi^*}\left(\tau\left(\bar{\omega},i|0\right)\right)\right\}.$$

Taking limit of both sides, we have

$$J_\beta^{\pi^*}(\bar{\omega}) \quad = \max_{a=i}\left\{\omega_i + \beta\omega_i J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|1\right)\right) + \beta\left(1-\omega_i\right)J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|0\right)\right)\right\}.$$

Note that (2.27) is nothing but the dynamic programming equation for the infinite horizon discounted reward problem given in (2.9). From the uniqueness of the dynamic programming solution, then, we have

$$J_\beta^{\pi^*}(\bar{\omega}) = V_\beta(\bar{\omega}) = \max_{\pi} J_\beta^{\pi}(\bar{\omega})$$

hence, the optimality of the myopic policy. □

Theorem 2.13. *Consider (P3) with the expected average reward and under the ergodicity assumption $|p_{11}-p_{00}|<1$. Myopic policy is optimal for problem (P3) if it is optimal for (P1).*

Proof. We consider the infinite horizon discounted cost for $\beta<1$ under the optimal policy denoted by π^*:

$$J_\beta^{\pi^*}(\bar{\omega}) = \max_{a=i}\left\{\omega_i + \beta\omega_i J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|1\right)\right) + \beta(1-\omega_i)J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|0\right)\right)\right\}.$$

This can be written as

$$\begin{aligned}&(1-\beta)J_\beta^{\pi^*}(\bar{\omega})\\ &= \max_{a=i}\left\{\omega_i + \beta\omega_i\left[J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|1\right)\right) - J_\beta^{\pi^*}(\bar{\omega})\right] + \beta(1-\omega_i)\left[J_\beta^{\pi^*}\left(\tau\left(\bar{\omega},i|0\right)\right) - J_\beta^{\pi^*}(\bar{\omega})\right]\right\}.\end{aligned}$$

Notice that the boundedness of the reward function and compactness of information state implies that the sequence of $\{(1-\beta)J_\beta^{\pi^*}(\bar{\omega})\}$ is bounded, i.e. for all $0 \leq \beta \leq 1$,

$$(1-\beta)J_\beta^{\pi^*}(\bar{\omega}) \leq 1. \tag{2.27}$$

Also, applying Lemma 2 from [8] (which provides an upper bound on the difference in value functions between taking two different actions followed by the optimal policy) and noting that $-1 < p_{11} - p_{00} < 1$, we have that there exists some positive constant $K := \frac{1}{1-|p_{11}-p_{01}|}$ such that

$$\left|J_\beta^{\pi^*}\left(\tau\left(\bar{\omega}, i|0\right)\right) - J_\beta^{\pi^*}(\bar{\omega})\right| \leq K. \tag{2.28}$$

By Bolzano-Weierstrass theorem, (2.27) and (2.28) guarantee the existence of a converging sequence $\beta_k \to 1$ such that

$$\lim_{k\to\infty}(1-\beta_k)J_{\beta_k}^{\pi^*}(\bar{\omega}^*) := J^*, \tag{2.29}$$

$$\text{and} \quad \lim_{k\to\infty}\left[J_{\beta_k}^{\pi^*}(\bar{\omega}) - J_{\beta_k}^{\pi^*}(\bar{\omega}^*)\right] := h^{\pi^*}(\bar{\omega}) \ , \tag{2.30}$$

where $\omega_i^* := \frac{p_{01}}{1-p_{11}+p_{01}}$ is the steady-state belief (the limiting belief when channel i is not sensed for a long time).

As a result, (2.29) can be written as

$$J^* = \lim_{k\to\infty}\left\{(1-\beta_k)J_{\beta_k}^{\pi^*}(\bar{\omega}^*) + (1-\beta_k)\left[J_{\beta_k}^{\pi^*}(\bar{\omega}) - J_{\beta_k}^{\pi^*}(\bar{\omega}^*)\right]\right\}.$$

In other words,

$$\begin{aligned} J^* = \lim_{k\to\infty}\max_{a=i}\Big\{&\omega_i + \beta_k\omega_i\left[J_{\beta_k}^{\pi^*}\left(\tau\left(\bar{\omega}, i|1\right)\right)\right. \\ &\left. - J_{\beta_k}^{\pi^*}(\bar{\omega})\right] + \beta_k\left(1-\omega_i\right)\left[J_{\beta_k}^{\pi^*}\left(\tau\left(\bar{\omega}, i|0\right)\right) - J_{\beta_k}^{\pi^*}(\bar{\omega})\right]\Big\}. \end{aligned}$$

From (2.30), we can write this as

$$J^* + h^{\pi^*}(\bar{\omega}) = \max_{a=i} \left\{ \omega_i + \omega_i h^{\pi^*}\left(\tau\left(\bar{\omega}, i|1\right)\right) + \right.$$
$$\left. (1-\omega_i) h^{\pi^*}\left(\tau\left(\bar{\omega}, i|0\right)\right) \right\}. \tag{2.31}$$

Note that (2.31) is nothing but the DP equation as given by (2.10). In addition, we know that the immediate reward as well as function h are both bounded by $\max(1, K)$. This implies that J^* is the maximum average reward, i.e. $J^* = \max_\pi J^\pi_\infty(\bar{\omega}(t))$ (see [19, Theorems 6.1-6.3]).

On the other hand, we know from Theorem 2.12 that the myopic policy is optimal for (P2) if it is for (P1), and thus we can take π^* in (2.27) to be the myopic policy. Rewriting (2.27) gives the following:

(2.32)

$$J^{\pi^*}_\beta(\bar{\omega}) = \omega_{\pi^*(\bar{\omega})} + \beta\omega_{\pi^*(\bar{\omega})} J^{\pi^*}_\beta\left(\tau\left(\bar{\omega}, \pi^*(\bar{\omega})|1\right)\right) + \beta(1-\omega_{\pi^*(\bar{\omega})}) J^{\pi^*}_\beta\left(\tau\left(\bar{\omega}, \pi^*(\bar{\omega})|0\right)\right) \ .$$

Repeating steps (2.29)-(2.31) we arrive at the following:

(2.33)

$$J + h^{\pi^*}(\bar{\omega}) = \omega_{\pi^*(\bar{\omega})} + \omega_{\pi^*(\bar{\omega})} h^{\pi^*}\left(\tau\left(\bar{\omega}, \pi^*(\bar{\omega})|1\right)\right) + (1-\omega_{\pi^*(\bar{\omega})}) h^{\pi^*}\left(\tau\left(\bar{\omega}, \pi^*(\bar{\omega})|0\right)\right),$$

which shows that (J^*, h^{π^*}, π^*) is a canonical triplet [19, Theorems 6.2]. This, together with boundedness of h^{π^*} and immediate reward, implies that the myopic policy π^* is optimal for (P3) [19, Theorems 6.3]. □

2.7 Discussion and Related Work

The problem studied in this chapter may be viewed as a special case of a class of MDPs known as the *restless bandit problems* [4]. In this class of problems, N controlled Markov chains (also called *projects* or *machines*) are activated (or played) one at a time. A machine when activated generates a state dependent reward and

transits to the next state according to a Markov rule. A machine not activated transits to the next state according to a (potentially different) Markov rule. The problem is to decide the sequence in which these machines are activated so as to maximize the expected (discounted or average) reward over an infinite horizon. To put our problem in this context, each channel corresponds to a machine, and a channel is activated when it is probed, and its information state goes through a transition depending on the observation and the underlying channel model. When a channel is not probed, its information state goes through a transition solely based on the underlying channel model [5].

In the case that a machine stays frozen in its current state when not played, the problem reduces to the *multi-armed bandit problem*, a class of problems solved by Gittins in his 1970 seminal work [20]. Gittins showed that there exists an *index* associated with each machine that is solely a function of that individual machine and its state, and that playing the machine currently with the highest index is optimal. This index has since been referred to as the *Gittins index* due to Whittle [21]. The remarkable nature of this result lies in the fact that it essentially decomposes the N-dimensional problem into N 1-dimensional problems, as an index is defined for a machine independent of others. The basic model of multi-armed bandit has been used previously in the context of channel access and cognitive radio networks. For example, in [22], Bayesian learning was used to estimate the probability of a channel being available, and the Gittins indices, calculated based on such estimates (which were only updated when a channel is observed and used, thus giving rise to a multi-armed bandit formulation rather than a restless bandit formulation), were used for channel selection.

[5] The standard definition of bandit problems typically assumes finite or countably infinite state spaces. While our problem can potentially have an uncountable state space, it is nevertheless countable for a given initial state. This view has been taken throughout the chapter.

On the other hand, relatively little is known about the structure of the optimal policies for the restless bandit problems in general. It has been shown that the Gittins index policy is not in general optimal in this case [4], and that this class of problems is PSPACE-hard in general [23]. Whittle, in [4], proposed a Gittins-like index (referred to as the Whittle's index policy), shown to be optimal under a constraint on the *average* number of machines that can be played at a given time, and asymptotically optimal under certain limiting regimes [24]. There has been a large volume of literature in this area, including various approximation algorithms, see for example [25] and [26] for near-optimal heuristics, as well as conditions for certain policies to be optimal for special cases of the restless bandit problem, see e.g., [27, 28]. The nature of the results derived in the present chapter is similar to that of [27, 28] in spirit. That is, we have shown that for this special case of the restless bandit problem an index policy is optimal under certain conditions. For the indexability (as defined by Whittle [4]) of this problem, see [29].

Recently Guha and Munagala [30, 31] studied a class of problems referred to as the *feedback multi-armed bandit* problems. This class is very similar to the restless bandit problem studied in the present chapter, with the difference that channels may have different transition probabilities (thus this is a slight generalization to the one studied here). While we identified conditions under which a simple greedy index policy is optimal in the present chapter, Guha and Munagala in [30, 31] looked for provably good approximation algorithms. In particular, they derived a $2 + \epsilon$-approximate policy using a duality-based technique.

2.8 Conclusion

The general problem of opportunistic sensing and access arises in many multi-channel communication contexts. For cases where the stochastic evolution of channels can be modelled as i.i.d. two-state Markov chains, we showed that a simple and robust myopic policy is optimal for the finite and infinite horizon discounted reward criteria as well as the infinite horizon average reward criterion, when the state transitions are positively correlated over time. When the state transitions are negatively correlated, we showed that the same policy is optimal when the number of channels is limited to 2 or 3, and presented a counterexample for the case of 4 channels.

CHAPTER 3

Opportunistic Spectrum Access as a Restless Bandit Problem With Multiple Plays

3.1 Introduction

In this chapter we study a *multiple-play* extension to the problem defined in the previous chapter, where the user can select more than one channel at a time to sense their availability and can use all available channels among those selected. More specifically, we consider the following stochastic control problem: As before, there are n uncontrolled Markov chains, each an independent, identically-distributed, two-state discrete-time Markov process. The two states will be denoted as state 1 and state 0 and the transition probabilities are given by $p_{ij},\ i,j = 0,1$. The system evolves in discrete time. In each time instance, a user selects exactly $k \geq 1$ out of the n processes and is allowed to observe their states. For each selected process that happens to be in state 1 the user gets a reward; there is no penalty for selecting a channel that turns out to be state 0 but each such occurrence represents a lost opportunity because the user is limited to selecting only k of them. The ones that the user does not select do not reveal their true states. The objective is to derive a selection strategy whose total expected discounted rewarded over a finite or infinite horizon is maximized.

This is again a partially observed MDP (or POMDP) problem [3] due to the fact

that the states of the underlying Markov processes are not fully observed at all times. This problem is also an instance of the restless bandit problem with multiple plays [4, 32, 33]. More discussion on this literature is provided in section 3.5. The problem studied in the previous chapter is a special case of the present one when $k = 1$.

The application of the above problem abstraction to multichannel opportunistic access is as follows. Each Markov process represents a wireless channel, whose state transitions reflect dynamic changes in channel conditions caused by fading, interference, and so on. Specifically, we will consider state 1 as the "good" state, in which a user (or transmitter) can successfully communicate with a receiver; state 0 is the "bad" state, in which communication will fail. The channel state is assumed to remain constant within a single discrete time step. A multichannel system consists of n distinct channels. A user who wishes to use a particular channel at the beginning of a time step must first sense or probe the state of the channel, and can only transmit in a channel probed to be in the "good" state in the same time step. The user cannot sense and access more than k channels at a time due to hardware limitations. If all k selected channels turn out to be in the "bad" state, the user has to wait till the beginning of the next time step to repeat the selection process.

This model captures some of the essential features of multichannel opportunistic access as outlined above. On the other hand, it has the following limitations: the simplicity of the iid two-state channel model; the implicit assumption that channel sensing is perfect and the lack of penalty if the user transmits in a bad channel due to imperfect sensing; and the assumption that the user can select an arbitrary set of k channels out of n (e.g., it may only be able to access a contiguous block of channels due to physical layer limitations). Nevertheless this model does allow us to obtain analytical insights into the problem, and more importantly, some insight into the

more general problem of restless bandits with multiple plays.

As mentioned earlier, this model has been used and studied quite extensively in the past few years, mostly within the context of opportunistic spectrum access and cognitive radio networks, see for example [7, 34, 30, 31]. [7] studied the same problem and proved the optimality of the greedy policy in the special case of $k = 1, n = 2$, [12] proved the optimality of the greedy policy in the case of $k = n - 1$, while [30, 31] looked for provably good approximation algorithms for a similar problem. Furthermore, the indexability (in the context of Whittle's heuristic index and indexability definition [4]) of the underlying problem was studied in [29].

Our previous chapter (as well as [34]) established the optimality of the greedy policy for the special case of $k = 1$ for arbitrary n and under the condition $p_{11} \geq p_{01}$, i.e., when a channel's state transitions are positively correlated. In this sense, the results reported in the present chapter is a direct generalization of results in [34], as we shall prove the optimality of the greedy policy under the same condition but for any $n \geq k \geq 1$. The main thought process used to prove this more general result derives from that used in [34]. However, there were considerable technical difficulties we had to overcome to reach the conclusion.

