Distributed Resource Allocation in Wireless Networks: A Game-Theoretical Learning Framework

vorgelegt von
M.Sc. Setareh Maghsudi
aus Teheran

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Ingenieurwissenschaften
-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Thomas Sikora
Berichter: Prof. Giuseppe Caire, Ph.D.
Berichter: Prof. Leandros Tassiulas, Ph.D. (Yale University, USA)
Berichter: PD Dr.-Ing. Slawomir Stanczak

Tag der wissenschaftlichen Aussprache: 23 March 2015

Berlin, 2015

D83

With all my love, to
Maman, Baba, and my little angel, Sobhan

Abstract

Device-to-device (D2D) communications underlaying a cellular infrastructure is regarded as one of the key technology enablers for future wireless networks. The main advantages of underlay D2D communications stem from the reuse-, proximity-, and hop gains, which can be utilized for enhanced coverage, capacity and quality-of-service in mobile networks. The basic idea consists in enabling suitably-selected nearby device pairs to reuse the cellular spectrum for direct data transfer, while ensuring that there is no detrimental impact on traditional cellular transmissions via base stations. Despite its great potential for performance gains, D2D communications poses some fundamental challenges to system designers. These challenges, which include D2D discovery, transmission mode selection, resource allocation and interference management, are exacerbated by the lack of timely and accurate channel state information for direct D2D links at the level of base stations and wireless devices. Therefore, in order to avoid a significant increase in the feedback and signaling overhead, there is a strong need for D2D resource allocation solutions that (i) are amenable to distributed implementation and (ii) can beneficially exploit some side-information made available at the level of D2D links through the network assistance. In addition, such D2D solutions must be capable of dealing with the following characteristics of mobile networks:

- uncertainty, which is caused by the random nature of the wireless environment (including channels and users' behavior), and is further aggravated by the lack of information at the user level, and
- competition between users that attempt to access strictly limited wireless resources.

To address these challenges, the core objective of this thesis is to develop and study a novel theoretical framework for network-assisted D2D resource allocation that incorporates *game theory* and *reinforcement learning*. We model a distributed D2D wireless network as a multi-agent system, in which a set of self-interested smart agents with bounded rationality share limited resources, by taking actions according to some decision making strategy. Every joint action profile is associated with some reward (or cost) for each agent, and the agents selfishly compete for access to resources in order to achieve a higher utility. By incorporating a learning model into a game-theoretical formulation, the agents' actions

evolve over time, in general as a function of the past outcomes and the (possibly) observed side-information. Therefore, for the learning agents which strive for optimality in some sense in the long run, the distributed resource allocation problem can be rephrased as a decision making problem, to be solved by developing decision making strategies whose outcomes are some sort of efficient equilibria. An appropriate decision making strategy is however not unique, and depends strongly on the basic characteristics of the underlying networks such as the type of randomness and information availability. We study selected types of networks in the context of the features mentioned above, and develop decision making strategies for distributed resource allocation problem.

Chapter 2 deals with a multi-agent decision making problem in an adversarial environment. The self-interested agents (players) have no initial information and intend to efficiently learn the optimal joint action profile through sequential interactions with the environment. A particular instance of this problem is a joint channel and power-level selection in an interference-limited D2D wireless network. Based on the concept of *weighted-average* and *follow the leader* allocation rules, we propose two selection strategies that not only yield small regret for all players, but also guarantee that the empirical joint frequencies of the game converge to the set of correlated equilibria. In addition we study the convergence rate and complexity issues.

In Chapter 3 the focus is on a multi-agent adaptive decision making problem, where selfish agents learn the optimal joint action profile from successive interactions with the environment, which, unlike the previous chapter, is assumed to be stochastic. We also assume that some side-information is revealed to the players in the course of the game. An example of this formulation is a channel selection problem in a D2D interference network underlaying conventional cellular infrastructures. Using the concept of *calibrated forecasting*, we propose a selection strategy that yields small regret for all players. Furthermore, the empirical joint frequencies of the game converge to the set of correlated equilibria. Finally, we discuss convergence rate and complexity.

Chapter 4 studies the decision making problem in a network with two types of agents, namely primary and secondary agents. We assume the existence of an authority that regularizes the network in favor of the primary agents, by assigning restricted resources to the secondary agents. Secondary agents therefore compete for access to the assigned limited resources. In wireless networks, an instance of this scenario is a joint channel allocation and power control problem in a hybrid D2D and cellular network regularized by a base station, where cellular users have a higher priority in using wireless resources compared to D2D users. We prove a lower-bound for the cellular aggregate utility in the downlink, which allows for decoupling the channel allocation and D2D power control problems. An efficient graph-theoretical approach is proposed to solve the former problem,

whereas the latter is modeled as a multi-agent learning game, which is shown to be an exact potential game. We then use *Q-learning better-reply* dynamics to achieve equilibrium.

Zusammenfassung

Als eine der wichtigsten Schlüsseltechnologien für die zukünftige Kommunikationstechnik werden die mit direkter Gerät-zu-Gerät Kommunikation unterlegten zellularen Mobilfunknetze (Engl. underly Device-to-Device communications, D2D) angesehen. Sie kann zur Verbesserung der Leistungsfähigkeit von zellularen Netzwerken beitragen. Die Hauptvorteile leiten sich aus der Wiederverwendung- (Engl. reuse-), Nachbarschaft- (Engl. Proximity) und der Hop- Gewinne ab. Diese können für erhöhte Abdeckung (Engl. coverage), Kapazität (Engl. capacity) und Quality-of-Service (QoS) in Mobilfunknetzen benutzt werden. Die Grundidee ist es, dass ausgewählte Gerätepaare zelluläre Spektren (Engl. cellular spectrum) wiederverwenden. Gleichzeitig ist darauf zu achten, dass solche direkten Übertragungen keine negativen Auswirkungen auf die traditionellen Zellnutzer haben. Trotz des großen Potenzials für Leistungssteigerungen, enthält die D2D Kommunikation einige Herausforderungen für die Systementwickler; dazu gehören Übertragungsmodus-Auswahl (Engl. transmission mode selection), Ressourcenallokation (Engl. resource allocation) und Interferenz-Management (Engl. interference management). Diese Herausforderungen verschärfen sich wegen des Mangels an genauen Kanalzustandsinformationen (Engl. channel state information) für die direkten D2D Verbindungen auf der Ebene der Basisstationen (Engl. base stations) und der drahtlosen Geräte (Engl. wireless devices). Um eine deutliche Erhöhung des Signalisierungsaufwandes (Engl. signaling overhead) zu vermeiden, besteht ein starkes Bedürfnis nach D2D Ressourcenallokationslösungen, die (i) in einer verteilten Weise implementiert werden können und (ii) einige Nebeninformationen (Engl. side-information), falls verfügbar, vorteilhaft nutzen können. Darüber hinaus müssen solche D2D Lösungen in der Lage sein, die folgenden Eigenschaften der Mobilfunknetzwerke zu behandeln:

- Unsicherheit (Engl. uncertainty), die durch das zufällige Verhalten von Mobilfunkkanälen ebenso wie durch das unberechenbare Verhalten von Mobilfunknutzern verursacht ist und sich im Falle geringer Informationen auf der Ebene der Basisstation und/oder den Mobilfunknutzern weiterhin verschärft.

- Konkurrenz zwischen den Mobilnutzern, die auf stark begrenzte drahtlose Ressourcen zuzugreifen versuchen.

Im Rahmen dieser Arbeit wurde ein neuartiges theoretisches Systemkonzept entwickelt, das die verteilte Ressourcenallokation (Engl. decentralized resource allocation) unter Berücksichtigung der Unsicherheit betrachtet sowie *Spieletheorie* (Engl. game theory) und *Verstärkendes Lernen* (Engl. reinforcement learning) mit einbezieht. Innerhalb dieser Arbeit werden verteilte Mobilfunknetze mittels Multiagentsystemen (Engl. multi-agent systems) modelliert. Eine Anzahl von eigennützigen Agenten mit eingeschränkter Rationalität (Engl. bounded rationality) teilen begrenzte Ressourcen nur durch eine Auswahl an Entscheidungsstrategien (Engl. decision making strategy). Jedes dieser gemeinsame Entscheidungsprofile (Engl. joint action profile) ist verbunden mit Belohnungen (Engl. utility) oder Kosten (Engl. cost) für die Agenten. Die Agenten kämpfen deshalb für höhere Utility. Durch die Einbindung von lernenden Algorithmen (Engl. learning algorithms) in die spieltheoretische Formation, verbessern sich die Entscheidungen der Agenten mit der Zeit. Generell kann dies als Funktion der Belohnungen aus der bereits vergangenen Zeit und den möglicherweise zusätzlich erhaltenen Informationen ausgedrückt werden. Die so lernenden Agenten wollen die optimale langfristige Utility erreichen, zusammen mit einer Art von effizientem Gleichgewicht (Engl. equilibrium). Deswegen, kann die verteilte Ressourcenallokation als Entwicklung von Entscheidungsstrategien formuliert werden. Die dazugehörigen Entscheidungsprofile sind jedoch nicht einzigartig. Sie hängen insbesondere von den Grundeigenschaften des Netzwerks ab. Diese sind z.B. die verfügbaren Informationen und/oder die Art der Zufälligkeit. Innerhalb dieser Arbeit werden verschiedene Netzwerke unter Berücksichtigung der oben gennannten Aspekte untersucht. Entsprechende Entscheidungsprofile werden für unterschiedliche Aspekte entwickelt, um das Problem der verteilten Ressourcenallokation zu lösen.

