Multi-channel opportunistic access : a restless multi-armed bandit perspective

Kehao Wang

UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE DE INFORMATIQUE
Laboratoire de Recherche en Informatique

DISPLINE Informatique

THÈSE DE DOCTORAT

soutenue le 22/06/2012

par

Kehao WANG

Multi-Channel Opportunistic Access—A Restless Multi-Armed Bandit Perspective

Directeur de thèse:	Khaldoun AL AGHA	Professeur, Université Paris-Sud, France
Co-directeur de thèse:	Lin CHEN	Mâitre de conférence, Université Paris-Sud, France

Composition du jury:

Président du jury:	Pierre DUHAMEL	Research Director, CNRS, France
Rapporteurs:	Mérouane DEBBAH	Professeur, Supèlec, France
	Bartłomiej Błaszczyszyn	Research Director, INRIA, France
Examinateurs:	Jean-Claude Belfiore	Professeur, Telecom ParisTech, France

Acknowledgements

I would like to thank my advisors Khaldoun AL AGHA and Lin CHEN for many suggestions and constant support during this research. I thank them for giving me the opportunity to struggle, nurturing me as an independent researcher, and showing me the joy of doing good research. I have learned about myself and research more than I expected when I first arrived at Universite of Paris Sud, and I am forever indebted to Dr. Chen for this.

I would like to thank Weijia Wang, Youghourta, Sara, Loufeng, Yang Weihua for their generous help that makes my life hassle-free at LRI.

I should also mention that my graduate studies in France were supported by the China Scholarship Council.

I am really grateful to my parents and parents-in-law for their understanding and support over the past years and the years to come.

Finally, I want to thank my wife, Zou han, for being the most supportive wife in the world. This dissertation is dedicated to her.

Résumé

La radio cognitive, premièrement envisagée par Mitola, est la technologie clé pour les futures générations de systèmes sans fil qui répond à des défis critiques en matière d'efficacité spectrale, la gestion des interférences, et de la coexistence de réseaux hétérogènes. Le concept de base dans les réseaux radio cognitifs est centré sur l'accès au spectre opportuniste, dont l'objectif est de résoudre le déséquilibre entre la rareté du spectre d'un côté, et la sous-utilisation du spectre de l'autre côté.

Dans cette thèse, nous abordons le problème fondamental de l'accès au spectre opportuniste dans un système de communication multi-canal. Plus précisément, nous considérons un système de communication dans lequel un utilisateur a accès à de multiples canaux, tout en étant limité à la détection et la transmission sur un sous-ensemble de canaux. Nous explorons comment l'utilisateur intelligent exploite ses observations passées et les propriétés stochastiques de ces canaux afin de maximiser son débit.

Formellement, nous fournissons une analyse générique sur le problème d'accès au spectre opportuniste en nous basant sur le probème de *restless multi-bandit* (RMAB), l'une des généralisations les plus connues du problème du probème classique de multi-armed bandit (MAB), un problème fondamental dans la théorie de décision stochastique. Malgré les importants efforts de la communauté de recherche dans ce domaine, le problème RMAB dans sa forme générique reste encore ouvert. Jusqu'à aujourd'hui, très peu de résultats sont connus sur la structure de la politique optimale. L'obtention de la politique optimale pour un problème RMAB général est intraçable dû la complexité de calcul exponentiel. Par conséquent, une alternative naturelle est de se focaliser sur la politique myopique qui maximise la récompense à immédiate, tout en ignorant celles du futur.

Nous commençons par effectuer une analyse générique dans le chapitre 3 sur l'optimalité de la politique de détection myopique, où l'utilisateur peut accéder plusieurs canaux chaque

fois et obtient une unité de récompense si au moins l'un parmi eux est *good*. Grace à l'analyse mathématique, nous montrons que la politique de détection myopique n'est optimale que pour un petit sous-ensemble de cas où l'utilisateur est autorisé à détecter deux canaux chaque slot. Dans le cas général, nous donnons des contre-exemples pour illustrer que la politique myopique n'est pas optimale.

Motivés par l'analyse ci-dessus, nous étudions ensuite la question naturelle mais fondamentalement dans le chapitre 4 (pour le système homogène constitué de canaux i.i.d.) et le chapitre 5 (pour le système hétérogène, composée de canaux non i.i.d.): sous quelles conditions la politique myopique est-elle optimale? Nous répondons à la question posée par une étude axiomatique. Plus spécifiquement, nous développons trois axiomes caractérisant une famille de fonctions que nous appelons fonctions régulières, qui sont génériques et pratiquement importantes. Nous établissons ensuite l'optimalité de la politique myopique lorsque la fonction de récompense peut être exprimée comme une fonction régulière et le facteur de discount est borné par un seuil déterminée par la fonction de récompense. Nous illustrons également l'application des résultats pour analyser une classe de problèmes RMAB dans l'accès opportuniste.

Dans le chapitre 6, nous étudions un problème plus difficile, où l'utilisateur doit configurer le nombre de canaux à accéder afin de maximiser son utilité (par exemple, le débit). Nous formulons le problème d'optimisation correspondant qui repose sur le compromis entre l'exploitation et l'exploration: la détection de plus de canaux peut aider à apprendre et à prédire l'état futur du canal, augmentant ainsi la récompense à long terme, mais au prix de sacrifier la récompense au slot actuel puisque la détection de plus de canaux réduit le temps de transmission de données, ce qui diminue le débit dans le slot courant. Par conséquent, trouver le nombre optimal de canaux à détecter consiste à trouver un équilibre entre l'exploitation et l'exploration. Après avoir montré la complexité exponentielle du problème, nous développons une stratégie heuristique ν-step look-ahead qui consiste à détecter des canaux d'une manière myopique et d'arrêter la détection lorsque l'utilité agrégée attendue du slot courant t au slot $t+\nu$ commence à diminuer. Dans la stratégie développée, le paramètre ν permet de parvenir à un compromis souhaité entre l'efficacité sociale et de la complexité de calcul. Nous démontrons les avantages de la stratégie proposée via des simulations numériques sur plusieurs scénarios typiques.

Le chapitre 7 conclut la thèse et décrit plusieurs importants axes de recherche futurs dans ce domaine. Notons que malgré l'objectif de cette thèse dans le domaine de la communication opportuniste, la formulation du problème est applicable dans de nombreux autres domaines

de l'ingénerie tels que le brouillage de communication, la planification et le poursuit d'objet. Par conséquent, les résultats présentés dans cette thèse sont génériquement applicables dans un large ensemble de domaines bien au-delà de l'accès au spectre opportuniste.

Mots-clés: Multi-canal d'accès opportuniste, Restless Multi-Armed Bandit, politique myope, l'optimisation stochastique

Abstract

Cognitive radio, first envisioned by Mitola, is the key enabling technology for future generations of wireless systems that addresses critical challenges in spectrum efficiency, interference management, and coexistence of heterogeneous networks. The core concept in cognitive radio networks is opportunistic spectrum access, whose objective is to solve the imbalance between spectrum scarcity and spectrum under-utilization.

In the thesis, we address the fundamental problem of opportunistic spectrum access in a multi-channel communication system. Specifically, we consider a communication system in which a user has access to multiple channels, but is limited to sensing and transmitting only on part of them at a given time. We explore how the smart user should exploit past observations and the knowledge of the stochastic properties of these channels to maximize its transmission rate by switching channels opportunistically.

Formally, we provide a generic analysis on the opportunistic spectrum access problem by casting the problem into the restless multi-armed bandit (RMAB) problem, one of the most well-known generalizations of the classic multi-armed bandit (MAB) problem, which is of fundamental importance in stochastic decision theory. Despite the significant research efforts in the field, the RMAB problem in its generic form still remains open. Until today, very little result is reported on the structure of the optimal policy. Obtaining the optimal policy for a general RMAB problem is often intractable due to the exponential computation complexity. Hence, a natural alternative is to seek a simple myopic policy maximizing the short-term reward.

We start by conducting a generic analysis in Chapter 3 on the optimality of the myopic sensing policy where the user senses more than one channel each time and gets one unit of reward if at least one of the sensed channels is in the good state. Through mathematical analysis, we show that the myopic sensing policy is optimal only for a small subset of cases where the user is allowed to sense two channels each slot. In the general case, we give counterexamples to

illustrate that the myopic sensing policy is not optimal.

Motivated by the above analysis, we then study the following natural while fundamentally important question in Chapter 4 (for the homogeneous system consisting of i.i.d. channels) and Chapter 5 (for the heterogeneous system consisting of non i.i.d. channels): under what conditions is the myopic policy guaranteed to be optimal? We answer the above posed question by performing an axiomatic study. More specifically, we develop three axioms characterizing a family of functions which we refer to as regular functions, which are generic and practically important. We then establish the optimality of the myopic policy when the reward function can be expressed as a regular function and the discount factor is bounded by a closed-form threshold determined by the reward function. We also illustrate how the derived results, generic in nature, are applied to analyze a class of RMAB problems arising from multi-channel opportunistic access.

In Chapter 6, we further investigate the more challenging problem where the user has to decide the number of channels to sense in each slot in order to maximize its utility (e.g., throughput). We formulate the corresponding optimization problem which hinges on the following tradeoff between exploitation and exploration: sensing more channels can help learn and predict the future channel state, thus increasing the long-term reward, but at the price of sacrificing the reward at current slot as sensing more channels reduces the time for data transmission, thus decreasing the throughput in the current slot. Therefore, to find the optimal number of channels to sense consists of striking a balance between the above exploitation and exploration. After showing the exponential complexity of the problem, we develop a heuristic ν-step look-ahead strategy which consists of sensing channels in a myopic way and stopping sensing when the expected aggregated utility from the current slot t to slot $t+\nu$ begins to decrease. In the developed strategy, the parameter ν allows to achieve a desired tradeoff between social efficiency and computation complexity. We demonstrate the benefits of the proposed strategy via numerical experiments on several typical settings.

Finally, Chapter 7 concludes the thesis and outlines several important future research directions in this field. Note that despite the focus of this thesis in the domain of opportunistic communication, the problem formulation is applicable in many other engineering fields such as communication jamming, scheduling and object tracking. Hence the results presented in this thesis are generically applicable in a large range of domains beyond the scope of opportunistic spectrum access.

Contents

Chapter 1

Introduction

1.1 General Context and Motivation

In the last decades, wireless network has enabled the deployment of a large set of advanced communication systems, such as mobile communication, mobile Internet access, and wireless sensor data harvesting. The density of wireless accessing devices and the demand for wireless communications are witnessed to increase dramatically, which brings us into the ubiquitous computing and communication environment, termed by EU as the Internet of Things [1].

Cognitive radio, first envisioned by Mitola [2] and then investigated by the DARPA XG program [3], is the key enabling technology for future generations of wireless systems that addresses critical challenges in spectrum efficiency, interference management, and coexistence of heterogeneous networks. The core concept in cognitive radio networks is opportunistic spectrum access (OSA), whose objective is to solve the imbalance between spectrum scarcity and spectrum under-utilization. The basic idea of OSA is to allow secondary users to search for, identify, and exploit instantaneous spectrum opportunities while limiting the interference perceived by primary users (or licensees). Built upon a hierarchical access structure with primary and secondary users, OSA resolves the inefficiency of the current command-and-control model of spectrum regulation while maintains compatibility with legacy wireless systems.

While conceptually simple, OSA in cognitive radio networks presents novel challenges not encountered in conventional networks, such as sensing over a wide frequency band, identifying the presence of primary users, determining the nature of opportunities, and coordinating the use of these opportunities with other nodes without interfering with the primary users.

In the thesis, we address the fundamental problem of opportunistic spectrum access in a

multi-channel communication system. Specifically, we consider a communication system in which a user has access to multiple channels subject to fading and noisy circumstance, but is limited to sensing and transmitting only on part of these channels at a given time. We explore how the smart user should exploit past observations and the knowledge of the stochastic properties of these channels to maximize its transmission rate by switching channels opportunistically.

1.2 Thesis Contributions and Organization

In this thesis, we provide a generic analysis on the opportunistic spectrum access problem by casting the problem into the restless multi-armed bandit (RMAB) problem [4], one of the most well-known generalizations of the classic multi-armed bandit (MAB) problem [5], which is of fundamental importance in stochastic decision theory.

1.2.1 Optimality of Myopic Channel Sensing Policy

Despite the vital research efforts in the field of RMAB, the RMAB problem in its generic form still remains open and very few results are reported on the structure of the optimal policy. Furthermore, obtaining the optimal policy for a general RMAB problem is often intractable due to the exponential computation complexity [6]. Hence, a natural alternative is to seek a simple myopic policy maximizing the short-term reward. Due to its simple and robust structure, the myopic policy has attracted significant research attention, especially on its social optimality.

We start by performing a generic analysis on the optimality of the myopic sensing policy[1] where a user can sense more than one channel each time and gets one unit of reward if at least one of the sensed channels is in the good state. Through mathematic analysis, we show that the myopic sensing policy is optimal only for a small subset of cases where the user is allowed to sense two channels each slot. For the general case, we present counterexamples to illustrate that the myopic sensing policy is not optimal.

Motivated by the above analysis, we then study the following natural while fundamentally important question: under what conditions is the myopic policy guaranteed to be optimal? In the following chapters, we will answer the posed question by performing an axiomatic study. More specifically, we develop three axioms characterizing a family of generic and important functions, referred to as regular function, and then establish the optimality of the myopic policy

[1]To make the presentation concise, in the thesis by sensing we mean the operation of sensing one or multiple channels and the subsequent operation of choosing one or multiple available channels to access.

when the reward function can be expressed as a regular function and the discount factor is bounded by a closed-form threshold determined by the reward function. Meanwhile, the derived optimal conditions, generic in nature, are applied to analyze a class of RMAB problems arising in multi-channel opportunistic access.

Compared with the existing literature addressing the optimality of the myopic policy of the RMAB problem (cf. Section 2.3.1), our contributions in the first half of the thesis are summarized as follows:

- *Generic analysis:* In contrast to the research line showing the optimality/non-optimality of the myopic policy in some given application scenarios, we make more generic efforts on the sufficient conditions ensuring the optimality of the myopic policy.
- *Homogeneous and heterogeneous channels:* We analyze both homogeneous and heterogeneous scenarios where the channels are characterized by homogeneous and heterogeneous Markov chains, respectively.
- *Sensing error:* The vast majority of studies in OSA assume perfect detection of channel state. However, sensing errors are inevitable in practical scenario (e.g., due to noise and system limitations), especially in wireless communication systems. Our work captures the sensing error and studies the optimality of the myopic policy under imperfect sensing.

From the methodological perspective, we adopt an axiomatic approach to streamline the analysis. On one hand, such axiomatic approach provides a hierarchical view of the addressed problem and leads to clearer and more synthetic analysis. On the other hand, the axiomatic approach also reduces the complexity of solving the RMAB problem and illustrates some important engineering implications behind the myopic policy.

1.2.2 Beyond Myopic Sensing: a Heuristic ν-step Lookahead Policy

In the first part of the thesis, we study the optimality of the myopic sensing policy in the case where the user is allowed to sense k out of N channels. In the second part, we further investigate the more challenging problem where the user has to decide the number of channels to sense in each slot in order to maximize its utility. This optimization problem hinges on the following tradeoff between exploitation and exploration: sensing more channels can help learn and predict the future channel state, thus increasing the long-term reward, but at the price

of sacrificing the reward at current slot as sensing more channels reduces the time for data transmission, thus decreasing the throughput in the current slot. Therefore, to find the optimal number of channels to sense consists of striking a balance between the above exploitation and exploration. After showing the exponential complexity of the problem, we develop a heuristic ν-step lookahead policy which consists of sensing channels in a myopic way and stopping sensing when the expected aggregated utility from the current slot t to slot $t+\nu$ begins to decrease. In the developed policy, the parameter ν allows to achieve a desired tradeoff between social efficiency and computation complexity. We demonstrate the benefits of the proposed strategy via numerical experiments on several typical settings.

The intrinsic design tradeoff hinging behind the proposed heuristic sensing policy is that between *gaining immediate access and gaining information for future use*. Due to hardware limitations and the energy cost of spectrum monitoring, a user may not be able to sense all the channels in the spectrum simultaneously. A sensing strategy is thus needed for intelligent channel selection to track the rapidly varying spectrum opportunities. The purpose of a sensing strategy is twofold: to find good channels for immediate access and to gain statistical information on the spectrum occupancy for better opportunity tracking in the future. The optimal sensing strategy should thus strike a balance between these two often conflicting objectives.

Despite the focus of this thesis in the domain of opportunistic communication, the problem formulation is applicable in many other engineering fields such as communication jamming, scheduling and object tracking. Hence the results presented in this thesis are generically applicable in a large range of domains beyond the scope of opportunistic spectrum access.

1.2.3 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we review the state-of-the-art on the RMAB problem with a particular focus on its application in wireless networks. Chapter 3 provides a motivating analysis on the optimality of the myopic policy when the user can access multiple channels at one time. In Chapter 4, we provide an axiomatic analysis on the optimality of the myopic policy with imperfect sensing in the case of homogeneous channels. In Chapter 5, we further extend our analysis on the optimality of myopic policy in the case of heterogenous channels. In Chapter 6, we investigate the more challenging problem where the user has to decide the number of channels to sense in order to maximize its utility and develop a heuristic ν-step lookahead policy. Finally, Chapter 7 concludes the thesis and discusses some future

research directions.

Part of our research work presented in this thesis is published or under submission in various venues. Specifically, our work of Chapter 3 on the optimality of the myopic policy is accepted by IEEE Wireless Communications Letters [7]. Our work of Chapter 4 on the axiomatic analysis on the optimality of the myopic policy in the case of homogeneous channel model is under submission to IEEE Transaction on Communications [8] and part of the content for the case of perfect sensing is accepted in IET Signal Processing [9]. Our work of Chapter 5 on the case of heterogeneous channel model with perfect sensing is published in IEEE Transaction on Signal Processing [10] and part of the content for the case of imperfect sensing is under review in IEEE Transaction on on Signal Processing [11].

Chapter 2

RMAB and its Application in Communication Networks: State-of-the-art Analysis

As introduced in the previous chapter, the focus of this thesis is to study the fundamental problem of opportunistic spectrum access in a multi-channel communication system. Mathematically, this problem can be cast into a Restless Multi-armed Bandit problem. The RMAB problem is one of the most well-known generalizations of the classic multi-armed bandit problem, a classical problem in stochastic optimization with a wide range of engineering applications.

In this chapter, we start by providing a literature review on the main theory developed for the classic MAB problem and its extension to the RMAB problem. We then focus on the recent works on the application of the RMAB problem in the field of communication networks.

2.1 MAB and Gittins Index

Multi-armed bandit, first posed in 1933, has become a classical problem in stochastic optimization with a wide range of engineering applications, including but not limited to, multi-agent systems, web search and Internet advertising, social networks, and queueing systems. Recently, it has found new applications in communication networks and dynamic systems.

2.1.1 Formulation of the MAB Problem

Consider a dynamic system consisting of a player and N independent arms. In each time slot t ($t = 1, 2, \cdots$), the state of arm k is denoted by $s_k(t)$ and completely observable to the player. At slot t, the player selects one arm, i.e., arm k, to activate based on the system state $\mathcal{S}(t) = [s_1(t), s_2(t), \cdots, s_N(t)]$ and accrues reward $R(s_k(t))$ determined by the state $s_k(t)$ of arm k. Meanwhile, the state of arm k will transmit to another state in the next slot according to certain transition probabilities, i.e., $p_{i,j}^k = P(s_k(t+1) = j | s_k(t) = i), i, j \in \Omega_k$, where Ω_k denotes the state space of arm k. The states of other arms which are not activated will remain frozen, i.e., $s_n(t+1) = s_n(t)\ \forall n \neq k$.

The player's selection policy $\pi = \{\pi(1), \pi(2), \cdots\}$ is a serials of mapping from the system state $\mathcal{S}(t)$ to the action $a(t)$ indicating which arm is activated, i.e., $\pi(t) : \mathcal{S}(t) \rightarrow a(t)$. The objective is to obtain the optimal policy π^* to maximize the expected total discounted reward in an infinite horizon:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\Big[\lim_{T\rightarrow\infty} \sum_{t=1}^{T} \beta^{t-1} R(s_{a(t)}(t))\Big],$$

where the discount factor $0 \leq \beta < 1$.

Since the size of the system states grows exponentially with the number of arms, the above problem, called the classic MAB problem, has an exponential complexity for its general numerical solutions.

2.1.2 Gittins index

This sequential decision problem was firstly proposed by Thompson in 1933 [5], but the theoretical structure of the optimal solution for the classic MAB has not been obtained until Gittins's seminal work [12] in 1974. Gittins showed that an index policy is optimal, called Gittins index later, and thus reduces the complexity of the problem from exponential to linear with the number N of arms.

Theorem 2.1 (Gittins, 1974)**.** *The optimal policy has an index form. Specially, for all $1 \leq k \leq N$, there exists an index function $G_k(\cdot)$ that maps the state $i \in \Omega_k$ of arm k to a real number. At each time, the optimal action is to activate the arm with the largest index.*

Gittins also gave a specific form of the index function $G_k(\cdot)$, referred as Gittins index, as given in the following definition.

Definition 2.1 (Gittins Index)**.** *For any state $i \in \Omega_k$ of arm k,*

$$G_k(i) = \limsup_{\sigma \geq 1} \frac{\mathbb{E}\Big[\sum_{t=1}^{\sigma} \beta^{t-1} R(s_k(t)) | s_k(1) = i\Big]}{\mathbb{E}\Big[\sum_{t=1}^{\sigma} \beta^{t-1} | s_k(1) = i\Big]},$$

where σ is a stopping time for activating the arm k.

Basically, Gittins index measures the maximum reward rate that can be achieved by focusing on activating one arm starting from its current state. Therefore, by Gittins index, the player can accrue reward as quickly as possible and thus maximize the total discounted reward.

2.2 RMAB and Whittle Index

Whittle [4] extended the MAB to a more general model where a set of K ($K > 1$) arms, denoted as $K(t)$, can be activated simultaneously and change their states in each slot and meanwhile the passive arms are also allowed to offer reward and change state, which makes it different from the classic MAB. If arm k is activated, then its state transits according to a transmitting rule P_{k1} and yields the immediate reward $g_{k1}(s_k(t))$ while it transits by another rule P_{k2} and yields the immediate reward $g_{k2}(s_k(t))$ when arm k isn't activated. A policy $\pi = \{\pi(t)\}_{t=1}^{\infty}$ is a serials of mappings where $\pi(t)$ maps the system state $\mathcal{S}(t)$ to the set of K arms $K(t)$ to be activated in slot t.

In [4], Whittle considered the above problem to maximize the average reward over an infinite horizon[1], which can be formulated as follows:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\Big\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \Big[\underbrace{\sum_{i \in K(t)} g_{i1}(s_i(t)) + \sum_{j=1, j\notin K(t)}^{N} g_{j2}(s_j(t))}_{R(t)} \Big] \Big\}.$$

We introduce some notations. Let γ_k denote the maximum expected average reward obtained by playing arm k without constraint:

$$\gamma_k = \max_{\pi} \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_{k a_k(t)}(s_k(t)) \Big], \text{ where } a_k(t) \in \{1 \text{ (active)}), 2 \text{ (passive)}\}.$$

Let $f_k(s_k(1))$ denote the differential reward caused by the transient effect of starting from

[1] The discounted reward can be similarly discussed.

state $s_k(1)$ rather than from an equilibrium situation:

$$f_k(s_k(1)) = \lim_{T\to\infty} \mathbb{E}_{\pi^*}\Big[\frac{1}{T}\sum_{t=1}^{T} g_{ka_k(t)}(s_k(t)) - \gamma_k\Big].$$

We have the following optimal equation for the maximum expected average reward γ_k:

$$\gamma_k + f_k(s_k(t)) = \max_{a=\{1,2\}} \Big[g_{ka}(s_k(t)) + \mathbb{E}[f_k(s_k(t+1))|s_k(t)]\Big].$$

We can rewrite the above formulation more compactly as

$$\gamma_k + f_k(s_k(t)) = \max\Big[L_{k1}f_k, L_{k2}f_k\Big].$$

We consider the following relaxed condition: K out of N arms are activated on average rather than exactly in all time slots, i.e.,

$$\mathbb{E}[|K(t)|] = K \text{ instead of } |K(t)| = K, \forall t.$$

Then the objective under the relaxed condition is the following:

$$\max \mathbb{E}\Big[\sum_{n=1}^{N} r_n\Big], \text{ s.t. } \mathbb{E}\Big[\sum_{n=1}^{N} I_n\Big] = N - M,$$

where r_n is the average reward obtained from arm n under the relaxed constraint, and $I_n = 1, 0$ according to whether arm n is activated or not.

We have the objective by the classic Lagrangian multiplier as follows:

$$\max \mathbb{E}\Big[\sum_{n=1}^{N} r_n + \nu\sum_{n=1}^{N} I_n\Big] = \max \mathbb{E}\Big[\sum_{n=1}^{N}(r_n + \nu I_n)\Big].$$

We thus have an ν-subsidy problem

$$\gamma_k(\nu) + f_k = \max\Big[L_{k1}f_k, \nu + L_{k2}f_k\Big],$$

where ν is referred as subsidy for passivity.

We define the index $W_k(i)$ of arm k in state $i \in \Omega_k$ as the value of ν which makes the active

and the passive phases equally attractive:

$$L_{k1}f_k = \nu + L_{k2}f_k.$$

Let $D_k(\nu)$ be the set of states for which arm k would be make passive under a ν-subsidy policy. Then the arm is indexable if $D_k(\nu)$ increases monotonically from $\emptyset$ to Ω_k as ν increases from $-\infty$ and $+\infty$.

Thus, if all arms are indexable, arm k will be activated in slot t if $W_k(s_k(t)) > \nu$. Therefore, we obtain the following Whittle index policy.

Definition 2.2 (Whittle Index Policy)**.** *If all the bandits are indexable, activate the K arms of the greatest indices in each slot.*

Conjecture 1 (Whittle Conjecture)**.** Suppose all arms are indexable, the index policy is optimal in terms of average yield per arm in the limit.

2.3 Application of RMAB in Communication Networks

As illustrated in the previous section, to obtain the Whittle index of an RMAB needs to prove the indexability firstly, which is dramatically difficult in applications, and in general, the Whittle index policy can only achieve the asymptotically optimal performance even under the indexability. Therefore, the research efforts on addressing the RAMB problem arising in various applications, especially communication systems and networks and dynamic systems, usually fall into the following three categories:

- The first one is to seek sufficient conditions for simple and robust policies (e.g., myopic policy, greedy policy) under which the optimality of such policies is guaranteed.
- The second one is to construct particular policy whose performance to the optimal is bounded.
- The third one, following the research line of Whittle, is to calculate the Whittle index and to derive policies based on the Whittle index.

In the following analysis, we provide an overview of the three research thrusts.

2.3.1 Myopic and Greedy Policy

This line of research consists of seeking simple myopic policies which maximize the short-term reward and studying their performance.

In [13], the authors studied the RMAB problem arising from a multi-channel communication system, where the channel states evolve as independent and statistically identical Markov chains. The objective was to design a sensing policy for channel selection to maximize the average reward. The optimality of the myopic policy was established for the two-channel case and conjectured for the general case based on numerical results. This model was further studied in [14] where the objective is to design a channel selection policy that maximizes the expected discounted or average reward accrued over a finite or infinite horizon. The authors showed that the myopic policy maximizing the immediate one-step reward is optimal when the state transitions of all channels are positively correlated over time and it is also optimal for two or three channels in the case of negatively correlated channel model. The optimality of the myopic policy was extended to the case of accessing arbitrary channels when the state transitions of all channels are positively correlated over time in [15]. When channel state detection is subject to errors, the same simple structure of the myopic policy was established in [16] under a certain condition on the false alarm probability of the channel state detector. The optimality of the myopic policy was proved for the case of two channels and conjectured for general cases. Lower and upper bounds on the performance of the myopic policy were obtained in closed-form, which characterize the scaling behavior of the achievable throughput of the multichannel opportunistic system.

In [17], an opportunistic channel access problem over multiple primary frequency bands was investigated by exploiting primary ACK/NAK packets overheard by the secondary user to overlay their communications on top of active primary channels. The conditions were derived to prove the optimality of the myopic policy with a simple decision structure without relying on a priori knowledge of channel state-transition probabilities by the secondary user.

In [18], the authors considered the downlink of a cellular system where the channel between the base station and each user is modeled by a two-state Markov chain and the ARQ feedback signal arrives at the scheduler with a random delay that is i.i.d. across users and time. The problem was formulated as an RMAB problem to maximize the throughput by indirectly estimating the channel via accumulated delayed-ARQ feedback. For the case of two users in the

system, a greedy policy was proved to achieve the optimal sum throughput for any distribution on the ARQ feedback delay. For the case of more than two users, the greedy policy is suboptimal and has near optimal performance.

2.3.2 Constant Factor Approximation

In [19] [20], the authors developed a novel and general duality-based algorithmic technique that yields a simple and intuitive $2+\epsilon$-approximate greedy policy to the problem and then defined a general sub-class of restless bandit problems called monotone bandits, for which the policy is a 2-approximation. The technique is robust enough to handle generalizations of these problems to incorporate various side constraints such as blocking plays and switching costs. This technique is also of independent interest for other restless bandit problems, and finally shows the policies are closely related to the Whittle index.

2.3.3 Whittle Index

In [21], the author considered a class of restless multi-armed bandit processes that arises in dynamic multichannel access, user/server scheduling, and optimal activation in multi-agent systems. For this class of RMAB problem, the indexability was established and Whittle index was obtained in closed form for both discounted and average reward. When arms are stochastically identical, Whittle index policy was shown to be optimal under certain conditions without knowing the Markov transition probabilities. For nonidentical arms, efficient algorithms were developed for computing a performance upper bound given by Lagrangian relaxation.

In [22], the authors developed efficient sampling policies—link sampling and node sampling—based on the Whittle's indices for tracking the topology of dynamic networks under sampling constraints, and proved its indexability under certain conditions. In [23], the authors investigated a discrete dynamic unmanned aerial vehicle routing problem with a potentially large number of targets and vehicles by regarding each target as an independent two-state Markov chain, and then formulated this problem as an RMAB problem and obtained its closed-form Whittle index. In [24], the real-time multicast scheduling of access point in wireless broadcast networks with strict deadlines was formulated as an RMAB problem in a finite state space, and the indexability was established and Whittle index was obtained in closed-form. In [25], the optimality of an index policy was studied for allocating a singer server to N parallel queues

when the queue size is not perfectly observed, the arrivals are bernoulli and the services are differentiated, and sufficient condition for the index was established to guarantee the optimality in both the finite horizon and infinite horizon cases.