In the remainder of this chapter we first formulate the problem in Section 3.2, present preliminaries in Section 3.3, and then prove the optimality of the greedy policy in Section 3.4. We discuss our work within the context of restless bandit problems in Section 3.5. Section 3.6 concludes the chapter.

3.2 Problem Formulation

As outlined in the introduction, we consider a user trying to access the wireless spectrum pre-divided into n independent and statistically identical channels, each

given by a two-state Markov chain. The collection of n channels is denoted by $\mathcal{N}$, each indexed by $i = 1, 2, \cdots, n$.

The system operates in discrete time steps indexed by t, $t = 1, 2, \cdots, T$, where T is the time horizon of interest. At time t^-, the channels go through state transitions, and at time t the user makes the channel selection decision. Specifically, at time t the user selects k of the n channels to sense, the set denoted by $a^k \subset \mathcal{N}$.

For channels sensed to be in the "good" state (state 1), the user transmits in those channels and collects one unit of reward for each such channel. If none is sensed good, the user does not transmit, collects no reward, and waits until $t+1$ to make another choice. This process repeats sequentially until the time horizon expires.

The underlying system (i.e., the n channels) is not fully observable to the user. Specifically, channels go through state transition at time t^- (or anytime between $(t-1, t)$), thus when the user makes the channel sensing decision at time t, it does not have the true state of any channel at time t. Furthermore, upon its action (at time t^+) only k channels reveal their true states. The user's action space at time t is given by the finite set $a^k(t) \subset \mathcal{N}$, where $a^k(t) = \{i_1, \ldots, i_K\}$.

We know (see e.g., [3, 16, 17]) that a sufficient statistic of such a system for optimal decision making, or the *information state* of the system [16, 17], is given by the conditional probabilities of the state each channel is in given all past actions and observations. Since each channel can be in one of two states, we denote this information state by $\bar{\omega}(t) = [\omega_1(t), \cdots, \omega_n(t)] \in [0, 1]^n$, where $\omega_i(t)$ is the conditional probability that channel i is in state 1 at time t given all past states, actions and observations [1]. Throughout the chapter $\omega_i(t)$ will be referred to as the information state of channel i at time t, or simply the channel probability of i at time t.

[1] Note that it is a standard way of turning a POMDP problem into a classic MDP problem by means of the information state, the main implication being that the state space is now uncountable.

Due to the Markovian nature of the channel model, the future information state is only a function of the current information state and the current action; i.e., it is independent of past history given the current information state and action. It follows that the information state of the system evolves as follows. Given that the state at time t is $\bar{\omega}(t)$ and action $a^k(t)$ is taken, $\omega_i(t+1)$ for $i \in a^k(t)$ can take on two values: (1) p_{11} if the observation is that channel i is in a "good" state; this occurs with probability $\omega_i(t)$; (2) p_{01} if the observation is that channel i is in a "bad" state; this occurs with probability $1-\omega_i$. For any other channel $j \notin a^k(t)$, with probability 1 the corresponding $\omega_j(t+1) = \tau(\omega_j(t))$ where the operator $\tau : [0,1] \to [0,1]$ is defined as

$$\tau(\omega) := \omega p_{11} + (1-\omega)p_{01},\ 0 \le \omega \le 1\ . \tag{3.1}$$

The objective is to maximize its total discounted expected reward over a finite horizon given in the following problem (P) (extension to infinite horizon is discussed in Section 3.5):

$$\text{(P):}\ \max_{\pi} J_T^{\pi}(\bar{\omega}) = \max_{\pi} E^{\pi}[\sum_{t=1}^{T} \beta^{t-1} R_{\pi_t}(\bar{\omega}(t)) | \bar{\omega}(1) = \bar{\omega}]$$

where $0 \le \beta \le 1$ is the discount factor, and $R_{\pi_t}(\bar{\omega}(t))$ is the reward collected under state $\bar{\omega}(t)$ when channels in the set $a^k(t) = \pi_t(\bar{\omega}(t))$ are selected.

The maximization in (P) is over the class of deterministic Markov policies [2]. An admissible policy π, given by the vector $\pi = [\pi_1, \pi_2, \cdots, \pi_T]$, is such that π_t specifies a mapping from the current information state $\bar{\omega}(t)$ to a channel selection action $a^k(t) = \pi_t(\bar{\omega}(t)) \subset \{1, 2, \cdots, n\}$. This is done without loss of optimality due to the Markovian nature of the underlying system, and due to known results on POMDPs [17, Chapter 6].

[2] A Markov policy is a policy that derives its action only depending on the current (information) state, rather than the entire history of states, see e.g., [17].

3.3 Preliminaries

The dynamic programming (DP) representation of problem (P) is given as follows:

$$
\begin{aligned}
V_T(\bar{\omega}) &= \max_{a^k \in \mathcal{N}, |a^k|=k} E[R_{a^k}(\bar{\omega})] \\
V_t(\bar{\omega}) &= \max_{a^k \in \mathcal{N}, |a^k|=k} \Big(\sum_{i \in a^k} \omega_i + \beta \cdot \sum_{l_i \in \{0,1\}, i \in a^k} \left(\prod_{i \in a^k} \omega_i^{l_i} (1-\omega_i)^{1-l_i} \right) \cdot \\
& \qquad V_{t+1}(p_{01}, \ldots, p_{01}, \tau(\omega_j), p_{11}, \ldots, p_{11})\Big), \qquad (3.2) \\
& \qquad t = 1, 2, \cdots, T-1.
\end{aligned}
$$

In the last term, the channel state probability vector consists of three parts: a sequence of p_{01}'s that represent those channels sensed to be in state 0 at time t and the length of this sequence is the number of l_i's equaling zero; a sequence of values $\tau(\omega_j)$ for all $j \notin a^k$; and a sequence of p_{11}'s that represent those channels sensed to be in state 1 at time t and the length of this sequence is the number of l_i's equaling one. Note that the future expected reward is calculated by summing over all possible realizations of the k selected channels.

The value function $V_t(\bar{\omega})$ represents the maximum expected future reward that can be accrued starting from time t when the information state is $\bar{\omega}$. In particular, we have $V_1(\bar{\omega}) = \max_\pi J_T^\pi(\bar{\omega})$, and an optimal deterministic Markov policy exists such that $a = \pi_t^*(\bar{\omega})$ achieves the maximum in (3.3) (see e.g., [18] (Chapter 4)).

For simplicity of representation, we introduce the following notations:

- $p_{01}[x]$: this is the vector $[p_{01}, p_{01}, \cdots, p_{01}]$ of length x;
- $p_{11}[x]$: this is the vector $[p_{11}, p_{11}, \cdots, p_{11}]$ of length x.
- We will use the notation:

$$q(l_1, \cdots, l_k) := \prod_{1 \le i \le k} \left(\omega_i^{l_i} (1-\omega_i)^{1-l_i}\right)$$

for $l_1, \cdots, l_k \in \{0, 1\}$. That is, given a vector of 0s and 1s (total of k elements), $q()$ is the probability that a set of k channels are in states given by the vector.

With the above notation, Eqn (3.3) can be written as

$$\begin{aligned} V_t(\bar{\omega}) &= \max_{a^k \in \mathcal{N}, |a^k|=k} \Big(\sum_{i \in a^k} \omega_i + \beta \cdot \\ &\qquad \sum_{l_i \in \{0,1\}, i \in a^k} q(l_1, \cdots, l_k) \cdot V_{t+1}(p_{01}[k - \sum l_i], \cdots, \tau(\omega_j), p_{11}[\sum l_i]) \Big) . \end{aligned}$$

Solving (P) using the above recursive equation can be computationally heavy, especially considering the fact that $\bar{\omega}$ is a vector of probabilities. It is thus common to consider suboptimal policies that are easier to compute and implement. One of the simplest such heuristics is a greedy policy where at each time step we take an action that maximizes the immediate one-step reward. Our focus is to examine the optimality properties of such a simple greedy policy.

For problem (P), the greedy policy under state $\bar{\omega} = [\omega_1, \omega_2, \cdots, \omega_n]$ is given by

$$a^k(\bar{\omega}) = \arg \max_{a^k \subset \mathcal{N}, |a^k|=k} \sum_{i \in a^k} \omega_i \ . \tag{3.3}$$

That is, the greedy policy seeks to maximize the reward *as if* there were only one step remaining in the horizon. In the next section we investigate the optimality of this policy. Specifically, we will show that it is optimal in the case of $p_{11} \geq p_{01}$. This extends the earlier result in [34] that showed this to be true for the special case of $k = 1$.

3.4 Optimality of the Greedy Policy

In this section we show that the greedy policy is optimal when $p_{11} \geq p_{01}$. The main theorem of this section is as follows.

Theorem 3.1. *The greedy policy is optimal for Problem (P) under the assumption that* $p_{11} \geq p_{01}$. *That is, for* $t = 1, 2, \cdots, T$, $k \leq n$, *and* $\forall \bar{\omega} = [\omega_1, \cdots, \omega_n] \in [0,1]^n$, *we have*

$$V_t^k(\bar{\omega}; z^k(\bar{\omega})) \geq V_t^k(\bar{\omega}; a^k), \ \forall a^k \subset \mathcal{N}, \tag{3.4}$$

where $z^k(\bar{\omega})$ *is the subset whose elements (indices) correspond to the* k *largest values in* $\bar{\omega}$, *and* $V_t^k(\bar{\omega}; a^k)$ *the expected value of action* a^k *followed by behaving optimally.*

Below we present a number of lemmas used in the proof of this theorem. The first lemma introduces a notation that allows us to express the expected future reward under the greedy policy.

Lemma 3.2. *There exist* T n*-variable functions, denoted by* $W_t^k(\bar{\omega})$, $t = 1, 2, \cdots, T$, *each of which is a polynomial of order 1*[3] *and can be represented recursively in the following form:*

$$\begin{aligned} W_T^k(\bar{\omega}) &= \sum_{n-1+1 \leq i \leq n} \omega_i \\ W_t^k(\bar{\omega}) &= \sum_{n-1+1 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_n, l_{n-1}, \cdots, l_{n+k-1} \in \{0,1\}} q(l_n, \cdots, l_{n+k-1}) \cdot \\ & W_{t+1}^k(p_{01}[k - \sum l_i], \tau(\omega_i), \cdots, \tau(\omega_{n-k}), p_{11}[\sum l_i]) \ . \end{aligned}$$

The proof is easily obtained using backward induction on t given the recursive equation and noting that the mapping $\tau()$ is linear. The detailed proof is thus omitted for brevity.

A few remarks are in order on this function $W_t^k(\bar{\omega})$.

1. Firstly, when $\bar{\omega}$ is given by an ordered vector $[\omega_1, \omega_2, \cdots, \omega_n]$ with $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_n$, $W_t^k(\bar{\omega})$ is the expected total discounted future reward (from t to T) by following the greedy policy.

[3] Each function W_t is affine in each variable, when all other variables are held constant.

This follows from how the greedy policy works in the special case of $p_{11} \geq p_{01}$. Note that in this case the conditional probability updating function $\tau(\omega)$ is a monotonically increasing function, i.e., $\tau(\omega_1) \geq \tau(\omega_2)$ for $\omega_1 \geq \omega_2$. Therefore the ordering of channel probabilities is preserved among those that are not observed.

If a channel has been observed to be in state "1" (respectively "0"), its probability at the next step becomes $p_{11} \geq \tau(\omega)$ (respectively $p_{01} \leq \tau(\omega)$) for any $\omega \in [0, 1]$. In other words, a channel observed to be in state "1" (respectively "0") will have the highest (respectively lowest) possible probability among all channels.

Therefore if we take the initial information state $\bar{\omega}(1)$, order the channels according to their probabilities $\omega_i(1)$, and sense the highest k channels (top k of the ordered list) with ties broken randomly, then following the greedy policy means that in subsequent steps we will keep a channel in its current position if it was sensed to be in state 1 in the previous slot; otherwise, it was observed to be in state 0 and gets thrown to the bottom of the ordered list. The policy then selects the next top most (or rightmost) k channels on this new ordered list. This procedure is essentially the same as that given in the recursive expression of $W()$.

2. Secondly, when $\bar{\omega}$ is not ordered, $W_t^k()$ reflects a policy that simply goes down the list of channels by the order fixed in $\bar{\omega}$, while each time tossing the ones observed to be 0 to the end of the list and keeing those observed to be 1 at the top of the list.

3. Thirdly, the fact that W_t^K is a polynomial of order 1 and affine in each of its

elements implies that

$$\begin{aligned} & W_t^K(\omega_1, \cdots, \omega_{n-2}, y, x) - W_t^K(\omega_1, \cdots, \omega_{n-2}, x, y) \\ = \;& (x-y)[W_t^K(\omega_1, \cdots, \omega_{n-2}, 0, 1) - \; W_t^K(\omega_1, \cdots, \omega_{n-2}, 1, 0)] \; . \end{aligned}$$

Similar results hold when we change the positions of x and y. To see this, consider the above as two functions of x and y, each having an x term, a y term, an xy term and a constant term. Since we are only swapping the positions of x and y in these two functions, the constant term remains the same, and so does the xy term. Thus the only difference is the x term and the y term, as given in the above equation. This linearity result is used later in our proof.

The next lemma establishes a sufficient condition for the optimality of the greedy policy.

Lemma 3.3. *Consider Problem (P) under the assumption that $p_{11} \geq p_{01}$. To show that the greedy policy is optimal at time t given that it is optimal at $t+1, t+2, \cdots, T$, it suffices to show that at time t we have*

$$W_t^k(\omega_1, \cdots, \omega_j, x, y, \cdots, \omega_n) \leq W_t^k(\omega_1, \cdots, \omega_j, y, x, \cdots, \omega_n), \tag{3.5}$$

for all $x \geq y$ and all $0 \leq j \leq n-2$, with $j = 0$ implying $W_t^k(x, y, \omega_3, \cdots, \omega_n) \leq W_t^k(y, x, \omega_3, \cdots, \omega_n)$.