Das Kapitel 2 beschäftigt sich mit dem Multiagent Entscheidungsproblem in einer feindlichen (Engl. adversarial) Umgebung. Die eigennützigen Agenten werden nicht mit Informationen versorgt. Sie beabsichtigen jedoch effiziente gemeinsame Entscheidungsprofile durch sequenziell, iterative Handlungen zu lernen. Ein Beispiel dieses Problems ist die richtige Auswahl eines gemeinsamen Funkkanals und der Sendeleistung in einem D2D Interferenznetzwerk. Basierend auf dem *gewichteten-Mittelwert* (Engl. weighted average) und *folge dem Anführer* (Engl. follow the leader) Strategien, entwickeln wir zwei Entscheidungsstrategien. Ziel soll dabei nicht nur eine geringe Ablehnungsquote der Nutzer sein, sondern auch zu garantieren, dass die empirisch gemeinsamen Frequenzen (Engl. empirical joint frequencies) zu dem Satz von korrelierten Gleichgewichten (Engl. correlated equilibria) konvergieren. Die Konvergenzrate sowie die Komplexität werden ebenfalls untersucht.

Im Kapitel 3 ist der Fokus auf das Multiagenten Entscheidungsproblem in einer stochastischen (Engl. stochastic) Umgebung. In diesem erlernen eigenständige Agenten das opti-

male gemeinsame Entscheidungsprofil basierend auf fortlaufenden Wechselwirkungen mit der Umgebung. Dieses Verhalten wird im Gegensatz zum vorherigen Kapitel als stochastisch angenommen. Ferner werden den Agenten einige weitere Informationen offenbart. Ein Beispiel dieses Problems stellt die Kanalwahl innerhalb eines unterlegten D2D Interferenznetzwerkes dar. D2D Nutzern ist es erlaubt das Mobilfunkspektrum erneut zu verwenden, vorausgesetzt, dass der negative Effekt der D2D Übertragung für den Zellnutzer minimiert wird. Mittels des Konzepts der *kalibrierten Vorhersage* (Engl. calibrated forecaster) wird eine Entscheidungsstrategie vorgeschlagen, die möglichst wenige Nutzer zurückweisen soll. Außerdem streben so die empirisch gemeinsamen Frequenzen zu dem Satz von korrelierten Gleichgewichten. Insbesondere die Konvergenzrate und die Komplexität werden hier diskutiert.

Innerhalb des Kapitels 4 wird das Entscheidungsproblem in einem Netzwerk mit zwei Arten von Agenten, den sogenannten primären und sekundären Agenten, untersucht. In diesem Zusammenhang wird von einer Autorität ausgegangen, die primäre Agenten zur Netzwerknutzung favorisiert und beschränkten Zugriff an sekundäre Agenten vergibt. Dementsprechend konkurrieren sekundäre Agenten untereinander um die beschränkten Ressourcen. In der Funkkommunikation kann ein solches Szenario beispielsweise in der gemeinsamen Kanal- und Sendeleistungszuweisung innerhalb eines D2D Interferenznetzes auftreten. Dort haben Zellnutzer bevorzugten Zugriff auf die Netzressource. Wir definieren eine untere Grenze der zellularen Gesamt-Utility (Engl. cellular aggregate utility) im Downlink, die die Entkopplung der Kanalallokation von der D2D Leistungsregelung ermöglicht. Ein effizienter graphentheoretischer Ansatz wird benutzt, um das Problem der Kanalallokation zu lösen. Das letztgenannte Problem wird mit einem Multiagenten Lernspiel (Engl. learning game) modelliert. Die *Q-learning better-reply* Strategie wird anschließend benutzt, um ein Gleichgewicht zu erreichen.

Acknowledgements

Dear Maman and Baba, I wanted to make these few lines very perfect and special for you; now, after playing over and over with words, I believe what you have done for me, as well as my love and gratitude towards you, are so astounding that cannot be described in words. So, I simply tell you this: no matter how far you are, my heart is always warm thinking about you, and, in every single moment, I feel your presence and support. Without doubt, having you is my biggest luck. I have an absolutely wonderful life, and I owe it all to you. I give you the biggest hug and kiss ever, and hope you are happy with me.

My dear Babak, you are just perfect. Thanks for many calls during the day: for being there for me to laugh, dance, sing, cry, complain or scream (!) behind the webcam or on the phone. Thanks for all your support when I was sad, disappointed, angry, or worried. Thanks for all the love. Thanks for asking me to think big, to pursue my dreams, to take care of myself, and to let it go if necessary. I have enjoyed our journey from the very first moment till today, and cannot wait for the rest. I want you to know that I treasure every single second of my life spent with you.

My dear Baharan, thanks for praying for me. Thanks for very personal and thoughtful gifts you buy for me, which makes me wonder how you always find things that seem to have been made just for me! Thanks for making my trips home a great pleasure and so much fun. Thanks for being such a nice and wise sister. I know you are always there for me, and I feel really lucky to have a sister like you. I want you to know how much I love you, and how sorry I am for not being around to support you when I should have been there for you.

Dear Andreas and Qiqi, I want to thank you for your friendship, for sharing your experiences, for bringing souvenirs for me from your trips, and for drinking tea or hot chocolate with me. Especially, dear Andreas, I am very thankful to you for your invaluable support through my rough times, for making me a wonderful graduation hat, and for proofreading the German abstract of my thesis. It was great to get to know you in Berlin. Thanks for many good memories.

I would like to thank Dr.-Ing. Slawomir Stanczak for giving me a chance to conduct research at the Technical University of Berlin, and for providing feedback on my scientific work. I am grateful to Prof. Giuseppe Caire and Prof. Leandros Tassiulas for accepting

to be on the examination committee, and providing review on my thesis. I sincerely thank Prof. Thomas Sikora for accepting to serve as the chair of examination committee, and for his invaluable patience during the defense session.

The financial support from German Ministry for Education and Research (BMBF) as well as that of German Research Foundation (DFG) is gratefully acknowledged.

Contents

List of Figures

List of Tables

List of Abbreviations

AR ..... Average Reward

a.s. ..... Almost Surely

AWGN ..... Additive White Gaussian Noise

BERTS ..... Bandit Experimental Regret-Testing Strategy

BS ..... Base Station

CBS ..... Calibrated Bandit Strategy

CSI ..... Channel State Information

CSMA ..... Carrier Sense Multiple Access

D2D ..... Device-to-Device

GLIE ..... Greedy in the Limit with Infinite Exploration

GPS ..... Global Positioning System

HRAS ..... Hybrid Resource Allocation Strategy

LMP ..... Larger Midpoint Property

MAB ..... Multi-Armed Bandits

MDP ..... Markov Decision Process

MP-MAB ..... Multi-Player Multi-Armed Bandit

MS ..... Mixed Strategy

NR-BEWAS ..... No-Regret Bandit Exponential-Based Weighted Average Strategy

NR-BFPLS ..... No-Regret Bandit Follow the Perturbed Leader Strategy

OSI ..... Open System Interconnection

o.w. ..... Otherwise

PO-MDP ..... Partially Observable Markov Decision Process

QoS ..... Quality of Service

SIR ..... Signal to Interference Ratio

List of Symbols

$|h_{uv,x}|^2$ ..... Average gain of channel x between u and v

$R_{\mathbf{Ext}}$ ..... Cumulative external regret

$R_{\mathbf{Int}}$ ..... Cumulative internal regret

R_n ..... Cumulative regret

n ..... Game horizon

d ..... Instantaneous loss

g ..... Instantaneous reward

r ..... Instantaneous regret

$\mathbf{i}^-$ ..... Joint action profile of opponents

$\mathbf{i}$ ..... Joint action profile of players

f ..... Mean reward process

i ..... Player action

P ..... Power

α ..... Power price factor

$\mathcal{M}$ ..... Set of actions

$\mathcal{C}$ ..... Set of correlated equilibria

$\mathbb{Z}$ ..... Set of integer numbers

$\mathcal{I}^-$ ..... Set of joint action profiles of opponents

$\mathcal{I}$ ..... Set of joint action profiles of players

$\mathbb{R}^+$ ..... Set of non-negative real numbers

$\mathcal{K}$ ..... Set of players

$\mathbb{R}$ ..... Set of real numbers

$\mathcal{S}$ ..... Set of states

$\mathbf{T}$ ..... State transition matrix

N_0 ..... Variance of zero-mean AWGN

Notation

In this work, sets and matrices are presented by calligraphic capital letters (such as $\mathcal{X}$), and non-italic bold capital letters (such as $\mathbf{X}$), respectively. The cardinality of any particular set is shown by the same letter as that set, but in italic. For instance, X is the cardinality of a set $\mathcal{X}$. As for matrices, the l-th column of a matrix $\mathbf{X}$ is denoted by $\mathbf{X}_l$. Moreover, the entry in the l-th row and k-th column of a matrix $\mathbf{X}$ is referred to as $\mathbf{X}[l, k]$. Non-italic bold lower case and italic lower case letters correspond to the column vectors and scalars, for instance $\mathbf{x}$ and x, respectively. Superscript k stands for player k. Whenever necessary, subscripts t and i are correspondingly used to denote time and action. For example, $g_t^{(k)}$ is interpreted as the instantaneous reward of player k at time t, while $f_i^{(k)}$ is the mean reward of action i to player k. For any given variable r, symbols $\tilde{r}$ and $\hat{r}$ are used to represent its estimated and experimental values, respectively.