In [26], a comprehensive modeling framework for the RMAB problem of scheduling a finite number of finite-length jobs where the available service rate is time-varying was introduced and, a priority index to minimize the mean waiting time was obtained to improve the flow-level performance and proved to achieve the maximum stability regions. And it was extended in [27] to multi-class job scheduling with user abandonment to minimize the sum of linear holding costs and abandonment penalties, and a simple index was obtained which is, under certain conditions, equivalent to the asymptotically optimal $c\mu/\theta$-rule in multi-server system with overload conditions. The simple index would degenerate to $c\mu$-rule without abandonment.

In [28], a closed-form index for Markovian time-varying channels was evaluated arising in opportunistic flow-level scheduling of wireless downlink systems where the index value of the bad channel takes into account both the one-period and the steady-state potential improvement of the service completion probability while the good channel gets an absolute priority with the $c\mu$-rule. In [29], the anticipative congestion control mechanism in the Internet with time-varying input flow was formulated as an RMAB problem, and the closed-form Whittle index was obtained and proved to be optimal.

In [30], a resource allocation in the security surveillance of an infrastructure consisting of various sectors modeled by a continuous-time Markov decision process was considered as an RMAB problem, and the index was implemented as a basis of a heuristics to define a suboptimal rule to reduce the required memory.

In [31], a novel sensing scheme for dynamic multi-channel access was formulated as an RMAB problem with flexible ratio between transmission period and sensing interval, and the Whittle index obtained was applied as a sorting standard for second users to choose which channel to sense. Throughput was proved to converge to a fixed bound when the ratio approaches to infinity while sensing cost diverged in some cases.

In [32], a distributed best-relay node selection scheme was proposed to maximize the achievable data rate for cooperative communications over underlay-paradigm based cognitive radio networks. The CR relay network was formulated as an RMAB problem where the time-varying channel state is characterized by the finite-state Markov chain, and then the optimal relay node selection policy was obtained by a primal-dual priority-index heuristic.

In [33], a distributed relay selection and power allocation was presented to investigate the channel states of all related links and residual energy state of the relay nodes for cooperative transmission in cognitive radio networks where the cognitive radio network is formulated as an RMAB problem with the channel state and residual energy state characterized by finite state Markov chains.

In [34], a novel multi-channel access scheme based on the TCP throughput in the transport layer in cognitive radio networks was formulated as an RMAB problem and then the optimal channel access policy was obtained to improve the TCP throughput with the cross-layer technology.

In [35], a distributed network selection scheme in heterogeneous wireless networks was proposed to study the multimedia application layer QoS by formulating the integrated network as an RMAB problem, and the index policy was obtained by a primal-dual heuristic.

In [36], opportunistic multiuser scheduling in downlink networks with Markov-modeled outage channels was considered where the scheduler does not have full knowledge of the channel state information, but instead estimates the channel state information by exploiting the memory inherent in the Markov channels along with ARQ-styled feedback from the scheduled users. The scheduling problem was formulated as a partially observable Markov decision process with the classic 'exploitation vs exploration' trade-off or Restless Multi-armed Bandit Processes and furthermore, indexability was proved and the closed-form Whittle index was derived for this kind of downlink scheduling under imperfect channel state information.

In [37], a cooperative opportunistic multiuser scheduling using ARQ feedback in multi-cell downlink systems was formulated as an RMAB problem where two typical scenarios are investigated. When the cooperation between the cells is asymmetric, the optimal scheduling policy was shown to have a greedy flavor and be simple to implement. Under symmetric cooperation, a low complexity index scheduling policy was proposed only if the scheduling problem is Whittle indexable, and then extensive numerical experiments were carried out to demonstrate that the proposed policy is near-optimal.

2.4 Non-Bayesian MAB

Another extension of MAB is the so-called non-bayesian MAB where the channels' availability statistics are not correlated in time as Markov chains and are initially unknown to the users and

need to be estimated via learning. This leads to a tradeoff between exploration—sensing new channels to obtain more statistical information, and exploitation—ensuring successful transmission in the current slot. Research using this approach seeks to optimize the asymptotic performance by minimizing the regret of the developed policy, given that the system regret under a policy π is defined as the accumulative expected reward loss up to time T under the policy π compared to the genie-aided policy where the available probabilities of channels are known to the users at each slot. In this thesis, we do not consider this kind of MAB, and readers can refer to the literatures [38–46].

Chapter 3

Optimality of Myopic Sensing Policy in OSA: a Motivating Analysis

3.1 Introduction

As introduced in Chapter 1, the basic idea of Opportunistic Spectrum Access (OSA) in multi-channel communication system is to exploit instantaneous spectrum availability by allowing users to access those good channels in an opportunistic fashion. In this context, a well-designed channel sensing and access policy is crucial to achieve efficient spectrum usage. In this chapter, we provide a primary study on the optimality of the myopic sensing policy, which serves as a motivating analysis of the subsequent chapters.

We consider a generic scenario where there are N slotted spectrum channels, each one evolving as an independent and identically distributed (i.i.d.), two-state discrete-time Markov chain. The two states for each channel, bad (state 0) and good (state 1), indicate whether the channel is free for a user to transmit its packet on that channel at a given slot. The state transition probabilities are given by $\{p_{ij}\}, i, j = 0, 1$. A user seeks a sensing policy to opportunistically exploit the good channels to transmit its packets. To this end, in each slot, the user selects a subset of channels to sense based on its prior observations, and obtains one unit reward if at least one of the sensed channel is in the good state, indicating that the user can effectively send one packet using the good channel (or one of the good channels) in the current slot. The objective of the user is to find the optimal sensing policy that maximizes the reward accrued over a finite or infinite horizon.

As stated in [13], the design of the optimal sensing policy can be formulated as a partially observable Markov decision process (POMDP), or a restless multi-armed bandit problem (RMAB), of which the application is far beyond the domain of cognitive radio systems. Unfortunately, obtaining the optimal policy for a general POMDP or RMAB is often intractable due to the exponential computation complexity. Hence, a natural alternative is to seek simple myopic policies for the user. In this line of research, a myopic sensing strategy is developed in [14] for the case where the user is limited to sensing only one channel at each slot. The myopic sensing policy in this case is proven to be optimal when $p_{11} \geq p_{01}$.

In this chapter, we naturally extend the proposed myopic policy in [14] to the generic case where the user can sense more than one channel in each slot and get one unit reward if at least one of the sensed channels transmits packet successfully. Theoretical analysis shows that the myopic policy is optimal only for a small subset of cases where the user senses two channels in each slot. In the generic cases, we give counterexamples to show that the myopic policy, despite its simple structure, is not optimal. It is insightful to compare our results obtained in this chapter with another parallel extension [15] on the similar problem. In [15], the authors show that when $p_{11} \geq p_{01}$ holds, the myopic sensing policy is optimal even for the case where the user senses more than one channel in each slot. However, that result seems to be contradictory to our conclusion. In fact, this contradiction is due to the fact that the objective of the user in [15] is to find as many good channels as possible so that the user can transmit over all the good channels. In contrast, our results are focused on the scenario where the user can successfully transmit on one good channel even though multiple good channels are sensed in the current slot. In another word, the user aims at maximizing the probability of successful transmission. It is insightful to notice that the nuance on the model (more precisely on the utility function) indeed leads to totally contrary results, indicating that more research efforts are required to understand the intrinsic relation between the myopic policy and the optimal policy, which motivates our work in this thesis.

The rest of the chapter is organized as follows: Section 3.2 formulates the system model and the myopic sensing policy; Section 3.3 analyzes the optimality of the myopic policy; Section 3.4 summarizes the results.

3.2 Problem Formulation

We are interested in a synchronously slotted multi-channel opportunistic communication system where a user can opportunistically access a set $\mathcal{N}$ of N i.i.d. channels. The state of each channel i in time slot t, denoted by $S_i(t)$, is modeled by a discrete time two-state Markov chain. At the beginning of each slot t, the user selects a set $\mathcal{A}(t)$ ($\mathcal{A}(t) \subset \mathcal{N}, |\mathcal{A}(t)| = k$) of channels to sense, and further obtains the observations $\{O_i(t) \in \{0,1\} : i \in \mathcal{A}(t)\}$, herein $O_i(t) = 1$ indicates channel i is sensed good while $O_i(t) = 0$ indicates channel i is sensed bad. If at least one of the sensed channels is in the good state, the user transmits its packet and collects one unit reward. Otherwise, the user cannot transmit, and thus no reward is obtained. The decision procedure is repeated for each slot t ($1 \leq t \leq T$).

Obviously, by sensing only k out of N channels, the user cannot observe the state information of the whole system. Hence, the user has to infer the channel states from its past decision and observation history so as to make its future decision. To this end, we define the *channel state belief vector* (hereinafter referred to as *belief vector* for briefness) $\Omega(t) \triangleq \{\omega_i(t), i \in \mathcal{N}\}$, where $0 \leq \omega_i(t) \leq 1$ is the conditional probability that channel i is in good state (i.e., $S_i(t) = 1$). Given the sensing action $\mathcal{A}(t)$ and the sensing observations $\{O_i(t) \in \{0,1\} : i \in \mathcal{A}(t)\}$, the belief vector in $t+1$ slot can be updated recursively using Bayes rule as shown in (3.1):

$$\omega_i(t+1) = \begin{cases} p_{11}, & i \in \mathcal{A}(t), O_i(t) = 1 \\ p_{01}, & i \in \mathcal{A}(t), O_i(t) = 0 \\ \mathcal{T}(\omega_i(t)), & i \notin \mathcal{A}(t) \end{cases} \tag{3.1}$$

where, $\mathcal{T}(\omega_i(t)) = \omega_i(t)p_{11} + [1 - \omega_i(t)]p_{01}$.

A sensing policy π specifies a sequence of functions $\pi = [\pi_1, \pi_2, \cdots, \pi_T]$ where π_t maps the belief vector $\Omega(t)$ to the action (i.e., the set of channels to sense) $\mathcal{A}(t)$ in each slot t: $\pi_t : \Omega(t) \to \mathcal{A}(t), |\mathcal{A}(t)| = k$.

We are interested in the user's optimization problem to find the optimal sensing policy π^* that maximizes the expected total reward over a finite horizon:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=1}^{T} R(\pi_t(\Omega(t)))\middle|\, \Omega(1)\right] \tag{3.2}$$

where $R(\pi_t(\Omega(t)))$ is the reward collected in slot t under the sensing policy π_t with the initial belief vector $\Omega(1)$[1].

We derive the dynamic programming formulation of (3.2) as follows:

$$V_T(\Omega(T)) = \max_{\mathcal{A}(T)} \mathbb{E}\Big[1 - \prod_{i\in\mathcal{A}(T)} (1-\omega_i(T))\Big]$$

$$V_t(\Omega(t)) = \max_{\mathcal{A}(t)} \mathbb{E}\Big[[1 - \prod_{i\in\mathcal{A}(t)} (1-\omega_i(t))] + \sum_{\mathcal{E}\in\mathcal{A}(t)} \prod_{i\in\mathcal{E}} \omega_i(t) \cdot \prod_{j\in\mathcal{A}(t)\backslash\mathcal{E}} (1-\omega_j(t)) \cdot V_{t+1}(\Omega(t+1))\Big],$$

where, $V_t(\Omega(t))$ is the value function corresponding to the maximal expected reward from time slot t to T $(1 \le t \le T)$ with the believe vector $\Omega(t+1)$ following the evolution described in (3.1) given that the channels in the subset $\mathcal{E}$ are sensed in good state and the channels in $\mathcal{A}(t)\backslash\mathcal{E}$ are sensed in bad state.

As argued in Introduction, the optimization problem (3.2) is by nature a POMDP, or RMAB, of which the optimal policy is in general intractable. Hence, a natural alternative is to seek simple myopic policy, i.e., the policy maximizing the immediate reward based on current believe vector.

The following definition gives the structure of the myopic sensing policy maximizing the reward for the current slot in the generic scenario.

Definition 3.1 (Myopic Policy under Homogeneous Channel Model with Perfect Sensing)**.** *Sort the elements of the belief vector in descending order such that $\omega_1(t) \ge \omega_2(t) \ge \cdots \ge \omega_N(t)$, the myopic sensing policy in the generic case, where the user is allowed to sense k channels, consists of sensing channel* 1 *to channel k.*

In the next section, we show that the myopic sensing policy is optimal for the case $k = 2$, $T = 2$ when $p_{11} \ge p_{01}$ and for the case $k = 2$, $T = 2$ and $N \le 4$ when $p_{11} < p_{01}$. Beyond this small subset, we show that the myopic policy, despite its simple structure, in general, is not optimal by giving representative counterexamples.

3.3 Optimality of Myopic Sensing Policy

In this section, we study the optimality of the myopic sensing policy. More specifically, we proceed our analysis in two cases: (1) $T = 2$, $k = 2$, (2) $T > 2$, $k > 2$.

[1]If no information on the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution $\omega_0 = \frac{p_{01}}{1+p_{01}-p_{11}}$.

3.3.1 Optimality of myopic policy in the case of $T = 2$, $k = 2$

This subsection is focused on the case where the user is allowed to sense two channels each slot and aims at maximizing the reward of the upcoming two slots. This case models the behavior of a short-sighted user. The following two theorems study the optimality of the myopic sensing policy in the setting with $p_{11} \geq p_{01}$ and $p_{11} < p_{01}$, respectively.

Theorem 3.1 (optimality of myopic sensing policy when $p_{11} \geq p_{01}$ for $T = 2$ and $k = 2$)**.** *In the case where $T = 2$ and $k = 2$, the myopic sensing policy is optimal when $p_{11} \geq p_{01}$.*

Proof. We sort the elements of the believe vector $[\omega_1(t), \omega_2(t), \cdots, \omega_N(t)]$ at the beginning of the slot t in descending order such that $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_N$[2]. Under this notation, we can write the reward of the myopic sensing policy (i.e., sensing channel 1 and 2), denoted as R^*, as

$$\begin{aligned} R^* = & \underbrace{1-(1-\omega_1)(1-\omega_2)}_{A} + \underbrace{\omega_1\omega_2[1-(1-p_{11})(1-p_{11})]}_{B} \\ & + \underbrace{\omega_1(1-\omega_2)[1-(1-p_{11})(1-\mathcal{T}(\omega_3))]}_{C} + \underbrace{(1-\omega_1)\omega_2[1-(1-p_{11})(1-\mathcal{T}(\omega_3))]}_{D} \\ & + \underbrace{(1-\omega_1)(1-\omega_2)[1-(1-\mathcal{T}(\omega_3)(1-F))]}_{E}, \end{aligned} \tag{3.3}$$

where $F = p_{01}$ when $N = 3$ and $F = \mathcal{T}(\omega_4)$ when $N \geq 4$. More specifically, the term A denotes the immediate reward in the current slot t; the term B denotes the expected reward of slot $t+1$ when both channels are sensed to be good; the term C (term D, respectively) denote the expected reward of slot $t+1$ when only channel 1 (channel 2) is sensed to be good; the term E denotes the expected reward of slot $t+1$ when both channels are sensed to be bad.

The proof consists of showing that sensing any subset of two channels $\{i, j\} \neq \{1, 2\}$ cannot lead to more reward. We proceed our proof for two cases:

- $\{i, j\}$ is partially overlapped with $\{1, 2\}$, i.e., $\{i, j\} \bigcap \{1, 2\} \neq \emptyset$;
- $\{i, j\}$ is totally distinct to $\{1, 2\}$, i.e., $\{i, j\} \bigcap \{1, 2\} = \emptyset$.

Case 1. When $\{i, j\}$ is partially overlapped with $\{1, 2\}$, without loss of generality, assume that $i = 1$ and $j \geq 3$, we can derive the upper bound of the expected reward of sensing channel $\{i, j\} = \{1, j\}$ in (3.4) ($j = 3$) or (3.6) ($j > 3$). Here by upper bound we mean that the user,

[2]For the simplicity of presentation, by slightly abusing the notations without introducing ambiguity, we drop the time slot index of $\omega_i(t)$.

first sensing channel i and j in slot t and then sensing the two channels with the largest available probability in slot $t+1$, cannot obtain the maximal reward that the user can achieve.

when $j = 3$,

$$\begin{aligned} R_1 =& 1-(1-\omega_1)(1-\omega_j)+\omega_1\omega_j[1-(1-p_{11})(1-p_{11})] \\ &+\omega_1(1-\omega_j)[1-(1-p_{11})(1-\mathcal{T}(\omega_2))]+(1-\omega_1)\omega_j[1-(1-p_{11})(1-\mathcal{T}(\omega_2))] \\ &+(1-\omega_1)(1-\omega_j)[1-(1-\mathcal{T}(\omega_2)(1-F))]. \end{aligned} \tag{3.4}$$

With some algebraic operations, we have

$$R^* - R_1 = (1-\omega_1)(\omega_2-\omega_3)(1-(1-p_{11})(F-p_{01})) \geq 0. \tag{3.5}$$

when $j > 3$,

$$\begin{aligned} R_2 =& 1-(1-\omega_1)(1-\omega_j)+\omega_1\omega_j[1-(1-p_{11})(1-p_{11})] \\ &+\omega_1(1-\omega_j)[1-(1-p_{11})(1-\mathcal{T}(\omega_2))]+(1-\omega_1)\omega_j[1-(1-p_{11})(1-\mathcal{T}(\omega_2))] \\ &+(1-\omega_1)(1-\omega_j)[1-(1-\mathcal{T}(\omega_2))(1-\mathcal{T}(\omega_3))]. \end{aligned} \tag{3.6}$$

Furthermore, we have

$$\begin{aligned} R^* - R_2 =& \omega_1(1-\omega_2)(\omega_3-\omega_j)(p_{11}-p_{01}) \\ &+(1-\omega_1)(\mathcal{T}(\omega_2)-\mathcal{T}(\omega_j))[\omega_2(1-p_{11})+(1-\mathcal{T}(\omega_3))(1-\omega_2)] \\ &+(1-\omega_1)(\mathcal{T}(\omega_2)-\mathcal{T}(\omega_j))[1-(1-p_{11})(\mathcal{T}(\omega_3)-p_{01})] \geq 0. \end{aligned} \tag{3.7}$$

Thus, (3.5) and (3.7) show that the myopic sensing policy achieves the maximal reward in this case.

Case 2. When $\{i, j\}$ is totally distinct to $\{1, 2\}$, implying $N \geq 4$, we can write the reward of sensing channel $\{i, j\}$ in (3.8):

$$\begin{aligned} R_3 =& 1-(1-\omega_i)(1-\omega_j)+\omega_i\omega_j[1-(1-p_{11})(1-p_{11})] \\ &+\omega_i(1-\omega_j)[1-(1-p_{11})(1-\mathcal{T}(\omega_1))]+(1-\omega_j)\omega_j[1-(1-p_{11})(1-\mathcal{T}(\omega_1))] \\ &+(1-\omega_i)(1-\omega_j)[1-(1-\mathcal{T}(\omega_1))(1-\mathcal{T}(\omega_2))]. \end{aligned} \tag{3.8}$$

With some algebraic operations, we have

$$\begin{aligned}
R_2 - R_3 =&(1-\omega_j)(\omega_1-\omega_i)+\omega_j(\omega_1-\omega_i)(p_{11}+p_{11}-(p_{11})^2)\\
&+(1-\omega_j)[p_{11}(\omega_1-\omega_i)+(1-p_{11})(\omega_1\mathcal{T}(\omega_2)-\mathcal{T}(\omega_1)\omega_i)]\\
&+\omega_j[(1-p_{11})((1-\omega_1)\mathcal{T}(\omega_2)-(1-\omega_i)\mathcal{T}(\omega_1))-p_{11}(\omega_1-\omega_i)]\\
&+(1-\omega_j)\big[-(\omega_1-\omega_i)+(1-\mathcal{T}(\omega_2))[(1-\mathcal{T}(\omega_1))(1-\omega_i)-(1-\omega_1)(\mathcal{T}(\omega_3))]\big]\\
\geq&(1-\omega_j)(\omega_1-\omega_i)+\omega_j(\omega_1-\omega_i)(p_{11}+p_{11}-(p_{11})^2)\\
&+(1-\omega_j)[p_{11}(\omega_1-\omega_i)+(1-p_{11})(\omega_1\mathcal{T}(\omega_i)-\mathcal{T}(\omega_1)\omega_i)]\\
&+\omega_j[(1-p_{11})((1-\omega_1)\mathcal{T}(\omega_i)-(1-\omega_i)\mathcal{T}(\omega_1))-p_{11}(\omega_1-\omega_i)]\\
&+(1-\omega_j)\big[-(\omega_1-\omega_i)+(1-\mathcal{T}(\omega_2))[(1-\mathcal{T}(\omega_1))(1-\omega_i)-(1-\omega_1)(\mathcal{T}(\omega_i))]\big]\\
=&(1-\omega_j)(\omega_1-\omega_i)[p_{11}+(1-p_{11})p_{01}+(1-p_{11})(1-\mathcal{T}(\omega_2))]\geq 0
\end{aligned}$$

Therefore, we have

$$R^* - R_3 = (R^* - R_2) + (R_2 - R_3) \geq 0 \tag{3.9}$$

meaning that the myopic sensing policy achieves the maximal reward in this case, too.

Combining the results of both cases completes the proof of the theorem. □

The following theorem studies the optimality of the myopic sensing policy when $p_{11} < p_{01}$. The proof follows the similar way as that of Theorem 3.1 and is thus omitted.

Theorem 3.2 (optimality of myopic sensing policy when $p_{11} < p_{01}$ for $T = 2$ and $k = 2$)**.** *In the case where $T = 2$ and $k = 2$, the myopic sensing policy is optimal when $p_{11} < p_{01}$ for the system consisting of at most 4 channels (i.e., $N \leq 4$).*

The optimality of the myopic sensing policy derived in this subsection, especially when $p_{11} \geq p_{01}$, hinges on the fact that the eventual loss of reward in slot $t+1$, if there is, is over compensated by the reward gain in the current slot t. However, this result cannot be iterated in the general cases. On the contrary, in the next subsection, we show that the myopic sensing policy may not be optimal by providing a serials of representative counterexamples.

3.3.2 Non-optimality of myopic sensing policy in general cases

Counterexample 1 ($k = 3$, $T = 2$, $N = 6$)**.** Consider a system with $k = 3$, $T = 2$, $N = 6$ and $p_{11} > p_{01}$, the reward generated by the myopic sensing policy (sensing the 3 channels with

highest elements in the believe vector at each slot, i.e., ω_1, ω_2, ω_3) is given by

$$\begin{aligned}
R_1^* =& 1-(1-\omega_1)(1-\omega_2)(1-\omega_3)+\omega_1\omega_2\omega_3[1-(1-p_{11})^3] \\
&+[\omega_1\omega_2(1-\omega_3)+\omega_1(1-\omega_2)\omega_3+(1-\omega_1)\omega_2\omega_3][1-(1-p_{11})^2(1-\mathcal{T}(\omega_4))] \\
&+[\omega_1(1-\omega_2)(1-\omega_3)+(1-\omega_1)\omega_2(1-\omega_3)+(1-\omega_1)(1-\omega_2)\omega_3]\cdot \\
&\quad [1-(1-p_{11})(1-\mathcal{T}(\omega_4))(1-\mathcal{T}(\omega_5))] \\
&+(1-\omega_1)(1-\omega_2)(1-\omega_3)[1-(1-\mathcal{T}(\omega_4))(1-\mathcal{T}(\omega_5))(1-\mathcal{T}(\omega_6))].
\end{aligned}$$

On the other hand, considering the sensing policy that senses the 2 highest elements and the forth highest element in the believe vector (i.e., ω_1, ω_2 and ω_4 according to our notation) in slot t and senses the highest 3 elements in the believe vector in slot $t+1$, the reward generated by this policy is

$$\begin{aligned}
R_1 =& 1-(1-\omega_1)(1-\omega_2)(1-\omega_4)+\omega_1\omega_2\omega_4[1-(1-p_{11})^3] \\
&+[\omega_1\omega_2(1-\omega_4)+\omega_1(1-\omega_2)\omega_4+(1-\omega_1)\omega_2\omega_4][1-(1-p_{11})^2(1-\mathcal{T}(\omega_3))] \\
&+[\omega_1(1-\omega_2)(1-\omega_4)+(1-\omega_1)\omega_2(1-\omega_4)+(1-\omega_1)(1-\omega_2)\omega_4]\cdot \\
&\quad [1-(1-p_{11})(1-\mathcal{T}(\omega_3))(1-\mathcal{T}(\omega_5))] \\
&+(1-\omega_1)(1-\omega_2)(1-\omega_4)[1-(1-\mathcal{T}(\omega_3))(1-\mathcal{T}(\omega_5))(1-\mathcal{T}(\omega_6))].
\end{aligned}$$

It is straightforward to verify that under the setting $p_{11}=0.5$, $p_{01}=0.3$, $[\omega_1,\omega_2,\cdots,\omega_6]=[0.99,0.50,0.40,0.39,0.25,0.25]$, it holds that $R_1-R_1^*=0.00005625>0$.

In the case of $p_{11}<p_{01}$, $k=3$, $T=2$ and $N=6$, we have the reward as follows:

$$\begin{aligned}
R_2^* =& 1-(1-\omega_1)(1-\omega_2)(1-\omega_3)+\omega_1\omega_2\omega_3[1-(1-\mathcal{T}(\omega_6))(1-\mathcal{T}(\omega_5))(1-\mathcal{T}(\omega_4))] \\
&+[\omega_1\omega_2(1-\omega_3)+\omega_1(1-\omega_2)\omega_3+(1-\omega_1)\omega_2\omega_3][1-(1-p_{01})(1-\mathcal{T}(\omega_6))(1-\mathcal{T}(\omega_5))] \\
&+[\omega_1(1-\omega_2)(1-\omega_3)+(1-\omega_1)\omega_2(1-\omega_3)+(1-\omega_1)(1-\omega_2)\omega_3][1-(1-p_{01})^2(1-\mathcal{T}(\omega_6))] \\
&+(1-\omega_1)(1-\omega_2)(1-\omega_3)[1-(1-p_{01})^3].
\end{aligned}$$

$$\begin{aligned}
R_2 =& 1-(1-\omega_1)(1-\omega_2)(1-\omega_4)+\omega_1\omega_2\omega_4[1-(1-\mathcal{T}(\omega_6))(1-\mathcal{T}(\omega_5))(1-\mathcal{T}(\omega_3))]
\end{aligned}$$

$$+ [\omega_1\omega_2(1-\omega_4) + \omega_1(1-\omega_2)\omega_4 + (1-\omega_1)\omega_2\omega_4][1-(1-p_{01})(1-\mathcal{T}(\omega_6))(1-\mathcal{T}(\omega_5))]$$
$$+ [\omega_1(1-\omega_2)(1-\omega_4) + (1-\omega_1)\omega_2(1-\omega_4) + (1-\omega_1)(1-\omega_2)\omega_4][1-(1-p_{01})^2(1-\mathcal{T}(\omega_6))]$$
$$+ (1-\omega_1)(1-\omega_2)(1-\omega_3)[1-(1-p_{01})^3].$$

Thus, we have $R_2 - R_2^* = 0.00002 > 0$ when the parameters are set as $p_{11} = 0.30$, $p_{01} = 0.50$ and $[\omega_1, \omega_2, \cdots, \omega_6] = [0.99, 0.50, 0.40, 0.39, 0.25, 0.25]$.

Counterexample 2 ($k = 2$, $T = 3$, $N = 6$)**.** Consider a system with $k = 2$, $T = 3$, $N = 6$ and $p_{11} > p_{01}$, the reward generated by the myopic sensing policy (sensing the 2 channels with highest elements in the believe vector at each slot, i.e., ω_1, ω_2) is given by

$$\begin{aligned}
R^* =& 1-(1-\omega_1)(1-\omega_2) + \omega_1\omega_2 R_A^* + [\omega_1(1-\omega_2) + (1-\omega_1)\omega_2]R_B^* + (1-\omega_1)(1-\omega_2)R_C^*.\\
R_A^* =& 1-(1-p_{11})(1-p_{11}) + p_{11}p_{11}[1-(1-p_{11})(1-p_{11})]\\
&+ (p_{11}(1-p_{11}) + (1-p_{11})p_{11})[1-(1-p_{11})(1-\mathcal{T}(\omega_3))]\\
&+ (1-p_{11})(1-p11)[1-(1-\mathcal{T}(\omega_3))(1-\mathcal{T}(\omega_4))].\\
R_B^* =& 1-(1-p_{11})(1-\mathcal{T}(\omega_3)) + p_{11}\mathcal{T}(\omega_3)[1-(1-p_{11})(1-p_{11})]\\
&+ (p_{11}(1-\mathcal{T}(\omega_3)) + (1-p_{11})\mathcal{T}(\omega_3))[1-(1-p_{11})(1-\mathcal{T}(\omega_4))]\\
&+ (1-p_{11})(1-\mathcal{T}(\omega_3))[1-(1-\mathcal{T}(\omega_4))(1-\mathcal{T}(\omega_5))].\\
R_C^* =& 1-(1-\mathcal{T}(\omega_3))(1-\mathcal{T}(\omega_4)) + \mathcal{T}(\omega_3)\mathcal{T}(\omega_4)[1-(1-p_{11})(1-p_{11})]\\
&+ (\mathcal{T}(\omega_3)(1-\mathcal{T}(\omega_4)) + (1-\mathcal{T}(\omega_3))\mathcal{T}(\omega_4))[1-(1-p_{11})(1-\mathcal{T}(\omega_5))]\\
&+ (1-\mathcal{T}(\omega_3))(1-\mathcal{T}(\omega_4))[1-(1-\mathcal{T}(\omega_5))(1-\mathcal{T}(\omega_6))].
\end{aligned}$$

On the other hand, considering the sensing policy that senses the highest element and the third highest element in the believe vector (i.e., ω_1, ω_3 according to our notation) in slot t and senses the highest 2 elements in the believe vector in slot $t+1$ and $t+2$, the reward, R_1, generated by this policy can be obtained by the similar induction of R^*.

It is straightforward to verify that under the setting $p_{11} = 0.5$, $p_{01} = 0.4$, $[\omega_1, \omega_2, \cdots, \omega_6] = [0.999, 0.800, 0.700, 0.600, 0.500, 0.400]$, it holds that $R_1 - R^* = 0.0001 > 0$.

In the case of $p_{11} < p_{01}$, $k = 2$, $T = 3$ and $N = 6$ under the setting $p_{11} = 0.4$, $p_{01} = 0.5$, $[\omega_1, \omega_2, \cdots, \omega_6] = [0.99, 0.50, 0.40, 0.39, 0.25, 0.25]$ with the similar policy as that of $p_{11} > p_{01}$, it holds that $R_1 - R^* = 0.0025 > 0$.

We are now ready to state the major result in this paper.