Proof. Since the greedy policy is optimal from $t+1$ on, it is sufficient to show that selecting the best k channels followed by the greedy policy is better than selecting any other set of k channels followed by the greedy policy. If channels are ordered $\omega_1 \leq \cdots \leq \omega_i \leq \cdots \leq \omega_n$ then the reward of the former is precisely given by $W_t^K(\omega_1, \ldots, \omega_n)$. On the other hand, the reward of selecting an arbitrary set a^k of k

channels followed by acting greedily can be expressed as $W_t^k(\overline{a^k}, a^k)$, where $\overline{a^k}$ is the (increasingly) ordered set of channels not included in a^k. It remains to show that if Eqn (3.5) is true then we have $W_t^k(\overline{a^k}, a^k) \leq W_t^K(\omega_1, \ldots, \omega_n)$. This is easily done since the ordered list $(\overline{a^k}, a^k)$ may be converted to $\omega_1, \ldots, \omega_n$ through a sequence of switchings between two neighboring elements that are not increasingly ordered. Each such switch invokes (3.5), thereby maintaining the "$\leq$" relationship. □

Lemma 3.4. *For $0 \leq \omega_1 \leq \omega_2 \leq \ldots \leq \omega_n \leq 1$, we have the following two inequalities for all $t = 1, 2, \cdots, T$:*

$$(A): \quad 1 + W_t^k(\omega_2, \cdots, \omega_n, \omega_1) \geq W_t^k(\omega_1, \cdots, \omega_n)$$

$$(B): \quad W_t^k(\omega_1, \cdots, \omega_j, y, x, \omega_{j+3}, \cdots, \omega_n) \geq W_t^k(\omega_1, \cdots, x, y, \omega_{j+3}, \cdots, \omega_n),$$

where $x \geq y$, $0 \leq j \leq n-2$, and $j = 0$ implies $W_t^k(y, x, \omega_3, \cdots, \omega_n) \geq W_t^k(x, y, \omega_3, \cdots, \omega_n)$.

This lemma is the key to our main result and its proof, which uses a sample path argument, highly instructive. It is however also lengthy, and for this reason has been relegated to the Appendix.

With the above lemmas, Theorem 1 is easily proven:

Proof of Theorem 1: We prove by induction on T. When $t = T$, the greedy policy is obviously optimal. Suppose it is also optimal for all times $t+1, t+2, \cdots, T$, under the assumption $p_{11} \geq p_{01}$. Then at time t, by Lemma 3.3, it suffices to show that $W_t^k(\omega_1, \cdots, \omega_j, x, y, \cdots, \omega_n) \leq W_t^k(\omega_1, \cdots, \omega_j, y, x, \cdots, \omega_n)$ for all $x \geq y$ and $0 \leq j \leq n-2$. But this is proven in Lemma 3.4.

3.5 Discussion

While the formulation (P) is a finite horizon problem, the same result applies to the infinite horizon discounted reward case using standard techniques as we have done in our previous work [8, 34].

In the case of infinite horizon, the problem studied in this chapter is closely associated with the class of multi-armed bandit problems [20] and restless bandit problems [4]. This is a class of problems where n controlled Markov chains (also called machines or arms) are activated (or played) one at a time. A machine when activated generates a state dependent reward and moves to the next state according to a Markov rule. A machine not activated either stays frozen in its current state (a rested bandit) or moves to the next state according to a possibly different Markov rule (a restless bandit). The problem is to decide the sequence in which these machines are activated so as to maximize the expected (discounted or average) reward over an infinite horizon.

The multi-armed bandit problem was originally solved by Gittins (see [20]), who showed that there exists an *index* associated with each machine that is solely a function of that individual machine and its state, and that playing the machine currently with the highest index is optimal. This index has since been referred to as the *Gittins index*. The remarkable nature of this result lies in the fact that it decomposes the n-dimensional problem into n 1-dimensional problems, as an index is defined for a machine independent of others. The restless bandit problem on the other hand was proven much more complex, and is PSPACE-hard in general [23]. Relatively little is known about the structure of its optimal policy in general. In particular, the Gittins index policy is not in general optimal [4].

When multiple machines are activated simultaneously, the resulting problem is referred to as multi-armed bandits with *multiple plays*. Again optimal solutions to this class of problems are not known in general. A natural extension to the Gittins index policy in this case is to play the machines with the highest Gittins indices (this will be referred to as the *extended Gittins index policy* below). This is not in general optimal for multi-armed bandits with multiple plays and an infinite horizon discounted reward criterion, see e.g., [35, 36]. However, it may be optimal in some cases, see e.g., [36] for conditions on the reward function, and [37] for an undiscounted case where the Gittins index is always achieved at time 1. Even less is known when the bandits are restless, though asymptotic results for restless bandits with multiple plays were provided in [4] and [24].

The problem studied in the present chapter is an instance of the restless bandits with multiple plays (in the infinite horizon case). Therefore what we have shown in this chapter is an instance of the restless bandits problem with multiple plays, for which the extended Gittins index policy is optimal.

3.6 Conclusion

In this chapter we studied a stochastic control problem that arose in opportunistic spectrum access. A user can sense and access k out of n channels at a time and must select judiciously in order to maximize its reward. We extend a previous result where a greedy policy was shown to be optimal in the special case of $k = 1$ under the condition that the channel state transitions are positively correlated over time. In this chapter we showed that under the same condition the greedy policy is optimal for the general case of $k \geq 1$. This result also contributes to the understanding of the class of restless bandit problems with multiple plays.

CHAPTER 4

Opportunistic Spectrum Sharing as a Congestion Game

4.1 Introduction

In this chapter we present a generalized form of the class of non-coopertive strategic games known as *congestion games* (CG) [38, 39], and study its properties as well as its application to spectrum sharing in a cognitive radio network.

In a classical congestion game, multiple users share multiple resources. The payoff [1] for any user to use a particular resource depends on the number of users using that resource concurrently. A detailed and formal description is provided in Section 4.2. The congestion game framework is well suited to model resource competition where the resulting payoff is a function of the level of congestion (number of active users). It has been extensively studied within the context of network routing, see for instance the congestion game studied in [40] [2], where source nodes seek minimum delay path to a destination and the delay of a link depends on the number of flows going through that link.

A congestion game enjoys many nice properties. For example, it always has a pure strategy Nash Equilibrium (NE), and any asynchronous improvement path is

[1] Sometimes people consider the cost of using a resource instead of the payoff. If we define the cost as the inverse of the payoff, then maximizing the payoff is equivalent to minimizing the cost. For simplicity of presentation, we will focus on the maximization of payoff in this chapter.

[2] Note that while also called "network congestion game", the game studied in [40] is essentially a classical congestion game with resources being links in a network. By contrast, the network congestion game defined in this chapter is a network game meaning that the relationship among players are given by a network.

finite and will lead to a pure strategy NE (referred to as the finite improvement property (FIP). In fact, while the system is decentralized and all players are selfish, by seeking to optimize their individual objectives they end up optimizing a global objective, also called the potential function, and doing so in a finite number of steps regardless of the updating sequence.

For these reasons, it is tempting to model resource competition in a wireless communication system as a congestion game. However, the standard congestion game fails to capture a critical aspect of resource sharing in wireless communication: *interference*. A key assumption underlying the congestion game model is that all users have an equal impact on the congestion, and therefore all that matters is the total number of users of a resource. This however is not true in wireless communication. Specifically, if we consider bandwidth or channels as resources, then sharing the same channel is complicated by interference; a user's payoff (e.g., channel quality, achievable rates, etc.) depends on *who* the other users are and how much interference it receives from them. If all other simultaneous users are located sufficiently far away, then sharing may not cause any performance degradation, a feature commonly known as spatial reuse.

The above consideration poses significant challenge in using the congestion game model depending on what type of user objectives we are interested in. In our recent work [41], we addressed the user-specific interference issue within the congestion game framework, by introducing a concept called *resource expansion*, where we define virtual resources as certain spectral-spatial unit that allows us to capture pair-wise interference. This approach was shown to be quite effective for user objectives like interference minimization.

In this chapter, we take a different approach where we generalize the standard

congestion games to directly account for the interference relationship and spatial reuse in wireless networks. Specifically, under this generalization, users are placed over a network representing an interference graph. An edge exists between two users that interfere with each other. In using a resource (a wireless channel), a user's payoff is a function of the total number of users *within its interference neighborhood* using it. Therefore, resources are *reusable* beyond a user's interference set. This extension is a generalization of the original congestion game definition, as the former reduces to the latter if the underlying network is complete (i.e., every user interferes with every other user). This class of generalized games will be referred to as *network congestion games* (NCG).

The applicability of this class of games to a multi-channel, multi-user wireless communication system can be easily understood. Specifically, we consider such a system where a user can only access one channel at a time, but can switch between channels. A user's principal interest lies in optimizing its own performance objective (i.e., its data rate) by selecting the best channel for itself. This and similar problems have recently captured increasing interest from the research community, particularly in the context of cognitive radio networks (CRN) and software defined ratio (SDR) technologies, whereby devices are expected to have far greater flexibility in sensing channel availability/condition and moving operating frequencies.

While directly motivated by resource sharing in a multi-channel, multi-user wireless communication system, the definition of a NCG is potentially more broadly applicable. It simply reflects the notion that in some application scenarios resources may be shared without conflict of interest. In subsequent sections we will examine what properties this class of games possesses. Our main findings are summarized as follows for undirected network graphs and non-increasing payoff functions:

1. The FIP property is preserved in an NCG with only two resources/channels. Counter examples exist for three or more resources.

2. A pure strategy NE exists in a NCG over a tree network and a loop.

3. A pure strategy NE exists when there is either a dominating resource (a channel with much larger bandwidth than the rest) or when all resources are identical (all channels are equal). Furthermore, in the latter case the FIP also exists.

In addition, we also show that an NE does not in general exist if the network graph is directed (meaning that the interference relationship between users is asymmetric), or that the user payoff functions are non-monotonic.

It has to be mentioned that game theoretic approaches have often been used to devise effective decentralized solutions to a multi-agent system. Within the context of wireless communication networks and interference modeling, different classes of games have been studied. An example is the well-known *Gaussian interference game* [42, 43]. In a Gaussian interference game, a player can spread a fixed amount of power arbitrarily across a continuous bandwidth, and tries to maximize its total rate in a Gaussian interference channel over all possible power allocation strategies. It has been shown [42] that it has a pure strategy NE, but the NE can be quite inefficient; playing a repeated game can improve the performance. In addition, previous work [44] investigated a market based power control mechanism via supermodularity, while previous work [45] studied the Bayesian form of the Gaussian interference game in the case of incomplete information. By contrast, in our problem the total power of a user is not divisible, and it can only use it in one channel at a time. This setup is more appropriate for scenarios where the channels have been pre-defined, and the users do not have the ability to access multiple channels simultaneously (which is

the case with many existing devices).

The organization of the remainder of this chapter is as follows. In Section 4.2 we present a brief review on the literature of congestion games, and formally define the class of network congestion games in Section 4.3. We then derive conditions under which this class of games possesses the finite improvement property (Section 4.4). We further show a series of conditions, on the underlying network graph in Section 4.5 and on the user payoff function in Section 4.6, under which these games have a pure strategy NE. We discuss extensions to our work and conclude the chapter in 4.7.

4.2 A Review of Congestion Games

In this section we provide a brief review on the definition of congestion games and their known properties[3]. We then discuss why the standard congestion game does not take into account spatial reuse and motivate our generalized network congestion games.

4.2.1 Congestion Games

Congestion games [38, 39] are a class of strategic games given by the tuple $(\mathcal{I}, \mathcal{R}, (\Sigma_i)_{i\in\mathcal{I}}, (g_r)_{r\in\mathcal{R}})$, where $\mathcal{I} = \{1, 2, \cdots, N\}$ denotes a set of users, $\mathcal{R} = \{1, 2, \cdots, R\}$ a set of resources, $\Sigma_i \subset 2^R$ the strategy space of player i, and $g_r : \mathbb{N} \to \mathbb{Z}$ a payoff (or cost) function associated with resource r. The payoff (cost) g_r is a function of the total number of users using resource r and in general assumed to be non-increasing (non-decreasing). A player in this game aims to maximize (minimize) its total payoff (cost) which is the sum total of payoff (cost) over all resources its strategy involves.

If we denote by $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_N)$ the strategy profile, where $\sigma_i \in \Sigma_i$, then

[3]This review along with some of our notations are primarily based on references [38, 39, 46].

user i's total payoff (cost) is given by

$$q^i(\boldsymbol{\sigma}) = \sum_{r \in \sigma_i} g_r(n_r(\boldsymbol{\sigma})) \tag{4.1}$$

where $n_r(\boldsymbol{\sigma})$ is the total number of users using resource r under the strategy profile $\boldsymbol{\sigma}$, with and $r \in \sigma_i$ denoting that user i selects resource r under $\boldsymbol{\sigma}$.

Rosenthal's potential function $\phi : \Sigma_1 \times \Sigma_2 \times \cdots \times \Sigma_N \to \mathbb{Z}$ is defined by

$$\phi(\boldsymbol{\sigma}) = \sum_{r \in \mathcal{R}} \sum_{i=1}^{n_r(\boldsymbol{\sigma})} g_r(i) = \sum_{i=1}^{N} \sum_{r \in \sigma_i} g_r(m_r^i(\boldsymbol{\sigma})) \ , \tag{4.2}$$

where the second equality comes from exchanging the two sums and $m_r^i(\boldsymbol{\sigma})$ denotes the number of players who use resource r under strategy $\boldsymbol{\sigma}$ and whose corresponding indices do not exceed i (i.e., in the set $\{1, 2, \cdots, i\}$).