We note that $f(x) \in o(g(x))$ if $\lim_{x\to\infty} \frac{g(x)}{f(x)} = 0$. Moreover, $f(x) \in O(g(x))$ if there exists $M \in \mathbb{R}^+$ and $a \in \mathbb{R}$ so that $|f(x)| \leq M\,|g(x)|$ for all $x \geq a$.

Further we use the following.

$|\cdot|$ ..... Absolute value

$\log_2$ ..... Base 2 logarithm

$\bigotimes$ ..... Cartesian product

$\mathbf{E}\{\cdot\}$ ..... Expected value

exp ..... Exponential function

$\forall$ ..... For all

$\mathbf{1}_{\{x\}} = \begin{cases} 1 & x \textbf{ holds} \\ 0 & \textbf{o.w.} \end{cases}$ Indicator function

∞ ..... Infinity

max ..... Maximum

min ..... Minimum

$\log$	Natural logarithm
$\|\cdot\|_p$	p-norm
$\mathbf{Pr}(\cdot)$	Probability
$\sup$	Supremum
$\exists$	There exists
$x \in \mathcal{X}$	x belongs to $\mathcal{X}$
$\mathcal{Z} \subseteq \mathcal{X}$	$\mathcal{Z}$ is a subset of $\mathcal{X}$ or is equal to $\mathcal{X}$

1. Introduction

Due to the ever-increasing need for mobile services, we expect a massive growth in demand for wireless access in the years to come. As a result, future mobile networks are expected to accommodate new communications and networking concepts, including device-to-device (D2D) communications underlaying cellular networks. The basic idea of underlay D2D communications is to replace traditional end-to-end connections via access points, base stations (BS) or relays by direct short-distance communications links between suitably-selected nearby wireless devices that reuse the cellular spectrum of the base stations. The major potential advantages of this approach stem from the proximity-, hop-, and reuse gains that can be translated to enhance the network performance with respect to coverage, capacity, energy efficiency and quality-of-service (QoS) [DRW+09], [FDM+12], [KA08].

In order to realize D2D communications as an underlay to cellular networks, system designers face some fundamental challenges. In particular, while some channel state information (CSI) for cellular users[1] is usually available at the serving BS, acquiring some CSI for D2D links at the network side is a challenging task since pilot-based measurements for D2D channels are costly in terms of communication and control overhead, and therefore highly undesired. As an immediate consequence, the allocation of resources to D2D links (users) has to be performed in a distributed manner. Thereby, it is of utmost importance that direct D2D transmissions are coordinated to ensure that there is no detrimental impact on the prioritized cellular transmissions. Such coordination must involve a careful power-controlled allocation of D2D users to available radio frequency channels, primarily used by cellular users. This problem, which is in general difficult to solve even in a centralized manner, is significantly aggravated in D2D settings by the aforementioned need for distributed solutions [14].

Against this background, it becomes evident that efficient and robust D2D communications design needs to deal with inherent uncertainty, as well as strong and abrupt variations under various conditions and different types of randomness in a network. The uncertainty and variations concern not only information availability but also the network topology/architecture and the overhead tolerance. In this thesis, we model an under-

[1] In this thesis, any wireless device that operates in the traditional cellular mode is referred to as a cellular user. Accordingly, any *pair* of wireless devices that communicate directly are counted as a D2D user.

lay D2D communications network as a distributed multi-agent system. From this point of view, independent from others, each device pair takes actions to access shared radio resources, which includes the selection of its frequency channel and/or its transmission power-level. The actions of each pair, however, not only impact its own transmission performance expressed in terms of some utility, but also influence the utility of others, for instance as a result of interference. In other words, each joint action profile corresponds to a utility vector so that for any given agent, each action is associated with a utility (or cost) that is a function of (possibly random) network characteristics as well as other agents' actions. In the absence of a centralized controller and given no prior information, device pairs are modeled as selfish learning agents that play a strategic game repeatedly. This gives rise to a distributed decision making problem in a multi-agent system. For the learning agents that strive for asymptotic optimality, the resource allocation problem can be rephrased as a problem of developing decision making strategies according to which the players' actions evolve sequentially by observing the past outcomes of the game and (probably) some side-information. The goal is the asymptotic convergence to some equilibrium that guarantees the satisfaction of (almost) all involved agents in some sense.[2] This thesis provides some solutions to this problem by connecting multiple areas of research, including wireless communications, game theory, graph theory and reinforcement learning.

1.1. Background and Motivation

In this section, we first describe the analogy between multi-agent systems and distributed wireless networks. We afterwards classify wireless networks based on some basic features, and describe why and how game theory is used to address the wireless resource allocation problem in certain network types. We then clarify its inadequacy for solving the resource management problem in the absence of prior information. We discuss reinforcement learning as a theoretical tool to deal with the lack of prior knowledge, and describe how it can be complemented by game theory to develop distributed resource allocation schemes.[3]

[2] We emphasize that although the convergence to equilibrium might be desired from the network's perspective, it is not necessarily achieved by all learning models. In fact, the main anti-equilibrium reasoning is the long time it might take to converge to equilibrium in a dynamic network [FL96]. Another argument is the difficulty of guiding a set of distributed agents to settle at the most efficient equilibrium in the games where multiple equilibriums exist. In fact, in games with multiple equilibriums, guiding agents to expect the same equilibrium (even not the most efficient one) might require some common knowledge, which is in general difficult to acquire. Moreover, in case of cooperation assumption, incentive compatibility issues arise.

[3] This part of the thesis should not be considered as a comprehensive survey that is intended to cover all the existing literatures. Our goal here is to briefly discuss the trend of using game theory and reinforcement learning to solve the resource allocation problems in wireless networks.

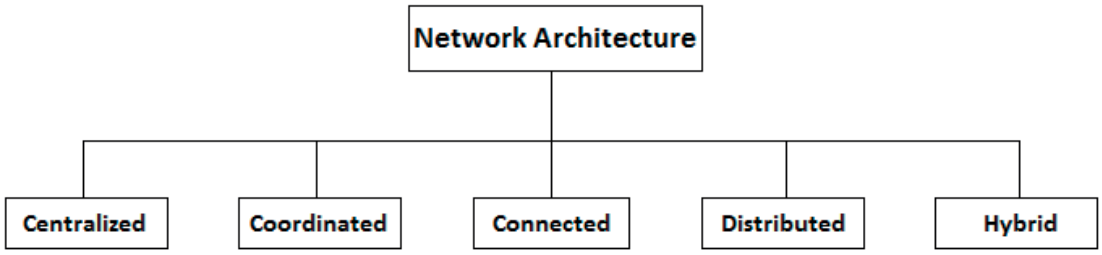


Figure 1.1.: Network classification based on architecture.

1.1.1. Wireless Networks as Multi-Agent Systems

As it is well-known, in wireless networks users affect each other by sharing limited resources in different layers of the OSI (open systems interconnection) model. In this context, any wireless network can be modeled as a multi-agent system, as summarized in Table 1.1. As the theory of multi-agent systems is wide and also well-developed, a great majority of

Table 1.1.: Analogy between Multi-Agent Systems and Distributed Wireless Networks

Multi-Agent Systems	Wireless Networks
Agents	Nodes in wireless networks that share wireless resources, e.g. D2D users
Resources	Radio and hardware resources depending on the OSI layer, e.g. power, frequency channel, relay
Actions	Decisions about the amount and type of resources to be used by the agent
Utility function	Any function based on signal, noise and interference, for instance throughput or delay

resource allocation problems can be addressed by some mathematical technique related to this field, depending on the characteristics of the network under consideration. In order to clarify this issue, in what follows, we first briefly classify wireless networks with respect to some basic features, namely architecture, information availability, randomness, and overheated tolerance.[4]

Regarding the architecture, networks can be divided into five main categories as briefly described in the following.

- Centralized: In a centralized network, there exists a central controller that governs the network by making decisions. Other network elements are therefore passive

[4]It should be emphasized that the classifications are mainly after the author's taste and from the perspective of this thesis; therefore they cannot be claimed as exhaustive and/or unique.

during the decision making process, and either accept the controller's suggestion or leave the network.

- Coordinated: In coordinated networks, there exists no central controller. Moreover, network nodes[5] are not in direct contact with each other; nevertheless, (almost) all of them are able to communicate with some coordinator. In such networks, every element actively takes part in decision making that is led by the coordinator.

- Connected: In this type of networks at least some of network nodes are able to communicate with each other in a pairwise manner. Decisions making therefore might include some sort of cooperation or information exchange provided that necessary incentives exist. The availability of a leader or coordinator is not required.