Theorem 3.3 (Non-optimality of Myopic Sensing Policy in General Cases)**.** *The myopic sensing policy is not guaranteed to be optimal in the general cases.*

To conclude this section, it is insightful to note that the optimality of the myopic sensing policy, stated in Theorem 3.1, Theorem 3.2 and Theorem 3.3, hinges on the fundamental trade-off between exploration, by sensing unexplored channels in order to learn and predict the future channel state, thus increasing the long-term reward (e.g., term B, C, D, E in (3.3)), and exploitation, by accessing the channel with the highest estimated good probability based on currently available information (the belief vector) which greedily maximizes the immediate reward (e.g., term A in (3.3)). For a short-sighted user ($T = 1$ and $T = 2$), exploitation naturally dominates exploration (i.e., the immediate reward overweighs the potential gain in future reward), resulting in the optimality of the myopic sensing policy in a subset of this scenario. In contrast, to achieve maximal reward for $T \geq 3$, the user should strike a balance between exploration and exploitation. In such context, the myopic sensing policy that greedily maximizes the immediate reward is no more optimal.

3.4 Conclusion

We study the optimality of the myopic policy in the generic scenario of opportunistic spectrum access of multi-channel communication system. We show that the myopic sensing policy is optimal only for a small subset of cases where a user is allowed to sense two channels each slot. In the generic case, we give counterexamples to show that the myopic sensing policy, despite its simple structure, is not optimal, which is contrary to the results [47] where the myopic policy is optimal when a user is permitted to accrue the reward on every channel sensed to be good. More research thus should be devoted to studying the intrinsic structure of the myopic policy and its optimality, which is the focus of the following chapters of this thesis.

Chapter 4

An Axiomatic Analysis on Optimality of Myopic Sensing Policy in OSA under Imperfect Sensing: the Case of Homogeneous Channels

4.1 Introduction

As illustrated in the chapter 3, the optimality of myopic policy is not always guaranteed. In such context, a natural while fundamentally important question arises: under what conditions is the myopic policy guaranteed to be optimal? In this chapter and the next chapter, we answer the above posed question by performing an axiomatic study on the optimality of the myopic policy for the the case of homogeneous channels and the case of heterogeneous channels, respectively. More specifically, we develop three axioms characterizing a family of functions which we refer to as regular functions, which are generic and practically important. We then establish the optimality of the myopic policy when the reward function can be expressed as a regular function and the discount factor is bounded by a closed-form threshold determined by the reward function. We also illustrate how the derived results, generic in nature, are applied to analyze a class of RMAB problems arising from multi-channel opportunistic access.

In our study, we also take into consideration the imperfect channel state sensing due to sensing error. Note that the vast majority of studies in the area assume perfect observation of

channel states. However, sensing or observation errors are inevitable in practical scenario (e.g., due to noise and system limitations), especially in wireless communication systems which is the focus of our work. More specifically, a good (bad, respectively) channel may be sensed as bad (good) and accessing a bad channel leads to zero reward. In such context, it is crucial to study the structure and the optimality of the myopic sensing policy with imperfect observation. We would like to emphasize that the presence of sensing error brings two difficulties when studying the myopic sensing policy in this new context.

- The belief vector evolves as a non-linear mapping instead of a linear one in the perfect sensing case;
- In the non-perfect sensing case, the belief value of a channel depends not only on the channel evolution itself, but also on the observation outcome, meaning that the transition is not deterministic.

Due to the above particularities[1], our problem requires an original study on the optimality of the myopic sensing policy that cannot draw on existing results in the perfect sensing case. We would like to report that despite its practical importance and particularities, very few work has been done on the impact of sensing error on the performance of the myopic sensing policy, or more generically, on the RMAB problem under imperfect observation. To the best of our knowledge, [48] is the only work in this area, where the optimality of the myopic policy is proved for the case of two channels with a particular utility function. In this chapter, we derive closed-form conditions to guarantee the optimality of the myopic sensing policy under imperfect sensing for arbitrary N and for a class of utility functions.

The rest of the chapter is organized as follows: Section 4.2 formulates the system model; Section 4.3 establishes a set of axioms characterizing a class of generic utility functions; Section 4.4 studies the optimality of the myopic policy and illustrates the application of the derived results via two typical examples; The chapter is concluded by Section 4.5.

[1]Please refer to the remark of (4.1) for a detailed analysis

4.2 Problem Formulation

4.2.1 System Model

We consider a multi-channel opportunistic communication system, in which a user is able to access a set $\mathcal{N}$ of N independent channels, each characterized by a Markov chain of two states, *good* (1) and *bad* (0). The channel state transition matrix $\mathbf{P}^{(i)}$ for channel i ($i \in \mathcal{N}$) is given as follows

$$\mathbf{P}^{(i)} = \begin{bmatrix} p_{11}^{(i)} & 1-p_{11}^{(i)} \\ p_{01}^{(i)} & 1-p_{01}^{(i)} \end{bmatrix},$$

where

$p_{01}^{(i)}$ = prob(channel i is *good* in the current slot given being *bad* in the previous slot),

$p_{11}^{(i)}$ = prob(channel i is *good* in the current slot given being *good* in the previous slot).

We assume that channels go through state transition at the beginning of each slot t. The system operates in a synchronous time slot fashion with the time slot indexed by t ($t = 1, 2, \cdots, T$), where T is the time horizon of interest.

Due to hardware constraints and energy cost, the user is allowed to sense only k ($1 \leq k \leq N$) of the N channels at each slot t. We assume that the user makes the channel selection decision at the beginning of each slot after the channel state transition. Once a channel is chosen, the user detects the channel state $S_i(t)$, which can be considered as a binary hypothesis test:

$$\mathcal{H}_0 : S_i(t) = 1 \text{ (good)} \quad vs. \quad \mathcal{H}_1 : S_i(t) = 0 \text{ (bad)}.$$

The performance of channel i state detection is characterized by the probability of false alarm ϵ_i and the probability of miss detection δ_i:

$$\begin{aligned} \epsilon_i &\triangleq \Pr\{\text{decide } \mathcal{H}_1 \mid \mathcal{H}_0 \text{ is true }\}, \\ \zeta_i &\triangleq \Pr\{\text{decide } \mathcal{H}_0 \mid \mathcal{H}_1 \text{ is true }\}. \end{aligned}$$

We denote the set of channels chosen by the user at slot t by $\mathcal{A}(t)$ where $\mathcal{A}(t) \subseteq \mathcal{N}$ and $|\mathcal{A}(t)| = k$.We assume that the user only transmit packets on the channels sensed to be good.

We also assume that when the receiver successfully receives a packet from a channel, it

sends an acknowledgement to the transmitter over the same channel at the end of the slot. The absence of an ACK (NACK) signifies that the transmitter does not transmit over this channel or transmitted but the channel is busy in this slot. We assume that acknowledgement are received without error since acknowledgements are always transmitted over idle channels [48].

4.2.2 Restless Multi-Armed Bandit Formulation

Obviously, by imperfectly sensing only k out of N channels, the user cannot observe the state information of the whole system. Hence, the user has to infer the channel states from its past decision and observation history so as to make its future decision. To this end, we define the *channel state belief vector* (hereinafter referred to as *belief vector* for briefness) $\Omega(t) \triangleq \{\omega_i(t), i \in \mathcal{N}\}$, where $0 \leq \omega_i(t) \leq 1$ is the conditional probability that channel i is in good state (i.e., $S_i(t) = 1$). As stated in [48], in order to ensure that the user and its intended receiver tune to the same channel in each slot, channel selections should be based on common observations $\{0 \text{ (NACk)}, 1 \text{ (ACK)}\}^k$ rather than the detection outcomes at the transmitter. Given the sensing action $\mathcal{A}(t)$ and the common observations $\{O_i(t) \in \{0,1\} : i \in \mathcal{A}(t)\}$, the belief vector in $t+1$ slot can be updated recursively using Bayes Rule as shown in (4.1):

$$\omega_i(t+1) = \begin{cases} p_{11}^{(i)}, & i \in \mathcal{A}(t), O_i(t) = 1 \\ \mathcal{T}_i(\varphi_i(\omega_i(t))), & i \in \mathcal{A}(t), O_i(t) = 0 \\ \mathcal{T}_i(\omega_i(t)), & i \notin \mathcal{A}(t) \end{cases} \tag{4.1}$$

where,

$$\mathcal{T}_i(\omega_i(t)) \triangleq \omega_i(t)p_{11}^{(i)} + (1-\omega_i(t))p_{01}^{(i)}, \tag{4.2}$$

$$\varphi_i(\omega_i(t)) \triangleq \frac{\epsilon_i \omega_i(t)}{1-(1-\epsilon_i)\omega_i(t)}. \tag{4.3}$$

Note that the belief update under $O_i(t) = 0$ results from the fact that the receiver i cannot distinguish a failed transmission (i.e., collides with the primary user with probability $\delta_i(1-\omega_i(t))$) from no transmission (with probability $\epsilon_i\omega_i(t) + (1-\delta_i)(1-\omega_i(t))$) [48].

Remark. *We would like to emphasize that the sensing error introduces further complications in the system dynamics (i.e., $\varphi_i(\omega_i(t))$ is non-linear with $\omega_i(t)$) compared with the perfect sensing case. Therefore, those results [10, 15, 49] obtained without sensing error cannot be trivially*

extended to the scenario with sensing error.

A sensing policy π specifies a sequence of functions $\pi = [\pi_1, \pi_2, \cdots, \pi_T]$ where π_t maps the belief vector $\Omega(t)$ to the action (i.e., the set of channels to sense) $\mathcal{A}(t)$ in each slot t: $\pi_t : \Omega(t) \to \mathcal{A}(t), |\mathcal{A}(t)| = k$.

Given the imperfect sensing context, we are interested in the user's optimization problem to find the optimal sensing policy π^* that maximizes the expected total discounted reward over a finite horizon:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R(\pi_t(\Omega(t)))\middle| \Omega(1)\right] \tag{4.4}$$

where $R(\pi_t(\Omega(t)))$ is the reward collected in slot t under the sensing policy π_t with the initial belief vector $\Omega(1)$[2], $0 \leq \beta \leq 1$ is the discount factor characterizing the feature that the future rewards are less valuable than the immediate reward. By treating the belief value of each channel as the state of each arm of a bandit, the user's optimization problem can be cast into a restless multi-armed bandit problem.

4.2.3 Myopic Sensing Policy

In order to get more insight on the structure of the optimization problem formulated in (4.4) and the complexity to solve it, we derive the dynamic programming formulation of (4.4) as follows:

$$V_T(\Omega(T)) = \max_{\mathcal{A}(T)} \mathbb{E}\big[R(\pi_t(\Omega(T)))\big],$$

$$V_t(\Omega(t)) = \max_{\mathcal{A}(t)} \mathbb{E}\Big[R(\pi_t(\Omega(t))) + \beta \sum_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod_{i\in\mathcal{E}} (1-\epsilon_i)\omega_i(t) \prod_{j\in\mathcal{A}(t)\backslash\mathcal{E}} [1-(1-\epsilon_j)\omega_j(t)] V_{t+1}(\Omega(t+1))\Big].$$

In the above Bellman equations, $V_t(\Omega(t))$ is the value function corresponding to the maximal expected reward from time slot t to T ($1 \leq t \leq T$) with the belief vector $\Omega(t+1)$ following the evolution described in (4.1) given that the channels in the subset $\mathcal{E}$ are sensed in good state (i.e., receiving ACK) and the channels in $\mathcal{A}(t)\backslash\mathcal{E}$ are sensed in bad state.

Solving (4.4) using the above recursive iteration is computationally heavy due to the fact that the belief vector $\{\Omega(t), t = 1, 2, \cdots, T\}$ is a Markov chain with uncountable state space when $T \to \infty$, resulting the difficulty in tracing the optimal sensing policy π^*. Hence, a natural

[2]If no information on the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution $\omega_0^{(i)} = \frac{p_{01}^{(i)}}{1+p_{01}^{(i)}-p_{11}^{(i)}}$, $1 \leq i \leq N$.

alternative is to seek simple myopic sensing policy which is easy to compute and implement that maximizes the immediate reward, formally defined as follows:

Definition 4.1 (Myopic Policy)**.** *Let the expected reward function* $F(\Omega_A(t)) \triangleq \mathbb{E}[R(\pi_t(\Omega(t)))]$ *denote the expected immediate reward obtained in slot* t *under the sensing policy* π_t *(i.e., sensing the channels in* $\mathcal{A}(t)$*). The myopic sensing policy* $\overline{\mathcal{A}}(t)$*, consists of sensing the* k *channels that maximizes* $F(\Omega_A(t))$*, i.e.,* $\overline{\mathcal{A}}(t) = \text{argmax}_{\mathcal{A}(t)} F(\Omega_A(t))$.

Despite its simple and robust structure, the optimality of the myopic sensing policy is not guaranteed. More specifically, when the channels are stochastically identical (i.e., all channels follow the same Markovian dynamics $\mathbf{P}^{(i)} = \mathbf{P}, \forall i \in \mathcal{N}$) and positively correlated, the myopic sensing policy is shown to be optimal when the user is limited to sensing one channel each slot ($k = 1$) and obtains one unit of reward when the sensed channel is good [13]. The analysis [49] and our work in the previous chapter further extend the study on the generic case where $k \geq 1$. However, the authors [49] show that the myopic sensing policy is optimal if the user gets one unit of reward for each channel sensed to be good[3], while our work shows that the myopic sensing policy is not guaranteed to be optimal when the user's objective is to find at least one good channel[4]. Given that such nuance on the reward function leads to totally contrary results, a natural while fundamentally important question arises: how does the expected slot reward function $F(\Omega_A(t))$ impact the optimality of the myopic sensing policy? Or more specifically, under what conditions on $F(\Omega_A(t))$ is the myopic sensing policy guaranteed to be optimal?

In the sequel analysis in Section 4.3-4.4 by performing an axiomatic study, we shall give affirmative answer to the above posed questions and study some important engineering implications behind the myopic sensing policy for the case of homogeneous channels in this chapter, while the case of the heterogeneous channels would be discussed in the next chapter.

In the following we summarize the assumptions in this chapter:

A1. $\mathbf{P}^{(i)} = \mathbf{P}, \forall i \in \mathcal{N}$ (Homogeneous Channels);

A2. $p_{11}^{(i)} > p_{01}^{(i)}, \forall i \in \mathcal{N}$ (Positively Correlated Channels);

A3. $\epsilon_i = \epsilon, \forall i \in \mathcal{N}$.

[3]Formally, in [49], the expected slot reward function is defined as $F(\Omega_A(t)) = \sum_{i \in \mathcal{A}(t)} \omega_i(t)$

[4]In the previous chapter, the expected slot reward function is defined as $F(\Omega(t)) = 1 - \prod_{i \in \mathcal{A}(t)} (1 - \omega_i(t))$

Note that the channel model under Assumption A2 corresponds to the realistic scenarios where the channel states are observed to evolve gradually over time. Under the assumptions A1–A3, we drop the channel index for notation simplicity in the rest of this chapter.

To conclude this subsection, we state some structural properties of $\mathcal{T}(\omega_i(t))$ and $\varphi(\omega_i(t))$ that are useful in the subsequent proofs.

Lemma 4.1. *For positively correlated channel, i.e., $p_{01} < p_{11}$, we have*

- *$\mathcal{T}(\omega_i(t))$ is monotonically increasing in $\omega_i(t)$;*
- *$p_{01} \leq \mathcal{T}(\omega_i(t)) \leq p_{11}$, $\forall\ 0 \leq \omega_i(t) \leq 1$.*

Proof. It follows from $\mathcal{T}(\omega_i(t)) = (p_{11} - p_{01})\omega_i(t) + p_{01}$ straightforwardly. □

Lemma 4.2. *If $0 \leq \epsilon \leq \frac{(1-p_{11})p_{01}}{p_{11}(1-p_{01})}$ and $p_{01} < p_{11}$, then*

- *$\varphi(\omega_i(t))$ increases monotonically in $\omega_i(t)$ with $\varphi(0) = 0$ and $\varphi(1) = 1$;*
- *$\varphi(\omega_i(t)) \leq p_{01}$, $\forall p_{01} \leq \omega_i(t) \leq p_{11}$.*

Proof. Noticing that $\varphi(\omega_i) = \frac{\epsilon\omega_i(t)}{\epsilon\omega_i(t)+1-\omega_i(t)}$, the lemma follows straightforwardly. □

4.3 Axioms

This section introduces a set of three axioms characterizing a family of generic and practically important functions, to which we refer as *regular* functions. The axioms developed in this section and the implied fundamental properties serve as a basis for the further analysis on the structure and the optimality of the myopic sensing policy in Section 4.4.

Throughout this section, for the convenience of presentation, we sort the elements of the believe vector $\Omega(t) = [\omega_1(t), \cdots, \omega_N(t)]$ for each slot t such that $\mathcal{A} = \{1, \cdots, k\}$ (i.e., the user senses channel 1 to channel k) and let $\Omega_A \triangleq \{\omega_i : i \in \mathcal{A}\} = \{\omega_1, \cdots, \omega_k\}$[5]. The three axioms derived in the following characterize a generic function f defined on Ω_A.

Axiom 1 (Symmetry)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is symmetrical if $\forall i, j \in \mathcal{A}$ it holds that*

$$f(\omega_1, \cdots, \omega_i, \cdots, \omega_j, \cdots, \omega_k) = f(\omega_1, \cdots, \omega_j, \cdots, \omega_i, \cdots, \omega_k).$$

[5]For presentation simplicity, by slightly abusing the notations without introducing ambiguity, we drop the time slot index t.

Axiom 2 (Monotonicity)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is monotonically increasing if it is monotonically increasing in each variable ω_i, i.e., $\forall i \in \mathcal{A}$*

$$\omega_i' > \omega_i \Longrightarrow f(\omega_1, \cdots, \omega_i', \cdots, \omega_k) > f(\omega_1, \cdots, \omega_i, \cdots, \omega_k).$$

Axiom 3 (Decomposability)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is decomposable if $\forall i \in \mathcal{A}$ it holds that*

$$f(\omega_1, \cdots, \omega_i, \cdots, \omega_k) = \omega_i f(\omega_1, \cdots, 1, \cdots, \omega_k) + (1-\omega_i) f(\omega_1, \cdots, 0, \cdots, \omega_k).$$

Axioms 4 and 5 are intuitive. Axiom 6 on the decomposability states that $f(\Omega_A)$ can always be decomposed into two terms that replace ω_i by 0 and 1, respectively. The three axioms introduced in this section are consistent and non-redundant. Moreover, they can be used to characterize a family of generic functions, referred to as *regular* functions, defined as follows:

Definition 4.2 (Regular Function)**.** *A function is called regular if it satisfies all the three axioms.*

The following definition studies the structure of the myopic sensing policy if the expected reward function is regular.

Definition 4.3 (Structure of Myopic Sensing Policy)**.** *Sort the elements of the belief vector in descending order such that $\omega_1 \geq \cdots \geq \omega_N$, if the expected reward function F is regular, then the myopic sensing policy $\overline{\mathcal{A}}$, where the user is allowed to sense k channels, consists of sensing channel 1 to channel k.*

Remark. *In case of tie, we sort the channels in tie in the descending order of $\omega_i(t+1)$ calculated in (4.1). The argument is that larger $\omega_i(t+1)$ leads to larger expected payoff in next slot $t+1$. If the tie persists, the channels are sorted by indexes.*

We would like to emphasize that the developed three axioms characterize a set of generic functions widely used in practical applications. To see this, we give two examples to get more insight: (1) The user gets one unit of reward for each channel that is sensed good and is indeed good. In this example, the expected reward function (for each slot), denoted as F, is the expected slot reward function is $F(\Omega_A) = \sum_{i=1}^{k}[(1-\epsilon)\omega_i]$; (2) The user gets one unit of reward if at least one channel is sensed good. In this example, the expected reward function

is $F(\Omega_A) = 1 - \prod_{i=1}^{k}[1-(1-\epsilon)\omega_i]$. It can be verified that in both examples, F is regular by satisfying the three axioms.

4.4 Optimality of Myopic Sensing Policy under Imperfect Sensing

The goal of this section is to establish closed-form conditions under which the myopic sensing policy, despite of its simple structure, achieves the system optimum under imperfect sensing. To this end, we set up by defining an auxiliary function and studying the structural properties of the auxiliary function, which serve as a basis in the study of the optimality of the myopic sensing policy. We then establish the main result on the optimality followed by the illustration on how the obtained result can be applied via two concrete application examples.

For the convenience of discussion, we firstly state some notations before presenting the analysis:

- $\mathcal{N}(m)$ denotes the first m channles in belief vector;
- Given $\mathcal{E} \subseteq \mathcal{M} \subseteq \mathcal{N}$, $Pr(\mathcal{M}, \mathcal{E}) \triangleq \prod_{i\in\mathcal{E}}(1-\epsilon)\omega_i(t) \prod_{j\in\mathcal{M}\setminus\mathcal{E}}[1-(1-\epsilon)\omega_j(t)]$;
- $\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}$ denotes the vector of length $|\mathcal{E}|$ with each element being p_{11};
- $\mathbf{\Phi}(l,m) \triangleq [\tau(\omega_i(t)) : l \le i \le m]$ where the components are sorted by belief value; $\mathbf{\Phi}_i(l,m) \triangleq [\tau(\omega_j(t)) : l \le j \le m, j \ne i, \omega_j(t) \ge \omega_i(t)]$; $\mathbf{\Phi}^j(l,m) \triangleq [\tau(\omega_i(t)) : l \le i \le m, i \ne j, \omega_j(t) > \omega_i(t)]$; $\mathbf{\Phi}_i^j(l,m) \triangleq [\tau(\omega_h(t)) : l \le h \le m, h \ne i, h \ne j, \omega_j(t) > \omega_h(t) \ge \omega_i(t)]$;
- Given $\mathcal{E} \subseteq \mathcal{M} \subseteq \mathcal{N}$, $\mathbf{Q}^{\mathcal{M},\mathcal{E}} \triangleq [\mathcal{T}(\varphi(\omega_i(t))) : i \in \mathcal{M} \setminus \mathcal{E}]$ where the components are sorted by belief value; $\overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}} \triangleq [\mathcal{T}(\varphi(\omega_i(t))) : i \in \mathcal{M} \setminus \mathcal{E} \setminus \{l\} \text{ and } \omega_i(t) \ge \omega_l(t)]$; $\underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}} \triangleq [\mathcal{T}(\varphi(\omega_i(t))) : i \in \mathcal{M} \setminus \mathcal{E} \setminus \{l\} \text{ and } \omega_i(t) < \omega_l(t)]$;
- Let $\omega_{-i} \triangleq \{\omega_j : j \in \mathcal{A}, j \ne i\}$ and

$$\begin{cases} \Delta_{max} \triangleq \max\limits_{\omega_{-i}\in[0,1]^{k-1}} \{F(1,\omega_{-i}) - F(0,\omega_{-i})\}, \\ \Delta_{min} \triangleq \min\limits_{\omega_{-i}\in[0,1]^{k-1}} \{F(1,\omega_{-i}) - F(0,\omega_{-i})\}. \end{cases}$$

4.4.1 Definition and Properties of Auxiliary Value Function

In this subsection, inspired by the form of the value function $V_t(\Omega(t))$ and the analysis in [47], we first define the auxiliary value function with imperfect sensing and then derive several fundamental properties of the auxiliary value function, which are crucial in the study on the optimality of the myopic sensing policy.

Definition 4.4 (Auxiliary Value Function under Imperfect Sensing)**.** *The auxiliary value function, denoted as $W_t(\Omega(t))$ ($1 \le t \le T$, $t+1 \le r \le T$) is recursively defined as follows:*

$$\begin{cases} W_T(\Omega(T)) = F(\Omega_{\overline{A}}(T)); \\ W_r(\Omega(r)) = F(\Omega_{\overline{A}}(r)) + \beta \sum_{\mathcal{E} \subseteq \overline{\mathcal{A}}(r)} Pr(\overline{\mathcal{A}}(r), \mathcal{E}) W_{r+1}(\Omega_{\mathcal{E}}(r+1)); \\ W_t(\Omega(t)) = F(\Omega_{\mathcal{N}(k)}(t)) + \beta \sum_{\mathcal{E} \subseteq \mathcal{N}(k)} Pr(\mathcal{N}(k), \mathcal{E}) W_{t+1}(\Omega_{\mathcal{E}}(t+1)), \end{cases} \tag{4.5}$$

where $\Omega_{\mathcal{E}}(t+1)$ and $\Omega_{\mathcal{E}}(r+1)$ are generated by $\langle \Omega(t), \mathcal{N}(k), \mathcal{E} \rangle$ and $\langle \Omega(r), \overline{\mathcal{A}}(r), \mathcal{E} \rangle$, respectively, according to (4.1)*, and then sorted by belief value.*

The above recursively defined auxiliary value function gives the expected discounted accumulated reward of the following sensing policy: in slot t sense the first k channels in the belief vector and then sense the channels in $\overline{\mathcal{A}}(r)$ $(t+1 \le r \le T)$ (i.e., adopt the myopic policy from slot $t+1$ to T). If $\mathcal{N}(k) = \overline{\mathcal{A}}(t)$, then the above sensing policy is the myopic sensing policy with $W_t(\Omega(t))$ being the total reward from slot t to T.

In the subsequent analysis of this subsection, we prove some structural properties of the auxiliary value function.

Lemma 4.3 (Symmetry)**.** *Given $0 \le \epsilon \le \frac{(1-p_{11})p_{01}}{p_{11}(1-p_{01})}$, if F is regular, the correspondent auxiliary value function $W_t(\Omega(t))$ is symmetrical in ω_i, ω_j where $i, j \in \mathcal{A}(t)$ or $i, j \notin \mathcal{A}(t)$ for all $t = 1, 2, \cdots, T$, i.e.,*

$$W_t(\omega_1, \cdots, \omega_i, \cdots, \omega_j, \cdots, \omega_N) = W_t(\omega_1, \cdots, \omega_j, \cdots, \omega_i, \cdots, \omega_N).$$

Proof. The proof is given in the appendix. □

Lemma 4.4 (Decomposability)**.** *Given $0 \le \epsilon \le \frac{(1-p_{11})p_{01}}{p_{11}(1-p_{01})}$, if F is regular, then the correspondent auxiliary value function $W_t(\Omega(t))$ is decomposable for all $t = 1, 2, \cdots, T$ and $\forall l \in \mathcal{N}$,*

i.e.,

$$W_t(\omega_1,\cdots,\omega_l,\cdots,\omega_N)=\omega_l W_t(\omega_1,\cdots,1,\cdots,\omega_N)+(1-\omega_l)W_t(\omega_1,\cdots,0,\cdots,\omega_N).$$

Proof. The proof is given in the appendix. □

To demonstrate the property of decomposability of the auxiliary function which is crucial to the study of the optimality, we provide an illustrative example in the following.

Lemma 4.4 can be applied one step further to prove the following corollary.

Corollary 4.1. *Given* $0\leq\epsilon\leq\frac{(1-p_{11})p_{01}}{p_{11}(1-p_{01})}$*, if* F *is regular, then for any* $l,m\in\mathcal{N}$*,* $t=1,2,\cdots,T$*, it holds*

$$\begin{aligned}&W_t(\omega_1,\cdots,\omega_l,\cdots,\omega_m,\cdots,\omega_N)-W_t(\omega_1,\cdots,\omega_m,\cdots,\omega_l,\cdots,\omega_N)\\&\qquad=(\omega_l-\omega_m)\Big[W_t(\omega_1,\cdots,1,\cdots,0,\cdots,\omega_N)-W_t(\omega_1,\cdots,0,\cdots,1,\cdots,\omega_N)\Big].\end{aligned}$$

4.4.2 Optimality of Myopic Sensing under Imperfect Sensing

In this section, we study the optimality of the myopic sensing policy under imperfect sensing. We start by showing the following important auxiliary lemmas (Lemma 4.5, 4.7 and 4.8) and then establish the sufficient condition under which the optimality of the myopic sensing policy is guaranteed.

Lemma 4.5. *Given that (1)* $\epsilon<\frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$*, (2)* $\beta\leq\frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]}$*, and (3)* F *is regular, if* $p_{11}\geq\omega_i\geq p_{01}, i\in\mathcal{N}$*,* $l<m$ *and* $\omega_l>\omega_m$*, for any* $1\leq t\leq T$*, it holds that*

$$W_t(\omega_1,\cdots,\omega_l,\cdots,\omega_m,\cdots,\omega_N)\geq W_t(\omega_1,\cdots,\omega_m,\cdots,\omega_l,\cdots,\omega_N).$$

Lemma 4.6. *Given that (1)* $\epsilon<\frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$*, (2)* $\beta\leq\frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]}$*, and (3)* F *is regular, if* $p_{11}\geq\omega_1\geq\cdots\geq\omega_N\geq p_{01}$*, for any* $1\leq t\leq T$*, it holds that*

$$W_t(\omega_1,\cdots,\omega_{k-1},\omega_k,\cdots,\omega_{N-1},\omega_N)-W_t(\omega_1,\cdots,\omega_{k-1},\omega_N,\omega_k,\cdots,\omega_{N-1})\leq(1-\omega_N)\Delta_{max},$$

Based on Lemma 4.3, $W_t(\omega_1,\cdots,\omega_{k-1},\omega_N,\omega_k,\cdots,\omega_{N-1})=W_t(\omega_N,\omega_1,\cdots,\omega_{N-1})$, combined with Lemma 4.6, we have the following Lemma 4.7:

Lemma 4.7. *Given that (1)* $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$*, (2)* $\beta \leq \frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]}$*, and (3) F is regular, if* $p_{11} \geq \omega_1 \geq \cdots \geq \omega_N \geq p_{01}$*, for any* $1 \leq t \leq T$*, it holds that*

$$W_t(\omega_1, \cdots, \omega_{k-1}, \omega_k, \cdots, \omega_{N-1}, \omega_N) - W_t(\omega_N, \omega_1, \cdots, \omega_{k-1}, \omega_k, \cdots, \omega_{N-1}) \leq (1-\omega_N)\Delta_{max},$$

Lemma 4.8. *Given that (1)* $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$*, (2)* $\beta \leq \frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]}$*, and (3) F is regular, if* $p_{11} \geq \omega_1 \geq \cdots \geq \omega_N \geq p_{01}$*, for any* $1 \leq t \leq T$*, it holds that*

$$W_t(\omega_1, \cdots, \omega_N) - W_t(\omega_N, \omega_2, \cdots, \omega_{N-1}, \omega_1) \leq (p_{11}-p_{01})\Delta_{max}\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t+1}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}.$$

Lemma 4.5 states that by swapping two elements in Ω with the former larger than the latter, the user does not increase the total expected reward. Lemma 4.7 and 4.8, on the other hand, give the upper bounds on the difference of the total reward of the two swapping operations, swapping ω_N and ω_j $(j = N-1, \cdots, 1)$ and swapping ω_1 and ω_N, respectively. For clarity of presentation, the detailed proofs of the three lemmas are deferred to the Appendix. From a technical point of view, it is insightful to compare the methodology in the proof with that in the analysis presented in [49] for the perfect sensing case with $k = 1$. The key point of the analysis in [49] lies in the coupling argument leading to Lemma 3 in [49]. This analysis, however, cannot be directly applied in the generic case with imperfect sensing due to the non-linear update of the belief vector as stated in the remark after equation (4.1). Hence, we base our analysis on the intrinsic structure of the auxiliary value function W and investigate the different 'branches' of channel realizations to derive the relevant bounds, which are further applied to study the optimality of the myopic sensing policy, as stated in the following theorem.