Next we show that the change in a user's payoff as a results off its unilateral move (i.e., all other users stay put) is exactly the same as the change in the potential, which may be viewed as a global objective function. Consider player i, who unilaterally moves from strategy σ_i (within the profile $\boldsymbol{\sigma}$) to strategy $\sigma_i^{'}$ (within the profile $\boldsymbol{\sigma}^{'}$). The potential changes by

$$\begin{aligned} \phi(\sigma_i', \sigma_{-i}) - \phi(\sigma_i, \sigma_{-i}) &= \sum_{r \in \sigma_i^{'}, r \notin \sigma_i} g_r(n_r(\boldsymbol{\sigma}) + 1) - \sum_{r \in \sigma_i, r \notin \sigma_i^{'}} g_r(n_r(\boldsymbol{\sigma})) \\ &= \sum_{r \in \sigma_i^{'}} g_r(n_r(\boldsymbol{\sigma}^{'})) - \sum_{r \in \sigma_i} g_r(n_r(\boldsymbol{\sigma})) = g^i(\sigma^{-i}, \sigma_i^{'}) - g^i(\sigma^{-i}, \sigma_i) \ , \end{aligned} \tag{4.3}$$

where the second equality comes from the fact that for resources that are used by both strategies σ_i and $\sigma_i^{'}$ there is no change in their total number of users. To see why the first equality is true, set $i = N$, in which case this equality is a direct consequence of the change of sums equation (4.2). To see why this is true for any $1 \leq i \leq N$, simply note that the ordering of users is arbitrary so any user making a change may be viewed as the Nth user.

Consider now a sequence of strategy changes made by users asynchronously in which each change improves the user's payoff (this is referred to as a sequence of improvement steps). The above result shows that upon each such change the potential also improves. Since the potential of any strategy profile is finite, it follows that every sequence of improvement steps is finite, and they converge to a pure strategy Nash Equilibrium. This is known as the finite improvement property (FIP). Furthermore, this NE is a local optimal point of the potential function ϕ, defined as a strategy profile where changing one coordinate cannot result in a greater value of ϕ.

It is not difficult to see why the standard definition of a congestion game does not capture the features of wireless communication. In particular, if we consider channels as resources, then the payoff $g_r(n)$ for using channel r when there are n simultaneous users does not reflect reality: the function $g_r(\cdot)$ in general takes a user-specific argument since different users experience different levels of interference even when using the same resource. This user specificity is also different from that studied in [2], where $g_r(\cdot)$ is a user-specific function $g_r^i(\cdot)$ but it takes the non-user specific argument n. To analyze and understand the consequence of this difference, we would need to extend and generalize the definition of the standard congestion game.

For the rest of this chapter, the term *player* or *user* specifically refers to a *pair* of transmitter and receiver in the network. Interference in this context is between one user's transmitter and another user's receiver. This is commonly done in the literature, see for instance [42]. We will also assume that each player has a fixed transmit power.

4.3 Problem Formulation

In this section we formally definite our generalized congestion game, the *network congestion games* (NCG). Specifically, a NCG is given by $(\mathcal{I}, \mathcal{R}, (\Sigma_i)_{i\in\mathcal{I}}, \{\mathcal{K}_i\}_{i\in\mathcal{I}}, \{g_r^i\}_{r\in\mathcal{R},i\in\mathcal{I}})$, where $\mathcal{K}_i$ is the interference set of user i, excluding itself, while all other elements maintain the same meaning as in a standard CG. The payoff user i receives for using resource r is given by $g_r^i(n_r^i(\boldsymbol{\sigma}) + 1)$ where $n_r^i(\boldsymbol{\sigma}) = |\{j : r \in \sigma_j, j \in \mathcal{K}_i\}|$. That is, user i's payoff for using r is a (user-specific) function of the number of users interfering with itself, plus itself. Here we have explicitly made the payoff functions user-specific, as evidenced by the index i in $g_r^i(\cdot)$. This is done in an attempt to capture the fact that users with different coding/modulation schemes may obtain different rates from using the same channel.

A user's payoff is the sum of payoffs from all the resources it uses. Note that if a user is allowed to simultaneously use all available resources, then its best strategy is to simply use all of them regardless of other users, provided that g_r^i is a non-negative function. If all users are allowed such a strategy, then the existence of an NE is trivially true.

In this chapter, we will limit our attention to the case where each user is allowed only one channel at a time, i.e., its strategy space $\Sigma_i \in \mathcal{R}$ consists of R single channel strategies. In this case the payoff user i receives for using a single channel r is given by $g_r^i(n_r^i + 1)$ where $n_r^i(\boldsymbol{\sigma}) = |\{j : r = \sigma_j, j \in \mathcal{K}_i\}|$. Our goal is to identify key properties of this game.

It's worth noting that due to this generalization, Rosenthal's definition of a potential function as given in the previous section no longer applies. To slightly simplify this problem, we make the extra assumption that $i \in \mathcal{K}_j$ if and only if $j \in \mathcal{K}_i$. This

has the intuitive meaning that if node i interferes with node j, the reverse is also true. This symmetry does not always hold in reality, but is nonetheless a useful one to help us obtain meaningful insight. It is easy to see that we can equivalently represent a more general problem on the following directed graph, where a node represents a user and a directed edge connects node i to node j if and only if $i \in \mathcal{K}_j$. The network congestion game can now be stated as a coloring problem, where each node picks a color and receives a value depending on the conflict (number of same colors neighboring to a node); the goal is to see whether an NE exists and whether a decentralized selfish scheme leads to an NE. For the special case that we consider in this chapter, the graph is undirected, where there is an undirected edge between nodes i and j if and only if $i \in \mathcal{K}_j$ and $j \in \mathcal{K}_i$.

For simplicity of exposition, in subsequent sections we will often present the problem in its coloring version, and will use the terms *resource*, *channel*, and *color* interchangeably.

4.4 Existence of the Finite Improvement Property

In this section we investigate whether the network congestion game as defined in the previous section possesses the FIP property. Once a game has the FIP, it immediately follows that it has an NE as we described in Section 4.2. Below we show that in the case of two resources (colors) this game indeed has the FIP property, and as a result an NE exists. We also show through a counter example that for the case of 3 or more colors the FIP property does not hold. This also implies that in such cases an exact potential function does not exist for this game, as the FIP is a direct consequence of the existence of a potential function.

4.4.1 The Finite Improvement Property for 2 resources

We shall establish this result by a contradiction argument. Suppose that we have a sequence of asynchronous[4] updates that starts and ends in the exact same color assignment (or state) for all users. We denote such a sequence by

$$U = \{u(1), u(2), \cdots u(T)\}, \tag{4.4}$$

where $u(t) \in \{1, 2, \cdots, N\}$ denotes the user making the change at time t, and T is the length of this sequence. The starting state (or the color choice) of the system is given by

$$S(1) = \{s_1(1), s_2(1), \cdots, s_N(1)\}, \tag{4.5}$$

where $s_i(1) \in \{r, b\}$, i.e., the state of each user is either "r" for Red, or "b" for Blue. A user's state of color is defined for time t^-, i.e., right before a color change is made by some user at time t. In other words, $s_i(t)$ denotes the color of user i at time t^-. Since there are only two colors, we use the notation $\bar{s}$ to denote the opposite color of a color s. The states after the last round of change at time T is denoted by $S(T)$.

Since this sequence of updates form a loop, i.e., $S(1^-) = S(T)$, we can naturally view these updates as being placed around a circle, starting at time 1^- and ending at T, when the system returns to its original state. This is shown in Figure 4.1. Note that traversing the circle starting from any point results in an improvement path; hence the notion of a starting point becomes inconsequential.

Since this sequence of updates is an improvement path, each change must increase the payoff of the user making the change [5]. For example, suppose user i changes from

[4] We will remove the word *asynchronous* in the following with the understanding that whenever we refer to updates they are assumed to be asynchronous updates, i.e., there will not be two or more users changing their strategies simultaneously at any time.

[5] Here we assume that a user only makes a change if there is *strict* increase in its payoff.

red to blue at time t, and i has x red neighbors and y blue neighbors at t[6]. Then we must have:

$$g_b^i(y+1) > g_r^i(x+1) \ . \tag{4.6}$$

Similarly, we can obtain one inequality for each of the T changes. Our goal is to show that these T inequalities cannot be consistent with each other. The challenge here is that this contradiction has to hold for arbitrary non-increasing functions $\{g_r^i, g_b^i\}$. The way we address this challenge is to show that the above inequality leads to another inequality that does *not* involve the payoff functions when we consider pairs of reverse changes by the same user. This is shown in Lemma 4.2.

Definition 4.1 (Reverse-change pairs)**.**

Consider an arbitrary user i's two reverse strategy/color changes in an improvement path, one from s to $\bar{s}$ at time t and the other from $\bar{s}$ to s at time t'. Let $\mathcal{SS}_{t,t'}^i$ denote the set of i's neighbors (not including i) who have the same color as i at both times of change (i.e., at t^- and t'^-, respectively). Let $\mathcal{OO}_{t,t'}^i$ denote the set of i's neighbors (not including i) who have the opposite color as i at both times of change. Similarly, we will denote by $\mathcal{SO}_{t,t'}^i$ (respectively $\mathcal{OS}_{t,t'}^i$) the number of i's neighbors whose color is the same as (opposite of, respectively) i's at the first update and the opposite of (same as, respectively) i's at the second update.

Lemma 4.2. (Reverse-change inequality) *Consider the network congestion game defined in the previous section with non-increasing payoff functions and two resources/colors. Suppose an arbitrary user i makes two reverse strategy/color changes in an improvement path, one from s to $\bar{s}$ at time t and the other from $\bar{s}$ to s at time*

[6]Since the users update their strategies in an asynchronous fashion, x and y do not change between t^- and t^+.

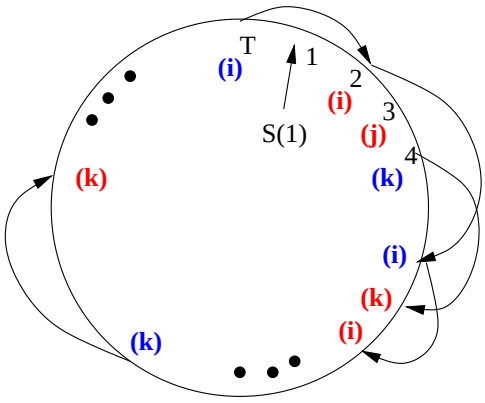

Figure 4.1: Representing an improvement loop on a circle: times of updates t and the updating user $(u(t))$ are illustrated along with their color right before a change. An arrow connects a single user's two consecutive color changes.

t'. *Then we have*

$$|\mathcal{SS}^i_{t,t'}| > |\mathcal{OO}^i_{t,t'}|, \ \forall i \in \mathcal{I} \ . \tag{4.7}$$

That is, among i's neighbors, there are strictly more users that have the same color as i at both times of change than those with the opposite color as i at both times of change.

Proof. Since this is an improvement path, whenever i makes a change it is for higher payoff. Thus we must have at the time of its first change and its second change, respectively, the following inequalities:

$$g^i_{\bar{s}}(|\mathcal{OS}^i_{t,t'}| + |\mathcal{OO}^i_{t,t'}| + 1) > g^i_s(|\mathcal{SO}^i_{t,t'}| + |\mathcal{SS}^i_{t,t'}| + 1) \ ; \tag{4.8}$$

$$g^i_s(|\mathcal{SO}^i_{t,t'}| + |\mathcal{OO}^i_{t,t'}| + 1) > g^i_{\bar{s}}(|\mathcal{OS}^i_{t,t'}| + |\mathcal{SS}^i_{t,t'}| + 1) \ . \tag{4.9}$$

We now prove the lemma by contradiction. Suppose that the statement is not true and that we have $|\mathcal{SS}^i_{t,t'}| \leq |\mathcal{OO}^i_{t,t'}|$. Then due to the non-increasing assumption on

the payoff functions we have

(4.10)

$$g^i_{\bar{s}}(|\mathcal{OS}^i_{t,t'}| + |\mathcal{SS}^i_{t,t'}| + 1) \geq g^i_{\bar{s}}(|\mathcal{OS}^i_{t,t'}| + |\mathcal{OO}^i_{t,t'}| + 1) > g^i_s(|\mathcal{SO}^i_{t,t'}| + |\mathcal{SS}^i_{t,t'}| + 1)$$

(4.11)

$$\geq g^i_s(|\mathcal{SO}^i_{t,t'}| + |\mathcal{OO}^i_{t,t'}| + 1)$$

where the second inequality is due to (4.8). However, this contradicts with (4.9), completing the proof. □

We point out that by the above lemma the payoff comparison is reduced to counting different sets of users. This greatly simplifies the process of proving the main theorem of this section. Below we show that it is impossible to have a finite sequence of asynchronous improvement steps ending in the same color state as it started with. At the heart of the proof is the repeated use of the above lemma to show that loops cannot form in a sequence of asynchronous updates.

Theorem 4.3. *Consider the network congestion game defined in the previous section with non-increasing payoff functions. For the special case when there are only two resources/colors to choose from and a user can only use one at a time, we have the finite improvement property.*

Proof. We prove this by contradiction. As illustrated by Figure 4.1, we consider a sequence of improvement updates that results in the same state.

Consider every two successive color changes, along this circle clockwise starting from time $t = 1$, that a user $u(t)$ makes at time t and t' from color $s = s_{u(t)}(t)$ to $\bar{s}$, and then back to s, respectively. Note that this will include the two "successive" changes formed by a user's last change and its first change (successive on this circle but not in terms of time). We have illustrated this in Figure 4.1 by connecting a pair

of successive color changes using an arrow. It is easy to see that there are altogether T such pairs (or arrows).