- Distributed: In this category, nodes do not communicate in pairs, but they might be able to hear a common signal. Each node therefore makes its own decisions, independent of others, possibly by using the information included in the common signal.

- Hybrid: This architecture is a combination of at least two models named before. For instance, in a partially connected network there might be also a coordinator. The hybrid type also includes hierarchical and multi-hop networks, where the network structures are not necessarily identical in all stages or hops.

In addition to the network architecture, an appropriate design of resource allocation scheme depends on some other factors, one of them being the information availability at node level. In this regard, networks can be divided into three categories, depicted in Figure 1.2. In this figure, it should be noted that side-information differs from prior-information in the sense that the former is revealed to nodes during or after each transmission round, while the latter is available a priori.

The next classification is performed with respect to the randomness. In a wireless network, the most important random variable is channel quality, which includes both fading and shadowing effects. Nodes' behavior is also random in general, which affects the utility, depending on the applied multiple access protocol or the agreed upon reward sharing contract. Another source of randomness is the availability of resources. For example, in a cognitive radio network, channels are available to secondary users only randomly. In general, random effects are either adversarial or stochastic. While in the former model random changes do not follow any specific rule, in the latter they can be attributed to a probability distribution that might be even time-variant. Figure 1.3 summarizes this discussion.

[5]Throughout, a node is any communicating element of a wireless network, for instance a user device.

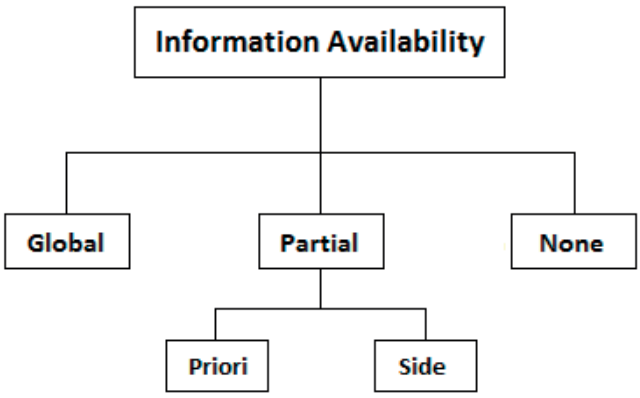


Figure 1.2.: Network classification based on information availability.

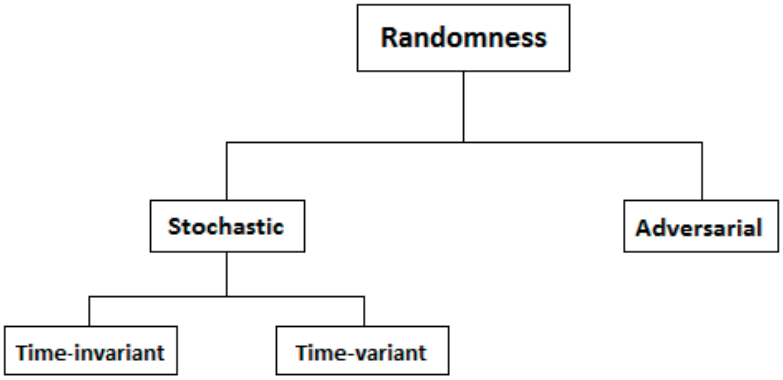


Figure 1.3.: Network classification based on randomness.

The final classification is related to the overhead tolerance, i.e. the amount of network resources that can be spent on procedures except for useful data transmission, for instance achieving an agreement among nodes that demonstrate a conflict of interests. Unlike network architecture, this aspect does not represent any physical characteristics of the network but rather reflects some sort of privilege, as clarified in the following example. Consider a wireless network in which any given node is able to hear a subset of other nodes. This network is therefore partially connected; nonetheless, local communications for decision making only take place if for instance some channels (i.e. resources) are reserved to accommodate such interactions. From this perspective, networks can be divided into the three following categories:

- Large overhead (pairwise communications): Nodes are allowed to exchange data with each other in a pairwise manner.
- Mild overhead (Control channel): There exists at least one control channel that is

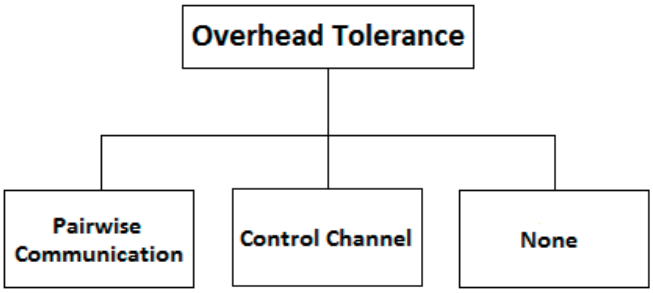


Figure 1.4.: Network classification based on overhead tolerance.

commonly used for instance in order to broadcast common signals which might be informative to all nodes.

- None: Nodes do not communicate and also no control channel exists.

1.1.2. Game Theory

Intuitively, the multi-agent model described before suggests game theory as a strong mathematical tool to manage wireless resources in a distributed manner. Formally, a game is defined as $\mathfrak{G} := \left\{\mathcal{K}, \mathcal{I}, \left\{g^{(k)}\right\}_{k\in\mathcal{K}}\right\}$, where

- $\mathcal{K} = \{1, ..., K\}$ is the set of players,
- $\mathcal{I}$ is the set of pure strategy joint action profile of agents, and
- $g^{(k)}$ is the of reward of player $k \in \mathcal{K}$.

Based on the network classifications described before, it is relatively straight-forward to map the resource allocation problem to a game-theoretical scenario. Several examples are given in the following.

- The resource allocation problem in coordinated networks can be solved by using *game-theoretical auctions* [NRTV07]. Auction models usually cause mild overhead as a result of communications among users (bidders) and the coordinator (auctioneer). Typical auction models include the second price auction, combinatorial auction and share auction. Examples are [CWWL10], [GXW11] and [HHCP08], among many others. Moreover, [XSH+12b] and [XSH+12a] use auction theory specifically for resource allocation in underlay D2D networks. Reference [ZLNW13] provides a comprehensive survey of auction games in wireless networks.

- *Cooperative game theory* [BDT08] is a branch of game theory that is widely used in order to solve the resource allocation problem in connected networks. Through pairwise communications, players construct coalitions, and users in each coalition enforce cooperative behavior so as to maximize the value of the coalition, which is considered as a utility measure. The players have to agree on some contract in order to share the utility achieved by the coalition. Coalitional games can be divided into games with and without transferable utility. Moreover, they are either strategic or of partition form. References [GVJ10], [BF11] and [ZYC12] are some examples. Application of cooperative games in D2D networks is discussed for instance in [SNHH14]. In addition, References [SHD+09] and [YFX12] provide comprehensive study on using cooperative game theory for wireless resource allocation.

- In connected networks with self-interested nodes that affect the utility of each other, *non-cooperative game theory* [NM44] is utilized to address resource allocation problems, specially power control. Moreover, models based on *exchange economy* [JR11], for instance virtual oligopoly market model, are conventionally used for solving channel allocation problems, mostly in conjunction with game theory such as Stackelberg or Cournot games. In such models users are regarded as buyers that are willing to purchase bandwidth. Buyers compete with each other while communicating with the seller (or sellers) in order to achieve an agreement on the price of resources. Therefore, at least a control channel must be reserved to accommodate the price-demand information flow. Research studies that apply these techniques include [LTH+07], [MKZ09] and [HJL07], to name just a few. References [WWJ+14] and [WSH+13] use non-cooperative games specifically for D2D communications, and [NH07] provides a survey.

It should be noted that the above-mentioned games can be played as a single-shot or repeated game, finitely or infinitely. Under the assumption that players are fully rational, an equilibrium (usually Nash equilibrium) is conventionally considered as the desired outcome of the game. An important merit of game-theoretical models is the rapid convergence, as in most of the models that include communications among nodes, such as price-demand market models or combinatorial auctions, a steady state is achieved after only few iterations [1], [WHL09].

Despite vast area of application, many game-theoretical models suffer from some shortcomings. Most importantly,

- Each player requires to know at least its own utility function a priori, which is sometimes very difficult or even impossible to acquire.

- It is a common practice to include some coordinator in the model, although its existence cannot be guaranteed; even if such entity is available, it is not clear how it comes to the common knowledge of distributed players. Moreover, in the event that the model includes no coordinator, at least partial connectivity is required.
- Some models necessitate frequent communications among players, which yields excessive overhead.
- The assumption of rationality in non-cooperative games is not always valid. Moreover, in cooperative games, it is difficult to establish the incentive-compatibility of cooperation model.
- Uncertain and time-varying nature of wireless medium and network is not taken into account.
- In a vast majority of models, game parties cannot be anonymous to each other.

Hence, with regards to the classifications described in Section 1.1.1, pure game-theoretical models do not span all possible scenarios that arise in wireless networks, in particular a scenario of conflicting users, where players do not have any prior knowledge and also the overhead should remain low.

1.1.3. Reinforcement Learning

In order to compensate for the lack of information and connectivity, and also to take the uncertainty into account, one can resort to the learning theory. Learning methods are conventionally divided into three groups, namely *supervised*, *unsupervised* and *reinforcement learning*. For brevity, in this thesis, by learning we refer only to the last category, i.e. reinforcement learning, as the other two models are not relevant to this thesis.