Theorem 4.1. *If* $p_{01} \leq \omega_i(1) \leq p_{11}, i \in \mathcal{N}$*, the myopic sensing policy is optimal if the following conditions hold: (1) F is regular; (2)* $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$*; (3)* $\beta \leq \frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]}$.

Proof. It suffices to show that for $t = 1, \cdots, T$, by sorting $\Omega(t)$ in decreasing order such that $\omega_1 \geq \cdots \geq \omega_N$, it holds that $W_t(\omega_1, \cdots, \omega_N) \geq W_t(\omega_{i_1}, \cdots, \omega_{i_N})$, where $(\omega_{i_1}, \cdots, \omega_{i_N})$ is any permutation of $(1, \cdots, N)$.

We prove the above inequality by contradiction. Assume, by contradiction, the maximum of W_t is achieved at $(\omega_{i_1^*}, \cdots, \omega_{i_N^*}) \neq (\omega_1, \cdots, \omega_N)$, i.e.,

$$W_t(\omega_{i_1^*}, \cdots, \omega_{i_N^*}) > W_t(\omega_1, \cdots, \omega_N). \tag{4.6}$$

However, run a bubble sort algorithm on $(\omega_{i_1^*}, \cdots, \omega_{i_N^*})$ by repeatedly stepping through it, comparing each pair of adjacent element $\omega_{i_l^*}$ and $\omega_{i_{l+1}^*}$ and swapping them if $\omega_{i_l^*} < \omega_{i_l^*+1}$. Note that when the algorithm terminates, the channel belief vector are sorted decreasingly, that is to say, it becomes $(\omega_1, \cdots, \omega_N)$. By applying Lemma 4.5 at each swapping, we have $W_t(\omega_{i_1^*}, \cdots, \omega_{i_N^*}) \leq W_t(\omega_1, \cdots, \omega_N)$, which contradicts to (4.6). Theorem 4.1 is thus proven. □

As noted in [48], when the initial belief $\omega_i(1)$ is set to $\frac{p_{01}}{p_{01}+1-p_{11}}$ as is often the case in practical systems, it can be checked that $p_{01} \leq \omega_i(1) \leq p_{11}$ holds. Moreover, even the initial belief value does not fall in $[p_{01}, p_{11}]$, all the the belief values are bounded in the interval from the second slot following Lemma 4.1. Hence our results can be extended by treating the first slot separately from the future slots.

4.4.3 Discussion

In this subsection, we illustrate the application of the result obtained above in two concrete scenarios and compare our work with the existing results.

Consider the channel access problem in which the user is limited to sense k channels and gets one unit of reward if a sensed channel is in the good state (i.e., receiving ACK), thus the utility function can be formulated as $F(\Omega_A) = (1-\epsilon)\sum_{i\in\mathcal{A}} \omega_i$. Note that the optimality of the myopic sensing policy under this model is studied in [48] for a subset of scenarios where $k = 1$, $N = 2$. We now study the generic case with $k, N \geq 2$. To that end, we apply Theorem 4.1. Notice in this example, we have $\Delta_{min} = \Delta_{max} = 1 - \epsilon$. We can then verify that when $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$, it holds that $\frac{\Delta_{min}}{\Delta_{max}[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}]} > 1$. Therefore, when the condition 1 and 2 holds, the myopic sensing policy is optimal for any β. This result in generic cases significantly extends the results obtained in [48] where the optimality of the myopic policy is proved for the case of two channels and only conjectured for general cases.

Next consider another special scenario where the user can sense and access all channels that are sensed as good, and gets one unit of reward if any of the channels has a successful transmission. Under this model, the user wants to maximize its expected throughput. More specifically, the slot utility function $F = F(\Omega_A) = 1 - \Pi_{i\in\mathcal{A}}[1-(1-\epsilon)\omega_i]$, which is regular. In this context, we have $\Delta_{max} = (1-\epsilon)^{k-1}p_{11}^{k-1}$ and $\Delta_{min} = (1-\epsilon)^{k-1}p_{01}^{k-1}$. The third condition on for the myopic policy to be optimal becomes $\beta \leq \frac{p_{01}^{k-1}}{p_{11}^{k-1}[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}]}$. Particularly, when $\epsilon = 0$, $\beta \leq \frac{p_{01}^{k-1}}{p_{11}^{k-1}(1-p_{01})}$. It can be noted that even when there is no sensing error, the

myopic policy is not ensured to be optimal for any β.

4.5 Conclusion

In this paper, we have investigated the problem of opportunistic channel access under imperfect channel state sensing. We have derived closed-form conditions under which the myopic sensing policy is ensured to be optimal. Due to the generic RMAB formulation of the problem, the obtained results and the analysis methodology presented in this paper are widely applicable in a wide range of domains.

4.6 Appendix

4.6.1 Proof of Lemma 4.3

Recall $W_t(\Omega(t)) = F(\Omega_{\mathcal{N}(k)}(t)) + \beta \sum_{\mathcal{E}\subseteq\mathcal{N}(k)} Pr(\mathcal{N}(k),\mathcal{E}) W_{t+1}(\Omega_{\mathcal{E}}(t+1))$, we prove the lemma by distinguishing the following two cases:

- *Case 1:* $i,j \in \mathcal{A}(t)$. Noticing that (1) both F and $\sum_{\mathcal{E}\subseteq\mathcal{N}(k)} Pr(\mathcal{N}(k),\mathcal{E}) = \sum_{\mathcal{E}\subseteq\mathcal{A}(t)} Pr(\mathcal{A}(t),\mathcal{E})$ are symmetrical w.r.t. ω_i and ω_j, (2) $(\omega_1,\cdots,\omega_i,\cdots,\omega_j,\cdots,\omega_N)$ and $(\omega_1,\cdots,\omega_j,\cdots,\omega_i,\cdots,\omega_N)$ generate the same belief vector $\Omega_{\mathcal{E}}(t+1)$ for any $\mathcal{E}$, and (3) myopic policy is adopted from slot $t+1$ to T, it holds that $W_{t+1}(\Omega_{\mathcal{E}}(t+1))$ is symmetrical w.r.t. ω_i and ω_j.
- *Case 2:* $i,j \notin \mathcal{A}(t)$. Noticing that (1) both F and $\sum_{\mathcal{E}\subseteq\mathcal{N}(k)} Pr(\mathcal{N}(k),\mathcal{E}) = \sum_{\mathcal{E}\subseteq\mathcal{A}(t)} Pr(\mathcal{A}(t),\mathcal{E})$ are unrelated to ω_i,ω_j, (2) $(\omega_1,\cdots,\omega_i,\cdots,\omega_j,\cdots,\omega_N)$ and $(\omega_1,\cdots,\omega_j,\cdots,\omega_i,\cdots,\omega_N)$ generate the same belief vector $\Omega_{\mathcal{E}}(t+1)$ for any $\mathcal{E}$, and (3) myopic policy is adopted from slot $t+1$ to T, it holds that $W_{t+1}(\Omega_{\mathcal{E}}(t+1))$ is symmetrical w.r.t. ω_i and ω_j.

Combing the analysis completes the proof.

4.6.2 Proof of Lemma 4.4

We prove the lemma by backward induction. Firstly, it can be checked that Lemma 4.4 holds for slot T.

Assume that Lemma 4.4 holds for slots $t+1,\cdots,T$, we now prove that it holds for slot t by distinguishing the following two cases.

- Case 1: l is not sensed in slot t, i.e. $l \geq k+1$. In this case, let $\mathcal{M} \triangleq \mathcal{N}(k) = \{1, \cdots, k\}$, we have

$$W_t(\omega_1, \cdots, \omega_l, \cdots, \omega_n) = F(\omega_1, \cdots, \omega_k) + \beta \sum_{\mathcal{E} \subseteq \mathcal{M}} Pr(\mathcal{M}, \mathcal{E}) W_{t+1}(\Omega_l^{\mathcal{E}}(t+1)),$$

where

$$\Omega_l^{\mathcal{E}}(t+1) = (\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), \tau(\omega_l), \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}).$$

Let $\omega_l = 0$ and 1, respectively, we have

$$W_t(\omega_1, \cdots, 0, \cdots, \omega_n) = F(\omega_1, \cdots, \omega_k) + \beta \sum_{\mathcal{E} \subseteq \mathcal{M}} Pr(\mathcal{M}, \mathcal{E}) W_{t+1}(\Omega_{l,0}^{\mathcal{E}}(t+1)),$$
$$W_t(\omega_1, \cdots, 1, \cdots, \omega_n) = F(\omega_1, \cdots, \omega_k) + \beta \sum_{\mathcal{E} \subseteq \mathcal{M}} Pr(\mathcal{M}, \mathcal{E}) W_{t+1}(\Omega_{l,1}^{\mathcal{E}}(t+1)),$$

where

$$\Omega_{l,0}^{\mathcal{E}}(t+1) = (\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), p_{01}, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}),$$
$$\Omega_{l,1}^{\mathcal{E}}(t+1) = (\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), p_{11}, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}).$$

To prove the lemma in this case, it is sufficient to prove

$$W_{t+1}(\Omega_l^{\mathcal{E}}(t+1)) = (1-\omega_l) W_{t+1}(\Omega_{l,0}^{\mathcal{E}}(t+1)) + \omega_l W_{t+1}(\Omega_{l,1}^{\mathcal{E}}(t+1)). \tag{4.7}$$

From the induction result, we have

$$\begin{aligned} W_{t+1}(\Omega_l^{\mathcal{E}}(t+1)) =& \tau(\omega_l) \cdot W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), 1, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}) \\ &+ (1 - \tau(\omega_l)) \cdot W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), 0, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}), \end{aligned} \tag{4.8}$$

$$\begin{aligned} W_{t+1}(\Omega_{l,0}^{\mathcal{E}}(t+1)) =& p_{01} \cdot W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), 1, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}) \\ &+ (1 - p_{01}) \cdot W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1, N), 0, \mathbf{\Phi}^l(k+1, N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}), \end{aligned} \tag{4.9}$$

$$\begin{aligned} W_{t+1}(\Omega^{\mathcal{E}}_{l,0}(t+1)) =& p_{11} \cdot W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}_l(k+1,N), 1, \mathbf{\Phi}^l(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}) \\ &+ (1-p_{11}) \cdot W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}_l(k+1,N1), 0, \mathbf{\Phi}^l(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}). \end{aligned} \tag{4.10}$$

Combing (4.8), (4.9), (4.10), we obtain (4.7).

- Case 2: l is sensed in slot t, i.e. $l \leq k$. In this case, let $\mathcal{M} \triangleq \mathcal{N}(k) \setminus \{l\} = \{1, \cdots, l-1, l+1, \cdots, k\}$, it follows (4.5) that

$$\begin{aligned} W_t(\Omega(t)) =& F(\omega_1, \cdots, \omega_l, \cdots, \omega_k) \\ &+ \beta(1-\epsilon)\omega_l \sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E}) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, p_{11}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &+ \beta[1-(1-\epsilon)\omega_l] \sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E}) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \tau(\varphi(\omega_l)), \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}). \end{aligned}$$

Let $\omega_l = 0$ and 1, respectively, we have

$$\begin{aligned} W_t(\omega_1, \cdots, 0, \cdots, \omega_n) =& F(\omega_1, \cdots, 0, \cdots, \omega_k) \\ &+ \beta \sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E}) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{01}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}), \\ W_t(\omega_1, \cdots, 1, \cdots, \omega_n) =& F(\omega_1, \cdots, 1, \cdots, \omega_k) \\ &+ \beta(1-\epsilon) \sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E}) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, p_{11}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &+ \beta\epsilon \sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E}) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{11}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}). \end{aligned}$$

To prove the lemma in this case, it is sufficient to show

$$\begin{aligned} &[1-(1-\epsilon)\omega_l] W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \tau(\varphi(\omega_l)), \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\quad = (1-\omega_l) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{01}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\qquad + \epsilon\omega_l W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{11}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}). \end{aligned} \tag{4.11}$$

From the induction result, we have

$$\begin{aligned} &W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \tau(\varphi(\omega_l)), \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\quad = \tau(\varphi(\omega_l)) W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 1, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \end{aligned}$$

$$+ (1-\tau(\varphi(\omega_l)))W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 0, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}), \quad (4.12)$$

$$\begin{aligned} &W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{01}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\quad = p_{01}W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 1, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\qquad + (1-p_{01})W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 0, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}), \end{aligned} \quad (4.13)$$

$$\begin{aligned} &W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{11}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\quad = p_{11}W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 1, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) \\ &\qquad + (1-p_{11})W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, 0, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}). \end{aligned} \quad (4.14)$$

Combing (4.12), (4.13), (4.14), we obtain (4.11).

Combing the above analysis completes our proof.

4.6.3 Proof of Lemma 4.5, Lemma 4.6, Lemma 4.7 and Lemma 4.8

Due to the dependency among the three lemmas, we prove them together by backward induction.

We first show that Lemma 4.5 – 4.8 hold for slot T. It is easy to verify that Lemma 4.5 holds.

We then prove Lemma 4.6, 4.7 and 4.8. Noticing the conditions $p_{01} \leq \omega_N \leq \omega_k \leq p_{11} \leq 1$ in Lemma 4.7 and $p_{01} \leq \omega_N \leq \omega_1 \leq p_{11}$ in Lemma 4.8, we have

$$\begin{aligned} &W_T(\omega_1,\cdots,\omega_N) - W_T(\omega_1,\cdots,\omega_{k-1},\omega_N,\omega_k,\cdots,\omega_{N-1}) = F(\omega_1,\cdots,\omega_k) - F(\omega_1,\cdots,\omega_{k-1},\omega_N) \\ &\qquad = (\omega_k-\omega_N)[F(\omega_1,\cdots,\omega_{k-1},1) - F(\omega_1,\cdots,\omega_{k-1},0)] \leq (1-\omega_N)\Delta_{max}, \\ &W_T(\omega_1,\cdots,\omega_N) - W_T(\omega_N,\omega_1,\cdots,\omega_{k-1},\omega_k,\cdots,\omega_{N-1}) = F(\omega_1,\cdots,\omega_k) - F(\omega_N,\omega_1,\cdots,\omega_{k-1}) \\ &\qquad = (\omega_k-\omega_N)[F(\omega_1,\cdots,\omega_{k-1},1) - F(\omega_1,\cdots,\omega_{k-1},0)] \leq (1-\omega_N)\Delta_{max}, \\ &W_T(\omega_1,\cdots,\omega_N) - W_T(\omega_N,\omega_2,\cdots,\omega_{N-1},\omega_1) = F(\omega_1,\cdots,\omega_k) - F(\omega_N,\omega_2,\cdots,\omega_k) \\ &\qquad = (\omega_1-\omega_N)[F(1,\omega_2,\cdots,\omega_k) - F(0,\omega_2,\cdots,\omega_k)] \leq (p_{11}-p_{01})\Delta_{max}. \end{aligned}$$

Lemma 4.6, 4.7 and 4.8 thus hold for slot T.

Assume that Lemma 4.5 – 4.8 hold for slots $T,\cdots,t+1$, we now prove that they

hold for slot t.

We first prove Lemma 4.5. We distinguish the following three cases:

Case 1: $l, m \notin \mathcal{N}(k)$. This case follows Lemma 3.

Case 2: $l \in \mathcal{N}(k)$ and $m \notin \mathcal{N}(k)$. In this case, $\mathcal{M} \triangleq \mathcal{N}(k) \setminus \{l\}$, it can be noted that $\mathbf{Q}^{\mathcal{M},\mathcal{E}} = (\underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}) = (\underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{m}}, \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{m}})$ and $(\mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N)) = (\mathbf{\Phi}_l(k+1,m-1), \mathbf{\Phi}_l(m+1,N), \mathbf{\Phi}^l(k+1,m-1), \mathbf{\Phi}^l(m+1,N))$. In this case, we have

$$
\begin{aligned}
&W_t(\omega_1,\cdots,\omega_l,\cdots,\omega_m,\cdots,\omega_N) - W_t(\omega_1,\cdots,\omega_m,\cdots,\omega_l,\cdots,\omega_N)\\
=&(\omega_l-\omega_m)[W_t(\omega_1,1,\cdots,0,\cdots,\omega_N) - W_t(\omega_1,\cdots,0,\cdots,1,\cdots,\omega_N)]\\
=&(\omega_l-\omega_m)\Big\{F(\omega_1,\cdots,1,\cdots,\omega_k) - F(\omega_1,\cdots,0,\cdots,\omega_k)+\\
&\beta\sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E})\Big[(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, p_{11}, \mathbf{\Phi}_m(k+1,N), p_{01}, \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}})\\
&+\epsilon W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_m(k+1,N), p_{01}, \mathbf{\Phi}^m(k+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}}, p_{11}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{l}})\\
&-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_l(k+1,m-1), \mathbf{\Phi}_l(m+1,N), p_{11}, \mathbf{\Phi}^l(k+1,m-1), \mathbf{\Phi}^l(m+1,N), \overline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{m}}, p_{01}, \underline{\mathbf{Q}}^{\mathcal{M},\mathcal{E},\mathbf{m}})\Big]\Big\}\\
\geq&(\omega_l-\omega_m)\Big\{\Delta_{min} + \beta\sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E})\cdot\Big[(1-\epsilon)W_{t+1}(p_{01}, \mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, p_{11}, \mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}})\\
&+\epsilon W_{t+1}(p_{01}, \mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}, p_{11})\\
&-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, p_{11}, \mathbf{\Phi}_l(k+1,m-1), \mathbf{\Phi}_l(m+1,N), \mathbf{\Phi}^l(k+1,m-1), \mathbf{\Phi}^l(m+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}, p_{01})\Big]\Big\}\\
=&(\omega_l-\omega_m)\Big\{\Delta_{min} + \beta\sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E})\cdot\Big[(1-\epsilon)W_{t+1}(p_{01}, \mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, p_{11}, \mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}})\\
&+\epsilon W_{t+1}(p_{01}, \mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, \mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}, p_{11})\\
&-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}}, p_{11}, \mathbf{\Phi}_m(k+1,N), \mathbf{\Phi}^m(k+1,N), \mathbf{Q}^{\mathcal{M},\mathcal{E}}, p_{01})\Big]\Big\}\\
\geq&(\omega_l-\omega_m)\Big[\Delta_{min} - \beta\sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E})\cdot\\
&\Big((1-\epsilon)(1-p_{01})\Delta_{max} + \epsilon(p_{11}-p_{01})\Delta_{max}\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}\Big)\Big]\\
\geq&(\omega_l-\omega_m)\sum_{\mathcal{E}\subseteq\mathcal{M}} Pr(\mathcal{M},\mathcal{E})\cdot\\
&\left[\Delta_{min} - \beta\left((1-\epsilon)(1-p_{01})\Delta_{max} + \epsilon(p_{11}-p_{01})\Delta_{max}\frac{1}{1-(1-\epsilon)(p_{11}-p_{01})}\right)\right] \geq 0,
\end{aligned}
$$

where the first inequality follows the induction result of Lemma 4.5, the second inequality follows the induction result of Lemma 4.7 and 4.8, the forth inequality follows the condition in the lemma.

Case 3: $l, m \in \mathcal{N}(k)$. This case follows Lemma 4.3.

Lemma 4.5 is thus proven for slot t.

We then proceed to prove Lemma 4.6. We start with the first inequality. We develop W_t w.r.t. ω_k and ω_N according to Lemma 4.4 as follows:

$$\begin{aligned}
&W_t(\omega_1,\cdots,\omega_{k-1},\omega_k,\cdots,\omega_{n-1},\omega_n) - W_t(\omega_1,\cdots,\omega_{k-1},\omega_n,\omega_k,...,\omega_{n-1})\\
=&\omega_k\omega_n[W_t(\omega_1,\cdots,\omega_{k-1},1,\omega_{k+1},\cdots,\omega_{n-1},1) - W_t(\omega_1,\cdots,\omega_{k-1},1,1,\omega_{k+1},\cdots,\omega_{n-1})]\\
&+\omega_k(1-\omega_n)[W_t(\omega_1,\cdots,\omega_{k-1},1,\omega_{k+1},\cdots,\omega_{n-1},0) - W_t(\omega_1,\cdots,\omega_{k-1},0,1,\omega_{k+1},\cdots,\omega_{n-1})]\\
&+(1-\omega_k)\omega_n[W_t(\omega_1,\cdots,\omega_{k-1},0,\omega_{k+1},\cdots,\omega_{n-1},1) - W_t(\omega_1,\cdots,\omega_{k-1},1,0,\omega_{k+1},\cdots,\omega_{n-1})]\\
&+(1-\omega_k)(1-\omega_n)[W_t(\omega_1,\cdots,\omega_{k-1},0,\omega_{k+1},\cdots,\omega_{n-1},0) - W_t(\omega_1,\cdots,\omega_{k-1},0,0,\omega_{k+1},\cdots,\omega_{n-1})].
\end{aligned} \tag{4.15}$$

We proceed the proof by upbounding the four terms in (4.15).

For the first term, we have

$$\begin{aligned}
&W_t(\omega_1,\cdots,\omega_{k-1},1,\omega_{k+1},\cdots,\omega_{n-1},1) - W_t(\omega_1,\cdots,\omega_{k-1},1,1,\omega_{k+1},\cdots,\omega_{n-1})\\
=&\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\Big[(1-\epsilon)W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}})\\
&+\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11})\\
&-(1-\epsilon)W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}})\\
&-\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11})\Big]\le 0,
\end{aligned}$$

where, the inequality follows the induction of Lemma 4.5.

For the second term, we have

$$\begin{aligned}
&W_t(\omega_1,\cdots,\omega_{k-1},1,\omega_{k+1},\cdots,\omega_{n-1},0) - W_t(\omega_1,\cdots,\omega_{k-1},0,1,\omega_{k+1},\cdots,\omega_{n-1})\\
=&F(\omega_1,\cdots,\omega_{k-1},1) - F(0,\omega_1,\cdots,\omega_{k-1})\\
&+\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\Big[(1-\epsilon)W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}})\\
&+\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11}) - W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\Big]\\
\le&F(\omega_1,\cdots,\omega_{k-1},1) - F(0,\omega_1,\cdots,\omega_{k-1})
\end{aligned}$$

$$
\begin{aligned}
&+\beta \sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\Big[(1-\epsilon)W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\\
&+\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11},p_{01})-W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\Big]\\
=&F(\omega_1,\cdots,\omega_{k-1},1)-F(0,\omega_1,\cdots,\omega_{k-1})+\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\\
&\Big[\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11},p_{01})-\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\Big]\\
\leq&\Delta_{max}
\end{aligned}
$$

following the induction of Lemma 4.5.

For the third term, we have

$$
\begin{aligned}
&W_t(\omega_1,\cdots,\omega_{k-1},0,\omega_{k+1},\cdots,\omega_{n-1},1)-W_t(\omega_1,\cdots,\omega_{k-1},1,0,\omega_{k+1},\cdots,\omega_{n-1})\\
=&F(\omega_1,\cdots,\omega_{k-1},0)-F(1,\omega_1,\cdots,\omega_{k-1})\\
&+\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\Big[W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\\
&-(1-\epsilon)W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},p_{01},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}})\\
&-\epsilon W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{01},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11})\Big]\\
\leq&-\Delta_{min}+\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\Big[W_{t+1}(\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\\
&-(1-\epsilon)W_{t+1}(p_{01},p_{11},\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}})\\
&-\epsilon W_{t+1}(p_{01},\mathbf{P}^{\mathcal{E}}_{\mathbf{11}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{11})\Big]\\
\leq&-\Delta_{min}+\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\\
&\Big[(1-\epsilon)(1-p_{01})\Delta_{max}+\epsilon(p_{11}-p_{01})\Delta_{max}\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}\Big]\\
\leq&\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)} Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\\
&\Big[-\Delta_{min}+\beta\Big[(1-\epsilon)(1-p_{01})\Delta_{max}+\epsilon(p_{11}-p_{01})\Delta_{max}\frac{1}{1-(1-\epsilon)(p_{11}-p_{01})}\Big]\Big]\leq 0,
\end{aligned}
$$

where the first inequality follows the induction result of Lemma 4.5, the second equality follows the induction result of Lemma 4.7 and 4.8, the forth inequality is due the condition in Lemma 4.7.

For the fourth term, we have

$$
\begin{aligned}
&W_t(\omega_1,\cdots,\omega_{k-1},0,\omega_{k+1},\cdots,\omega_{n-1},0)-W_t(\omega_1,\cdots,\omega_{k-1},0,0,\omega_{k+1},\cdots,\omega_{n-1})\\
=&\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)}Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\\
&\Big[W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},p_{01},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\Big]\\
\leq&\beta\sum_{\mathcal{E}\subseteq\mathcal{N}(k-1)}Pr(\mathcal{N}(k-1),\mathcal{E})\cdot\\
&\Big[W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01},p_{01})-W_{t+1}(p_{01},\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{N}(\mathbf{k-1}),\mathcal{E}},p_{01})\Big]\\
\leq&\beta(1-p_{01})\Delta_{max},
\end{aligned}
$$

where the first inequality follows Lemma 4.5, the second follows the induction result of Lemma 4.7.

Combing the above results of the four terms, we have

$$
\begin{aligned}
&W_t(\omega_1,\cdots,\omega_N)-W_t(\omega_1,\cdots,\omega_{k-1},\omega_N,\omega_k,\cdots,\omega_{N-1})\\
\leq&\omega_k(1-\omega_N)\cdot\Delta_{max}+(1-\omega_k)(1-\omega_N)\cdot(1-p_{01})\beta\Delta_{max}\\
\leq&\omega_k(1-\omega_N)\Delta_{max}+(1-\omega_k)(1-\omega_N)\Delta_{max}\leq(1-\omega_N)\Delta_{max},
\end{aligned}
$$

which completes the proof of Lemma 4.6.

Based on Lemma 4.3, $W_t(\omega_1,\cdots,\omega_{k-1},\omega_N,\omega_k,\cdots,\omega_{N-1})=W_t(\omega_N,\omega_1,\cdots,\omega_{k-1},\omega_k,\cdots,\omega_{N-1})$, combined with Lemma 4.6, **we conclude the proof of Lemma 4.7**.

Finally, we prove Lemma 4.8. To this end, denote $\mathcal{M}\triangleq\{2,\cdots,k\}$, we have

$$
\begin{aligned}
&W_t(\omega_1,\cdots,\omega_N)-W_t(\omega_N,\omega_2,\cdots,\omega_{N-1},\omega_1)\\
=&(\omega_1-\omega_N)[W_t(1,\omega_2,\cdots,\omega_{N-1},0)-W_t(0,\omega_2,\cdots,\omega_{N-1},1)]\\
=&(\omega_1-\omega_N)\Big\{F(1,\omega_2,\cdots,\omega_k)-F(0,\omega_2,\cdots,\omega_k)+\beta\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})\cdot\\
&\Big[(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},p_{11},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{M},\mathcal{E}})+\epsilon W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{01},p_{11},\mathbf{Q}^{\mathcal{M},\mathcal{E}})\\
&-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{01})\Big]\Big\}\\
\leq&(\omega_1-\omega_N)\Big\{\Delta_{max}+\beta\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})\Big[(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},p_{11},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{M},\mathcal{E}})
\end{aligned}
$$

$$
\begin{aligned}
&+\epsilon W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{01})-W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{01})\Big]\Big\}\\
=&(\omega_1-\omega_N)\Big\{\Delta_{max}+\beta\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})\Big[(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},p_{11},\mathbf{\Phi}(k+1,N-1),p_{01},\mathbf{Q}^{\mathcal{M},\mathcal{E}})\\
&-(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),p_{11},\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{01})\Big]\Big\}\\
\leq&(\omega_1-\omega_N)\Big\{\Delta_{max}+\beta\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})\Big[(1-\epsilon)W_{t+1}(\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},p_{11},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{01})\\
&-(1-\epsilon)W_{t+1}(p_{01},\mathbf{P}_{\mathbf{11}}^{\mathcal{E}},\mathbf{\Phi}(k+1,N-1),\mathbf{Q}^{\mathcal{M},\mathcal{E}},p_{11})\Big]\Big\}\\
\leq&(p_{11}-p_{01})\Big[\Delta_{max}+\beta\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})(1-\epsilon)\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}(p_{11}-p_{01})\Delta_{max}\Big]\\
=&\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})(p_{11}-p_{01})\Big[\Delta_{max}+\beta(1-\epsilon)\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}(p_{11}-p_{01})\Delta_{max}\Big]\\
=&\sum_{\mathcal{E}\subseteq\mathcal{M}}Pr(\mathcal{M},\mathcal{E})\Big[1+\beta(1-\epsilon)(p_{11}-p_{01})\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}\Big](p_{11}-p_{01})\Delta_{max}\\
=&\frac{1-[\beta(1-\epsilon)(p_{11}-p_{01})]^{T-t+1}}{1-\beta(1-\epsilon)(p_{11}-p_{01})}(p_{11}-p_{01})\Delta_{max},
\end{aligned}
$$

where the first two inequalities follows the induction result of Lemma 4.5, the third inequality follows the induction result of Lemma 4.8.

We thus complete the whole process of proving Lemma 4.5–4.8.

Chapter 5

An Axiomatic Analysis on Optimality of Myopic Sensing Policy in OSA under Imperfect Sensing: the Case of Heterogeneous Channels

In the previous chapter, we have studied the optimality of the myopic policy under imperfect sensing for the case of homogeneous channels. In this chapter, we further consider the more challenging scenario of heterogeneous channels.

5.1 System Model and Problem Formulation

We are interested in the user's optimization problem to find the optimal sensing policy π^* that maximizes the expected total discounted reward over a finite horizon. More specifically, we establish closed-form conditions under which the myopic sensing policy is guaranteed to be optimal.

In this chapter, we adopt the same system setting as previous chapter. Hence for this part, readers can refer to the System Model and Restless Multi-Armed Bandits Formulation of Chapter 4. Moreover, some results are quoted from the previous chapter and extended to the case of heterogeneous channels.

In the following, we summarize the assumptions of this chapter:

A2. $p_{11}^{(i)} > p_{01}^{(i)}, \forall i \in \mathcal{N}$;

A3. $\epsilon_i = \epsilon, \forall i \in \mathcal{N}$.

We would like to point out that compared with the previous Chapter 4, the assumption A1 is dropped in this chapter to cover the heterogeneous channels.

We next state some structural properties of $\mathcal{T}_i(\omega_i(t))$ and $\varphi(\omega_i(t))$ that are useful in the subsequent proofs.