For each arrow in Figure 4.1, or equivalently each pair of successive color changes by the same user, we consider the two sets $\mathcal{SS}_{t,t'}^{u(t)}$ and $\mathcal{OO}_{t,t'}^{u(t)}$ in Definition 4.1. Due to the user association, we will also refer to these sets as *perceived* by user $u(t)$. By Lemma 1, given an updating sequence with the same starting and ending states, we have for each pair of successive reverse changes by the same user, at time t and time t', respectively:

$$|\mathcal{SS}_{t,t'}^{u(t)}| > |\mathcal{OO}_{t,t'}^{u(t)}|,\ t = 1, 2, \cdots, T\ . \tag{4.12}$$

That is, the $\mathcal{SS}$ sets are strictly larger than the $\mathcal{OO}$ sets.

This gives us a total of T inequalities, one for each update in the sequence and each containing two sets. Equivalently there is one inequality per arrow illustrated in Figure 4.1. We next consider how many users are in each of these $2T$ sets (note that by keeping the same "$>$" relationship, the $\mathcal{SS}$ sets are always on the LHS of these inequalities and the $\mathcal{OO}$ sets are always on the RHS). To do this, we will examine users by pairs – we will take a pair of users and see how many times they appear in each other's sets in these inequalities. In Claim 1 below, we show that they collectively appear the same number of times in the LHS sets and in the RHS sets. We then enumerate all user pairs. What this result says is that these users collectively contribute to an equal number of times to the LHS and RHS of the set of inequalities given in Eqn. (4.12). Adding up all these inequalities, this translates to the fact that the total size of the sets on the LHS and those on the RHS must be equal. This however contradicts the strict inequality, thus completing the proof. □

Claim 1. *Consider a pair of users A and B in an improvement updating loop, and*

consider how they are perceived in each other's set. Then A and B collectively appear the same number of times in the LHS sets (the $\mathcal{SS}$ sets) and in the RHS sets (the $\mathcal{OO}$ sets).

Proof. First note that A and B have to be in each other's interference set for them to appear in each other's $\mathcal{SS}$ and $\mathcal{OO}$ sets. Since we are only looking at two users and how they appear in each other's sets, without loss of generality we can limit our attention to a subsequence of the original updating sequence involving only A and B, given by

$$U_{AB} = \{u(t_1), u(t_2), \cdots, u(t_l)\} \tag{4.13}$$

where $u(t_j) \in \{A, B\}$, $t_j \in \{1, 2, \cdots, T\}$, and l is the length of this subsequence, i.e., the total number of updates between A and B. As before, this subsequence can also be represented clockwise along a circle.

It helps to consider an example of such a sequence, say, $ABAABBABAA$, also shown in Figure 4.2. In what follows we will express an odd train as the odd number of consecutive changes of one user sandwiched between the other user's changes, e.g., the odd train ABA in the above subsequence. To avoid ambiguity, we will further write this sequence as $A_1B_2A_3A_4B_5B_6A_7B_8A_9A_{10}$.

A few things to note about such a sequence:

1. Since the starting and ending states are the same, each user must appear an even number of times in the sequence. Since each user appears an even number of times, there must be an even number of odd trains along the circle for any user.

2. A user (say A) only appears in the other's (say B's) $\mathcal{SS}$ or $\mathcal{OO}$ sets if it has an odd train between the other user's two successive appearances. This means

that there is an even number of relevant inequalities where A appears in B's inequalities (either on the LHS or the RHS), and vice versa.

3. Consider the collection of all relevant inequalities discussed above, one for each odd train, in the order of their appearance on the circle (all four such inequalities are illustrated in Figure 4.2). Then A and B contribute to each other's inequalities on alternating sides along this updating sequence/circle. That is, suppose the first inequality is A's and B goes into its LHS, then in the next inequality (could be either A's or B's) the contribution (either A to B's inequality or B to A's inequality) is on the RHS. Take our running example, for instance, the first inequality is due to the odd train marked by the sequence $A_1B_2A_3$, and the second $B_6A_7B_8$. Suppose A and B start with different colors, then in the first inequality, B appears in the RHS; in the second, A appears in the LHS.

We now explain why the third point above is true. The reason is because for one user (B) to appear in the other's (A's) LHS, they must start by having the same color and again have the same color right before A's second change (see e.g. the subsequence $A_1B_2A_3$ in the running example). Until the next odd train ($B_6A_7B_8$), both will make an even number of changes including A's second change ($A_3A_4B_5B_6$). The next inequality belongs to the user who makes the last change before the next odd train (B) . As perceived by this user (B) right before this change, the two must now have different colors. This is because as just stated A will have made an even number of changes from the last time they are of the same color (by the end of A_1B_2), while B is exactly one change away from an even number of changes (by the end of $A_1B_2A_3A_4B_5$). Therefore, the contribution from the other user (A) to this inequality must be to the RHS.

To summarize, one can see that essentially the color relationship between A and

B reverses upon each update, and there is an odd number of updates between the starting points of two consecutive odd trains (e.g., 5 updates between A_1 and B_6, or 1 update between B_6 and A_7) so the color relationship flips for each inequality in sequence.

The above argument establishes that as we go down the list of inequalities and count the size of the sets on the LHS vs. that on the RHS, we alternate between the two sides. Since there are exactly even number of such inequalities, we have established that A and B collectively appear the same number of times in the LHS sets and in the RHS sets. □

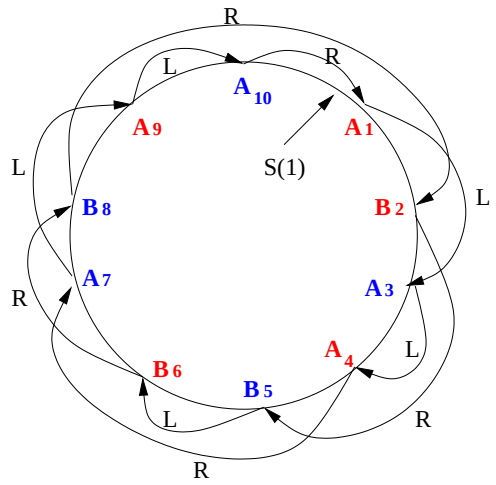

Figure 4.2: Example of an updating sequence "$A_1B_2A_3A_4B_5B_6A_7B_8A_9A_{10}$" illustrated on a circle. The color coding denotes the color of a user right before the indicated change. Each arrow connecting two successive changes by the same user induces an inequality perceived by this user. The labels "L" and "R" on an arrow indicate to which side of this inequality (LHS and RHS respectively) the other user contributes to. As can be seen the labels alternates in each subsequent inequality.

4.4.2 Counter-Example for 3 Resources

The above theorem establishes that when there are only two resources, the FIP property holds, and consequently an NE exists. This holds for the general case of user-specific payoff functions. Below we show a counter-example that the FIP property does not necessarily hold for 3 resources/colors or more.

time step	A	B	C	D
0	b	p	p	b
1	b → r			
2		p → r		
3				b → r
4			p → r	
5	r → p			
6				r → b
7		r → b		
8			r → b	
9	p → b			
10			b → p	
11		b → p		

Table 4.1: 3-color counter example.

Example 4.4. Suppose we have three colors to assign, denoted by r (red), p (purple), and b (blue). Consider a network topology shown in Figure 4.3, where we will primarily focus on nodes A, B, C and D. In addition to node C, node A is also connected to A_r, A_p and A_b nodes of colors red, green and blue, respectively. B_r, B_p, B_b, C_r, C_p, C_b, and D_r, D_p, D_b and similarly defined and illustrated in Figure 4.3. Note that these sets may not be disjoint, e.g., a single node may contribute to both A_r and B_r, and so on.

Consider now the following sequence of improvement updates involving only nodes A, B, C, and D, i.e., within this sequence none of the other nodes change color (note that this is possible in an asynchronous improvement path), where the notation $s_1 \to s_2$ denotes a color change from s_1 to s_2 and at time 0 the initial color assignment is given.

We see that this sequence of color changes form a loop, i.e., all nodes return to the same color they had when the loop started. For this to be an improvement loop such that each color change results in improved payoff, it suffices for the following

sets of conditions to hold (here we assume all users have the same payoff function and have suppressed the superscript i in $g_r^i()$, and the notation "$>_k$" denotes that the improvement occurs at time k):

$$g_r(A_r+1) >_1 g_b(A_b+1) > g_b(A_b+2) >_9 g_p(A_p+1) >_5 g_r(A_r+2) \ ;$$

$$g_r(B_r+1) >_2 g_p(B_p+2) >_{11} g_b(B_b+1) >_7 g_r(B_r+2) \ ;$$

$$g_b(C_b+3) >_8 g_r(C_r+1) > g_r(C_r+4) >_4 g_p(C_p+1) >_{10} g_b(C_b+4) \ ;$$

$$g_r(D_r+1) >_3 g_b(D_b+1) >_6 g_r(D_r+2)$$

It is straightforward to verify the sufficiency of these conditions by following a node's sequence of changes. To complete this counter example, it remains to show that the above set of inequalities are feasible given appropriate choices of A_x, B_x, C_x and D_x, $x \in \{r, p, b\}$. There are many such choices; one example is $A_x = 5, B_x = 3, C_x = 7, D_x = 1$, for all $x \in \{r, p, b\}$. With such a choice, and substituting them into the earlier set of inequalities and through proper reordering, we obtain the following single chain of inequalities:

$$\begin{aligned} & g_r(2) > g_b(2) > g_r(3) > g_r(4) > g_p(5) > g_b(4) > g_r(5) > g_r(6) \\ > \ & g_b(6) > g_b(7) > g_p(6) > g_r(7) > g_b(10) > g_r(8) > g_r(11) > g_p(8) > g_b(11) \end{aligned}$$

It should be obvious that this chain of inequalities can be easily satisfied by the right choices of non-increasing payoff functions.

It is easy to see how if we have more than 3 colors, this loop will still be an improving loop as long as the above inequalities hold. This means that for 3 colors or more the FIP property does not hold in general. Note that the updates in this example are not always best response updates.

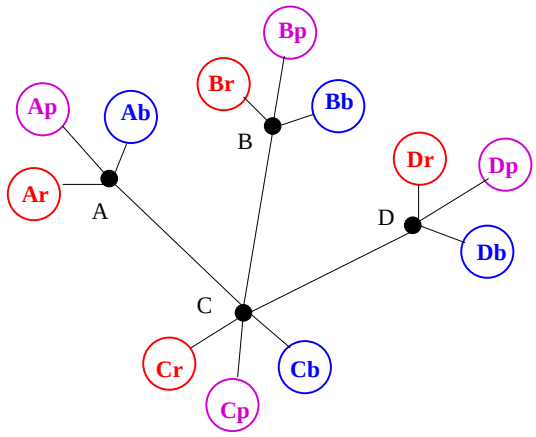

Figure 4.3: A counter example of 3 colors: nodes A, B, C, and D are connected as shown; in addition, node W, $W \in \{A, B, C, D\}$, is connected to W_x other nodes of color $x \in \{r, p, b\}$ as shown.

4.5 Existence of a Pure Strategy Nash Equilibrium

In this section we examine what graph properties will guarantee the existence of a NE. Specifically, we show that for a network congestion game defined on graphs that are (1) complete, (2) in the form of a tree, or (3) in the form of a loop, a pure strategy NE always exists with user-specific payoff functions that are non-increasing in the number of interfering users. Below we present these results in sequence. We will also give counter examples to show that a pure strategy NE does not generally exist in such a game when the network graph is directed or when the payoff functions are non-monotonic. In addition to the results presented here, we believe a pure strategy NE exists in a general undirected graph (i.e., Theorem 4.8 holds for a general undirected graph, not just trees). Unfortunately, a formal proof remains elusive to this point. We continue to pursue this in our on-going research.

4.5.1 Existence of pure strategy NE on an undirected graph

Theorem 4.5. *When the graph is complete, a NE always exists for the network congestion game defined on this graph.*

This theorem is trivially true. It is simply a direct consequence of known results on the standard CG: in a complete graph every node is every other node's neighbor, therefore a NCG reduces to the original CG, thus this result. Furthermore, for the same reason when the graph is complete the FIP property also holds.

We next show that a pure strategy NE exists when the underlying network graph is given by a tree. Let Γ_N denote a network congestion game with N players; G_N the underlying N-player network, where players are indexed by $1, 2, ..., N$ and the payoff functions $g_r^i(n_r^i)$ are nonincreasing. Recall that $n_r^i(\boldsymbol{\sigma})$ denotes the number of neighbors of user i (excluding i) playing strategy r.

Lemma 4.6. *If every network congestion game with N players (Γ_N) and user specific non-increasing payoff functions has at least one pure strategy NE, then the $(N+1)$-player network congestion game (denoted as Γ_{N+1}) formed by connecting a new player with index $N+1$ to a single player in the N-player network G_N, has at least one pure strategy NE.*

Proof. By assumption Γ_N has a pure strategy NE denoted by $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_N\}$. Suppose Γ_N is in such an NE. Now connect player $N+1$ to a single player in G_N. Call this player j and the resulting network G_{N+1}. This is illustrated in Figure 4.4. Let player $N+1$ select its best response strategy:

$$\sigma_{N+1} = r_o = \text{argmax}_{r \in \mathcal{R}} g_r^{N+1}(n_r^{N+1}(\boldsymbol{\sigma}) + 1),$$

where n_r^{N+1} is defined on the extended network G_{N+1}, and takes on the value of 1 or 0 depending on whether player j selects strategy r or not. We now consider three cases depending on j's strategy change in response to the change in network from G_N to G_{N+1}.

Case 1: $\sigma_j \neq r_o$. In this case $N+1$ selected a resource different from j's, so j

has no incentive to change its strategy in response to the addition of player $N+1$. In turn player $N+1$ will remain in r_o as this is its best response, and no other players are affected by this single-link network extension. Thus the strategy profile $(\sigma_1, \cdots, \sigma_N, r_o)$ is a pure strategy NE for the game Γ_{N+1}.