In the area of wireless networking, reinforcement learning is mainly concerned with sequential online decision making in an unknown and uncertain environment with different possible states. In a typical form of this problem, an agent (the decision maker) repeatedly selects an action from a finite set of actions in order to receive some a priori unknown reward. Based on the achieved rewards, system state, and also by using other (possibly) revealed information, agent's actions are expected to evolve over time so as to satisfy some long run optimality condition, conventionally defined in terms of cumulative or average reward. A typical learning model can be formally defined as $\mathfrak{L} \coloneqq \{\mathcal{S}, \mathcal{M}, \mathbf{T}, \{f_m\}_{m\in\mathcal{M}}, \mathbf{x}\}$, where

- $\mathcal{S}$ is the set of states,

- $\mathcal{M}$ is the set of actions,
- $\mathbf{T}$ is the state transition matrix,
- f_m is the mean reward process of action $m \in \mathcal{M}$, and
- $\mathbf{x}$ is the vector of observations.

When addressed by learning models, the online decision making problem is mostly formulated in one of the two main frameworks, either *partially observable Markov decision processes* with unknown rewards and/or state transition matrix [Put94] or *multi-armed bandits* [CBL06]. In this thesis we only use the bandit model; as a result, in what follows, Markov decision processes are mentioned only briefly for the sake of completeness, whereas bandit models are described in more details.

- Markov decision processes (MDPs): A Markov decision process is a discrete time stochastic control process. At each time step, the process is in some state, and the decision maker (agent) may choose any action that is available in the current state. After selecting an action, the agent receives some reward associated to the played action in that state, and the process randomly moves to a new state according to some transition matrix. If the state is unknown while making the decision, the problem is called a partially observable Markov decision process (PO-MDP). In the absence of prior knowledge about rewards or state transition matrix, the problem is mostly solved using learning methods.

- Multi-armed bandits (MAB): Multi-armed bandit problem was first introduced in [Rob52] and [Bel56]. In the most basic setting, the problem models an agent that faces the challenge of sequentially selecting an arm from a set of arms in order to receive an a priori unknown reward. This problem corresponds to the exploration-exploitation dilemma, i.e. the conflict between taking actions that yield immediate rewards and taking actions that would result in reward only in future, for instance activating an inferior arm only to acquire information. Multi-armed bandit model benefits from a wide range of variations in the setting, and hence can be considered as an appropriate model for wireless resource allocation problem under uncertainty and in the absence of prior knowledge. Some important variations of the bandit problem are described in the following.
  - Bandit models are either *stateful* or *stateless*. Stateful models are similar to MDPs in the sense that the reward of each arm depends on the current state, which changes over time according to some transition matrix. Therefore they are also referred to as *Markovian bandits* [BCB12].

- Stateful bandits are conventionally divided into two groups. In *rested bandit* models, at each round, only the state of the played arm changes, while in *restless* models, the state of every arm is subject to change [GMS10].
- Based on the random nature of arms' reward functions, stateless bandit problems are divided into two classes, namely *adversarial* and *stochastic* [ACBFS03], [BCB12].
- *Bandits with side-information* is a class of bandit problems where in addition to the reward, some side-information is revealed to the learning agent at each round of decision making [WKP05]. In the literature this setting is also called *contextual bandits* [BCB12] and *covariate bandits* [YZ02], [Cla89].
- *Mortal bandits* or *sleeping bandits* refer to the bandit problems in which each arm is only available for a finite number of trials [CKRU08].

In addition to the variants mentioned above, the set of arms can be finite or infinite [WAM08], and arms can be either independent or dependent [PCA07].[6] In computer science and mathematics, multi-armed bandits is considered as a classic theory, giving rise to a large body of related literature. In essence, numerous solutions are developed for each type of bandit problems. We describe some important solution concepts briefly in the following.

- Markovian bandit problems are often solved using *indexing policies*, pioneered by the Gittins index [Git79]. Roughly speaking, at each round, a real scalar value, referred to as index, is associated to each arm. The index of any arm is counted as a measure of the reward that can be achieved by activating that arm in the current state. The arm with the highest index is played at each round. Before using an indexing policy, the indexability of the problem has to be verified that is in general not trivial (see for instance [NM01]). In addition to the Gittins indices, Whittle's index policy [Whi80] is well-known and often used. Other works include [Web92] and [KJ87]. A survey can be found in [GGW11].
- In order to solve the stochastic bandit problem, many methods are developed so far that are based on the *upper confidence bound* policy, first proposed in the seminal work of Lai and Robbins [LR85]. In this method, in order to deal with the exploration-exploitation dilemma, an upper-bound of the mean reward of each arm is estimated at some fixed confidence level. The arm with the highest estimated bound is then played. Some important works that adopt this basic concept are [ACBF02], [AO10] and [BCB12], to name a few.

[6] In the basic setting that we consider in this thesis the set of arms is finite and arms are independent.

- Adversarial bandits are mainly solved by using *potential-based* or *weighted average* approaches [ACBFS03]. In the most basic form of these methods, at each trial, a mixed strategy is calculated over the set of arms. The selection probability of each arm is proportional to its average performance in the past, possibly weighted by a specific potential function (for example the exponential function). Accordingly, the actions with better past performance are more likely to become activated in the future, and vice versa. Examples can be found in [CBFH$^+$97] and [CBL06].

As a result of emerging new networking paradigms such as self-organized, cognitive and distributed systems, decision making by using reinforcement learning finds numerous applications in wireless networking scenarios. In what follows we provide some examples. It should be however emphasized that the list provided below might not be exhaustive, and the application of learning models, in particular PO-MDPs and MABs, is by no means limited to the following problems.

- Distributed channel allocation: Application of bandit theory in distributed channel allocation has been the topic of numerous research studies. For example, in [GKJ10], a combinatorial MAB model is used for multi-user channel allocation. Reference [DL13] provides some solutions for channel selection problem in cognitive radio networks, by using a restless MAB model. In [LJP08], the authors propose a Markovian bandit model for channel selection, where the availability of each channel is a Markov chain. Moreover, a two-by-two channel allocation problem is considered in [Li09], and Q-learning is utilized for solving the formulated problem. Dynamic channel assignment by using Q-learning is addressed also in [NH99]. Reference [HGF13] studies distributed channel access with limited information, where switching between channels is costly. Similar examples include [LZ09], [LPT13], [LZ10b], [LZ10a] and [FYX13].

- Relay selection: In [5] and [6], a relay selection problem in two-hop wireless networks is addressed, where given limited information, every transmitter-receiver pair selects one of the available relays in order to improve data transmission. The problem is solved using stochastic and adversarial bandit games. Reference [CJL11], on the other hand, utilizes a restless MAB model in order to develop a cross-layer approach for joint relay selection and physical-layer adaptive modulation and coding, which maximizes the throughput of a cognitive radio network.

- Channel/Transmission scheduling: In [JMMM13], the wireless system is modeled as a network with parallel queues. Each channel can have one of the two states, which evolve according to a Markov process. Provided with no state information but given

the queues' lengths, the scheduler selects one queue at a time. This problem is then solved by using a PO-MDP model. Similarly, by applying a restless bandit model, [OMES11] exploits the channel memory for joint estimation and downlink scheduling. In [CZKD06a] and [CZKD07], the authors use a stochastic bandit formulation to cast the transmission scheduling in wireless sensor networks as a shortest path problem.

- Sensor scheduling: References [KD07] and [CZKD06b] analyze the sensor scheduling problem, where the number of sites to be surveyed is larger than that of available sensors. The problem is solved by using PO-MDP or MAB games.

- Transmission mode selection: In [9], an underlay D2D network is considered, where users are allowed to select one of the two transmission modes, namely direct mode and cellular mode. The problem of selecting the optimal mode is formulated as a two-armed bandit game with one risky arm and one safe arm. The formulated problem is solved by means of statistical hypothesis testing.

- User selection: Reference [SYJL10] proposes a sender selection scheme by using restless bandit models, with the goal of maximizing the receiving data rate and minimizing the energy consumption. Similarly, in [SYJL08], the authors utilize indexing policies to solve the bandit-theoretical formulation of a sender selection problem.

- Power control: In [VH04a], reinforcement learning is used to solve the power control problem. More precisely, a Q-learning algorithm is developed to adjust the power to the system states by means of reward values. In [CZZ13], multi-user Q-learning is used for power control in a cognitive radio network. Joint relay selection and power control is formulated as a restless bandit problem in [LMY+13].

- Channel sensing: Reinforcement learning has also been utilized for solving channel sensing problems in cognitive radio networks. For instance, channel sensing using fuzzy Q-learning is proposed in [PO13]. In [HLF11], a channel sensing problem in cognitive radio networks is modeled by a PO-MDP, and an approach is proposed for optimal decision making. Moreover, non-Bayesian channel sensing using bandit theory is addressed in [LYW+11]. Reference [OKP12] is another example of channel sensing using bandit models.

- In addition to the areas mentioned above, bandit models are also used in social queries [BPSF13], dynamic pricing [ZTL+11] and reconfigurable antennas [GD14].