Lemma 5.1. *For any positively correlated channel i (i.e., $p_{01}^{(i)} < p_{11}^{(i)}$), the following structural properties of $\mathcal{T}_i(\omega_i(t))$ hold:*

- *$\mathcal{T}_i(\omega_i(t))$ is monotonically increasing in $\omega_i(t)$;*
- *$p_{01}^{(i)} \leq \mathcal{T}_i(\omega_i(t)) \leq p_{11}^{(i)}$, $\forall\ 0 \leq \omega_i(t) \leq 1$.*

Proof. Noticing that $\mathcal{T}_i(\omega_i(t))$ can be written as $\mathcal{T}_i(\omega_i(t)) = (p_{11}^{(i)} - p_{01}^{(i)})\omega_i(t) + p_{01}^{(i)}$, Lemma 5.1 holds straightforwardly. □

Lemma 5.2. *$\varphi(\omega_i(t))$ monotonically increases with $\omega_i(t)$ when $0 \leq \epsilon < 1$.*

Proof. Noticing that $\varphi(\omega_i) = \frac{\epsilon\omega_i(t)}{\epsilon\omega_i(t)+1-\omega_i(t)}$, Lemma 5.2 follows straightforwardly. □

5.2 Axioms

This section defines three axioms characterizing a family of generic and practically important functions referred to as *g-regular* functions, which serve as a basis for the further analysis on the structure and the optimality of the myopic sensing policy[1].

Axiom 4 (Symmetry)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is symmetrical if for any two distinct channels i and j, it holds that*

$$f(\omega_1, \cdots, \omega_i, \cdots, \omega_j, \cdots, \omega_k) = f(\omega_1, \cdots, \omega_j, \cdots, \omega_i, \cdots, \omega_k).$$

[1]Throughout this section, for the convenience of presentation, we sort the elements of the believe vector $\Omega(t) = [\omega_1(t), \omega_2(t), \cdots, \omega_N(t)]$ in the descending order for each slot t such that $\mathcal{A} = \{1, 2, \cdots, k\}$ (i.e., the user senses channel 1 to channel k) and let $\Omega_A \triangleq \{\omega_1, \omega_2, \cdots, \omega_k\}$[2]. The three axioms derived in the following characterize a generic function f defined on Ω_A.

Axiom 5 (Monotonicity)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is monotonically increasing if it is monotonically increasing in each variable ω_i, i.e.,*

$$\omega_i' > \omega_i \Longrightarrow f(\omega_1, \cdots, \omega_i', \cdots, \omega_k) > f(\omega_1, \cdots, \omega_i, \cdots, \omega_k), \quad \forall i \leq k.$$

The above axioms are the intuitive with Axiom 4 stating that once the sensing set $\mathcal{A}$ is given, the sensing order will not change the final reward under a symmetrical function f. The following axiom, however, significantly extends the axiom of decomposability in chapter 4 and [10] so as to cover a much larger range of utility functions.

Axiom 6 (g-Decomposability)**.** *A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is decomposable if there exists a continuous and increasing function $g : [0,1] \to [0,\infty)$ and a constant c such that for any $i \leq k$ it holds that*

$$f(\omega_1, \cdots, \omega_{i-1}, \omega_i, \omega_{i+1}, \cdots, \omega_k) = c \cdot g(\omega_i) f(\omega_1, \cdots, \omega_{i-1}, 1, \omega_{i+1}, \cdots, \omega_k) + c \cdot (1 - g(\omega_i)) f(\omega_1, \cdots, \omega_{i-1}, 0, \omega_{i+1}, \cdots, \omega_k).$$

Axiom 6 on the g-decomposability states that $f(\Omega_A)$ can always be decomposed into two terms by introducing the function g and replacing ω_i by 0 and 1, respectively. It is insightful to note that Axiom of g-decomposability significantly extends Axiom of decomposability in chapter 4 by covering a much larger range of utility functions which cannot be covered by formal, particularly the logarithmic function (e.g., $f(\Omega_A) = \sum_{i=1}^k \log_a(1+\omega_i)$ $(a > 1)$, where $c = \frac{1}{\log_2 a}$, $g(\omega_i) = \log_2(1+\omega_i)$) and the power function (e.g., $f(\Omega_A) = \sum_{i=1}^k \omega_i^a, a > 0$, where $c = 1$, $g(\omega_i) = \omega_i^a$) that are widely used in engineering problems. By setting $g(\omega_i) = \omega_i$ and $c = 1$, Axiom 6 degenerates to the Axiom of decomposability in chapter 4.

In the following, we use the above axioms to characterize a family of generic functions, referred to as *g-regular* functions, defined as follows.

Definition 5.1 (g-Regular Function)**.** *A function is called g-regular if it satisfies all the three axioms.*

If the expected reward function F is g-regular, the myopic sensing policy, defined in Definition 4.3, consists of sensing the k channels with the largest belief values. In case of tie, we can sort the channels in tie in the descending order of $\omega_i(t+1)$ calculated in (4.1). The argument

is that larger $\omega_i(t+1)$ leads to larger expected payoff in next slot $t+1$. If the tie persists, then the channels are sorted by their indexes.

5.3 Optimality of Myopic Sensing Policy under Imperfect Sensing

In this section, we establish the closed-form conditions under which the myopic sensing policy achieves the system optimum under imperfect sensing. To this end, we study its structural property which is then used to establish the main result on the optimality.

5.3.1 Auxiliary Value Function

Armed with the three axioms, this section first derives a fundamental property of auxiliary value function, which is crucial in the study on the optimality of the myopic sensing policy.

Definition 5.2 (Auxiliary Value Function)**.** *The auxiliary value function, denoted as* $W_t(\Omega)$ *(*$t = 1, 2, \cdots, T$ *and* $t < r < T$*) is recursively defined as follows:*

$$\begin{cases} W_T(\Omega(T)) = F(\Omega_{\overline{A}}(T)); \\ W_r(\Omega(r)) = F(\Omega_{\overline{A}}(r)) + \beta \sum\limits_{\mathcal{E} \subseteq \overline{\mathcal{A}}(r)} Pr(\overline{\mathcal{A}}(r), \mathcal{E}) W_{r+1}(\Omega_{\mathcal{E}}(r+1)); \\ W_t(\Omega(t)) = F(\Omega_A(t)) + \beta \underbrace{\sum\limits_{\mathcal{E} \subseteq \mathcal{A}(t)} Pr(\mathcal{A}(t), \mathcal{E}) W_{t+1}(\Omega_{\mathcal{E}}(t+1))}_{\Gamma(\Omega(t))}. \end{cases} \tag{5.1}$$

where $\Omega_{\mathcal{E}}(t+1)$ *and* $\Omega_{\mathcal{E}}(r+1)$ *are generated by* $\langle \Omega(t), \mathcal{A}(t), \mathcal{E} \rangle$ *and* $\langle \Omega(r), \overline{\mathcal{A}}(r), \mathcal{E} \rangle$*, respectively, according to* (4.1)*. If* $\mathcal{A}(t) = \overline{\mathcal{A}}(t)$*, then* $W_t(\Omega(t))$ *is the total reward generated by the myopic sensing policy.*

Lemma 5.3. *If the expected reward function* $F(\Omega_A)$ *is g-regular, the correspondent auxiliary value function* $W_t(\Omega)$ *is symmetric in any two channel* $i, j \in \mathcal{A}$ *or* $i, j \notin \mathcal{A}$ *for all* $t = 1, 2, \cdots, T$*, i.e.,*

$$W_t(\omega_1, \cdots, \omega_i, \cdots, \omega_j, \cdots, \omega_N) = W_t(\omega_1, \cdots, \omega_j, \cdots, \omega_i, \cdots, \omega_N).$$

Proof. The proof follows the similar way as the proof of Lemma 2 in [10] and is omitted for briefty. □

Lemma 5.3 further implies that under the condition of Lemma 5.3, the auxiliary value function is robust against channel permutation given that all the permutated channels are sensed or none of them are sensed. Hence, it can be defined on the set Ω_A instead of the channel belief vector Ω, as stated in Corollary 5.1.

Corollary 5.1 (Robustness against Channel Permutation)**.** *Let Ω_A denote any permutation of the belief values of the elements in $\mathcal{A}$, if $F(\Omega_A)$ is symmetrical, it holds that $W_t(\Omega_A)$ has a unique value. In other words, $W_t(\Omega)$ is robust against channel permutation and thus can be defined as a function of Ω_A or $\mathcal{A}$.*

5.3.2 Myopic Sensing Policy: Condition of Optimality

In this subsection, we study the optimality of the myopic sensing policy. For the convenience of discussion, we firstly state some notation before presenting the analysis.

- $p_{11}^{max} \triangleq \max_{i\in\mathcal{N}} \{p_{11}^{(i)}\}$, $p_{01}^{min} \triangleq \max_{i\in\mathcal{N}} \{p_{01}^{(i)}\}$;
- $\delta_p^{max} \triangleq \max_{i\in\mathcal{N}} \{p_{11}^{(i)} - p_{01}^{(i)}\}$, $\delta_p^{min} \triangleq \min_{i\in\mathcal{N}} \{p_{11}^{(i)} - p_{01}^{(i)}\}$;
- $g'_{min} \triangleq \min_{p_{01}^{min}\le\omega\le p_{11}^{max}} \left\{\frac{\partial[g(\omega)]}{\partial\omega}\right\}$, $g'_{max} \triangleq \max_{p_{01}^{min}\le\omega\le p_{11}^{max}} \left\{\frac{\partial[g(\omega)]}{\partial\omega}\right\}$;
- Let $\omega_{-i} \triangleq \{\omega_j : j \in \mathcal{A}, j \neq i\}$ denote the believe vector except ω_i, and

$$\begin{cases} \Delta_{max} \triangleq \max\limits_{\omega_{-i}\in[0,1]^{N-1}} \big\{F(1,\omega_{-i}) - F(0,\omega_{-i})\big\}, \\ \Delta_{min} \triangleq \min\limits_{\omega_{-i}\in[0,1]^{N-1}} \big\{F(1,\omega_{-i}) - F(0,\omega_{-i})\big\}. \end{cases}$$

We start by showing the following important lemma (Lemma 5.4) and then establish the sufficient condition under which the optimality of the myopic sensing policy is ensured. In Lemma 5.4, we consider $\Omega_l = [\omega_1, \cdots, \omega_l, \cdots, \omega_N]$ and $\Omega'_l = [\omega_1, \cdots, \omega'_l, \cdots, \omega_N]$ which differ only in one element $\omega'_l \geq \omega_l$. Let $\mathcal{A}'$ and $\mathcal{A}$ denote the largest k elements in Ω'_l and Ω_l, respectively[3], Lemma 5.4 gives the upper and lower bounds of $W_t(\Omega_{A'}) - W_t(\Omega_A)$.

Lemma 5.4. *If the expected reward function F is g-regular, $\forall l \in \mathcal{N}$, $\omega_l \leq \omega'_l$ and $1 \leq t \leq T$, we have*

[3]The tie, if exists, is resolved in the way as stated in remark after Definition 4.3

1. *if* $l \in \mathcal{A}'$ *and* $l \in \mathcal{A}$, *then*

$$c \cdot (\omega_l' - \omega_l) g'_{min} \Delta_{min} \leq W_t(\Omega_{A'}) - W_t(\Omega_A) \leq c \cdot (\omega_l' - \omega_l) g'_{max} \Delta_{max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{max})^i;$$

2. *if* $l \notin \mathcal{A}'$ *and* $l \notin \mathcal{A}$, *then*

$$0 \leq W_t(\Omega_{A'}) - W_t(\Omega_A) \leq c \cdot (\omega_l' - \omega_l) g'_{max} \Delta_{max} \sum_{i=1}^{T-t} \beta^i (\delta_p^{max})^i;$$

3. *if* $l \in \mathcal{A}'$ *and* $l \notin \mathcal{A}$, *then*

$$0 \leq W_t(\Omega_{A'}) - W_t(\Omega_A) \leq c \cdot (\omega_l' - \omega_l) g'_{max} \Delta_{max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{max})^i.$$

Proof. The proof is given in the appendix. □

Remark. *It can be noted that there does not exist the case* $l \notin \mathcal{A}'$ *and* $l \in \mathcal{A}$ *according to the definition of the myopic sensing policy.*

In the following lemma, we consider $W_t(\Omega_{A_l})$ and $W_t(\Omega_{A_m})$ where $\mathcal{A}_l$ and $\mathcal{A}_m$ differ in one element ($l \in \mathcal{A}_l$ and $m \in \mathcal{A}_m$ and $\omega_l > \omega_m$). Lemma 5.5 establishes the sufficient condition under which $W_t(\Omega_{A_l}) > W_t(\Omega_{A_m})$ when F is g-regular.

Lemma 5.5. *If* $F(\Omega)$ *is* g*-regular and* $\dfrac{g'_{min}\Delta_{min}}{g'_{max}\Delta_{max}} \geq \sum_{i=1}^{T-1} \beta^i (\delta_p^{max})^i$, *then* $W_t(\Omega_{A_l}) \geq W_t(\Omega_{A_m})$ *holds for* $1 \leq t \leq T$.

Proof. Let Ω' denote the set of channel belief values in $\mathcal{A}_l$ with $\omega_l' = \omega_m$ and $\omega_i' = \omega_i$ for $\forall i \neq l$, apply Lemma 5.4, we have

$$\begin{aligned} & W_t(\Omega_{A_l}) - W_t(\Omega_{A_m}) = [W_t(\Omega_{A_l}) - W_t(\Omega')] - [W_t(\Omega_{A_m}) - W_t(\Omega')] \\ \geq \;& c \cdot (\omega_l - \omega_m) g'_{min} \Delta_{min} - c \cdot (\omega_l - \omega_m) g'_{max} \Delta_{max} \sum_{i=1}^{T-t} \beta^i (\delta_p^{max})^i \\ \geq \;& c \cdot (\omega_l - \omega_m) g'_{max} \Delta_{max} \cdot \left[\frac{g'_{min}}{g'_{max}} \cdot \frac{\Delta_{min}}{\Delta_{max}} - \sum_{i=1}^{T-1} \beta^i (\delta_p^{max})^i \right] \geq 0 \end{aligned}$$

if the conditions in the lemma hold. □

The following theorem studies the optimality of the myopic sensing policy under imperfect sensing.

Theorem 5.1. *The myopic sensing policy is optimal if the following two conditions hold: (1) the expected slot reward function F is g-regular; (2)* $\frac{g'_{min}\Delta_{min}}{g'_{max}\Delta_{max}} \geq \sum_{i=1}^{T-1} \beta^i (\delta_p^{max})^i$.

Proof. We prove the theorem by backward induction. The theorem holds trivially for $t = T$. Assume that it holds for $T, T-1, \cdots, t+1$, i.e., the optimal sensing policy is to sense the best k channels from time slot $t+1$ to T. We now show that it holds for t.

To this end, assume, by contradiction, that given the belief vector $\Omega \triangleq \{\omega_{i_1}, \cdots, \omega_{i_N}\}$, the optimal sensing policy is to sense the best k channels from time slot $t+1$ to T and at slot t to sense channels $\{i_1, \cdots, i_k\} \neq \{1, \cdots, k\}$, given that the latter contains the best k channels in terms of belief values at slot t. There must exist i_m and i_l where $m \leq k < l$ such that $\omega_{i_m} < \omega_k \leq \omega_{i_l}$. It then follows from Lemma 5.5 that

$$W_t^{\{i_1, i_2, \cdots, i_k\}}(\Omega) < W_t^{\{\omega_{i_1}, \omega_{i_2}, \cdots, \omega_{i_{m-1}}, \omega_{i_l}, \omega_{i_{m+1}}, \omega_{i_k}\}}(\Omega),$$

implying that sensing $\{i_1, \cdots, i_{m-1}, i_l, i_{m+1}, \cdots, i_k\}$ at slot t and then following the myopic sensing policy is better than sensing channels $\{i_1, \cdots, i_k\}$ at slot t and then following the myopic sensing policy, which contradicts with the assumption that the latter is the optimal sensing policy. This contradiction completes our proof. □

The following theorem further establishes the optimality conditions in asymptotic case $T \to \infty$. The proof follows straightforwardly from Theorem 5.1 by noticing that $\sum_{i=1}^{\infty} x^i = x/(1-x)$ for any $x \in (0, 1)$.

Theorem 5.2. *In the infinite horizon case $T \to \infty$, the myopic sensing policy is optimal if the following conditions hold: (1) the expected slot reward function F is g-regular; (2)* $\beta \leq \frac{g'_{min}\Delta_{min}}{(g'_{min}\Delta_{min} + g'_{max}\Delta_{max})\delta_p^{max}}$.

5.3.3 Discussion

For the technical perspective, compared with [10] and the previous chapter, we extend the third axiom in [10] to cover a much larger class of reward functions including the logarithmic and power functions. Moreover, the imperfect sensing leads to the non-linearity of system dynamic update, and thus the closed-form conditions of the optimality are non-trivially derived in the more general form. In essence, the RMAB with imperfect sensing is the same with that [10] (RMAB with perfect sensing) since the closed-form optimal conditions don't relate

with the parameter characterized by the imperfect sensing. This can be explained as follows: although the sensing error brings the non-linearity of the system dynamic update, this kind of non-linearity does not change the *decomposability* characteristics of the value function which serves as the cornerstone in deriving the closed-formed condition of optimality and therefore, the RMAB in [10] is isomorphic with our considered RMAB from the perspective of *decomposability*.

From the perspective of channel model, we consider the channel access problem where a user is limited to sensing k of N independently identical channels and gets one unit of reward if the sensed channel is in the good state, i.e., the utility function can be formulated as $F(\Omega_A) = (1-\epsilon)\sum_{i\in\mathcal{A}}\omega_i$. To that end, we apply Theorem 4.1 and have $\Delta_{min} = \Delta_{max} = 1-\epsilon$. We can then verify that when $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$, it holds that $\frac{\Delta_{min}}{\Delta_{max}\left[(1-\epsilon)(1-p_{01})+\frac{\epsilon(p_{11}-p_{01})}{1-(1-\epsilon)(p_{11}-p_{01})}\right]} > 1$. Therefore, when the condition 1 and 2 of Theorem 4.1 hold, the myopic sensing policy is always optimal for $0 \leq \beta \leq 1$, which significantly extends the result obtained in [48]. For the same scenario, we have $c = 1$, $g(\omega) = \omega$ and $\Delta_{min} = \Delta_{max} = 1-\epsilon$, and furthermore know that the myopic policy is optimal for $0 \leq \beta \leq 1$ without any constraint on ϵ if $\delta_p^{max} \leq 0.5$ according to Theorem 5.2. Compared with the optimal conditions in Theorem 4.1 with the independently identical channels, although both focusing on the optimality of the myopic policy, the closed-form conditions of optimality derived in this chapter are much stricter with respect to the transmission probabilities ($\delta_p^{max} \leq 0.5$, $0 \leq \epsilon < 1$) but much looser in the sensing error ($0 \leq \delta_p^{max} < 1$, $\epsilon < \frac{p_{01}(1-p_{11})}{P_{11}(1-p_{01})}$ in Theorem 4.1). The stricter constraint on the transmission probabilities is due to the proposed method itself which sacrifices part of the optimality to cover the case of the heterogeneous channels. On the other hand, since the analysis in this chapter does not rely on any particular sorting order of the channel list as in previous chapter (which itself relies on Lemma 4.2 with the condition on ϵ), the condition on ϵ is no more present here.

5.4 Conclusion

We have investigated the optimality of the myopic policy in the RMAB problem, which is of fundamental importance in many engineering applications. We have developed three axioms characterizing a family of generic and practically important functions which we refer to as g-regular functions. By performing a mathematical analysis based on the developed axioms, we have characterized the closed-form conditions under which the optimality of the myopic policy

is guaranteed.

5.5 Appendix

5.5.1 Proof of Lemma 5.4

We prove the lemma by backward induction.

For slot T, noticing that $W_T(\Omega) = F(\Omega_A)$ and that $g'_{min} \leq \frac{g(\omega)-g(\omega')}{\omega-\omega'} \leq g'_{max}$ for any $p_{01}^{min} \leq \omega' \leq \omega \leq p_{11}^{max}$, we have

1. For $l \in \mathcal{A}'$, $l \in \mathcal{A}$, it holds that

$$c\cdot(\omega'_l-\omega_l)g'_{min}\Delta_{min} \leq W_T(\Omega'_l)-W_T(\Omega_l) \leq c\cdot[g(\omega'_l)-g(\omega_l)]\Delta_{max} \leq c\cdot(\omega'_l-\omega_l)g'_{max}\Delta_{max};$$

2. For $l \notin \mathcal{A}'$, it holds that $l \notin \mathcal{A}$, $W_T(\Omega'_l) - W_T(\Omega_l) = 0$;

3. For $l \in \mathcal{A}'$, $l \notin \mathcal{A}$, it exists at least one channel m such that $\omega'_l \geq \omega_m \geq \omega_l$. It then holds that

$$\begin{aligned} 0 \leq c\cdot(\omega'_l-\omega_l)g'_{min}\Delta_{min} \leq W_T(\Omega'_l) - W_T(\Omega_l) &\leq c\cdot[g(\omega'_l)-g(\omega_m)]\Delta_{max} \\ &\leq c\cdot[g(\omega'_l)-g(\omega_l)]\Delta_{max} \leq c\cdot(\omega'_l-\omega_l)g'_{max}\Delta_{max}; \end{aligned}$$

Therefore, Lemma 5.4 holds for slot T.

Assume that Lemma 5.4 holds for $T, \cdots, t+1$. We now prove the lemma for slot t.

We first prove the first case: $l \in \mathcal{A}'$ and $l \in \mathcal{A}$. By rewriting $\Gamma(\Omega(t))$[4] in (5.1) and developing $\omega_l(t+1)$ in $\Omega(t+1)$, we have:

$$\Gamma(\Omega_{A'}) = (1-\epsilon)\omega'_l(t)\Gamma(\Omega^1_{A'}) + (1-(1-\epsilon)\omega'_l(t))\Gamma(\Omega_{A'}^{\varphi(\omega'_l)}) \tag{5.2}$$

$$\Gamma(\Omega_A) = (1-\epsilon)\omega_l(t)\Gamma(\Omega^1_A) + (1-(1-\epsilon)\omega_l(t))\Gamma(\Omega_A^{\varphi(\omega_l)}) \tag{5.3}$$

where, $\Omega^1_{A'}$ and $\Omega_{A'}^{\varphi(\omega'_l)}$ denote $\Omega_{A'}$ with $\omega'_l(t) = 1$ and $\varphi(\omega'_l)$, respectively, while Ω^1_A and $\Omega_A^{\varphi(\omega_l)}$ denote Ω_A with $\omega_l(t) = 1$ and $\varphi(\omega_l)$, respectively.

[4]Following Corollary 5.1, $\Gamma(\Omega(t))$ can also be expressed as a function of Ω_A.

Noticing $\Omega^1_{A'} = \Omega^1_A$, we have

$$\begin{aligned}
&\Gamma(\Omega_{A'}) - \Gamma(\Omega_A) \\
=&(1-\epsilon)(\omega_l'(t) - \omega_l(t))\big[\Gamma(\Omega^1_{A'}) - \Gamma(\Omega^{\varphi(\omega_l')}_{A'})\big] + (1-(1-\epsilon)\omega_l(t))\big[\Gamma(\Omega^{\varphi(\omega_l')}_{A'}) - \Gamma(\Omega^{\varphi(\omega_l)}_{A})\big]
\end{aligned}$$

Considering the whole realization of the belief vector, we further have

$$\begin{aligned}
&\Gamma(\Omega_l'(t)) - \Gamma(\Omega_l(t)) \\
=&\sum_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod_{i\in\mathcal{E}} (1-\epsilon)\omega_i(t) \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}} [1-(1-\epsilon)\omega_j(t)](\Gamma(\Omega_{A'}) - \Gamma(\Omega_A)) \\
=&\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}} \prod_{i\in\mathcal{E}} (1-\epsilon)\omega_i(t) \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}} [1-(1-\epsilon)\omega_j(t)] \\
&\cdot\Big[(1-\epsilon)(\omega_l'(t)-\omega_l(t))\big[W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1))\big] \\
&+(1-(1-\epsilon)\omega_l(t))\big[W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l)}(t+1))\big]\Big]
\end{aligned} \tag{5.4}$$

where, $\Omega_{l=a}(t+1)$ ($a \in \{1, \varphi(\omega_l'), \varphi(\omega_l)\}$) denotes the belief vector at slot $t+1$ under $\Omega(t)$ with $\omega_l(t+1) = \mathcal{T}_l(a)$.

Next, we derive the bound of $W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1))$ through three cases[5]:

Case 1: if $l \in \mathcal{A}'(t+1)$ and $l \in \mathcal{A}(t+1)$, according to the induction hypothesis, we have

$$\begin{aligned}
0 \le c\cdot(p_{11}^{(l)} - \mathcal{T}_l(\varphi(\omega_l')))g'_{min}\Delta_{min} \le W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) \\
\le c\cdot(p_{11}^{(l)} - \mathcal{T}_l(\varphi(\omega_l')))g'_{max}\Delta_{max}\sum_{i=0}^{T-t-1}\beta^i(\delta_p^{max})^i
\end{aligned}$$

Case 2: if $l \notin \mathcal{A}'(t+1)$ and $l \notin \mathcal{A}(t+1)$, according to the induction hypothesis, we have

$$0 \le W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) \le c\cdot(p_{11}^{(l)} - \mathcal{T}_l(\varphi(\omega_l')))g'_{max}\Delta_{max}\sum_{i=1}^{T-t-1}\beta^i(\delta_p^{max})^i$$

Case 3: if $l \in \mathcal{A}'(t+1)$ and $l \notin \mathcal{A}(t+1)$, according to the induction hypothesis, we have

$$0 \le W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) \le c\cdot(p_{11}^{(l)} - \mathcal{T}_l(\varphi(\omega_l')))g'_{max}\Delta_{max}\sum_{i=0}^{T-t-1}\beta^i(\delta_p^{max})^i$$

[5]It can be noted that the case $l \notin \mathcal{A}'(t+1)$ and $l \in \mathcal{A}(t+1)$ is impossible.

Combining the three cases, we obtain

$$\begin{aligned}
0 &\leq W_{t+1}(\Omega_{l=1}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) \\
&\leq c \cdot (p_{11}^{(l)} - \mathcal{T}_l(\varphi(\omega_l')))g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\
&= c \cdot \Big[1 - \frac{\epsilon\omega_l'}{1-(1-\epsilon)\omega_l'}\Big](p_{11}^{(l)} - p_{01}^{(l)})g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \qquad (5.5)
\end{aligned}$$

According to Lemma 4.1 and 4.2, we have $\mathcal{T}_l(\varphi(\omega_l')) \geq \mathcal{T}_l(\varphi(\omega_l))$ when $\omega_l' \geq \omega_l$. Thus we have the bounds of $W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l)}(t+1))$ by the similar induction as follows:

$$\begin{aligned}
0 &\leq W_{t+1}(\Omega_{l=\varphi(\omega_l')}(t+1)) - W_{t+1}(\Omega_{l=\varphi(\omega_l)}(t+1)) \\
&\leq c \cdot (\mathcal{T}_l(\varphi(\omega_l')) - \mathcal{T}_l(\varphi(\omega_l)))g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\
&= c \cdot \frac{\epsilon(\omega_l' - \omega_l)}{[1-(1-\epsilon)\omega_l'][1-(1-\epsilon)\omega_l]}(p_{11}^{(l)} - p_{01}^{(l)})g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i. \qquad (5.6)
\end{aligned}$$

Combining equations (5.4), (5.5) and (5.6) and recalling $p_{11}^{(l)} - p_{01}^{(l)} \leq \delta_p^{max}$, we have

$$0 \leq \Gamma(\Omega_l'(t)) - \Gamma(\Omega_l(t)) \leq c \cdot (\omega_l' - \omega_l)\delta_p^{max} g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i.$$

Since $W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) = F(\Omega_l'(t)) - F(\Omega_l(t)) + \beta(\Gamma(\Omega_l'(t)) - \Gamma(\Omega_l(t)))$ and $c \cdot (\omega_l' - \omega_l)g'_{min}\Delta_{min} \leq F(\Omega_l'(t)) - F(\Omega_l(t)) \leq c \cdot (\omega_l' - \omega_l)g'_{max}\Delta_{max}$, we have

$$\begin{aligned}
c \cdot (\omega_l' - \omega_l)g'_{min}\Delta_{min} &\leq W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) \\
&\leq c \cdot (\omega_l' - \omega_l)g'_{max}\Delta_{max} + \beta \cdot c \cdot (\omega_l' - \omega_l)g'_{max}\delta_p^{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\
&= c \cdot (\omega_l' - \omega_l)g'_{max}\Delta_{max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{max})^i.
\end{aligned}$$

We thus complete the proof of the first part ($l \in \mathcal{A}'$ and $l \in \mathcal{A}$) of Lemma 5.3.

Secondly, **we prove the second case $l \notin \mathcal{A}'$ and $l \notin \mathcal{A}$**. To this end, we have:

$$\begin{cases} \Gamma(\Omega_l(t)) = \sum\limits_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod\limits_{i\in\mathcal{E}}(1-\epsilon)\omega_i(t) \prod\limits_{j\in\mathcal{A}(t)\setminus\mathcal{E}} [1-(1-\epsilon)\omega_j(t)] W_{t+1}(\Omega_l(t+1)) \\ \Gamma(\Omega'_l(t)) = \sum\limits_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod\limits_{i\in\mathcal{E}}(1-\epsilon)\omega_i(t) \prod\limits_{j\in\mathcal{A}(t)\setminus\mathcal{E}} [1-(1-\epsilon)\omega_j(t)] W_{t+1}(\Omega'_l(t+1)) \end{cases},$$

where $\Omega_l(t+1)$ and $\Omega'_l(t+1)$ are the belief vector for slot $t+1$ generated by $\Omega_l(t)$ and $\Omega'_l(t)$ based on the belief update equation (4.1).