Case 2: $\sigma_j = \sigma_{N+1} = r_o$, and player j's best response to the change from G_N to G_{N+1} with player $N+1$ selecting r_o remains $\sigma_j = r_o$. That is, even with the additional interfering neighbor $N+1$, the best choice for j remains r_o. In this case again we reach a pure strategy NE for the game Γ_{N+1} with the same argument as in Case 1.

Case 3: $\sigma_j = \sigma_{N+1} = r_o$, and player j's best response to this change is to move away from strategy r_o. In this case more players may in turn change strategies. Suppose we hold player $(N+1)$'s strategy fixed at r_o. Consider now a new network congestion game $\bar{\Gamma}_N$, defined on the original network G_N, but with the following modified payoff functions for $r \in \mathcal{R}$ and $i \in \mathcal{I}$:

$$\bar{g}_r^i(n_r^i + 1) = \begin{cases} g_r^i(n_r^i + 2) & \text{if } i = j, r = r_o \\ g_r^i(n_r^i + 1) & \text{otherwise} \end{cases}.$$

In words, the game $\bar{\Gamma}_N$ is almost the same as the original game Γ_N, the only difference being that the addition of player $(N+1)$ and its strategy r_o is built into player j's modified payoff function.

By assumption of the lemma, this game with N players has a pure strategy NE and denote that by $\bar{\boldsymbol{\sigma}}$. Suppose $\bar{\boldsymbol{\sigma}}$ is reached in the network G_N with player $(N+1)$ fixed at $\sigma_{N+1} = r_o$. If we have $\bar{\sigma}_j = r_o$, then obviously player $(N+1)$ has no incentive to change its strategy because as far as it is concerned its environment has not changed at all. In turn no player in G_N will change its strategy because they are already in an NE with $(N+1)$ held at r_o. If $\bar{\sigma}_j \neq r_o$, then player $(N+1)$ has

even less incentive to change its strategy because its payoff for using r_o is no worse than before since j moved away with payoff functions being non-increasing, and at the same time its payoff for using any other resource is no better. Again r_o is player $(N+1)$'s best response.

In either case a new NE, the strategy profile $(\bar{\boldsymbol{\sigma}}, r_o)$, is reached for the game Γ_{N+1}. □

Remark 4.7. Note that in the above lemma, the network G_N itself does not have to be a tree. The lemma states that as long as a NE exists for one network, then by adding one more node through a single link, an NE exists in the new network.

Theorem 4.8. *A network congestion game defined over a* tree *network with non-increasing player-specific payoff functions has at least one pure strategy NE.*

Proof. The proof is easily obtained by noting that any tree can be constructed by starting from a single node and adding one node (connected through a single link) at a time. Formally, we prove this by induction on N. Start with a single player indexed by 1. This game has a pure strategy NE, in which the player selects $\sigma_1 = \text{argmax}_{r\in\mathcal{R}} g_r^1(1)$ for any payoff functions. Assume that any N-player game Γ_N over a tree G_N with any set of non-increasing payoff functions has at least one pure strategy NE. Any tree G_{N+1} may be constructed by adding one more leaf node to some other tree G_N by connecting it to only one of the players in G_N. Lemma 1 guarantees that such a formation will result in a game with at least one pure strategy NE. □

Theorem 4.9. *If the network is in the form of a loop and user payoff functions are identical for a given resource, then there always exists a pure strategy Nash equilibrium, involving no more than 3 resources/colors.*

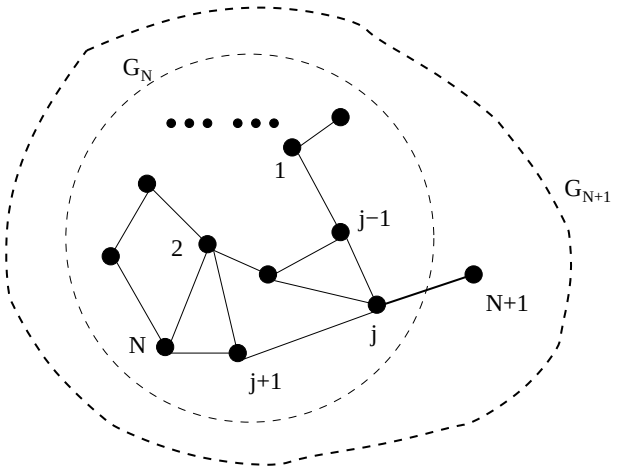

Figure 4.4: Adding one more player to the network G_N with a single link.

Proof. We know from Theorem 4.3 that when there are only 2 colors an NE always exists. Assume there are at least 3 colors to choose from. As payoff functions are non-user specific, we will suppress the superscript i in the function $g_r^i()$. Note that there always exist three colors r, b, p that have the highest single-user occupancy payoff values; suppose we have

$$g_r(1) \geq g_b(1) \geq g_p(1)$$

If the loop has an even number of nodes then compare $g_r(3)$ with $g_b(1)$. If $g_r(3) \geq g_b(1)$, then assigning r to all nodes will result in an NE; if $g_b(1) \geq g_r(3)$ then assigning r and b alternately will result in an NE.

Now consider the case where the loop has an odd number of nodes, labeled from 1 to $2n+1$, where node i is connected to node $i+1$ and node $2n+1$ is connected to node 1. Again we see that if $g_r(3) \geq g_b(1)$ then assigning r to all nodes results in an NE.

Assume now $g_b(1) \geq g_r(3)$ and consider the following assignment. Assign r and b alternately to nodes from 2 up to $2n$, so that nodes 2 and $2n$ are both colored r. It remains to determine the coloring of nodes 1 and $2n+1$. We have the following

User 3 /User 1,2	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
1	5, 5, 3	5, 4, 5	4, 5, 5	4, 4, 2
2	2, 2, 4	2, 6, 6	6, 2, 6	6, 6, 1

Table 4.2: 3-color counter example.

four cases (under the condition $g_b(1) \geq g_r(3)$): (1) $g_b(2) \geq g_r(2)$ and $g_b(2) \geq g_p(1)$: in this case $(b,\ b)$ assignment to nodes 1 and $2n+1$ will result in an overall NE. (2) $g_b(2) \geq g_r(2)$ and $g_b(2) < g_p(1)$: in this case either $(b,\ p)$ or $(p,\ b)$ for nodes 1 and $2n+1$ will result in an overall NE. (3) $g_b(2) < g_r(2)$ and $g_r(2) \geq g_p(1)$: in this case either $(b,\ r)$ or $(r,\ b)$ for node 1 and $2n+1$ will result in an overall NE. (4) $g_b(2) < g_r(2)$ and $g_r(2) < g_p(1)$: in this case either $(b,\ p)$ or $(p,\ b)$ for nodes 1 and $2n+1$ will result in an overall NE.

Therefore in all cases we have shown an NE exists. □

4.5.2 Counter Example of Non-monotonic Payoff Functions

Below we show that a pure strategy NE may not exist when the network graph is undirected but the payoff function is non-monotonic, even when they are non-user specific.

Example 4.10. Consider a 3-user, 2-channel network given in Figure 4.5. The payoff functions have the following property

$$g_2(2) > g_1(2) > g_2(1) > g_1(3) > g_1(1) > g_2(3)\ .$$

One example of this is when $g_1(1) = 2, g_1(2) = 5, g_1(3) = 3, g_2(1) = 4, g_2(2) = 6, g_2(3) = 1$. The game matrix corresponding to these payoff functions are given below. It is easy to verify that there exists no pure strategy NE.

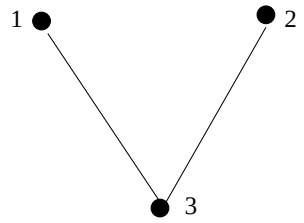

Figure 4.5: Counter example of non-monotonic payoff functions

4.5.3 Counter Example of a Directed Graph

Below we show that a pure strategy NE may not exist when the network graph is directed.

Example 4.11. Consider a 4-user, 3-channel network whose graph is given in Figure 4.6. It can be shown that a pure strategy NE does not exist when the payoff functions are non-increasing and have the following property.

$$g_3(1) > g_2(1) > g_2(2) > g_3(2) > g_1(1) > g_1(2) > g_2(3) > g_1(3) > g_2(4) > g_1(4) > g_3(3) > g_3(4) \ .$$

We do not include an example game matrix for brevity. We invite an interested reader to verify this example.

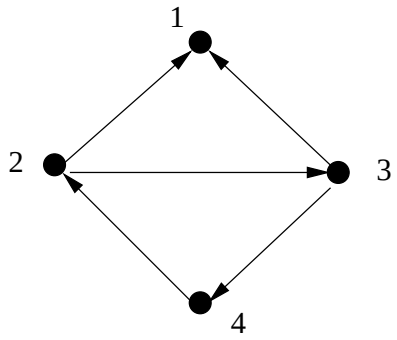

Figure 4.6: Counter-example for directed graphs

4.6 Sufficient Conditions on User Payoff Functions

In this section we examine what properties on the user payoff functions will guarantee the existence of an NE. Specifically, we show that for general network graphs,

an NE always exists if (1) there is one resource with a dominating payoff function (much larger than the others), or (2) different resources present the same type of payoff for users. Moreover, in the case of (2) the game has the FIP property. We note that case (2) is of particular practical interest and relevance, as this case in the context of spectrum sharing translates to evenly dividing a spectrum band into sub-bands, each providing users with the same bandwidth and data rate. Below we present and prove these results.

Theorem 4.12. *For a general network graph, if there exists a resource r and its payoff function is such that $g_r^i(K_d+1) \geq g_s^i(1)$, where $K_d = \max\{|\mathcal{K}_i|, i = 1, 2, \cdots, N\}$, for all $s \in 1, 2, \cdots, R$ and all $i \in \mathcal{I}$, then a Nash Equilibrium exists.*

Here K_d is the maximum node degree in the network, i.e., the maximum possible number of users sharing the same resource. In words, this theorem says that if there exists a resource whose payoff "dominates" all other resources, then an NE exists. This is a rather trivial result; an obvious NE is when all users share the dominating resource.

Theorem 4.13. *For a general undirected network graph, if all resources have identical non-increasing payoff functions for any given user, i.e., for all $r \in \mathcal{R}$ and $i \in \mathcal{I}$, we have $g_r^i(n) = g^i(n)$ for $n = 1, 2, \cdots, N$ and some non-increasing function $g^i(\cdot)$, then there exists a Nash Equilibrium, and the game has the finite improvement property. Note that the payoffs can remain user-specific.*

Proof. We prove this theorem by using a potential function argument.

Recall that user i's total payoff under the strategy profile $\boldsymbol{\sigma}$ is given by (here we have suppressed the subscript r since all resources are identical):

$$g^i(\boldsymbol{\sigma}) = g(n^i(\boldsymbol{\sigma}) + 1),\ n^i(\boldsymbol{\sigma}) = |\{j : \sigma_j = \sigma_i, j \in \mathcal{K}_i\}| \tag{4.14}$$

where $\sigma_i \in \mathcal{R}$ since we have limited our attention to the case where each user can select only one resource at a time.

Now consider the following function defined on the strategy profile space:

$$\phi(\boldsymbol{\sigma}) = \sum_{i,j \in \mathcal{K}} 1(i \in \mathcal{K}_j) 1(\sigma_i = \sigma_j) = \frac{1}{2} \sum_{i \in \mathcal{K}} n^i(\boldsymbol{\sigma}) \ , \tag{4.15}$$

where the indicator function $1(A) = 1$ if A is true and 0 otherwise. For a particular strategy profile $\boldsymbol{\sigma}$ this function ϕ is the sum of all pairs of users that are connected (neighbors of each other) and have chosen the same resource under this strategy profile. Viewed in a graph, this function is the total number of edges connecting nodes with the same color.

We see that every time user i improves its payoff by switching from strategy σ_i to $\sigma_i^{'}$, and reducing $n^i(\boldsymbol{\sigma}^{-i}, \sigma_i)$ to $n^i(\boldsymbol{\sigma}^{-i}, \sigma_i^{'})$ (g^i is a non-increasing function), the value of $\phi()$ strictly decreases accordingly[7]. As this function is bounded from below, this means that in this case the game has the FIP property so this process eventually converges to a fixed point which is a Nash Equilibrium. □

4.7 Conclusion

In this chapter we have considered an extension to the classical definition of congestion games by allowing resources to be reused among non-interfering users. This is a much more appropriate model to use in the context of wireless networks and spectrum sharing where due to decay of wireless signals over a distance, spatial reuse is frequently exploited to increase spectrum utilization.

The resulting game, congestion game with resource reuse, is a generalization to the original congestion game. We have shown that when there are only two resources

[7] It's easily seen that a non-increasing function $G(\sum_{i,j \in \mathcal{K}} 1(i \in \mathcal{K}_j) 1(\sigma_i = \sigma_j))$ is an ordinal potential function of this game as its value improves each time a user's individual payoff is improved thereby decreasing the value of its argument.

and users can only use one at a time, then the game has the finite improvement property; the same is shown to be false in general when there are three or more resources. We further showed a number of conditions on the network graph as well as the user payoff functions under which the game has an NE. Perhaps most relevant to spectrum sharing is the result that when all resources present the same payoff to users (e.g., all channels are of the same bandwidth and data rate for all users), then the game has the finite improvement property and an NE exists.

CHAPTER 5

Conclusion and Future Work

5.1 Conclusion and Future Work

In this dissertation, we studied a few problems regarding resource allocation problems within the contexts of game theory and stochastic control. These problems were motivated by applications within communication networks. We now summarize the main results before providing some potential future research on these problems.