However, in most of the research works mentioned above, at least one of the following questions is left open: i) *How to apply and analyze the developed learning model in a*

conflicting, reactive environment, and ii) *If the developed learning model is applied to a multi-agent scenario, does it yield a steady state (equilibrium)?* In other words, it is not clear what would be the outcome of a repeated non-cooperative game in which players (or some of them) play according to the proposed learning models.

1.1.4. Game Theory Meets Reinforcement Learning

In a non-cooperative multi-agent system, the reward of each action not only depends on the (probably) random environment, but also on the joint action profile of players. Therefore the learning model is generalized to $\mathfrak{L}_{\mathfrak{G}} \coloneqq \left\{\mathcal{K}, \mathcal{I}, \mathcal{S}, \mathbf{T}, \left\{f_m^{(k)}\right\}_{k\in\mathcal{K}, m\in\mathcal{M}^{(k)}}, \left\{\mathbf{x}^{(k)}\right\}_{k\in\mathcal{K}}\right\}$, where

- $\mathbf{x}^{(k)}$ is the observations vector of player $k \in \mathcal{K}$,
- $\mathcal{M}^{(k)}$ is the set of actions available to player $k \in \mathcal{K}$,[7] and
- $f_m^{(k)}$ is the mean reward process of action $m \in \mathcal{M}^{(k)}$ for player $k \in \mathcal{K}$ [14].

Given no prior information, every agent requires to interact with the random environment and other agents in order to solve the problems that arise in an unknown reactive model, for instance the long-term accumulated reward maximization, or average regret minimization. As a result of competition, which is inherent in non-cooperative multi-agent systems, learning models that ignore the presence of multiple agents might not yield satisfactory outcomes; in particular, equilibrium is not guaranteed to be achieved. This is when game theory and reinforcement learning meet and complement each other. Thereby, equilibrium arises as an asymptotic outcome of repeated interactions in a random environment among learning agents with bounded rationality that aim at achieving long run optimality in some sense. This problem is currently under intensive investigation, mainly in computer science and mathematics. For example, [CLRJ13] proposes convergent learning algorithms especially for potential games. References [KF08] and [FV97] discuss the relation of calibration with Nash and correlated equilibria, respectively. Convergence to Nash equilibrium in unknown games is studied in [FY03] and [GL07]. Moreover, the two seminal books [CBL06] and [FL96] comprehensively explore the problem of learning and prediction in games from the theoretical point of view.

In the theory of wireless communications there is also some ongoing research on game-theoretical learning. However, many of the developed solutions are of strictly limited applicability, as the algorithms are mostly designed for a specific game identified by some

[7] Throughout the thesis, we assume that all players have access to one action set, i.e. $\mathcal{M}^{(k)} = \mathcal{M}$, for all $k \in \mathcal{K}$. Note, however, that all results also hold for the case where every player k has its own action set $\mathcal{M}^{(k)}$.

utility function that is independent of users (not user-specific). More precisely, some works such as [XWW+12], [XWS+13] and [XWW+13], consider a game with some specific (but unknown to players) utility function, and then establish the convergence for the defined game. In these works, it is also assumed that any given action pays equal rewards to all players, which is certainly not realistic in wireless networks.

1.2. Contributions of the Thesis

In this thesis we develop some game-theoretical learning models with the goal of solving a resource allocation problem in distributed D2D wireless networks. The developed strategies are general in the sense that their convergence characteristics does not depend on any specific game model, and the reward generating functions of arms are assumed to be user-specific. Moreover, the strategies are efficient from the users' point of view and also require a relatively low overhead. Detailed description of contributions is presented in the following.

Chapter 2 deals with the problem of efficient resource allocation in dynamic infrastructureless wireless networks. In a reactive interference-limited scenario, at each transmission trial, every transmitter selects a frequency channel from some common pool, together with a power-level. As a result, for all transmitters, not only the fading gain, but also the number and the power of interfering transmissions vary over time. Due to the absence of a central controller and time-varying network characteristics, it is highly inefficient for transmitters to acquire global channel and network knowledge. Therefore, given no information, each transmitter selfishly intends to maximize its average reward, which is a function of the channel quality as well as the joint selection profile of all transmitters. This scenario is modeled as an adversarial multi-player multi-armed bandit game, where players attempt to minimize their so-called regret, while at the network side desired is to achieve equilibrium in some sense. Based on this model and in order to solve the resource allocation problem, we develop two joint power-level and channel selection strategies. We prove that the gap between the average reward achieved by our approaches and that based on the best fixed strategy converges to zero asymptotically. Moreover, the empirical joint frequencies of the game converge to the set of correlated equilibria.
The results presented in this chapter are partially published in [5] and [8].

In Chapter 3 we consider the distributed channel selection problem in the context of D2D communications as an underlay to a cellular network. Underlaid D2D users communicate directly by utilizing the cellular spectrum but their decisions are not governed by any central controller. Selfish D2D users that compete for access to the resources form a distributed system where the transmission performance depends on channel availability

and quality. This information, however, is difficult to acquire. Moreover, the adverse effects of D2D users on cellular transmissions should be minimized. This scenario is modeled as a stochastic multi-player multi-armed bandit game with side-information, for which a distributed algorithmic solution is proposed. The solution is a combination of no-regret learning and calibrated forecasting, and can be applied to a broad class of multi-player stochastic learning problems, in addition to the formulated channel selection problem. Theoretical analysis shows that the proposed approach not only yields vanishing regret, but also guarantees that the empirical joint frequencies of the game converge to the set of correlated equilibria.
The materials of this chapter are partially published in [6], [7] and [10].

Chapter 4 studies a resource allocation problem in a single-cell wireless network with multiple D2D users sharing the available radio frequency channels with cellular users. The priority of using limited wireless resources is however given to the cellular users. We consider a rather realistic scenario where the BS is provided with strictly limited channel knowledge while D2D and cellular users have no information. We prove a lower-bound for the cellular aggregate utility in the downlink with fixed BS power, which allows for decoupling the channel allocation and D2D power control problems. Channel allocation is performed by using a suboptimal, but efficient, heuristic approach that consists of bipartite matching and graph partitioning. Depending on the defined criterion (aggregate utility maximization, fairness, quality of service guarantee), the approach assigns one cellular user and (possibly) multiple D2D users to each channel. In each channel, every D2D user therefore gropes for optimality in the sense of utility maximization by adjusting its transmit power. We model this scenario as a game with unknown rewards, defined on a discrete strategy set. The game is shown to be an exact potential game, and the set of Nash equilibria is characterized. We further use a Q-learning better-reply dynamics to achieve Nash equilibrium.
The results of this chapter are partially published in [12] and [11].

1.3. Further Results

In order to keep the consistency, some parts of the results that are obtained and published during my Ph.D. studies are not included in this thesis. Below is a brief description.

The research work [9] studies D2D communications underlaying cellular infrastructure. In such networks, each device pair is provided with two transmission modes: indirect and direct. Indirect transmission is a two-hop interference-free transmission via a base station. Despite being interference-free, this transmission type might be inefficient in communications scenarios where short-distance connections can be established. Moreover, the need

for centralized resource allocation and deployment of extra hardware may lead to excessive complexity and unacceptable costs. In such scenarios, direct transmissions can utilize the proximity- and hop gains to achieve higher rates and lower end-to-end latencies. While having a potential for huge performance gains, direct D2D communications poses some fundamental challenges resulting from the absence of a devoted controller such as uncoordinated interference and unavailability of permanent direct channels. Roughly speaking, in an average sense, indirect transmission pays safe and steady reward, whereas direct transmission is risky, yielding a stochastic reward. While this random reward can be larger than the guaranteed reward of indirect transmission, there is also the possibility that it becomes lower than that, despite the proximity- and hop gains. Transmitters should therefore choose the most efficient transmission mode in the presence of limited information. The paper characterizes the reward process for each transmission mode to model the mode selection problem as a two-armed Levy-bandit game. Accordingly, the reward of the risky arm (direct mode) is considered to be a pure-jump Levy process, following compound Poisson distribution. Mathematical results from bandit and learning theories are used to solve the selection problem. Numerical results are also included.

In [13] we consider a distributed channel selection problem for D2D transmission underlaying conventional cellular networks. Specifically, underlaying devices exploit the (possibly) idle licensed cellular spectrum in order to establish direct communication links, and transmit using rateless coding under energy constraint. While the quality of each channel is assumed to be stochastic, the availability is non-stochastic (adversarial). Moreover, cellular channels are idle only for some finite time. As there might be numerous such primary channels, acquiring prior information about channel quality and/or availability yields excessive cost; therefore we assume that D2D devices do not possess any prior information. Device pairs face the problem of selecting a suitable channel so that a successful data delivery under the energy constraint is guaranteed. We model this problem as a multi-armed bandit game with mortal arms, and provide an algorithmic solution. The proposed game model and solution are evaluated analytically and numerically.

In [1], a hybrid centralized-decentralized resource allocation scheme for the downlink of two-hop relay-enhanced transmission is proposed. In the first-hop, centralized joint power and subcarrier allocation is performed. The second-hop is modeled as a virtual supply-demand market, and a two-level game is designed that converges to Nash equilibrium. The proposed scheme reduces the feedback overhead and exploits the resources efficiently, while taking the fairness into account.