We distinguish the following four cases:

Case I. If channel l is never chosen for $\Omega_l(t+1)$ and $\Omega'_l(t+1)$ from the slot $t+1$ to the end of time horizon of interest T, that is to say, $l \notin \mathcal{A}'(r)$ and $l \notin \mathcal{A}(r)$ for $t+1 \leq r \leq T$, it is easy to know $\Gamma(\Omega'_l(t)) - \Gamma(\Omega_l(t)) = 0$, furthermore $W_t(\Omega'_l(t)) - W_t(\Omega_l(t)) = 0$;

Case II. There exists t^0 ($t+1 \leq t^0 \leq T$) such that $l \notin \mathcal{A}'(r)$ and $l \notin \mathcal{A}(r)$ for $t+1 \leq r \leq t^0-1$ while $l \notin \mathcal{A}'(t^0)$ and $l \in \mathcal{A}(t^0)$. For this case, it holds $\mathcal{A}'(r) = \mathcal{A}(r)$ for $t+1 \leq r \leq t^0-1$ while $\mathcal{A}'(r)$ and $\mathcal{A}(r)$ differ in one element, assume that $m \in \mathcal{A}'(t^0)$ and $m \notin \mathcal{A}(r)$. According to the definition of the myopic policy, it follows $\omega_l(t^0) \geq \omega_m(t^0)$ and $\omega'_l(t^0) \leq \omega_m(t^0)$, which leads to contradiction since $\omega'_l(t+1) = p_{11}^{(l)} > \omega_l(t+1) = p_{01}^{(l)}$ leads to $\omega'_l(t^0) > \omega_l(t^0)$ following Lemma 4.2. This case is thus impossible to happen;

Case III. There exists t^0 ($t+1 \leq t^0 \leq T$) such that $l \notin \mathcal{A}'(r)$ and $l \notin \mathcal{A}(r)$ for $t+1 \leq r \leq t^0-1$ while $l \in \mathcal{A}'(t^0)$ and $l \in \mathcal{A}(t^0)$. For this case, according to the hypothesis ($l \in \mathcal{A}'$ and $l \in \mathcal{A}$), we have

$$\begin{aligned} 0 \leq W_{t^0}(\Omega'_l(t^o)) - W_{t^0}(\Omega_l(t^o)) &\leq c\cdot(\omega'_l(t^o) - \omega_l(t^o)) g'_{max}\Delta_{max} \sum_{i=0}^{T-t^o} \beta^i (\delta_p^{max})^i \\ &= c\cdot(p_{11}^{(l)} - p_{01}^{(l)})^{t^o-t}(\omega'_l(t) - \omega_l(t)) g'_{max}\Delta_{max} \sum_{i=0}^{T-t^o} \beta^i (\delta_p^{max})^i. \end{aligned}$$

Noticing $t^0 \geq t+1$, we have

$$0 \leq W_{t+1}(\Omega'_l(t+1)) - W_{t+1}(\Omega_l(t+1)) \leq c\cdot(p_{11}^{(l)} - p_{01}^{(l)})(\omega'_l(t) - \omega_l(t)) g'_{max}\Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i.$$

Furthermore,

$$\begin{aligned}0 \leq W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) &\leq \beta \cdot c \cdot (p_{11}^{(l)} - p_{01}^{(l)})(\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\ &\leq \beta \cdot c \cdot \delta_p^{max} (\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\ &= c(\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=1}^{T-t} \beta^i (\delta_p^{max})^i.\end{aligned}$$

Case IV. There exists t^0 $(t+1 \leq t^0 \leq T)$ such that $l \notin \mathcal{A}'(r)$ and $l \notin \mathcal{A}(r)$ for $t+1 \leq r \leq t^0 - 1$ while $l \in \mathcal{A}'(t^0)$ and $l \notin \mathcal{A}(t^0)$. For this case, by the induction hypothesis ($l \in \mathcal{A}'$ and $l \notin \mathcal{A}$), we have

$$\begin{aligned}0 \leq W_{t^0}(\Omega_l'(t^o)) - W_{t^0}(\Omega_l(t^o)) &\leq c \cdot (\omega_l'(t^o) - \omega_l(t^o)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t^o} \beta^i (\delta_p^{max})^i \\ &= c \cdot (p_{11}^{(l)} - p_{01}^{(l)})^{t^o - t} (\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t^o} \beta^i (\delta_p^{max})^i.\end{aligned}$$

Noticing that $t+1 \leq t^o$, we have

$$0 \leq W_{t+1}(\Omega_l'(t+1)) - W_{t+1}(\Omega_l(t+1)) \leq c \cdot (\omega_l'(t) - \omega_l(t))(p_{11}^{(l)} - p_{01}^{(l)}) g'_{max} \Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i.$$

Therefore, we have

$$\begin{aligned}0 \leq W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) &\leq \beta \cdot c \cdot (\omega_l'(t) - \omega_l(t))(p_{11}^{(l)} - p_{01}^{(l)}) g'_{max} \Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\ &\leq \beta \cdot c \cdot (\omega_l'(t) - \omega_l(t)) \delta_p^{max} g'_{max} \Delta_{max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{max})^i \\ &= c(\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=1}^{T-t} \beta^i (\delta_p^{max})^i.\end{aligned}$$

Combining the above results, we complete the proof of the second part ($l \notin \mathcal{A}'$ and $l \notin \mathcal{A}$) of Lemma 5.3.

Last, **we prove the third case** $l \in \mathcal{A}'(t)$ **and** $l \notin \mathcal{A}(t)$. In this case, there must exist a

channel m such that $\omega_l' \geq \omega_m \geq \omega_l$ and $\omega_l' \in \mathcal{A}'$ and $\omega_m \in \mathcal{A}$. We then have

$$\begin{aligned} W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) =& W_t(\omega_1, \cdots, \omega_l', \cdots, \omega_N) - W_t(\omega_1, \cdots, \omega_l, \cdots, \omega_N) \\ =& W_t(\omega_1, \cdots, \omega_l', \cdots, \omega_N) - W_t(\omega_1, \cdots, \omega_l = \omega_m, \cdots, \omega_N) \\ &+ W_t(\omega_1, \cdots, \omega_l = \omega_m, \cdots, \omega_N) - W_t(\omega_1, \cdots, \omega_l, \cdots, \omega_N) \end{aligned} \quad (5.7)$$

According to the induction hypothesis ($l \in \mathcal{A}'$ and $l \in \mathcal{A}$), the first term of the right hand of (5.7) can be bounded as follows:

$$\begin{aligned} 0 \leq W_t(\omega_1, \cdots, \omega_l', \cdots, \omega_N) - W_t(\omega_1, \cdots, \omega_l = \omega_m, \cdots, \omega_N) \\ \leq c \cdot (\omega_l'(t) - \omega_m(t)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{max})^i \end{aligned} \quad (5.8)$$

Meanwhile, the second term of the right hand of (5.7) is bounded by induction hypothesis ($l \notin \mathcal{A}'$ and $l \notin \mathcal{A}$) as:

$$\begin{aligned} 0 \leq W_t(\omega_1, \cdots, \omega_l = \omega_m, \cdots, \omega_N) - W_t(\omega_1, \cdots, \omega_l, \cdots, \omega_N) \\ \leq c \cdot (\omega_m(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=1}^{T-t} \beta^i (\delta_p^{max})^i \end{aligned} \quad (5.9)$$

Therefore, we have, combining (5.7), (5.8) and (5.9),

$$0 \leq W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) \leq c \cdot (\omega_l'(t) - \omega_l(t)) g'_{max} \Delta_{max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{max})^i,$$

which completes the third part ($l \in \mathcal{A}'$ and $l \notin \mathcal{A}$) of Lemma 5.3.

Lemma 5.3 is thus proven.

Chapter 6

Beyond Myopic Sensing: a Heuristic ν-step Lookahead Policy

6.1 Introduction

In the previous chapters, we have studied the optimality of the myopic sensing policy in the case where the user is allowed to sense k out of N channels. In this chapter, we further investigate the more challenging problem where the user has to decide the number of channels to sense in order to maximize its utility [1]. This optimization problem hinges on the following tradeoff between exploitation and exploration: sensing more channels can help learn and predict the future channel state, thus increasing the long-term reward, but at the price of sacrificing the reward at current slot as sensing more channels reduces the time for data transmission, thus decreasing the throughput in the current slot. Therefore, to find the optimal number of channels to sense consists of striking a balance between the above exploitation and exploration. After showing the exponential complexity of the problem, we develop a heuristic ν-step lookahead policy which consists of sensing channels in a myopic way and stopping sensing when the expected aggregated utility from the current slot t to slot $t+\nu$ begins to decrease. In the developed policy, the parameter ν allows to achieve a desired tradeoff between social efficiency and computation complexity. We demonstrate the benefits of the proposed strategy via numerical experiments on several typical settings.

[1] 'Sense' should be understood generally, i.e., detecting, or probing etc. Herein, we use the terminology 'sense' for the consistency of narrative

6.2 Problem Formulation

6.2.1 System Model

We consider the same scenario as the previous chapters in which a user tries to access a multi-channel opportunistic communication system consisting of a set $\mathcal{N}$ of N channels, each given by a two state Markov chain with transition probabilities $\{p_{ij}^{(k)}\}_{i,j=0,1}$ $(1 \leq k \leq N)$. The system operates in a slotted fashion where the slots are indexed by t $(1 \leq t \leq T)$, where T is the time horizon of interest (or the user gives up accessing the system). Specifically, we assume that channels go through state transition at the beginning of a slot. The length of each time slot is denoted as Δ, which is further divided into two parts: the sensing phase and the transmission phase. Let $\delta = a\Delta$ $(a \leq 1)$ denote the time needed to sense one channel, the sensing phase lasts $na\Delta$ if the user senses n channels and the transmission phase consists of the rest of the time $(1 - an)\Delta$.

The user's objective is to maximize its throughput by choosing the appropriate set of channels to sense. Let $\mathcal{A}(t)$, $\mathcal{O}_A(t)$ denote the set of channels sensed and the set of sensing results $\mathcal{O}_A(t) = \{O_i(t) \in \{1, 0\}, i \in \mathcal{A}(t)\}$ by the user at slot t who can sense at most M $(1 \leq M < N$ and $aM \leq 1)$ channels for the limit of hardware and sensing constraint. If at least one of the sensed channel is in the good state, the user can successfully transmit one packet.[2] In our study, we also take into consideration the imperfect sensing which is characterized by the missed detection (the channel is sensed good but is in fact bad) rate denoted as ζ and the false alarm rate denoted as ϵ (the channel is sensed bad but is in fact good).

Obviously, by imperfectly sensing only $|\mathcal{A}(t)|$ out of N channels at each slot t, the user cannot observe the state information of the whole system. Hence, the user has to infer the channel states from its past decision and observation history so as to make its future decision. Moreover, the current sensing outcome further serves as statistics for future decision. To this end, we define the *channel state belief vector* (hereinafter referred to as *belief vector* for briefness) $\Omega(t) \triangleq \{\omega_i(t), i \in \mathcal{N}\}$, where $0 \leq \omega_i(t) \leq 1$ is the conditional probability that channel i is in good state. Given the sensing set $\mathcal{A}(t)$ and the detection outcomes $\{O_i(t) \in \{0, 1\} : i \in \mathcal{A}(t)\}$,

[2] Our work can be extended to the case where the user is equipped with more than one radio and can access multiple channels at a time.

the belief vector in $t+1$ slot can be updated recursively using Bayes Rule as shown in (6.1):

$$\omega_i(t+1) = \begin{cases} p_{11}^{(i)}, & i \in \mathcal{A}(t), O_i(t) = 1 \\ \mathcal{T}_i(\varphi(\omega_i(t))), & i \in \mathcal{A}(t), O_i(t) = 0 \\ \mathcal{T}_i(\omega_i(t)), & i \notin \mathcal{A}(t), \end{cases} \tag{6.1}$$

where,

$$\mathcal{T}_i(\omega_i(t)) \triangleq \omega_i(t)p_{11}^{(i)} + (1-\omega_i(t))p_{01}^{(i)}, \tag{6.2}$$

$$\varphi(\omega_i(t)) \triangleq \frac{\epsilon\omega_i(t)}{1-(1-\epsilon)\omega_i(t)}. \tag{6.3}$$

6.2.2 Optimal Sensing Problem Formulation: Exploitation vs Exploration

We are interested in the user's optimization problem to find a channel sensing policy π^* that maximize the expected total discounted reward over a finite horizon. Mathematically, a sensing policy π_t is defined as a mapping from the belief vector $\Omega(t)$ to $\mathcal{A}(t)$ in slot t:

$$\pi_t : \Omega(t) \to \mathcal{A}(t), 1 \leq |\mathcal{A}(t)| \leq M, t = 1, 2, \cdots, T.$$

The formal definition of the optimal sensing problem $\mathbf{P}$ is given as follows:

$$\mathbf{P}: \ \pi^* = \underset{\pi_t}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R(\pi_t(\Omega(t)), \mathcal{O}_A(t)) \middle| \Omega(1)\right], \tag{6.4}$$

where the slot reward function $R(\pi_t(\Omega(t)), \mathcal{O}_A(t)) = R(\mathcal{A}(t), \mathcal{O}_A(t))$ is the user throughput in slot t under the sensing policy π_t with the initial belief vector $\Omega(1)$[3], $0 \leq \beta \leq 1$ is the discount factor characterizing the feature that the future rewards are less valuable than the immediate reward.

Solving $\mathbf{P}$ is computationally heavy due to the fact that the belief vector $\{\Omega(t), t = 1, 2, \cdots, T\}$ is a Markov chain with uncountable state space when $T \to \infty$, resulting the difficulty in tracing the optimal sensing policy π^*. More specifically, $\mathbf{P}$ can be cast into a class of the RMAB problem with unknown number of arms to be activated. It is worth noting that the RMAB

[3]If no information on the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution $\omega_0^{(i)} = \frac{p_{01}^{(i)}}{1+p_{01}^{(i)}-p_{11}^{(i)}}$, $1 \leq i \leq N$.

problem is proved to be PSPACE-hard. Hence, a natural alternative to tackle **P** is to seek myopic sensing policy that maximizes the immediate reward. The motivation of focusing on the myopic sensing policy is two-fold:

- As demonstrated in our previous work, under certain conditions, the myopic sensing policy is ensured to be optimal.
- The myopic sensing policy has a simple and robust structure that makes it easy to implement in practice.

Existing studies on the myopic policy in the RMAB problem implicitly assume that the number of arms to activate (in the context of our work, the number of channels to sense) is fixed. A natural while crucial research problem is how many channels to sense at each time slot so as to maximize the expected total reward, which is the focus of our work presented in this chapter. We sort $\Omega(t)$ for each slot t in the descending order such that $\omega_1(t) \geq \omega_2(t) \geq \cdots \geq \omega_N(t)$ and thus form a channel list $l^0(t) = (1, 2, \cdots, N)$[4], the optimization problem on the number of channels to sense at each slot is formalized as follows:

$$\mathbf{P_1}: n_t^* = \underset{|\mathcal{A}(t)|}{\operatorname{argmax}} \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R(\mathcal{A}(t), \mathcal{O}_A(t)) \middle| \Omega(1)\right], \tag{6.5}$$

where, in slot t, the first $|\mathcal{A}(t)|$ channels are sensed, i.e., $\mathcal{A}(t) = \{1, \cdots, |\mathcal{A}(t)|\}$.

It is insightful to note that the optimization problem $\mathbf{P_1}$ on the number of channels to sense hinges on the following tradeoff between exploitation and exploration: sensing more channels can help learn and predict the future channel state, thus increasing the long-term reward, but at the price of sacrificing the reward at current slot as sensing more channels reduces the time for data transmission, thus decreasing the throughput in the current slot.

To conclude this subsection, we would like to point out that despite our focus on the opportunistic access problem of multi-channels communication system, the model formulation and the consequent analysis to solve the optimization problem can be generalized in the context of the RMAB problem and are readily applied in a variety of engineering fields such as object tracking, communication jamming and opportunistic packet scheduling. Therefore, the following description and the use of terms in the context of opportunistic spectrum access should be

[4]The initial order of list is determined by the initial availability probability of each channel: $\omega_1(1) \geq \omega_2(1) \geq \cdots \geq \omega_N(1) \Rightarrow l^0(1) = (1, 2, \cdots, N)$.

understood generically. Moreover, the slot reward function $R(\mathcal{A}(t), \mathcal{O}_A(t))$ that we adopt can be generically expressed in the normalized form as follows:

$$R(\mathcal{A}(t), \mathcal{O}_A(t)) = \begin{cases} 1 - C(|\mathcal{A}(t)|), & \text{if } \prod_{i \in \mathcal{A}(t)} (1 - O_i(t)) = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

where $C(|\mathcal{A}(t)|)$ is the cost function monotonously increasing in $|\mathcal{A}(t)|$, representing the time associated to channel sensing and frequency switching. The first line of the right hand side of (6.6) indicates that by sensing the channels in $\mathcal{A}(t)$ that contains at least one channel sensed good, the user obtains a payoff $1 - C(|\mathcal{A}(t)|)$. The second line indicates the case where none of the channels in $\mathcal{A}(t)$ is sensed good, the user obtains 0 as payoff. In the channel access model depicted in Subsection 6.2.1, by normalizing $\Delta = 1$, we have $C(|\mathcal{A}(t)|) = a|\mathcal{A}(t)|$.

6.3 When to Stop Sensing New Channels: the ν-step Lookahead Policy

It can be noticed that given a policy $\{n(t), 1 \leq t \leq T\}$ (i.e., the number of channels to sense at each slot, given the myopic sensing order), the belief vectors $\{\Omega(t), 1 \leq t \leq T\}$ form a Markov process with an uncountable state space, which makes the optimization problem $\mathbf{P_1}$ intractable. Therefore, we turn to the following heuristic strategy referred to as ν-step lookahead policy: at each slot t, the user senses the channels in the decreasing order of $\Omega(t)$ and estimates the expected accumulated payoff from slot $t+1$ to slot $t+\nu$ ($t+\nu \leq T$), assuming that in slots $t+1, \cdots, t+\nu$, the user stops exploring new channels once an available one is found (or the maximal number of channels to be sensed, M, is reached); the user stops sensing new channels when the sum of the reward in the current slot plus that from slot $t+1$ to $t+\nu$ decreases.

We now give the mathematic description of the ν-step lookahead policy. Let $l^k(t)$ and $\Omega^k(t)$ ($k \leq M$) denote the channel list and belief vector formed in the descending order of $\omega_i(t)$ ($1 \leq i \leq N$) after sensing the first k best channels in slot t, and $l^i_j(t)$ denote the jth channel in $l^i(t)$. Moreover, the key mathematical symbols used in our analysis are summarized in Table 6.1.

To better streamline our presentation, we introduce the pseudo cost function defined as

Table 6.1: Key mathematical symbols

β	discount factor
N	number of total channels
$\mathcal{N}$	channel set $\{1, 2, \cdots, N\}$
M	the maximum number of channel allowed to sense in a slot
$\mathcal{A}(t)$	the set of channels chosen at slot t
T	the horizon of time
Δ	the length of a slot
δ	the length of a mini-slot
ϵ	false alarm rate
ζ	miss detection rate
$O_i(t)$	sensing outcome of channel i at slot t
$\omega_i(t)$	probability of channel i in state 1 at slot t
$f_i(t)$	probability of channel i in state 0 at slot t
$\Omega^0(t)$	belief vector formed at the beginning of slot t
$\Omega^i(t)$	belief vector formed after sensing i channels at slot t
$l^0(t)$	channel list in descendent order of ω at the beginning of slot t
$l^i(t)$	channel list in descendent order of ω after sensing i channels at slot t
$\widetilde{l}^i(t)$	the reverse channel list of $l^i(t)$
$l^i_j(t)$	the jth channel in channel list $l^i(t)$
$\overleftarrow{l}^i(t)$	channel list formed in the descendent order of ω if the ith channel of $l^0(t)$ would be sensed good at slot t
$\overrightarrow{l}^i(t)$	channel list formed in the descendent order of ω if the ith channel of $l^0(t)$ would be sensed bad at slot t
$R(\mathcal{A}(t), \mathcal{O}_A(t))$	immediate reward under choosing $\mathcal{A}(t)$ channels at slot t and obtaining the sensing outcomes $\mathcal{O}_A(t)$
$q(\mathcal{A}(t), \mathcal{O}_A(t))$	immediate cost under choosing $\mathcal{A}(t)$ channels at slot t and obtaining the sensing outcomes $\mathcal{O}_A(t)$
$Q^{t+\nu}_{t+1}(\mathbb{T}(\Omega^i(t)))$	expected accumulative cost from slot $t+1$ to $t+\nu$ when i channels are sensed at slot t
$\overline{Q}^{t+\nu}_{t+1}(\Omega^i(t))$	expected accumulative cost from slot $t+1$ to $t+\nu$ if $i+1$ channel is sensed at slot t

follows:

$$q(\mathcal{A}(t), \mathcal{O}_A(t)) \triangleq 1 - R(\mathcal{A}(t), \mathcal{O}_A(t)) = \begin{cases} C(|\mathcal{A}(t)|) = a|\mathcal{A}(t)|, & \text{if } \prod_{i\in\mathcal{A}(t)}(1 - O_i(t)) = 0 \\ C_0 = 1, & \text{otherwise.} \end{cases} \tag{6.7}$$

The optimization problem $\mathbf{P_1}$ can be written as the following optimization problem $\mathbf{P_2}$:

$$\mathbf{P_2}: n_t^* = \underset{|\mathcal{A}(t)|}{\operatorname{argmin}} \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} q(\mathcal{A}(t), \mathcal{O}_A(t)) \middle| \Omega(1)\right]. \tag{6.8}$$

Given the initial belief vector $\Omega^0(t+1)$ at the beginning of slot $t+1$ (with the correspondent

channel list $l^0(t+1)$), let $Q_{t+1}^{t+\nu}(\Omega^0(t+1))$ denote the expected accumulative pseudo cost accrued from the beginning of slot $t+1$ to slot $t+\nu$, given that the user stops sensing once a channel is sensed good or M is reached, i.e.,

$$Q_{t+1}^{t+\nu}(\Omega^0(t+1)) \triangleq \sum_{i=1}^{M} \underbrace{\Big[\big[\omega_{l_i^0(t+1)}(t+1) \prod_{j=1}^{i-1} (1-\omega_{l_j^0(t+1)}(t+1))\big][C(i)+\beta \cdot Q_{t+2}^{t+\nu}(\mathbb{T}(\Omega_1^i(t+1)))] \Big]}_{A}$$
$$+ \underbrace{\prod_{j=1}^{M} (1-\omega_{l_j^0(t+1)}(t+1))[C_0+\beta \cdot Q_{t+2}^{t+\nu}(\mathbb{T}(\Omega_0^M(t+1)))]}_{B},$$

where term A denotes the pseudo cost when $l_i^0(t+1)$ channel is sensed good while $l_1^0(t+1), \cdots, l_{i-1}^0(t+1)$ channels are sensed bad; term B denotes the pseudo cost when the first M channels of $l^0(t+1)$ are sensed bad; $\Omega_1^i(t+1)$ and $\Omega_0^i(t+1)$ denote the belief vectors where the channel $l_i^0(t+1)$ is sensed good and bad, respectively; $\mathbb{T}$ denotes the mapping from $\Omega^k(t)$ to $\Omega^0(t+1)$ according to (6.1) at the beginning of slot $t+1$, i.e., $\mathbb{T}: \Omega^k(t) \to \Omega^0(t+1)$.

At each slot t, the ν-step lookahead policy solving $\mathbf{P_2}$ can be implemented in a heuristic approach by transforming it into an optimal stopping problem, i.e., the user stops sensing new channels when the sum of the reward in the current slot plus that from slot $t+1$ to $t+\nu$ decreases. Mathematically, the number of channels to sense in the ν-step lookahead policy, denoted as $\overline{n}(t)$, is as follows:

$$\overline{n}(t) = \inf \Big\{ |\mathcal{A}(t)| : C(\mathcal{A}(t), \mathcal{O}_A(t)) + \beta Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t)))$$
$$< C(\mathcal{A}'(t), \mathcal{O}_{A'}(t)) + \beta \overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t)), 1 \le |\mathcal{A}(t)| \le M \Big\}, \quad (6.9)$$

where $\mathcal{A}'(t) = \mathcal{A}(t) \cup \{l_{|\mathcal{A}(t)|+1}^0(t)\}$ denotes the best $|\mathcal{A}(t)|+1$ channels in $l^0(t)$, $Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t)))$ is the expected accumulative pseudo cost from slot $t+1$ to $t+\nu$ when the best $|\mathcal{A}(t)|$ channels of $l^0(t)$ are sensed, and $\overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t))$ denotes the expected accumulative pseudo cost from slot $t+1$ to $t+\nu$ when the $|\mathcal{A}(t)|+1$th channel of $l^0(t)$ is sensed good with probability $(1-\epsilon)\omega_{l_{|\mathcal{A}(t)|+1}^0(t)}(t)$ and bad with probability $1-(1-\epsilon)\omega_{l_{|\mathcal{A}(t)|+1}^0(t)}(t)$, i.e.,

$$\overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t)) \triangleq (1-\epsilon)\omega_{l_{|\mathcal{A}(t)|+1}^0(t)}(t) Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega_1^{|\mathcal{A}(t)|+1}(t)))$$
$$+ (1-(1-\epsilon)\omega_{l_{|\mathcal{A}(t)|+1}^0(t)}(t)) Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega_0^{|\mathcal{A}(t)|+1}(t))). \quad (6.10)$$

The following theorem further studies the structure of the ν-step lookahead policy by developing an optimal stopping algorithm to implement it.

Theorem 6.1. *The ν-step lookahead policy can be implemented by Algorithm 1. In each iteration of algorithm,*

- *the user continues to sense new channel if all the sensed channels are bad (exploration);*
- *if at least one channel is sensed good, the user stops sensing new channels if the expected pseudo cost increases by sensing a new channel (exploration).*

Algorithm 1 ν-step lookahead policy: executed for each slot t

Input: $\Omega^0(t)$, $l^0(t)$
Output: $\mathcal{A}(t)$
Initialization: $\mathcal{A}(t) = \emptyset$
while $|\mathcal{A}(t)| < M$ **do**
  Sense the $(|\mathcal{A}(t)|+1)$th channel in $l^0(t)$
  Add the sensed channel in $\mathcal{A}(t)$, i.e., $\mathcal{A}(t) \leftarrow \mathcal{A}(t) + l^0_{|\mathcal{A}(t)|+1}(t)$
  if At least one channel in $\mathcal{A}(t)$ is sensed good and the following inequality holds:

$$C(|\mathcal{A}(t)|) + \beta Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t)))) < C(|\mathcal{A}(t)|+1) + \beta \overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t)) \tag{6.11}$$

  then
    Terminate the algorithm by outputting $\mathcal{A}(t)$
  end if
end while

Proof. It suffices to show that to solve $\overline{n}(t)$ in (6.9), the user should:

- continue to sense new channel if all the sensed channels are bad;
- if at least one channel is sensed good, stop sensing new channels if the expected pseudo cost increases by sensing a new channel.

The first action is trivial to prove by noticing that by sensing a new channel

- the pseudo cost for the current slot t will remain the same if the new channel is sensed bad and will be smaller if the new channel is sensed good;
- the user gets better payoff in the future by exploring the system state.

We now show the second action. If the user stops at the current channel, it holds that

$$C(\mathcal{A}(t), \mathcal{O}_A(t)) + \beta Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t))) = C(|\mathcal{A}(t)|) + \beta Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t)))).$$

By sensing a new channel, on the expected pseudo cost, it holds that

$$C(\mathcal{A}'(t), \mathcal{O}_{A'}(t)) + \beta \overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t)) = C(|\mathcal{A}(t)|+1) + \beta \overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t)),$$

where $\mathcal{A}'(t) = \mathcal{A}(t) \cup \{l^0_{|\mathcal{A}(t)|+1}(t)\}$. It can be noted that (6.9) is equivalent to the condition

$$C(|\mathcal{A}(t)|) + \beta Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^{|\mathcal{A}(t)|}(t)))) < C(|\mathcal{A}(t)|+1) + \beta \overline{Q}_{t+1}^{t+\nu}(\Omega^{|\mathcal{A}(t)|}(t))$$

in Algorithm 1, which completes our proof. □

Remark. *It is insightful to note that the proposed ν-step lookahead policy can be decomposed into two steps:*

- *Exploitation: the user exploits the current available information $\Omega(t)$ in a greedy way so as to find a good channel;*
- *Exploration: once a good channel secured, the user proceeds to explore the system state space for long term gain.*

The second step (exploration) can be absent if all the M best channels are sensed bad or if exploring does not increase gain in the long term (i.e., the condition in Algorithm 1 does not hold even once).

To conclude this subsection, we note that the complexity of the algorithm implementing the ν-step lookahead policy lies in the computation of (6.11), whose complexity is exponential with ν. On the other hand, a larger ν leads to better performance of the lookahead policy. Hence, the user can tune the parameter ν to achieve a desired tradeoff between complexity and efficiency.

6.4 One-step Lookahead Policy

Having derived the algorithm implementing the proposed ν-step lookahead policy, in this section we focus on the system of i.i.d. channels and provide an mathematical analysis on the case where $\nu = 1$, i.e., the one-step lookahead policy. Our motivation of investigating this particular policy is two-fold:

- the study on the one-step lookahead policy can provide structural insights on the computation of the expected pseudo cost, which is the foundation of the ν-step lookahead policy. The general case $\nu > 1$ can be extended iteratively from the case $\nu = 1$;
- through extensive numerical experiments (please refer to Section 6.5), we observe that the benefit of the ν-step lookahead policy is most important in the case of $\nu = 1$ and then decreases gradually with the increase of ν; this observation, combined with the fact that the complexity of the ν-step lookahead policy increases exponentially with ν, motivates a more focused analysis on the one-step lookahead policy, which seems to be the most practical strategy in many scenarios;

Given the system model presented in Subsection 6.2.1, assume that the user has sensed k channels with at least one of them is in state good, recalling Algorithm 1, the condition to decide whether to sense channel $k+1$ in the channel list can be written as:

$$a > \beta\big[Q_{t+1}^{t+\nu}(\mathbb{T}(\Omega^k(t))) - \overline{Q}_{t+1}^{t+\nu}(\Omega^k(t))\big]. \tag{6.12}$$

In the subsequent analysis, we show how to compute $Q_{t+1}^{t+1}(\mathbb{T}(\Omega^k(t)))$ and $\overline{Q}_{t+1}^{t+1}(\Omega^k(t))$ in an efficient way. Before presenting the detailed analysis, the following lemma studies how the channel list should be updated when a new channel is sensed.

Lemma 6.1. *For a system with positively correlated homogeneous i.i.d. channels, if* $0 \leq \epsilon \leq \frac{(1-p_{11})p_{01}}{p_{11}(1-p_{01})}$*, the channel sensed good (bad) should be moved to the head (tail) of the old channel list to form the new channel list.*

Proof. Assume the old channel list is $l^k(t) = (\sigma_1, \cdots, \sigma_k, \cdots, \sigma_N,)$ at slot t. We thus have $p_{11} \geq \omega_{\sigma_1}(t) \geq \cdots \geq \omega_{\sigma_k}(t) \geq \cdots \geq \omega_{\sigma_N}(t) \geq p_{01}$. If channel σ_{k+1} is sensed good, then $\omega_{\sigma_{k+1}}(t) = 1$, and further $l^k(t) = (\sigma_{k+1}, \sigma_1, \cdots, \sigma_k, \sigma_{k+2}, \cdots, \sigma_N)$ according to the descending order of ω. If channel σ_{k+1} is sensed bad, then $\omega_{\sigma_{k+1}}(t) = \varphi(\omega_{\sigma_{k+1}}(t)) \leq p_{01}$, and further $l^k(t) = (\sigma_1, \cdots, \sigma_k, \sigma_{k+2}, \cdots, \sigma_N, \sigma_{k+1})$. □

Assume that the channel list at the beginning of slot t before sensing any channels is $l^0(t) = (1, 2, \cdots, N)$, sorted in the decreasing order of the belief values. Assume that among the k sensed channels $\{1, \cdots, k\}$, m ($m \geq 1$) channels are sensed good while $k-m$ are bad. It follows from Lemma 6.1 that m channels are moved to the head of the channel list and others to the

tail, thus forming the new channel list $l^k(t)$. We now show how to compute $Q_{t+1}^{t+1}(\mathbb{T}(\Omega^k(t)))$, $Q_{t+1}^{t+1}(\mathbb{T}(\Omega_1^{k+1}(t)))$ and $Q_{t+1}^{t+1}(\mathbb{T}(\Omega_0^{k+1}(t)))$ in the case of $m \geq 1$ so as to decide whether to sense channel $k+1$.