For problem (OP) analyzed in Chapter 2 , we considered the case that the probability of each channel being in an idle state is known a priori to the secondary user. We formulated a dynamic programming problem and we searched for stochastic control techniques which would help us to solve the problem. The goal was to maximize expected reward during the horizon whether infinite or finite. Our strategy was fairly robust in case that the a priori distribution is not exact. We proved that in case that channels are independent , totally equal two-state Markov models with transition probabilities independent of time and at the same time bursty, the optimal policy for the secondary user is to probe the channels in a greedy way meaning at each step picking the arm which has the highest probability of being idle (highest instantaneous expected return). This result largely improves on the previous results where results were established for the cases that either there are only two arms or

if there is a tighter condition than arms being bursty imposed on the problem. We also found a counter-example to show that in the case that arms are not bursty, the greedy policy is not optimal effectively solving a problem which was open for a while.

In Chapter 3, we examined problem (MC-OP) which is a direct generalization to the first problem. In this problem we assumed the controller can pick exactly a fixed number of arms to play at every time moment. We studied finding optimal policy and strategies which maximize expected reward in the specific horizon subject to the constraint of selecting a specific number of arms at a time.

For the network congestion game problem in Chapter 4, we examined a problem which was a generalization to congestions games where several players in a game have to choose their resource in such a way that to maximize their reward. The reward function is a utility function defined for each user , and is decreasing in the number of users who have been using that same resource. Many problems can be studied on such games, the most significant of them is to study whether or not there always exists a Nash Equilibrium on these graphs. We proved in certain scenarios, specially when the number of resources is only 2 , this game always has a Nash Equilibrium point. For certain geometries and restrictions on the utility functions we also succeeded in proving existence of Nash equilibrium points. The proof was based on sequential change of colors in the nodes so that the graph will eventually be colored in a Nash equilibrium state.

5.1.1 Future Work

Here we discuss some potential future work on the two main problems that were examined in this thesis.

For problems (OP) and (MC-OP), some open problems include the following:

- Given that the optimality results were proven for the case that channels are independent, equal and bursty, many important questions remain here. One of the most important and natural questions to be addressed is how would the optimal policy change when the channels are not bursty. This case is a very realistic and important case which its study will pave the way for future progress on the solution to the general case of multi-armed restless bandit problem, a problem which has been proved to be PSPACE-hard while at the same time it has always been a subject of attention to the stochastic control community.

- Another exciting possible line of research for future is a special case of imperfect sensing for secondary user. Zhao and Krishnamachari consider the channels to be simple gilbert-elliot channels and also study a more specific problem where there are only 2 channels available to secondary user. In this problem which we call it "Sensing with Side Information" they consider absolutely no penalty for transferring through a busy channel interfering with primary user so the secondary user , would prefer to transfer through the channel he probes no matter the outcome of the probation shows the channel to be busy or available. In their work , they have proved subject to certain conditions on probability of false alarm and also missed detection , one can prove greedy policy (myopic) in choosing channels (with respect to probability of their availability) to be optimal. It is an interesting research to see if in this scenario , when we have multiple channels , myopic policy will still be optimal.

  There will be two exciting problems to be targeted here. First one is that in such case , what would be the optimal policy for probing and transmitting for secondary user. The complication is that at each time slot , not only he has to choose a channel to probe , but depending on the outcome of probing

and the current state of prob- ability vector (again it is a well-known fact that probability of occupancy vector is sufficient statistic for the secondary user to decide on which channel to probe), he has to choose between transmitting signal and taking the risk of an interference penalty or rather skipping transmitting through that slot and also giving up on the potential award he could have got if the channel was available and he transmitted through it.

- the Other exciting problem is to assume secondary user selects one channel and probes it in the beginning of each time slot and will transmit signal if and only if the channel is sensed to be available. This assumption reduces complexity of the problem as secondary user does not need to choose whether he needs to transmit or not. We consider a certain cost according to interfering with primary user

- Another exciting case is when secondary user tries to transmit a message while his message arrival is bursty meaning arrival of information packets at each time moment happens subject to a certain distribution. The main question here is that in case that there is certain penalty/cost for probing, what would his optimal policy for probing be. This means depending on the occupancy vector of channels which represents the probability of each channel being available at that time moment, what his optimal probing policy would be which consists of: 1)when to choose a channel to probe and 2)In case that he chooses to probe a channel which channel should he select to probe in order to maximize his expected long-run return. The idea for this problem is totally new and given that there is no previous work on this problem we believe it poses a host of rich properties to be explored. In reality , it would be naive to consider probing a

channel being free. At the very least , every time a secondary user probes a channel , there is a cost associated with tuning to that channel and also the time wasted to make sure the channel is available or not. When the secondary user does not have a message waiting he has the choice of choosing a channel and probing it , which will provide him with further information regarding the status of that channel in the future time slots as well as a cost associated with this probing and also a choice of not probing any channel which would not cost him anything at all and as well it provides him with no further information regarding occupancy probability in any of the channels. On the other hand , even if he has a message waiting , he might opt not to probe any channel as the channels might have such little occupancy probability that cost of probing one of them would not be justified with the respective chance to transfer a message and knowledge acquired regarding further probability of occupancy for that probed channel.

Potential future work on the network congestion game problem follows:

- An exciting possibility is to find out existence of Nash equilibrium in more general cases where there are more than two resources available for the users. While our intelligent method for proving existence of Nash equilibrium was creative , it is not applicable to the case that there are more than two channels as FIP property does not hold without major change for the case of more than two resources available.

- Another exciting possibility for research on this problem is to consider different geometries which existence of Nash equilibrium could be proven and at the same time , it's existence is useful. While we focused on specific geometries like

graphs in the form of a simple loop or a tree , there are more diverse geometries to be studied which arise from real life situations. Rather than studying each of them as a special and isolated case , creating a smart method that addresses multiple scenarios simultaneously is of great importance.

- Another important question is to study quality of Nash equilibrium point and how good it is indeed for the players of the game. While we worked on establishing results regarding existence of Nash equilibrium , finding a measure that describes how good an equilibrium is and assessing different equilibrium points with respect to this measure are of great importance for other applications of this problem.

APPENDICES

APPENDIX A

Appendix for Chapter 2

Proof of Lemma 2.6:

The LHS of the inequality represents the expected reward of a policy (referred to as L below) that probes in the sequence of channels 1 followed by n, $n-1$, $\cdots$, and then 1 again, and so on, plus an extra reward of 1; the RHS represents the expected reward of a policy (referred to as R below) that probes in the sequence of channels n followed by $n-1$, $\cdots$, and 1 and then n again, and so on. It helps to imagine lining up the n channels along a circle in the sequence of n, $n-1$, $\cdots$, 1, clock-wise, and thus L's starting position is 1, R's starting position is n, exactly one spot ahead of L clock-wise. Each will cycle around the circle till time T.

Now for any realization of the channel conditions (or any sample path of the system), consider the sequence of "0"s and "1"s that these two policies see, and consider the position they are on the circle. The reward a policy gets along a given sample path is $R_l = \sum_{j=t}^{T} \beta^j I_{h(j)=1}$ for policy L, where $I_{h^l(j)=1} = 1$ if L sees a "1" at time j, and 0 otherwise; the reward for R is $R_r = \sum_{j=t}^{T} \beta^j I_{h^r(j)=1}$ with $I_{h^r(j)=1}$ similarly defined. There are two cases.

Case (1): the two eventually catch up with each other at some time $K \leq T$, i.e., at some point they start probing exactly the same channel. From this point on the two policies behave exactly the same way along the same sample path, and the

reward they obtain from this point on is exactly the same. Therefore in this case we only need to compare the rewards (L has an extra 1) leading up to this point.

Case (2): The two never manage to meet within the horizon T. In this case we need to compare the rewards for the entire horizon (from t to T).

We will consider Case (1) first. There are only two possibilities for the two policies to meet: (Case 1.a) either L has seen exactly one more "0" than R in its sequence, or (Case 1.b) R has seen exactly $n-1$ more "0"s than L. This is because the moment we see a "0" we will move to the next channel on the circle. L is only one position behind R, so one more "0" will put it at exactly the same position as R. The same with R moving $n-1$ more positions ahead to catch up with L.

Case (1.a): L sees exactly one more "0" than R in its sequence. The extra "0" necessarily occurs at exactly time K, $t \leq K \leq T$, meaning that at K, L sees a "0" and R sees a "1". From t to K, if we write the sequence of rewards (zeros and ones) under L and R, we observe the following: between t and K both L and R have equal number of zeros, while for $\forall t' = t, t+1, \ldots, K-1$, the number of zeros up to time t' is less (or no more) for L than for R. In other words, L and R see the same number of "0"s, but L's is always lagging behind (or no earlier). That is, for every "0" R sees, L has a matching "0" that occurs no earlier than R's "0." This means that if we denote by $R_l(t_1, t_2)$ the rewards accumulated between t_1 and t_2, then for the rewards in $[t, K-1]$, we have $R_l(t, t') \geq R_r(t, t')$, for $\forall t' \leq K-1$, while $R_l(K, K) = \beta^K$ and $R_r(K, K) = 0$. Finally by definition we have $R_l(K+1, T) = R_r(K+1, T)$. Therefore overall we have $1 + R_l(t, T) \geq R_r(t, T)$, proving the above inequality.

Case (1.b): R sees $n-1$ more "0"s than L does. The comparison is simpler. We only need to note that R's "0"s must again precedes (or be no later than) L's since otherwise we will return to Case (1.a). Therefore we have $R_l \geq R_r$, and thus

$1 + R_l \geq R_r$ is also true.

We now consider Case (2). The argument is essentially the same. In this case the two don't get to meet, but they are on their way, meaning that either L has exactly the same "0"s as R and their positions are no earlier (corresponding to Case (1.a)), or R has more "0"s than L (but not up to $n - 1$) and their positions are no later than L's (corresponding to Case (1.b)). So either way we have $1 + R_l \geq R_r$.

The proof is thus complete.

APPENDIX B

Appendix for Chapter 3

Proof of Lemma 3: We would like to show

$$(A): \quad 1 + W_t^k(\omega_2, \cdots, \omega_n, \omega_1) \geq W_t^k(\omega_1, \cdots, \omega_n)$$

$$(B): \quad W_t^k(\omega_1, \cdots, \omega_j, y, x, \omega_{j+3}, \cdots, \omega_n) \geq$$

$$W_t^k(\omega_1, \cdots, x, y, \omega_{j+3}, \cdots, \omega_n),$$

where $x \geq y$, $0 \leq j \leq n-2$, and $j = 0$ implies $W_t^k(y, x, \omega_3, \cdots, \omega_n) \geq W_t^k(x, y, \omega_3, \cdots, \omega_n)$.

The two inequalities (A) and (B) will be shown together using an induction on t. For $t = T$, part (A) is true because $LHS = 1 + \omega_1 + \sum_{i=n-k+2}^{n} \omega_i \geq \omega_{n-k+1} + \sum_{i=n-k+2}^{n} \omega_i = RHS$. Part (B) is obviously true for $t = T$ since $x \geq y$.

Suppose (A) and (B) are both true for $t+1, \cdots, T$. Consider time t, and we will prove (A) first. Note that in the next step, channel 1 is selected by the action on the LHS of (A) but not by the RHS, while channel $n-k+1$ is selected by the RHS of (A) but not by the LHS. Other than this difference both sides select the same set of channels indexed $n-k+2, \cdots, n$. We now consider four possible cases in terms of the realizations of channels 1 and $n-k+1$.

Case (A.1): channels 1 and $n-k+1$ have the state realizations "0" and "1", respectively.

We will use a sample-path argument. Note that while these two channels are not both observed by either side, the realizations hold for the underlying sample path regardless. In particular, even though the LHS does not select channel $n-k+1$ and therefore does not get to actually observe the realization of "1", the fact remains that channel $n-k+1$ is indeed in state 1 under this realization, and therefore its future expected reward must reflect this. It follows that under this realization channel $n-k+1$ will have probability p_{11} for the next time step even though we did not get to observe the state 1. The same is true for the RHS. This argument applies to the other three cases and is thus not repeated.

Conditioned on this realization, the LHS and RHS are evaluated as follows (denoted as $\{LHS|_{(0,1)}\}$ and $\{RHS|_{(0,1)}\}$, respectively):

$$\{LHS|_{(0,1)}\} = 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n)\cdot$$
$$W^k_{t+1}(p_{01}[k-\sum l_i], \tau(\omega_2),\cdots,\ \tau(\omega_{n-k+1})=p_{11}, p_{11}[\sum l_i])\ ;$$

$$\{RHS|_{(0,1)}\} = 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n)\cdot$$
$$W^k_{t+1}(p_{01}[k-\sum l_i - 1], \tau(\omega_1)=p_{00},\ \tau(\omega_2),\cdots,\tau(\omega_{n-k}), p_{11}[\sum l_i + 1])$$
$$= \{LHS|_{(0,1)}\}$$

Case (A.2): channels 1 and $n-1+1$ have the state realizations "1" and "1", respectively.

$$\{LHS|_{(1,1)}\} = 1 + 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n)\cdot$$
$$W^k_{t+1}(p_{01}[k-\sum l_i - 1], \tau(\omega_2),\cdots,\ \tau(\omega_{n-k+1})=p_{11}, p_{11}[\sum l_i + 1])\ ;$$

$$
\begin{aligned}
& \{RHS|_{(1,1)}\} = 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i - 1], \tau(\omega_1) = p_{11},\ \tau(\omega_2),\cdots,\tau(\omega_{n-k}), p_{11}[\sum l_i + 1]) \\
\le\ & 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i - 1], \tau(\omega_2),\cdots,\tau(\omega_{n-k}),\ p_{11}, p_{11}[\sum l_i + 1]) \\
=\ & \{LHS|_{(1,1)}\} - 1 \le \{LHS|_{(1,1)}\}
\end{aligned}
$$

where the first inequality is due to the induction hypothesis of (B).