In [2], we consider a delay-constrained application in relay-enhanced cognitive radio networks, where secondary users interfere with primary users. We assume that the secondary users utilize rateless coding, and develop a relay selection scheme. The proposed

scheme provides joint benefits of two well-known relaying strategies, namely relay subset selection and incremental relaying. Both analytically and numerically, we show that the scheme not only reduces the feedback overhead significantly, but also minimizes the outage probability. Moreover, resources are exploited efficiently.

Research study [3] proposes a new transmission protocol for two-hop transmission employing rateless coding in the first hop and rateless-network coding in the second hop. In this protocol, device pairs and also relays are partitioned into small clusters; afterwards, each cluster of device pairs is assigned one relay cluster, to be used in order to improve the transmission performance. The scheme increases the diversity order (in comparison to the case where each pair utilizes its own relay with no cooperation), without increasing the transmission cost (such as delay or energy), provided that the clustering-assignment form is optimized in the sense of cost minimization. We show that finding the optimal clustering-assignment can be relaxed to the cascade of two graph-theoretical problems, namely graph partitioning and weighted matching, which can be solved efficiently. Moreover, we model the delay as cost, and analyze the expected transmission time.

Finally, [4] develops a joint power allocation and relay selection scheme for bidirectional relay-enhanced transmission utilizing network coding. The power allocation scheme aims at minimizing the required transmit power while providing the desired quality of service expressed in terms of outage probability. The relay selection task is formulated as a weighted matching problem, targeting at either minimizing some cost (wasted power or delay) or maximizing some utility (goodput). Our approach relies only on statistical channel knowledge, thereby reducing the feedback and signaling overhead compared to the approaches that require instantaneous channel knowledge.

Copyright Information

2. Distributed Resource Allocation in Adversarial Networks

In this chapter our focus is on a multi-agent decision making problem in an adversarial environment. The self-interested agents are provided with no information and intend to efficiently learn the optimal joint action profile by successive interactions. A particular instance of this scenario is the resource allocation problem in an infrastructureless wireless network, as considered here. In Section 2.1, we briefly summarize some important definitions and results in the area of adversarial bandit theory that are used in our model and analysis. Afterwards, in Section 2.2, we describe the system model and formulate the joint power control and channel allocation problem as an adversarial multi-player multi-armed bandit game. With the aim of efficient resource management and interference mitigation, we follow an approach suggested in [BM07] to develop two joint power control and channel selection algorithms. The first strategy, described in Section 2.3, is an adapted version of *exponential-based weighted average* allocation rule [ACBFS03], while the other, described in Section 2.4, is based on *follow the leader* strategy [KE05]. The developed strategies not only result in small (that is, with sublinear growth in time) regret for each individual player, but also guarantee that the empirical joint frequencies of the play converge to the set of correlated equilibria. Moreover, in Section 2.5, we implement the *experimental regret-testing procedure*, which is known to converge to the set of Nash equilibria of the game [GL07].

Our work extends the state-of-the-art in this area significantly since it differs from the existing research studies in at least the following crucial aspects:

- Unlike many previous works including [GLM04] and [FYX13] that study the single-agent learning problem, we analyze the multi-agent setting, while taking the selfishness of players into account.

- Our model and algorithms do not rely on the assumption that the reward generating process of every action is time-invariant. In fact, reward functions might vary arbitrarily; as a result the model captures the dynamic nature of wireless channels and distributed networks. This is in contrast to a great majority of previous works,

including [XWW+12], [XWS+13] and [KNJ14], where the reward generating process of every arm is assumed to be time-invariant.

- In contrast to [KNJ12] and [XWW+12], we neither allow pairwise information exchange, nor use a control channel. Therefore the overhead is minimized.

- Moreover, players do *not* observe the actions of each other. As a result, the algorithms are incentive-compatible and do not require any cooperation, thereby offering high applicability. An exemplary application is a power control problem with unknown power-levels used by other transmitters.

- In our system model, channel qualities are taken into account so that channels pay different rewards to different users; that is, contrary to [XWW+12] and [XWS+13], the reward generating functions are user-specific. In addition, we impose no limitation on the interference pattern.

- The convergence analysis is valid for a wide range of games. This is in contrast to many previous works where the game should be necessarily potential for the convergence analysis to hold. Example of such works are [XWW+12], [XWS+13] and [XWW+13], among many others.

2.1. Adversarial Multi-Player Multi-Armed Bandit Games

2.1.1. Notions of Regret

Multi-player multi-armed bandit problem (MP-MAB, hereafter) is a class of sequential decision making problems with limited information. In this game, there exists a set of players $\mathcal{K} = \{1, ..., K\}$. Each player $k \in \mathcal{K}$ is assigned M actions that are indexed by integer numbers; the action set therefore yields $\mathcal{M} = \{1, ..., M\}$. Every player selects an action at successive trials in order to receive an initially unknown reward that depends not only on its own actions, but also on those of other players. The action set, the played action and the reward achieved by each player are regarded as private information. The reward generating processes of arms are independent. Let $\mathcal{I}$ be the set of joint action profile. Accordingly, $\mathbf{i}_t = \left(i_t^{(1)}, ..., i_t^{(k)}, ..., i_t^{(K)}\right) \in \bigotimes_{k=1}^{K} \{1, ..., M\}$ denotes the joint action profile of players at time t, with $i_t^{(k)} \in \mathcal{M}$ being the action of player k. Clearly, $\mathbf{i}_t = \left(i_t^{(k)}, \mathbf{i}_t^{(-k)}\right)$, where $\mathbf{i}_t^{(-k)} \in \bigotimes_{k=1}^{K-1} \{1, ..., M\}$ is the the joint action profile of all players except for k, at time t. Moreover, let $g_t^{(k)}(\mathbf{i}_t) \in [0, 1]$ be the reward achieved by some player k at

time t.[1] The instantaneous regret of any player k is defined as the difference between the reward of the optimal action,[2] and that of the played action. Based on this definition, the *cumulative regret* of player k is formally defined in the following.

Definition 1 (Cumulative Regret)**.** *The cumulative regret of player k up to time n is defined as*

$$R_n^{(k)} = \max_{i=1,\dots,M} \sum_{t=1}^{n} g_t^{(k)} \left(i, \mathbf{i}_t^{(-k)}\right) - \sum_{t=1}^{n} g_t^{(k)} \left(i_t^{(k)}, \mathbf{i}_t^{(-k)}\right). \tag{2.1}$$

Each player aims at minimizing its accumulated regret, which is an instance of the well-known exploitation-exploration dilemma: Find a balance between exploiting actions that have exhibited well performance in the past (control) on the one hand, and exploring actions which might lead to a better performance in the future (learning) on the other hand.

Now, suppose that players use mixed strategies and let $\mathcal{P}$ denote the set of all possible probability distributions over M actions. This means that, at each trial t, player k selects a probability distribution $\mathbf{p}_t^{(k)} = \left(p_{1,t}^{(k)}, ..., p_{i,t}^{(k)}, ..., p_{M,t}^{(k)}\right)$ over arms, and plays arm i with probability $p_{i,t}^{(k)}$. In this case, we resort to the expected regret, also called *external regret* [CBL06], defined as follows.

Definition 2 (External Regret)**.** *The external cumulative regret of player k up to time n is defined as*

$$\begin{aligned} R_{\text{Ext}}^{(k)} \coloneqq R_{\text{Ext}}^{(k)}(n) &= \max_{i=1,\dots,M} \sum_{t=1}^{n} g_t^{(k)} \left(i, \mathbf{i}_t^{(-k)}\right) - \sum_{t=1}^{n} \bar{g}_t^{(k)} \left(\mathbf{p}_t^{(k)}, \mathbf{i}_t^{(-k)}\right) \\ &= \max_{i=1,\dots,M} \sum_{t=1}^{n} \sum_{j=1}^{M} p_{j,t}^{(k)} \left(g_t^{(k)} \left(i, \mathbf{i}_t^{(-k)}\right) - g_t^{(k)} \left(j, \mathbf{i}_t^{(-k)}\right)\right), \end{aligned} \tag{2.2}$$

where $\bar{g}_t^{(k)}(\cdot)$ denotes the expected reward at round t by using mixed strategy $\mathbf{p}_t^{(k)}$, defined as $\bar{g}_t^{(k)}(\cdot) = \sum_{j=1}^{M} g_t^{(k)}(j, \cdot) p_{j,t}^{(k)}$.

By definition, external regret compares the expected reward of the current mixed strategy with that of the best fixed action in the hindsight, but fails to compare the rewards achieved by changing actions in a pairwise manner. In order to compare actions in pairs, *internal regret* [CBL06] is introduced that is closely related to the concept of equilibrium in games.

[1] Note that all results can be also expressed in terms of loss (d), provided that the loss is related to the gain by $d = 1 - g$, $g \in [0, 1]$.

[2] Optimality is defined in the sense of the highest reward.