To make our analysis more tractable, we first define an auxiliary vector $\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$:

$$\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m) \triangleq \begin{pmatrix} 1 \\ X_1(\mathbb{T}(\Omega^k(t)), m) \\ X_2(\mathbb{T}(\Omega^k(t)), m) \\ X_3(\mathbb{T}(\Omega^k(t)), m+2) \\ X_4(\mathbb{T}(\Omega^k(t)), m+2) \end{pmatrix} \triangleq \begin{pmatrix} 1 \\ \prod_{j=1}^{m}(1-\omega_{l_j^k(t)}(t+1)) \\ 1+\sum_{i=1}^{m}\prod_{j=1}^{i}(1-\omega_{l_j^k(t)}(t+1)) \\ \prod_{j=m+2}^{M}(1-\omega_{l_j^k(t)}(t+1)) \\ \sum_{i=m+2}^{M}\prod_{j=m+2}^{i}(1-\omega_{l_j^k(t)}(t+1)) \end{pmatrix}.$$

The following lemma establishes an important structural property of $\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$ by showing that $\boldsymbol{X}(\mathbb{T}(\Omega^{k+1}(t)), m+1)$ can be recursively derived based on $\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$ in both cases where the channel $k+1$ is sensed good and bad, respectively.

Lemma 6.2. *The following recursive update on the auxiliary vector holds:*

- *If $k+1$ channel is sensed good, $\boldsymbol{X}(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = \boldsymbol{H_1} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$;*
- *If $k+1$ channel is sensed bad, $\boldsymbol{X}(\mathbb{T}(\Omega_0^{k+1}(t)), m+1) = \boldsymbol{H_2} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$,*

where

$$\boldsymbol{H_1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1-(1-\epsilon)p_{11} & 0 & 0 & 0 \\ 1 & 0 & 1-(1-\epsilon)p_{11} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} & 0 \\ -1 & 0 & 0 & 0 & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} \end{pmatrix},$$

$$\boldsymbol{H_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)} & 0 \\ -1 & 0 & 0 & \frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)} & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} \end{pmatrix}.$$

Proof. Please refer to the appendix for the detailed demonstration. □

The following theorem further shows that $Q_{t+1}^{t+1}(\mathbb{T}(\Omega^k(t)))$, $Q_{t+1}^{t+1}(\mathbb{T}(\Omega_1^{k+1}(t)))$ and $Q_{t+1}^{t+1}(\mathbb{T}(\Omega_0^{k+1}(t)))$ can be easily computed by using the auxiliary vector. Consequently, the one-step lookahead policy can be implemented in an efficient fashion by using the auxiliary vector, which can be updated recursively.

Theorem 6.2. *It holds that*

$$Q_{t+1}^{t+1}(\mathbb{T}(\Omega^k(t))) = a\Big[\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)+\boldsymbol{A_2}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\cdot\boldsymbol{A_3}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\Big] \quad (6.13)$$

$$Q_{t+1}^{t+1}(\mathbb{T}(\Omega_1^{k+1}(t))) = a\Big[\boldsymbol{A_4}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)+\boldsymbol{A_5}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\cdot\boldsymbol{A_6}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\Big] \quad (6.14)$$

$$Q_{t+1}^{t+1}(\mathbb{T}(\Omega_0^{k+1}(t))) = a\Big[\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)+\boldsymbol{A_7}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\cdot\boldsymbol{A_8}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\Big], \quad (6.15)$$

where,

$$\begin{aligned}
&\boldsymbol{A_1} = [0,0,1,0,0], && \boldsymbol{A_2} = [0,1-\omega_{l_{m+1}^k(t)}(t+1),0,0,0]\\
&\boldsymbol{A_3} = [1,0,0,\tfrac{1}{a}-M-1,1], && \boldsymbol{A_4} = [1,0,1-(1-\epsilon)p_{11},0,0]\\
&\boldsymbol{A_5} = [0,1-(1-\epsilon)p_{11},0,0,0], && \boldsymbol{A_6} = [0,0,0,\tfrac{1}{a}-M-1,1]\\
&\boldsymbol{A_7} = [0,1,0,0,0], && \boldsymbol{A_8} = [0,0,0,(\tfrac{1}{a}-M)(1-(1-\epsilon)\mathcal{T}(\varphi(\omega_{l_{m+1}^k(t)}(t)))),0].
\end{aligned}$$

Proof. Please refer to the appendix for the detailed demonstration. □

Recall Algorithm 1 and (6.13)–(6.15), it can be verified that the one-step lookahead policy has a linear computational complexity $O(M)$.

6.5 Numerical Experiments

In this section, we demonstrate some of the theoretical results derived in this chapter and gain further insight on the developed ν-step lookahead policy as well as the performance tradeoff hinging behind via a set of numerical experiments. Specifically, we present two typical scenarios, the homogeneous case with i.i.d. channels and the heterogeneous case with non i.i.d. channels. In both scenarios, we are interested in the performance in terms of average reward (throughput) of both the myopic policy discussed in previous chapters with fixed number of channels to sense and the ν-step lookahead policy. The results in this section provide a complementary

quantitative study on the performance of the ν-step lookahead policy, which is not explicitly addressed in the analytical part.

6.5.1 Homogeneous Case with i.i.d. Channels

We first consider a homogeneous system with $N = 8$ i.i.d. channels. The false alarm rate is set to $\epsilon = 0.02$. Each slot the user is allowed to sense at most $M = 3$ channels. The cost coefficient $a = 0.02$. The discount factor β is set to 1. The following two representative parameter settings, corresponding to a strongly and weakly correlated channel model respectively, are studied:

- Case 1: $p_{11} = 0.8$, $p_{01} = 0.2$;
- Case 2: $p_{11} = 0.5$, $p_{01} = 0.4$.

Figure 6.1 compares the average throughput of the myopic policy and the one-step lookahead policy for Case 1. For the myopic policy, the cases of $k = 1, 2, 3$ are simulated. From the results, it can be observed that after the stabilization, the one-step lookahead policy can further increase the throughput by approximately 8% w.r.t. the myopic policy with $k = 3$. The performance gain is more significant when compared to the myopic policy with $k = 1$ and 2. As analyzed in the theoretic analysis, this gain is due to the fact that the one-step lookahead policy can achieve a desired tradeoff between exploration and exploitation. This benefit in throughput is especially attractive given the low computation complexity of the one-step lookahead policy.

Figure 6.2 illustrates the same comparison for Case 2. It can be noticed from the results that the performance gain in Case 2 is less significant compared to Case 1. This can be explained by the fact that the channel correlation in Case 2 is less significant in time than Case 1 and consequently, the effect of prediction is less important.

We also run the same simulation by setting $\beta = 0.95$ in order to simulate a more conservative user. The results are shown in the following figures 6.3 and 6.4. Basically, we obtain the same finding.

We then proceed to study the performance of the ν-step lookahead policy in the case of $\nu > 1$. Figure 6.5–Figure 6.8 study the average throughput with $\nu = 1, 2, 3$ for Case 1 and Case 2, respectively. The following effects can be observed:

- For Case 1: statistically, the increase of ν does not enhance the performance gain;

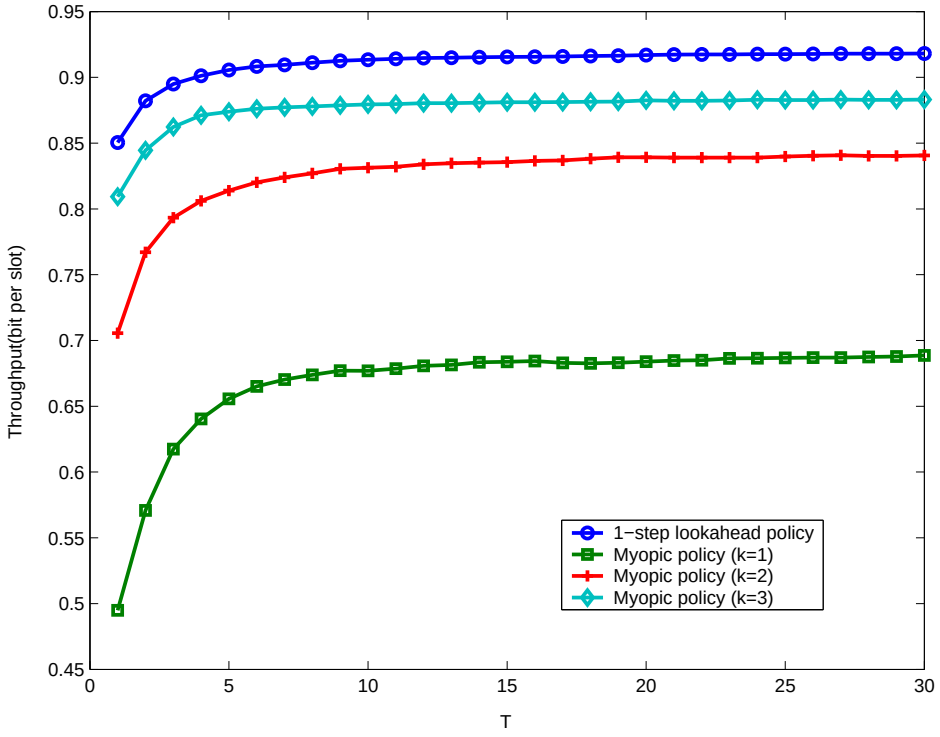

Figure 6.1: Throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 1, a = 0.02, \epsilon = 0.02, p_{11} = 0.8, p_{01} = 0.2$)

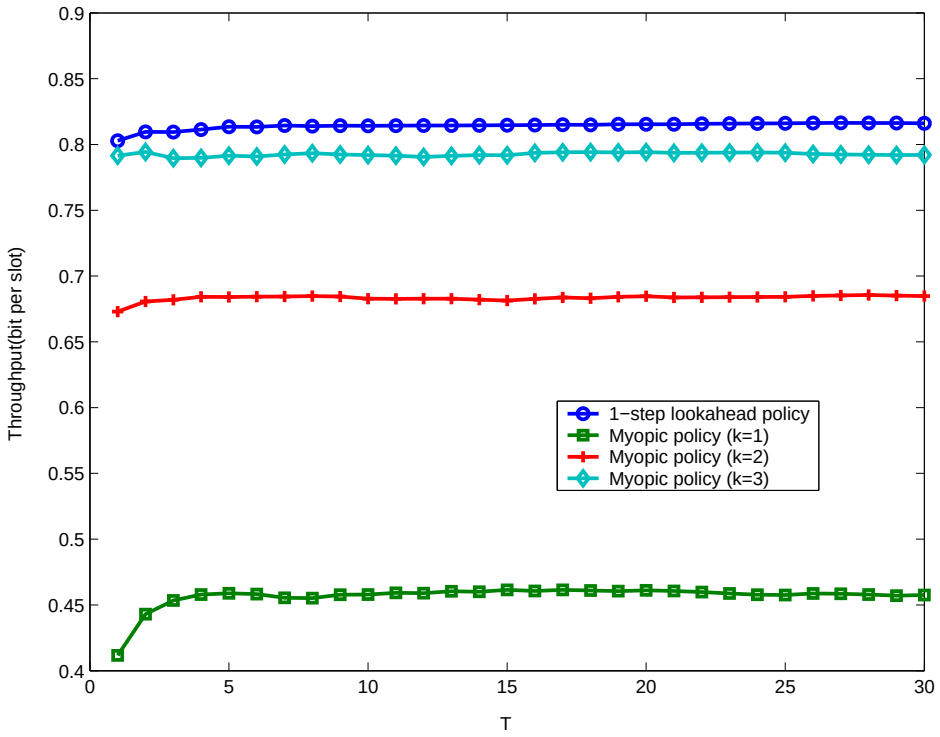

Figure 6.2: Throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 1, a = 0.02, \epsilon = 0.02, p_{11} = 0.5, p_{01} = 0.4$)

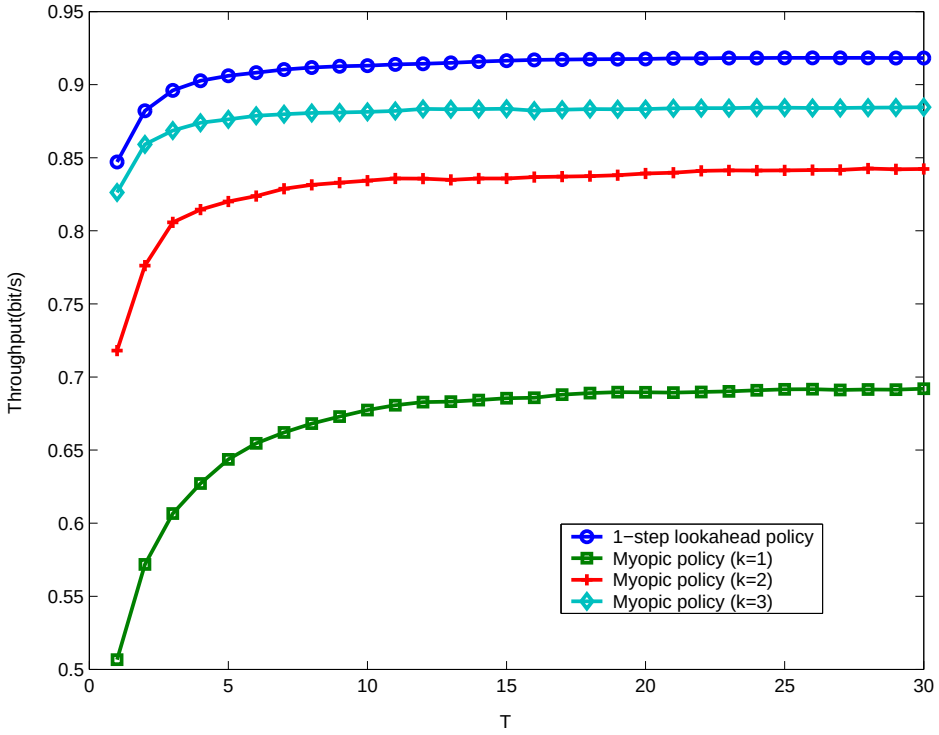

Figure 6.3: Throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 0.95, a = 0.02, \epsilon = 0.02, p_{11} = 0.8, p_{01} = 0.2$)

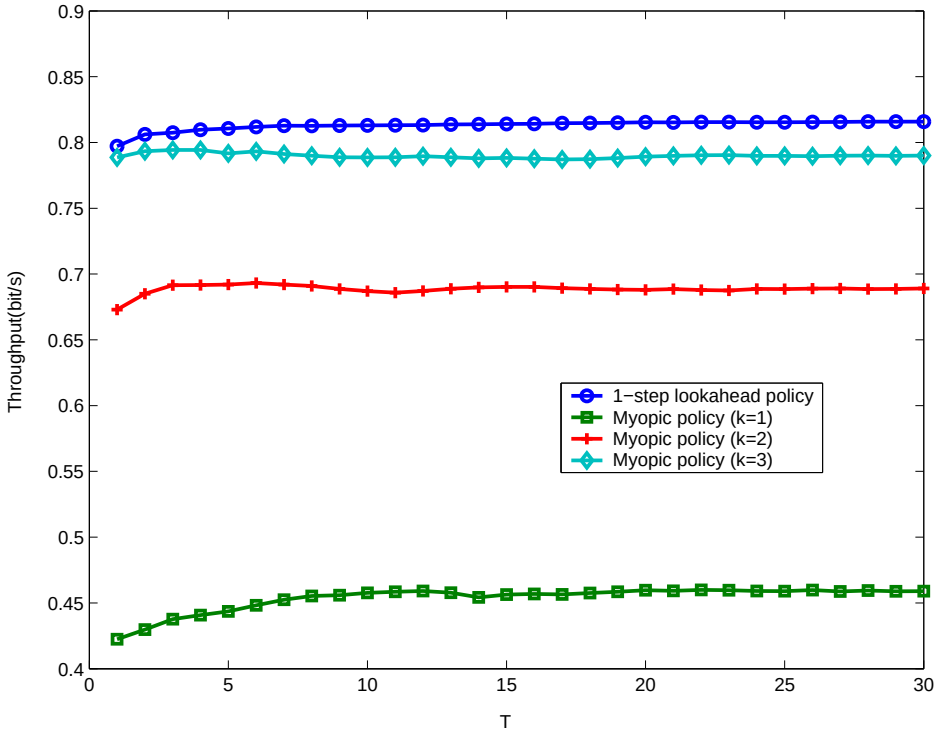

Figure 6.4: Throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 0.95, a = 0.02, \epsilon = 0.02, p_{11} = 0.5, p_{01} = 0.4$)

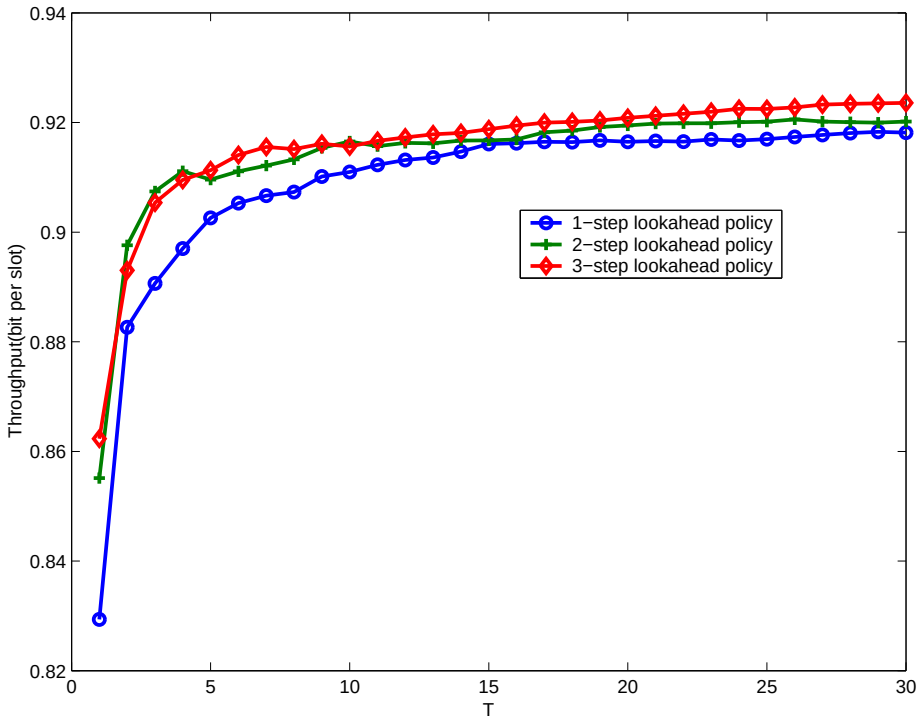

Figure 6.5: Throughput comparison of 1-,2-,3-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 1, a = 0.02, \epsilon = 0.02, p_{11} = 0.8, p_{01} = 0.2$)

- For Case 2: increasing ν from 1 to 2 slightly improve the average throughput (less than 0.5%), but further increasing ν cannot bring more benefit.

The above findings justify our focus on the one-step lookahead policy. More generically, by taking the complexity into account, we recommend to set $\nu = 1$ in a large variate of parameter settings.

6.5.2 Heterogeneous Case with non i.i.d. Channels

We now proceed to evaluate the performance of the ν-step lookahead policy in heterogeneous systems with non i.i.d. channels. To this end, we randomly generate 100 heterogeneous systems with the following parameter setting: $N = 8$, $M = 3$, $a = 0.02$, $\epsilon = 0.02$, and $p_{11}^{(i)} > p_{01}^{(i)}$ $(1 \leq i \leq N)$. We plot the average throughput in Figure 6.9 for $\beta = 1$ and Figure 6.10 for $\beta = 0.95$. Again, we observe similar results as that of homogenous systems, namely, the ν-step lookahead policy statistically outperforms the myopic policy in all parameter settings, and increasing ν does not bring significant throughput gain.

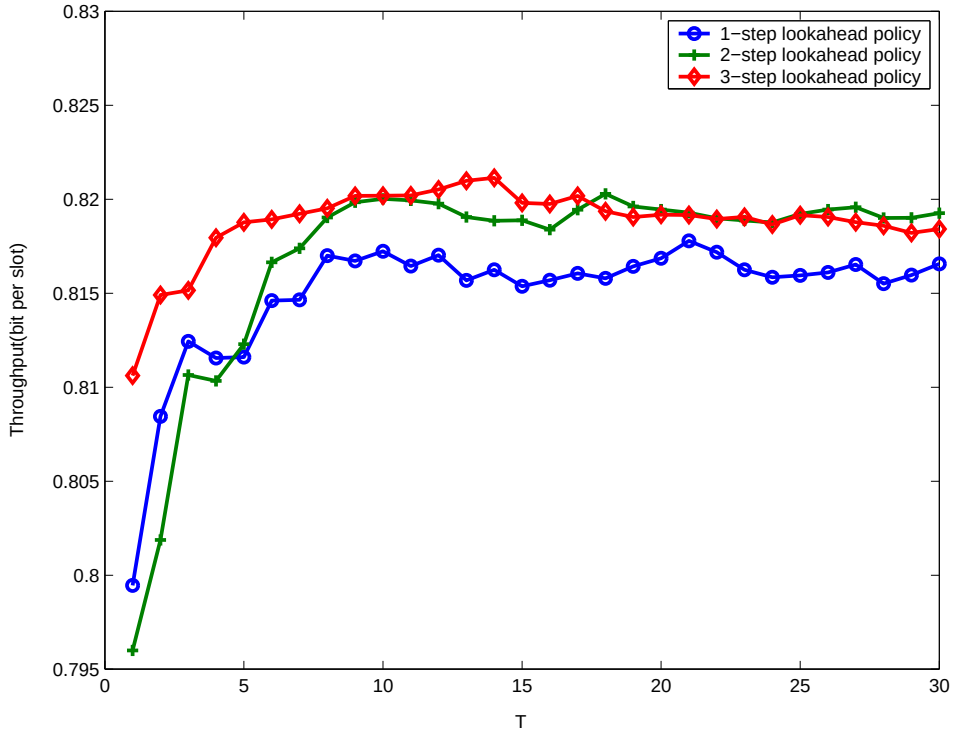

Figure 6.6: Throughput comparison of 1-,2-,3-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 1, a = 0.02, \epsilon = 0.02, p_{11} = 0.5, p_{01} = 0.4$)

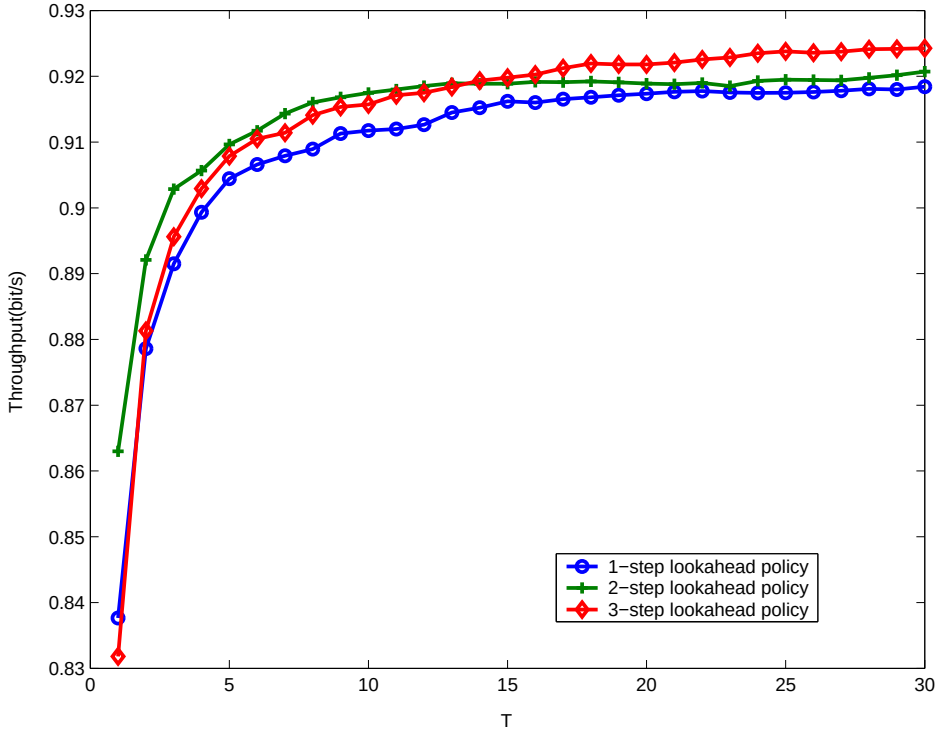

Figure 6.7: Throughput comparison of 1-,2-,3-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 0.95, a = 0.02, \epsilon = 0.02, p_{11} = 0.8, p_{01} = 0.2$)

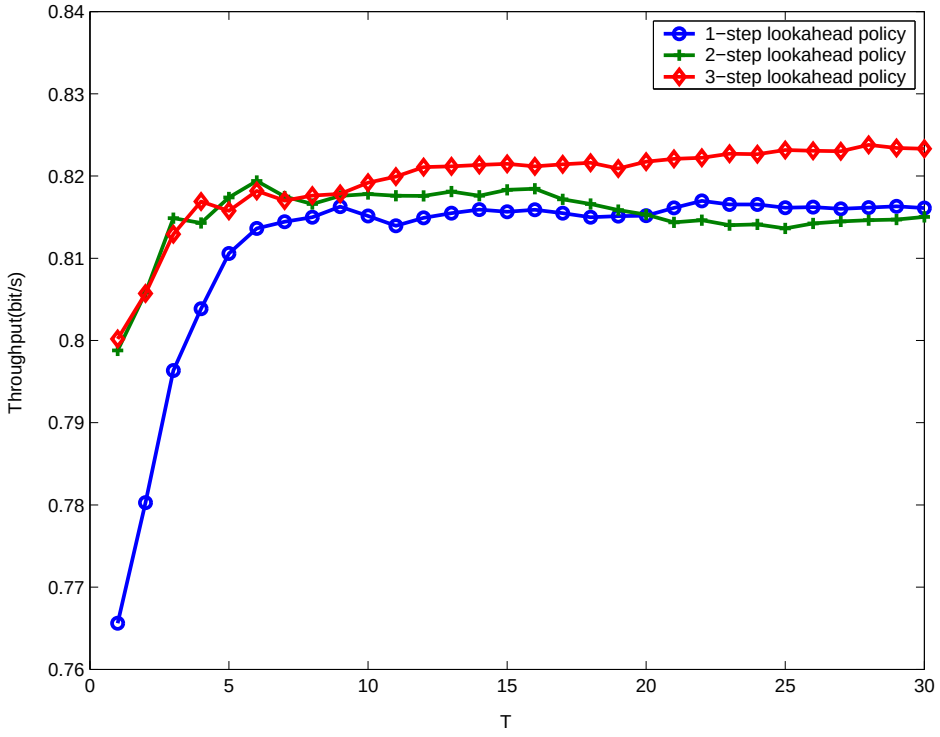

Figure 6.8: Throughput comparison of 1-,2-,3-step lookahead policy for homogeneous channels ($N = 8, M = 3, \beta = 0.95, a = 0.02, \epsilon = 0.02, p_{11} = 0.5, p_{01} = 0.4$)

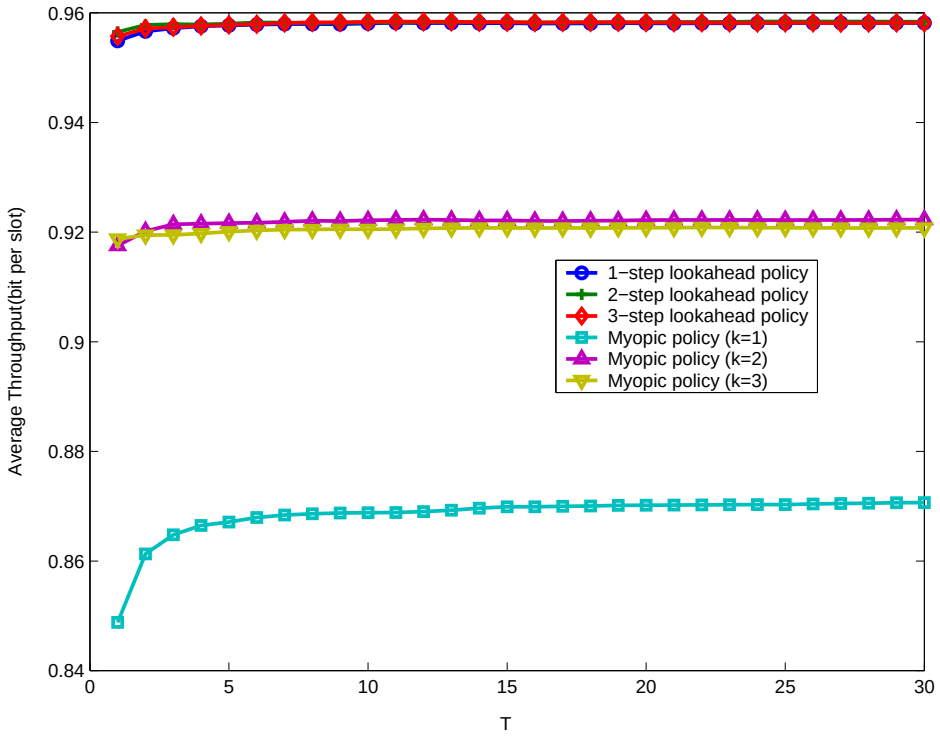

Figure 6.9: Average throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-,2-,3-step lookahead policy for heterogeneous channels ($N = 8, M = 3, \beta = 1, a = 0.02, \epsilon = 0.02$)

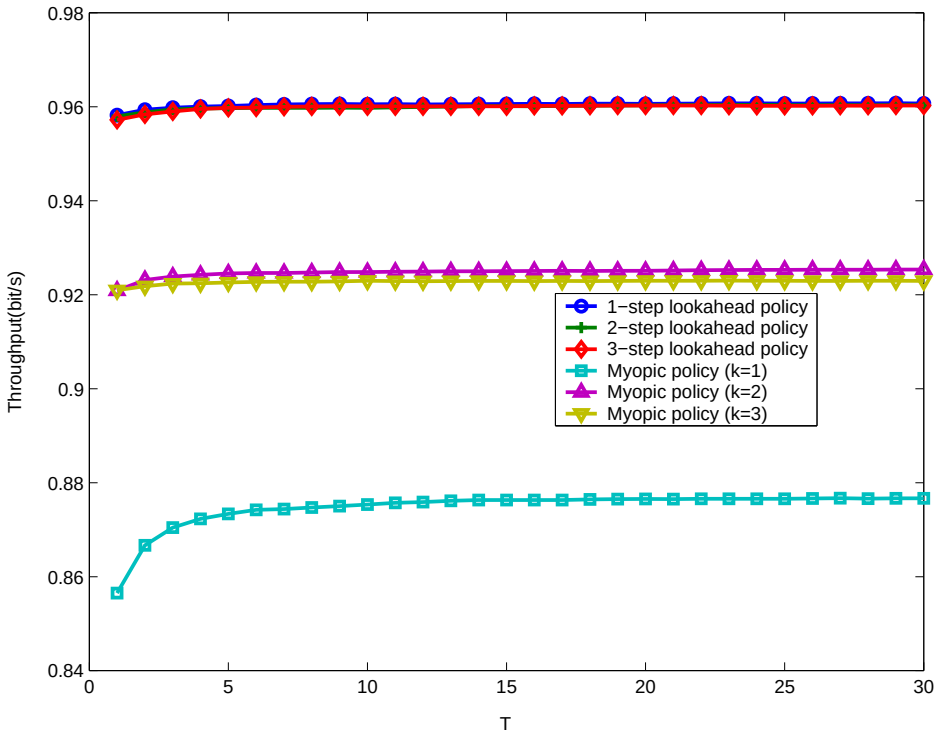

Figure 6.10: Average throughput comparison of myopic policies ($k = 1, 2, 3$) and 1-,2-,3-step lookahead policy for heterogeneous channels ($N = 8, M = 3, \beta = 0.95, a = 0.02, \epsilon = 0.02$)

6.6 Conclusion

In this chapter, we study the optimization problem where the user has to decide the number of channels to sense in order to maximize its utility. Given the exponential complexity of the problem, we develop a heuristic ν-step lookahead policy which consists of sensing channels in a myopic way and stopping sensing when the expected aggregated utility from the current slot t to slot $t + \nu$ begins to decrease. In the developed policy, the parameter ν allows to achieve a desired tradeoff between social efficiency and computation complexity. We demonstrate the benefits of the proposed strategy via numerical experiments on several typical settings.