Case (A.3): channels 1 and $n-1+1$ have the state realizations "0" and "0", respectively.

$$
\begin{aligned}
& \{RHS|_{(0,0)}\} = \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_1) = p_{01}, \tau(\omega_2),\cdots,\tau(\omega_{n-k}),\ p_{11}[\sum l_i])\ ;
\end{aligned}
$$

$$
\begin{aligned}
& \{LHS|_{(0,0)}\} = 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_2),\cdots,\tau(\omega_{n-k}),\ \tau(\omega_{n-k+1}) = p_{01}, p_{11}[\sum l_i]) \\
\ge\ & 1 + \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_2),\cdots,\tau(\omega_{n-k}),\ p_{11}[\sum l_i], p_{01}) \\
\ge\ & \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& \Big(1 + W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_2),\cdots,\tau(\omega_{n-k}),\quad p_{11}[\sum l_i], p_{01})\Big) \\
\ge\ & \sum_{n-k+2\le i\le n} \omega_i + \beta \cdot \sum_{l_{n-k+2},\cdots,l_n\in\{0,1\}} q(l_{n-k+2},\cdots,l_n) \cdot \\
& W^k_{t+1}(p_{01}, p_{01}[k - \sum l_i], \tau(\omega_2),\cdots,\tau(\omega_{n-k}), p_{11}[\sum l_i]) = \{RHS|_{(0,0)}\}
\end{aligned}
$$

where the first inequality is due to the induction hypothesis of (B), the last inequality due to the induction hypothesis of (A). Also, the second inequality utilizes the total probability over the distribution $q(l_{n-k+2}, \cdots, l_n)$ and the fact that $\beta \leq 1$.

Case (A.4): channels 1 and $n-1+1$ have the state realizations "1" and "0", respectively.

$$\begin{aligned}
\{RHS|_{(1,0)}\} = & \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_1) = p_{11}, \tau(\omega_2), \cdots, \ \tau(\omega_{n-k}), p_{11}[\sum l_i])
\end{aligned}$$

$$\begin{aligned}
& \{LHS|_{(1,0)}\} = 1 + 1 + \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i - 1], \tau(\omega_2), \cdots, \tau(\omega_{n-k}), \ \tau(\omega_{n-k+1}) = p_{01}, p_{11}[\sum l_i + 1]) \\
\geq \ & 1 + 1 + \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i - 1], \tau(\omega_2), \cdots, \tau(\omega_{n-k}), \ p_{11}[\sum l_i + 1], p_{01}) \\
\geq \ & 1 + \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& \Big(1 + W^k_{t+1}(p_{01}[k - \sum l_i - 1], \tau(\omega_2), \cdots, \tau(\omega_{n-k}), p_{11}[\sum l_i + 1], p_{01})\Big) \\
\geq \ & 1 + \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_2), \cdots, \tau(\omega_{n-k}), p_{11}[\sum l_i + 1]) \\
\geq \ & 1 + \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W^k_{t+1}(p_{01}[k - \sum l_i], p_{11}, \tau(\omega_2), \cdots, \tau(\omega_{n-k}), p_{11}[\sum l_i]) \\
= \ & 1 + \{RHS|_{(1,0)}\} \geq \{RHS|_{(1,0)}\}
\end{aligned}$$

where the first and last inequalities are due to the induction hypothesis of (B), the third due to the induction hypothesis of (A).

With these four cases, we conclude the induction step of proving (A). We next prove the induction step of (B). We consider three cases in terms of whether x and y are among the top k channels to be selected in the next step.

Case (B.1): both x and y belong to the top k positions on both sides. In this case there is no difference between the LHS and RHS along each sample path, since both channels will be selected and the result will be the same.

Case (B.2): neither x nor y is among the top k positions on either side. This implies that $j \leq n-k-2$. We have:

$$LHS = \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot$$
$$W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_1), \cdots, \tau(\omega_j), \tau(y), \tau(x), \tau(\omega_{j+3}), \cdots, p_{11}[\sum l_i]) \ ;$$

$$RHS = \sum_{n-k+2 \leq i \leq n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot$$
$$W^k_{t+1}(p_{01}[k - \sum l_i], \tau(\omega_1), \cdots, \tau(\omega_j), \tau(x), \tau(y), \tau(\omega_{j+3}), \cdots, p_{11}[\sum l_i]) \geq LHS$$

where the last inequality is due to the monotonicity of $\tau()$ and the induction hypothesis of (B).

Case (B.3): exactly one of the two belongs to the the top k channels on each side. This implies that $j = n-k-1$. By the linearity of the function W^k_t we have the following:

$$\begin{aligned} & W^k_t(\omega_1, \cdots, \omega_{n-k-1}, y, x, \omega_{n-k+2}, \cdots, \omega_n) - W^k_t(\omega_1, \cdots, \omega_{n-k-1}, x, y, \omega_{n-k+2}, \cdots, \omega_n) \\ = \ & (x-y)(W^k_t(\omega_1, \cdots, \omega_{n-k-1}, 0, 1, \omega_{n-k+2}, \cdots, \omega_n) \\ & - W^k_t(\omega_1, \cdots, \omega_{n-k-1}, 1, 0, \omega_{n-k+2}, \cdots, \omega_n)) \end{aligned}$$

However, we have

$$
\begin{aligned}
& W_t^k(\omega_1, \cdots, \omega_{n-k-1}, 1, 0, \omega_{n-k+2}, \cdots, \omega_n) \\
= & \sum_{n-k+2 \le i \le n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W_{t+1}^k(p_{01}[k - \sum l_i], \tau(\omega_1), \cdots, \tau(\omega_{n-k-1}), p_{11}, p_{11}[\sum l_i]) \le \sum_{n-k+2 \le i \le n} \omega_i + \beta \cdot \\
& \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot (1 + W_{t+1}^k(p_{01}[k - \sum l_i - 1], \tau(\omega_1), \cdots, \\
& \tau(\omega_{n-k-1}), p_{11}[\sum l_i + 1], p_{01})) \\
\le & \sum_{n-k+2 \le i \le n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& \Big(1 + W_{t+1}^k(p_{01}[k - \sum l_i - 1], \tau(\omega_1), \cdots, \tau(\omega_{n-k-1}), p_{01}, p_{11}[\sum l_i + 1])\Big) \\
\le & 1 + \sum_{n-k+2 \le i \le n} \omega_i + \beta \cdot \sum_{l_{n-k+2}, \cdots, l_n \in \{0,1\}} q(l_{n-k+2}, \cdots, l_n) \cdot \\
& W_{t+1}^k(p_{01}[k - \sum l_i - 1], \tau(\omega_1), \cdots, \tau(\omega_{n-k-1}), p_{01}, p_{11}[\sum l_i + 1]) \\
= & W_t^k(\omega_1, \cdots, \omega_{n-k-1}, 0, 1, \omega_{n-k+2}, \cdots, \omega_n)
\end{aligned}
$$

Since $x \ge y$, we have $LHS \ge RHS$ in the first equation. This concludes the induction step of (B).

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Q.Zhao and B.Krishnamachari "Structure and Optimality of Myopic Sensing for Opportunistic Spectrum Access" in *Proc. of IEEE International Conference on Communications (ICC) Workshops* , June , 2007.

[2] I. Milchtaich, "Congestion Games with Player-Specific Payoff Functions," *Games and Economic Behavior*, vol. 13, no. 1, pp. 111–124, 1996.

[3] R. Smallwood and E. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations Research*, pp. 1071–1088, 1971.

[4] P. Whittle, "Restless bandits: Activity allocation in a changing world," *A Celebration of Applied Probability, ed. J. Gani, Journal of applied probability*, vol. 25A, pp. 287–298, 1988.

[5] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing magazine*, vol. 24, pp. 79–89, May 2007.

[6] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework," *IEEE Journal on Selected Areas in Communications: Special Issue on Adaptive, Spectrum Agile and Cognitive Wireless Networks*, April 2007.

[7] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Communications*, vol. 7, pp. 5431–5440, December 2008.

[8] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," in *IEEE International Conference on Communications (ICC)*, May 2008. Beijing, China.

[9] A. Ganti, E. Modiano, and J. N. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Trans. Information Theory*, vol. 53, pp. 998–1008, March 2007.

[10] A. Ganti, *Transmission scheduling for multi-beam satellite systems.* Ph.D. dissertation, Dept. of EECS, MIT, 2003. Cambridge, MA.

[11] G. Koole, "Monotonicity in markov reward and decision chains: Theory and applications," *Foundations and Trends in Stochastic Systems*, 2006.

[12] K. Liu and Q. Zhao, "Channel probing for opportunistic access with multi-channel sensing," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, October 2008.

[13] N. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *International Conference on Mobile Computing and Networking (MOBICOM)*, September 2007. Montreal, Canada.

[14] A. T. Hoang, Y.-C. Liang, D. Tung Chong Wong, R. Zhang, and Y. Zeng, "Opportunistic spectrum access for energy-constrained cognitive radios," *IEEE Transactions on Wireless Communication*, March 2009.

[15] Q. Zhao, B. Krishnamachari, and K. Liu, "Low-complexity approaches to spectrum opportunity tracking," in *the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, August 2007.

[16] E. Fernandez-Gaucherand, A. Arapostathis, and S. I. Marcus, "On the average cost optimality equation and the structure of optimal policies for partially observable markov decision processes," *Annals of Operations Research*, vol. 29, December 1991.

[17] P. Kumar and P. Karaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, Inc, 1986. Englewood Cliffs, NJ.

[18] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Mathematical Statistics, Wiley Interscience, 1994.

[19] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. K. Gosh, and S. I. Marcus, "Discrete-time controlled markov processes with average cost criterion: A survey," *Siam Journal of Control and Optimization*, vol. 31, March 1993.

[20] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society*, vol. B14, pp. 148–167, 1972.

[21] P. Whittle, "Multi-armed bandits and the gittins index," *Journal of the Royal Statistical Society*, vol. 42, no. 2, pp. 143–149, 1980.

[22] A. Motamedi and A. Bahai, "Mac protocol design for spectrum-agile wireless networks: Stochastic control approach," in *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2007.

[23] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Mathematics of Operations Research*, vol. 24, pp. 293–305, May 1999.

[24] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, vol. 27, pp. 637–648, 1990.

[25] D. Bertsimas and J. E. Niño-Mora, "Restless bandits, linear programming relaxations, and a primal-dual heuristic," *Operations Research*, vol. 48, January-February 2000.

[26] J. E. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, vol. 33, pp. 76–98, 2001.

[27] C. Lott and D. Teneketzis, "On the optimality of an index rule in multi-channel allocation for single-hop mobile networks with multiple service classes," *Probability in the Engineering and Informational Sciences*, vol. 14, pp. 259–297, July 2000.

[28] N. Ehsan and M. Liu, "Server allocation with delayed state observation: Sufficient conditions for the optimality of an index policy," *IEEE Transactions on Wireless Communication*, 2008. to appear.

[29] K. Liu and Q. Zhao, "A restless multiarmed bandit formulation of opportunistic access: indexability and index policy," in *the 5th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, June 2008. a complete version submitted to *IEEE Transactions on Information Theory* and available at http://arxiv.org/abs/0810.4658.

[30] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with markovian rewards," in *48th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2007.

[31] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," in *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.

[32] V. Anantharam, P. Varaiya, and J. Walrand, Asymptotically efficient allocation rules for multi-armed bandit problems with multiple plays. Part I: I.I.D. rewards, Part II: Markovian rewards, IEEE Transactions on Automatic Control, vol. 32, pp. 968982, 1987.

[33] R. Agrawal, M. Hegde, and D. Teneketzis, Multi-armed bandit problems with multiple plays and switching cost, Stochastics and Stochastic Reports, vol. 29, pp. 437459, 1990.

[34] S. H. Ahmad, M. Liu, T. Javidi, Q. Zhao and B. Krishnamachari , "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access," *IEEE IEEE Trans. on Information Theory*, vol. 55, no. 9, pp. 4040-4050, September 2009.

[35] T. Ishikida, Informational aspects of decentralized resource allocation, Ph. D. Thesis, 1992. University of California, Berkeley.

[36] D. G. Pandelis and D. Teneketzis, On the topimality of the Gittins index rule for multi-armed bandits with multiple plays, Mathematical Methods of Operations Research, vol. 50, pp. 449461, 1999.

[37] N.-O. Song and D. Teneketzis, Discrete search with multiple sensors, Journal of Mathematical Methods of Operations Research, vol. 60, no. 1, pp. 113, 2004.

[38] R. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *International Journal of Came Theory*, vol. 2, pp. 65–67, 1973.

[39] B. Vöcking and R. Aachen, "Congestion games: Optimization in competition," *Proceedings of the 2nd Algorithms and Complexity in Durham Workshop*, H. Broersma, S. Dantchev, M. Johnson, and S. Szeider, Eds. Kings College Publications, London, 2006.

[40] A. Fabrikant, C. Papadimitriou, and K. Talwar, "The complexity of pure Nash equilibria," *Proc. 36th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 604–612, 2004.

[41] M. Liu and Y. Wu, "Spectrum sharing as congestion games," *Annual Allerton Conference*, September 2008.

[42] R. Etkin, A. Parekh, and D. Tse, "Spectrum Sharing for Unlicensed Bands," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, April 2007.

[43] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed Multiuser Power Control for Digital Subscriber Lines," *IEEE Journal on Selected Areas in Communications*, vol 20, no. 5, June 2002.

[44] J. Huang, R. A. Berry, and M. L. Honig, "Distributed Interference Compensation for Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, May 2006.

[45] S. Adlakha, R. Johari, and A. Goldsmith, "Competition in wireless systems via Bayesian interference games," submitted for publication, 2008.

[46] D. Monderer and L. S. Shapley, "Potential Games," *Games and Economic Behavior*, vol. 14, no. 0044, pp. 124–143, 1996.