6.7 Appendix

6.7.1 Proof of Lemma 6.2

Case 1. When channel $l^k_{m+1}(t)$ is sensed good, we have $\omega_{l^k_{m+1}(t)}(t+1) = (1-\epsilon)p_{11}$ according to (6.1) and false alarm rate. Recalling the definition of X_i $(i = 1, 2, 3, 4)$, we have

$$\begin{cases} X_1(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = [1 - \omega_{l^k_{m+1}(t)}(t+1)]X_1(\mathbb{T}(\Omega^k(t)), m), \\ X_2(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = 1 + [1 - \omega_{l^k_{m+1}(t)}(t+1)]X_2(\mathbb{T}(\Omega^k(t)), m), \\ X_3(\mathbb{T}(\Omega_1^{k+1}(t)), m+3) = \frac{X_3(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l^k_{m+2}(t)}(t+1)}, \\ X_4(\mathbb{T}(\Omega_1^{k+1}(t)), m+3) = \frac{X_4(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l^k_{m+2}(t)}(t+1)} - 1. \end{cases}$$

It is straightforward to verify that

$$\boldsymbol{X}(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = \boldsymbol{H_1} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m),$$

where

$$\boldsymbol{H_1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1-(1-\epsilon)p_{11} & 0 & 0 & 0 \\ 1 & 0 & 1-(1-\epsilon)p_{11} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} & 0 \\ -1 & 0 & 0 & 0 & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} \end{pmatrix}.$$

Case 2. When channel $l^k_{m+1}(t)$ is sensed bad, we have $\omega_{l^k_{m+1}(t)}(t+1) = (1-\epsilon)\mathcal{T}(\varphi(\omega_{l^k_{m+1}(t)}(t)))$ according to (6.1) and false alarm rate. Note, if $M = N$, we have $\omega_{l^k_{M+1}(t)}(t+1) = \omega_{l^k_{m+1}(t)}(t+1)$

according to Lemma 6.1. Recalling the definition of X_i $(i = 1, 2, 3, 4)$, we have

$$\begin{cases} X_1(\mathbb{T}(\Omega_0^{k+1}(t)), m) = X_1(\mathbb{T}(\Omega^k(t)), m), \\ X_2(\mathbb{T}(\Omega_0^{k+1}(t)), m) = X_2(\mathbb{T}(\Omega^k(t)), m), \\ X_3(\mathbb{T}(\Omega_0^{k+1}(t)), m+2) = X_3(\mathbb{T}(\Omega^k(t)), m+2)\frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)}, \\ X_4(\mathbb{T}(\Omega_0^{k+1}(t)), m+2) = \frac{X_4(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l^k_{m+2}(t)}(t+1)} - 1 + X_3(\mathbb{T}(\Omega^k(t)), m+2)\frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)}. \end{cases}$$

It is straightforward to verify that

$$\boldsymbol{X}(\mathbb{T}(\Omega_0^{k+1}(t)), m+1) = \boldsymbol{H_2} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m),$$

where

$$\boldsymbol{H_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)} & 0 \\ -1 & 0 & 0 & \frac{1-\omega_{l^k_{M+1}(t)}(t+1)}{1-\omega_{l^k_{m+2}(t)}(t+1)} & \frac{1}{1-\omega_{l^k_{m+2}(t)}(t+1)} \end{pmatrix}.$$

6.7.2 Proof of Theorem 6.2

Assume that m and $k-m$ channels of k channels are sensed good and bad respectively, and thus $l^k(t)$ is obtained.

Case 1. If the user does not sense channel $l^k_{m+1}(t)$, we have by separating the channel $l^k_{m+1}(t)$ from others

$$\begin{aligned} &Q^{t+1}_{t+1}(\mathbb{T}(\Omega^k(t))) = \sum_{i=1}^{M} C(i)\omega_{l^k_i(t)}(t+1)\prod_{j=1}^{i-1} f_{l^k_j(t)}(t+1) + \prod_{j=1}^{M} f_{l^k_j(t)}(t+1) \\ =&\frac{\delta}{\Delta}\sum_{i=1}^{m} i\omega_{l^k_i(t)}(t+1)\prod_{j=1}^{i-1} f_{l^k_j(t)}(t+1) + \frac{\delta}{\Delta}(m+1)\omega_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m} f_{l^k_j(t)}(t+1) \\ &+ \frac{\delta}{\Delta} f_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m} f_{l^k_j(t)}(t+1)\sum_{i=m+2}^{M} i\omega_{l^k_i(t)}(t+1)\prod_{j=m+2}^{i-1} f_{l^k_j(t)}(t+1) + \prod_{j=1}^{M} f_{l^k_j(t)}(t+1) \\ =&\frac{\delta}{\Delta}\Big[1 + f_{l^k_1(t)}(t+1) + \cdots + f_{l^k_1(t)}(t+1) \times \cdots \times f_{l^k_{m-1}(t)}(t+1) - m f_{l^k_1(t)}(t+1) \times \cdots \times f_{l^k_m(t)}(t+1)\Big] \\ &+ \frac{\delta}{\Delta}(m+1)\omega_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m} f_{l^k_j(t)}(t+1) \end{aligned}$$

$$
\begin{aligned}
&+\frac{\delta}{\Delta}f_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m}f_{l^k_j(t)}(t+1)\times\Big[(m+2)+f_{l^k_{m+2}(t)}(t+1)+\cdots\\
&+f_{l^k_{m+2}(t)}(t+1)\times\cdots\times f_{l^k_{M-1}(t)}(t+1)-Mf_{l^k_{m+2}(t)}(t+1)\times\cdots\times f_{l^k_M(t)}(t+1)\Big]\\
&+\prod_{j=1}^{M}f_{l^k_j(t)}(t+1)\\
=&\frac{\delta}{\Delta}X_2(\mathbb{T}(\Omega^k(t)),m)+\frac{\delta}{\Delta}(m+1)(\omega_{l^k_{m+1}(t)}(t+1)-1)\prod_{j=1}^{m}f_{l^k_j(t)}(t+1)\\
&+\frac{\delta}{\Delta}f_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m}f_{l^k_j(t)}(t+1)\times\Big[(m+2)+X_4(\mathbb{T}(\Omega^k(t)),m+2)\\
&+(\frac{\Delta}{\delta}-M-1)X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big]\\
=&\frac{\delta}{\Delta}X_2(\mathbb{T}(\Omega^k(t)),m)-\frac{\delta}{\Delta}(m+1)f_{l^k_{m+1}(t)}(t+1)X_1(\mathbb{T}(\Omega^k(t)),m)\\
&+\frac{\delta}{\Delta}f_{l^k_{m+1}(t)}(t+1)X_1(\mathbb{T}(\Omega^k(t)),m)\times\Big[(m+2)+X_4(\mathbb{T}(\Omega^k(t)),m+2)\\
&+(\frac{\Delta}{\delta}-M-1)X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big]\\
=&\frac{\delta}{\Delta}X_2(\mathbb{T}(\Omega^k(t)),m)+\frac{\delta}{\Delta}f_{l^k_{m+1}(t)}(t+1)X_1(\mathbb{T}(\Omega^k(t)),m)\times\Big[1\\
&+X_4(\mathbb{T}(\Omega^k(t)),m+2)+(\frac{\Delta}{\delta}-M-1)X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big]\\
=&\frac{\delta}{\Delta}\Big\{\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)+\boldsymbol{A_2}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\cdot\boldsymbol{A_3}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\Big\}.
\end{aligned}
$$

Case 2. If the channel $l^k_{m+1}(t)$ ($l^0_{k+1}(t)$) is sensed good, then we have by separating the channel $l^k_{m+1}(t)$ from others

$$
\begin{aligned}
&Q^{t+1}_{t+1}(\mathbb{T}(\Omega^{k+1}_1(t)))=\sum_{i=1}^{M}C(i)\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=1}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)+\prod_{j=1}^{M}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\\
=&\frac{\delta}{\Delta}\Big[1-f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)+f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)\sum_{i=2}^{M}i\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=2}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\Big]+\prod_{j=1}^{M}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\\
=&\frac{\delta}{\Delta}[1-f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)]+\frac{\delta}{\Delta}f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)*\Big[\sum_{i=2}^{m+1}i\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=2}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\\
&+\sum_{i=m+2}^{M}i\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=2}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\Big]+\prod_{j=1}^{M}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\\
=&\frac{\delta}{\Delta}[1-f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)]+\frac{\delta}{\Delta}f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)*\sum_{i=2}^{m+1}i\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=2}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\\
&+\frac{\delta}{\Delta}f_{\overleftarrow{l}^{k+1}_1(t)}(t+1)*\prod_{j=2}^{m+1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)*\Big[\sum_{i=m+2}^{M}i\omega_{\overleftarrow{l}^{k+1}_i(t)}(t+1)\prod_{j=m+2}^{i-1}f_{\overleftarrow{l}^{k+1}_j(t)}(t+1)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&+ f_{\overleftarrow{l}_1^{k+1}(t)}(t+1)\prod_{j=2}^{m+1} f_{\overleftarrow{l}_j^{k+1}(t)}(t+1) * \prod_{j=m+2}^{M} f_{\overleftarrow{l}_j^{k+1}(t)}(t+1)\\
=&\frac{\delta}{\Delta}[1 - f_{\overleftarrow{l}_1^{k+1}(t)}(t+1)] + \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * \sum_{i=1}^{m}(i+1)\omega_{l_i^k(t)}(t+1)\prod_{j=1}^{i-1} f_{l_j^k(t)}(t+1)\\
&+ \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * \prod_{j=1}^{m} f_{l_j^k(t)}(t+1) * \Big[\sum_{i=m+2}^{M} i\omega_{l_i^k(t)}(t+1)\prod_{j=m+2}^{i-1} f_{l_j^k(t)}(t+1)\Big]\\
&+ f_{\overleftarrow{l}_1^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{l_j^k(t)}(t+1) * \prod_{j=m+2}^{M} f_{l_j^k(t)}(t+1)\\
=&\frac{\delta}{\Delta}[1 - f_{\overleftarrow{l}_1^{k+1}(t)}(t+1)] + \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * \Big[1 + X_2(\mathbb{T}(\Omega^k(t)), m) - (m+2)X_1(\mathbb{T}(\Omega^k(t)), m)\Big]\\
&+ \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * X_1(\mathbb{T}(\Omega^k(t)), m) * \Big[(m+2) + X_4(\mathbb{T}(\Omega^k(t)), m+2)\\
&+ (\frac{\Delta}{\delta} - M - 1)X_3(\mathbb{T}(\Omega^k(t)), m+2)\Big]\\
=&\frac{\delta}{\Delta} + \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * X_2(\mathbb{T}(\Omega^k(t)), m) + \frac{\delta}{\Delta} f_{\overleftarrow{l}_1^{k+1}(t)}(t+1) * X_1(\mathbb{T}(\Omega^k(t)), m) *\\
&\Big[X_4(\mathbb{T}(\Omega^k(t)), m+2) + (\frac{\Delta}{\delta} - M - 1)X_3(\mathbb{T}(\Omega^k(t)), m+2)\Big]\\
=&\frac{\delta}{\Delta}\Big\{\boldsymbol{A_4}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m) + \boldsymbol{A_5}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\cdot\boldsymbol{A_6}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big\}
\end{aligned}
$$

Case 3. If the channel $l_{m+1}^k(t)$ ($l_{k+1}^0(t)$) is sensed bad, then we have by separating the channel $l_{m+1}^k(t)$ from others

$$
\begin{aligned}
&Q_{t+1}^{t+1}(\mathbb{T}(\Omega_0^{k+1}(t))) = \sum_{i=1}^{M} C(i)\omega_{\overrightarrow{l}_i^{k+1}(t)}(t+1)\prod_{j=1}^{i-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) + \prod_{j=1}^{M} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
=&\frac{\delta}{\Delta}\sum_{i=1}^{m} i\omega_{\overrightarrow{l}_i^{k+1}(t)}(t+1)\prod_{j=1}^{i-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) + \frac{\delta}{\Delta}\sum_{i=m+1}^{M-1} i\omega_{\overrightarrow{l}_i^{k+1}(t)}(t+1)\prod_{j=1}^{i-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
&+ \frac{\delta}{\Delta} M * \omega_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
&+ f_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
=&\frac{\delta}{\Delta}\sum_{i=1}^{m} i\omega_{\overrightarrow{l}_i^{k+1}(t)}(t+1)\prod_{j=1}^{i-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
&+ \frac{\delta}{\Delta}\prod_{j=1}^{m} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\sum_{i=m+1}^{M-1} i\omega_{\overrightarrow{l}_i^{k+1}(t)}(t+1)\prod_{j=m+1}^{i-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
&+ \frac{\delta}{\Delta} M * \omega_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)
\end{aligned}
$$

$$
\begin{aligned}
&+ f_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{\overrightarrow{l}_j^{k+1}(t)}(t+1)\\
=&\frac{\delta}{\Delta}\sum_{i=1}^{m} i\omega_{l_i^k(t)}(t+1)\prod_{j=1}^{i-1} f_{l_j^k(t)}(t+1) + \frac{\delta}{\Delta}\prod_{j=1}^{m} f_{l_j^k(t)}(t+1) * \sum_{i=m+1}^{M-1} i\omega_{l_{i+1}^k(t)}(t+1)\prod_{j=m+1}^{i-1} f_{l_{j+1}^k(t)}(t+1)\\
&+ \frac{\delta}{\Delta}M * \omega_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{l_j^k(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{l_{j+1}^k(t)}(t+1)\\
&+ f_{\overrightarrow{l}_M^{k+1}(t)}(t+1)\prod_{j=1}^{m} f_{l_{j+1}^k(t)}(t+1) * \prod_{j=m+1}^{M-1} f_{l_{j+1}^k(t)}(t+1)\\
=&\frac{\delta}{\Delta}\Big[X_2(\mathbb{T}(\Omega^k(t)),m) - (m+1)X_1(\mathbb{T}(\Omega^k(t)),m)\Big]\\
&+ \frac{\delta}{\Delta}X_1(\mathbb{T}(\Omega^k(t)),m) * \Big[m+1+X_4(\mathbb{T}(\Omega^k(t)),m+2) - M * X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big]\\
&+ \frac{\delta}{\Delta}M * \omega_{\overrightarrow{l}_M^{k+1}(t)}(t+1) * X_1(\mathbb{T}(\Omega^k(t)),m) * X_3(\mathbb{T}(\Omega^k(t)),m+2)\\
&+ f_{\overrightarrow{l}_M^{k+1}(t)}(t+1) * X_1(\mathbb{T}(\Omega^k(t)),m) * X_3(\mathbb{T}(\Omega^k(t)),m+2)\\
=&\frac{\delta}{\Delta}\Big[X_2(\mathbb{T}(l^k(t)),m) - (m+1)X_1(\mathbb{T}(\Omega^k(t)),m)\Big] + \frac{\delta}{\Delta}X_1(\mathbb{T}(\Omega^k(t)),m) * \Big[m+1+X_4(\mathbb{T}(\Omega^k(t)),m+2)\\
&- M * X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big] + \frac{\delta}{\Delta}M * \omega_{\overrightarrow{l}_M^{k+1}(t)}(t+1) * X_1(\mathbb{T}(\Omega^k(t)),m) * X_3(\mathbb{T}(\Omega^k(t)),m+2)\\
&+ f_{\overrightarrow{l}_M^{k+1}(t)}(t+1)X_1(\mathbb{T}(\Omega^k(t)),m) * X_3(\mathbb{T}(\Omega^k(t)),m+2)\\
=&\frac{\delta}{\Delta}X_2(\mathbb{T}(\Omega^k(t)),m) + \frac{\delta}{\Delta}X_1(\mathbb{T}(\Omega^k(t)),m) * \Big[X_4(\mathbb{T}(\Omega^k(t)),m+2)\\
&+ (\frac{\Delta}{\delta} - M)f_{\overrightarrow{l}_M^{k+1}(t)}(t+1) * X_3(\mathbb{T}(\Omega^k(t)),m+2)\Big]\\
=&\frac{\delta}{\Delta}\Big\{\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m) + \boldsymbol{A_7}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\cdot\boldsymbol{A_8}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)),m)\Big\}
\end{aligned}
$$

Chapter 7

Conclusion and Perspective

7.1 Thesis Summary

This thesis has presented a systematic study on a class of RMAB problems arising from the context of opportunistic spectrum access. More specifically, we focus on the myopic policy, a natural strategy with simple and robust structure that seeks to maximize the short-term reward.

In Chapter 3, we provide a generic analysis on the optimality of the myopic sensing policy where a user can sense more than one channel each time and gets one unit of reward if at least one of the sensed channels is in the good state. Through mathematic analysis, we show that the myopic sensing policy is optimal only for a small subset of cases where the user is allowed to sense two channels each slot. In the general case, we give counterexamples to illustrate that the myopic sensing policy is not optimal.

Motivated by the above analysis, we then study the following natural while fundamentally important question: under what conditions is the myopic policy guaranteed to be optimal? We answer the above posed question by performing an axiomatic study in Chapter 4 and Chapter 5. More specifically, we develop three axioms characterizing a family of functions which we refer to as regular functions, which are generic and practically important. We then establish the optimality of the myopic policy when the reward function can be expressed as a regular function and the discount factor is bounded by a closed-form threshold determined by the reward function.

Chapter 3, 4, 5 study the optimality of the myopic sensing policy in the case where the user is allowed to sense k out of N channels. In Chapter 6, we further investigate a more challenging problem where the user has to decide the number of channels to sense in order

to maximize its utility. This optimization problem hinges on the following tradeoff between exploitation and exploration: sensing more channels can help learn and predict the future channel state, thus increasing the long-term reward, but at the price of sacrificing the reward at current slot as sensing more channels reduces the time for data transmission, thus decreasing the throughput in the current slot. Therefore, to find the optimal number of channels to sense consists of striking a balance between the above exploitation and exploration. After showing the exponential complexity of the problem, we develop a heuristic ν-step lookahead policy which consists of sensing channels in a myopic way and stopping sensing when the expected aggregated utility from the current slot t to slot $t+\nu$ begins to decrease. In the developed policy, the parameter ν allows to achieve a desired tradeoff between social efficiency and computation complexity. We demonstrate the benefits of the proposed strategy via numerical experiments on several typical settings.

From the system perspective, our analysis presented in this thesis provides insight on the following design tradeoff in opportunistic spectrum access: *gaining immediate access (exploitation) versus gaining information for future use (exploration).* Due to hardware limitations and the energy cost of spectrum monitoring, a user may not be able to sense all the channels in the spectrum simultaneously. A sensing strategy is thus needed for intelligent channel selection to track the rapidly varying spectrum opportunities. The purpose of a sensing strategy is twofold: to find good channels for immediate access and to gain statistical information on the spectrum occupancy for better opportunity tracking in the future. The optimal sensing strategy should thus strike a balance between these two conflicting objectives.

7.2 Open Issues and Directions for Future Research

In this section, we discuss some key open issues and outline some potential directions for further research.

7.2.1 RMAB-based Channel Access with Multiple Users

In this thesis, we mainly focus on the decision making process and different tradeoff within a single user. A natural research direction is to take the results obtained in the thesis as a building block to further study the scenario of multiple users accessing opportunistically a multi-channel communication system. Here the key research challenge is how to coordinate the users to access

different channels in a distributed fashion without or with little explicit network-level feedback. A natural way to tackle this problem is to model the situation as a non-cooperative game among users and to see how the results obtained in this thesis can further be tailored in the new context. We are now beginning to perform numerical experiments on the channel sensing and accessing strategies developed in the thesis in the context of multiple users and exploring the problem from the perspective of signaling game.

7.2.2 Incorporating Channel Switching Cost

Another aspect that may limit the performance of the channel access mechanism is the channel switching cost. In the current wireless devices, channel switching introduces a cost in terms of delay, packet loss and protocol overhead. Hence, an efficient channel access policy should avoid frequent channel switching, unless necessarily. In the context of RMAB, this problem can be mapped into the generic RMAB problem with switching cost between arms. Our analysis in Chapter 6 is a primary step toward taking the channel switching cost into account, but more systematical works are called for so as to provide more in-depth insight on this problem.

It is important to note that the generic MAB with switching cost is NP-hard and there does not exist any optimal index policy [50]. More specifically, the introduction of switching cost renders not only the Gittins index policy suboptimal, but also makes the optimal policy computationally prohibitive. Given such difficulties, we envision to tackle the problem from the following aspects:

- Looking for suboptimal policy with bounded efficiency loss compared to the optimal policy;
- Developing heuristic policy achieving a tradeoff between optimality and complexity, as that presented in Chapter 6;
- Deriving optimal policy in a subset of scenarios or designing asymptotically optimal policy.

7.2.3 RMAB with Correlated Arms

Another practical extension is to consider the correlated channels, i.e., the Markov chains of different channels can be correlated. This problem can be cast into the RMAB problem with correlated arms. The introduction of the correlation among arms makes the tradeoff between exploration and exploitation more sophisticated as sensing a channel can not only reveal the

state of the sensed channel, but also provide information on other channels as they are not entirely independent. How to characterize the tradeoff in this new context and how to design efficient channel access policy are of course pressing research topics in this direction.

Bibliography

[1] Commission of the European Communities. Internet of things—an action plan for europe. Jun. 2009.

[2] J. Mitola. Cognitive radio for flexible mobile multimedia communications. In *IEEE Int. Workshop on Mobile Multimedia Communications (MoMuC)*, Nov. 1999.

[3] DARPA XG Working Group. The xg vision, request for comments. Jun. 2003.

[4] P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, (Special Vol. 25A):287–298, 1988.

[5] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:275–294, Dec. 1933.

[6] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queueing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[7] K. Wang, L. Chen, Khaldoun Al Agha, and Quan Liu. On the optimality of myopic sensing in multi-channel opportunistic access: the case of sensing multiple channels. *In submission to IEEE Wireless Communications Letter, available on Computing Research Repository (CoRR) arXiv:1103.1784v1*, 2011.

[8] K. Wang, L. Chen, Q. Liu and Khaldoun Al Agha. On optimality of myopic sensing policy with imperfect sensing in multi-channel opportunistic access. *In submission to IEEE Transactions on Communications*, 2011.

[9] K. Wang Q. Liu and L. Chen. On optimality of greedy policy for a class of standard reward function of restless multi-armed bandit problem. *In submission to IET Signal Processing*, 2011.

[10] K. Wang and L. Chen. On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach. *IEEE Transactions on Signal Processing*, 60(1):300–309, 2012.

[11] K. Wang, L. Chen, Q. Liu and Khaldoun Al Agha. On optimality of myopic policy for restless multi-armed bandit problem with non i.i.d. arms and imperfect detection. *In submission to IEEE Transactions on Signal Processing*, 2012.

[12] J. C. Gittins and D.M. Jones. A Dynamic Allocation Index For the Sequential Design of Experiments. *Progress in Statistics*, pages 241–266, 1974.

[13] Q. Zhao, and B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance. *IEEE Transactions Wireless Communication*, 7(3):5413–5440, Dec. 2008.

[14] S. Ahmand, and M. Liu, and T. Javidi, and Q. zhao and B. Krishnamachari. Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory*, 55(9):4040–4050, Sep. 2009.

[15] S. Ahmad and M. Liu. Multi-channel opportunistic access: A case of restless bandits with multiple players. In *Proc. Allerton Conf. Commun. Control Comput*, pages 1361–1368, Oct. 2009.

[16] K. Liu and Q. Zhao. Distributed Learning in Multi-Armed Bandit with Multiple Players. *Arxiv 0910.2065*, 2009.

[17] Fabio E. Lapiccirella, Keqin Liu and Zhi Ding. Multi-channel opportunistic access based on primary arq messages overhearing. In *Proceedings of IEEE ICC 2011*, Kyoto, Jun. 2011.

[18] S. Murugesan, P. Schniter, N. B. Shroff. multi-user scheduling in makov-modeled downlink using randomly delayed arq feedback. *Appear in IEEE Transactions on Information theory*, 2011.

[19] S. Guha and K. Munagala. Approximation algorithms for partial-information based stochastic control with markovian rewards. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence, RI, Oct. 2007.

[20] S. Guha and K. Munagala. Approximation algorithms for restless bandit problems. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, Jan. 2009.

[21] K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, Nov. 200.

[22] Ting He, Anandkumar A., Agrawal D. Index-based sampling policies for tracking dynamic networks under sampling constraints. In *INFOCOM 2011*, pages 1233–1241, Shanghai, China, April 2011.

[23] Ny, J.L., Dahleh, M., Feron, E. Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *Proceedings of American Control Conference*, pages 4220 – 4225, Seattle, WA, 2008.

[24] Raghunathan V., Borkar V., Min Cao, Kumar P.R. Index policies for real-time multicast scheduling for wireless broadcast systems. In *INFOCOM 2008*, pages 1570–1578, Phoenix, AZ, 2008.

[25] Mingyan Liu Ehsan N. On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services. In *INFOCOM 2004*, pages 1974–1983, Hong Kong, March 2004.

[26] Peter J. Urtzi A., Martin E. A modeling framework for optimaizing the flow-level scheduling with time-varing channels. *Performance Evaluation*, 67(11):1024–1029, Aug. 2010.

[27] Peter J. Vladimir N. Urtzi A. A nearly-optimal index rule for scheduling of users with abandonment. In *INFOCOM 2011*, pages 2849–2857, Shanghai, China, April 2011.

[28] Peter J. Value of information in optimal flow-level scheduling of users with markovian time-varing channels. *Performance Evaluation*, 68(11):1022–1036, 2011.

[29] Peter J., Brunilde S. Optimal anticipative congestion control of flows with time-varying input stream. *Performance Evaluation*, 69(2):86–101, 2011.

[30] Jonathan O. A continuous-time markov decision process for infrastructure surveillance. *Operations Research Proceedings*, pages 327–332, 2010.

[31] X.Y. Gan, Bo Chen. A novel sensing scheme for dynamic multichannel access. *IEEE Transactions on Vechiel Technology*, 61(1):208–221, 2011.

[32] H. Ji D. Chen and Xi Li. Distributed best-relay node selection in underlay cognitive radio networks a restless bandits approach. In *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, pages 1208–1212, Cancun, Quintana Roo, Mar. 2011.

[33] H. Ji C. Luo, F. Yu and V. Leung. Distributed relay selection and power control in cognitive radio networks with cooperative transmission. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5, Cape Town, May 2010.

[34] H. Ji C. Luo, F. Yu and V. Leung. Optimal channel access for tcp performance improvement in cognitive radio networks. *Wireless Networks*, 17:479–492, 2010.

[35] H. Ji P. Si, F. Yu. Optimal network selection in heterogeneous wireless multimedia networks. *Wireless Networks*, 16:1277–1288, 2010.

[36] W. Ouyan, S. Murugesan, A. Eryilmaz, and N. B. Shroff. Exploiting channel memory for joint estimation and scheduling in downlink networks. In *IEEECOM2011*, 2011.

[37] S. Murugesan, and P. Schniter, and N. B. Shroff. Opportunistic scheduling using arq feedback in multi-cell downlink. In *Asilomar Conference*, Pacific Grove, CA, Nov. 2010.

[38] T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Probability*, 6(1), 1985.

[39] V. Anantharam, P. Varaiya and J. Walrand. Asymptotically Efficient Adaptive Allocation Rules for the Multiarmed Bandit Problem with Switching. *IEEE Transactions on Automatic Control*, 32(11):968–976, Nov. 1987.

[40] R. Agrawal. Sample Mean Based Index Policies with O(logn) Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*, 27(4), 1995.

[41] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, 2002.

[42] A. Anandkumar, N. Michael and Ao Tang. Index-based sampling policies for tracking dynamic networks under sampling constraints. In *INFOCOM 2010*, pages 1–9, San Diego, CA, Mar. 2010.

[43] K. Liu, Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Wireless Communications*, 58(11):5667–5681, Nov. 2010.

[44] C. Tekin and M. Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *INFOCOM 2011 Proceedings IEEE*, pages 2462–2470, Shanghai, China, Apr. 2011.

[45] H. Liu, K. Liu and Q. Zhao. Learning and sharing in a changing world: Non-bayesian restless bandit with multiple players. In *Information Theory and Applications Workshop (ITA), 2011*, pages 1–7, La Jolla, CA, Feb. 2011.

[46] H. Liu, K. Liu and Q. Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1968–1971, Prague, May 2011.

[47] S. Ahmad and M. Liu. Multi-channel opportunistic access: a case of restless bandits with multiple plays. In *Allerton Conference*, Monticello, Il, Spet.-Oct. 2009.

[48] K. Liu, and Q. Zhao, and B. Krishnamachari. Dynamic multichannel access with imperfect channel state detection. *IEEE Transactions on Signal Processing*, 58(5):2795–2807, May 2010.

[49] T. Javidi S. H. Ahmad, M. Liu, Q. Zhao, and B. Krishnamachari. Optimality of myopic sensing in multi-channel opportunistic access. *IEEE Transactions on Information Theory*, 55(9):4040–4050, 2009.

[50] J. S. Banks and R. K. Sundaram. Swithing costs and the gittins index. *Economica*, 62:687–694, May 1994.