Definition 3 (Internal Regret)**.** *The internal cumulative regret of player k up to time n is defined as*

$$\begin{aligned} R_{\text{Int}}^{(k)} \coloneqq R_{\text{Int}}^{(k)}(n) &= \max_{i,j=1,\dots,M} R_{(i\to j),n}^{(k)} \\ &= \max_{i,j=1,\dots,M} \sum_{t=1}^{n} p_{i,t}^{(k)} \left(g_t^{(k)} \left(j, \mathbf{i}_t^{(-k)} \right) - g_t^{(k)} \left(i, \mathbf{i}_t^{(-k)} \right) \right). \end{aligned} \tag{2.3}$$

Notice that on the right-hand side of (2.3), $r_{(i\to j),t}^{(k)} = p_{i,t}^{(k)} \left(g_t^{(k)}(j,\cdot) - g_t^{(k)}(i,\cdot) \right)$ denotes the expected regret caused by pulling arm i instead of arm j. By comparing (2.2) and (2.3), the external regret can be bounded above by the internal regret as [SL05]

$$R_{\text{Ext}}^{(k)} = \max_{i=1,\dots,M} \sum_{j=1}^{M} R_{(i\to j),n}^{(k)} \leq M \max_{i,j=1,\dots,M} R_{(i\to j),n}^{(k)} = M R_{\text{Int}}^{(k)}. \tag{2.4}$$

Remark 1. *Throughout this chapter, vanishing (zero-average) external and internal regret means that $\lim_{n\to\infty} \frac{1}{n} R_{\text{Ext}} = 0$ and $\lim_{n\to\infty} \frac{1}{n} R_{\text{Int}} = 0$, respectively. In other words, we have $R_{\text{Ext}} \in o(n)$ and $R_{\text{Int}} \in o(n)$. Note that by (2.4), $R_{\text{Int}} \in o(n)$ implies $R_{\text{Ext}} \in o(n)$. We call any strategy with $R_{\text{Int}} \in o(n)$ as "no-regret strategy".*

2.1.2. Equilibrium

From the view point of each player k, an MP-MAB is a game with two agents: player k itself, and the *set* of all other $K-1$ players (referred to as the opponent), whose joint action profile affects the reward achieved by player k. We consider here the most general framework, where the opponent is non-oblivious, i.e. its series of actions depends on the actions of player k. It is known that a game against a non-oblivious opponent can be modeled by adversarial bandit games [BCB12], where similar to other game-theoretical formulations, desired is to achieve an efficient equilibrium, most importantly Nash and correlated equilibria [NRTV07], which are defined in Section 1 of Appendix A.[3]

In the context of game-theoretical bandits, an important result is the following theorem.

Theorem 1 ([CBL06])**.** *Consider a K-player bandit game, where each player k is provided with an action set of cardinality M. Denote the internal regret of player k by $R_{\text{Int}}^{(k)}$, and the set of correlated equilibria by $\mathcal{C}$. At time n, define the empirical joint distribution of*

[3] As a general rule, all standard definitions that are used frequently in this thesis are stated in the appendix.

the game as

$$\hat{\pi}_n(\mathbf{j}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{\{\mathbf{i}_t=\mathbf{j}\}},\ \mathbf{j} = \left(j^{(1)}, ..., j^{(K)}\right) \in \bigotimes_{k=1}^{K} \{1, ..., M\}, \tag{2.5}$$

where $\mathbf{1}_{\{x\}}$ *is the indicator function that returns one if* x *holds and zero otherwise. If all players* $k \in \{1, ..., K\}$ *play according to any strategy so that*

$$\lim_{n\to\infty} \frac{1}{n} R_{\text{Int}}^{(k)} = 0, \tag{2.6}$$

then the distance $\inf_{\pi\in\mathcal{C}} \sum_{\mathbf{j}} |\hat{\pi}_n(\mathbf{j}) - \pi(\mathbf{j})|$ *between the empirical joint distribution of plays and the set of correlated equilibria converges to* 0 *almost surely.*

Theorem 1 simply states that in an MP-MAB game, if all players play according to a strategy with vanishing internal regret (no-regret), then the empirical joint distribution of plays converges to the set of correlated equilibria. Note that the strategies used by players are not required to be identical. Since a rational player is interested in minimizing its regret, the assumption that every player plays according to some no-regret strategy is reasonable.

2.1.3. From Vanishing External Regret to Vanishing Internal Regret

In [SL05], an approach is proposed for converting any selection strategy with vanishing external regret to another version with vanishing internal regret. We describe this approach briefly in what follows. The player index (k) is omitted for convenience.

Consider some selection strategy κ that at each time t selects one of the M actions according to some probability distribution $\mathbf{p}_t$. Let $\mathbf{p}_1$ be the uniform distribution. In order to calculate $\mathbf{p}_t$ for $t > 1$, κ constructs a meta-strategy κ' with $M(M-1)$ virtual actions $(i \to j)$, $(i, j \in \{1, ..., M\}, i \neq j)$. Assume that κ' uses some mixed strategy $\mathbf{w}_t$ over $M(M-1)$ virtual actions, where the probability of the virtual action $(i \to j)$, i.e. $w_{(i\to j),t}$, depends on its past performance in some way.[4] Given $\mathbf{w}_t$ and $\mathbf{p}_{t-1} = (p_{1,t-1}, ..., p_{i,t-1}, ..., p_{j,t-1}, ..., p_{M,t-1})$, κ defines $\mathbf{p}_{(i\to j),t-1}$, which has 0 and $p_{j,t-1} + p_{i,t-1}$ at the place of $p_{i,t-1}$ and $p_{j,t-1}$, respectively, and all other elements remain unchanged; that is, $\mathbf{p}_{(i\to j),t-1} = (p_{1,t-1}, ..., 0, ..., p_{j,t-1} + p_{i,t-1}, ..., p_{M,t-1})$. Then $\mathbf{p}_t = \sum_{(i,j):i\neq j} \mathbf{p}_{(i\to j),t} w_{(i\to j),t}$. As a result, κ has the characteristic that its internal regret is upper-bounded by the external regret of κ'. Thus, if κ' exhibits vanishing external

[4] Note that the gains of virtual actions cannot be calculated explicitly. Later we see that the gain achieved by any virtual action $(i \to j)$ is calculated based on the gain achieved by playing true actions i and j.

regret, then κ results in vanishing internal regret. In Section 2.3 and 2.4, we use this property in order to design no-regret selection strategies.

2.2. Bandit-Theoretical Model of Infrastructureless Wireless Networks

2.2.1. System Model

We consider a network consisting of a set $\mathcal{K} = \{1, ..., K\}$ of transmitter-receiver pairs. Each pair is referred to as a D2D user and is denoted either by just k or by the pair (k, k'). Every user $k \in \mathcal{K}$ can access a set $\mathcal{M}' = \{1, ..., M'\}$ of mutually orthogonal channels. The transmission power can be selected from a set $\mathcal{M}''$ of M'' quantized power-levels. This implies that the strategy set includes $M = M' \times M''$ actions, where at time t each action $i_t^{(k)} = \left(i_t^{\prime(k)}, i_t^{\prime\prime(k)}\right)$ consists of one channel index $i_t^{\prime(k)}$ (which corresponds to some channel quality), and one power-level $i_t^{\prime\prime(k)}$. Therefore, the joint action profile of users, $\mathbf{i}_t$, is to be understood here as the pair $(\mathbf{i}'_t, \mathbf{i}''_t)$, where $\mathbf{i}'_t = \left(i_t^{\prime(1)}, ..., i_t^{\prime(K)}\right)$ and $\mathbf{i}''_t = \left(i_t^{\prime\prime(1)}, ..., i_t^{\prime\prime(K)}\right)$. As each channel might be accessible by multiple users, co-channel interference (collision, interchangeably) is likely to arise. Since users are allowed to select a new channel and to adapt their power-levels at each transmission trial, interference pattern in general changes over time. In addition, the distribution of fading coefficients might be also time-varying so that acquiring channel and/or network information at the level of autonomous transmitters would be extremely challenging and inefficient. Therefore, we assume the following.

Assumption (A1). *Throughout this chapter, we assume that:*

a) Transmitters have no channel knowledge or any other side-information such as the number of users or their selected actions.

b) Users do not coordinate their actions that can be chosen completely asynchronously by each user.

Note that as users do not observe the actions of each other, it might be in their interest to select their actions at the beginning of trials, thereby using the remaining time for data transmission.

2.2.2. Bandit-Theoretical Problem Formulation

In this chapter, we model the joint channel and power-level selection problem as a K-player adversarial bandit game, where player k decides for one of the M actions. For some

joint action profile $\mathbf{i} = \left(i^{(1)}, ..., i^{(k)}, ..., i^{(K)}\right)$, we define the bounded mean reward function of player k to be

$$f_t^{(k)}(\mathbf{i}) = \log_2 \left(\frac{i''^{(k)} |h_{kk',t,i'^{(k)}}|^2}{\sum_{q \in \mathcal{Q}^{(k)}} i''^{(q)} |h_{qk',t,i'^{(k)}}|^2 + N_0} \right) - \alpha i''^{(k)}, \tag{2.7}$$

for some given joint action profile $\mathbf{i} = (\mathbf{i}', \mathbf{i}'')$. In (2.7), $\mathcal{Q}^{(k)}$ is the set of players that interfere with user k in channel $i'^{(k)}$. Throughout the chapter, $|h_{uv,t,x}|^2 \in \mathbb{R}^+$ is used to denote the *average* gain of some channel x between u and v at time t, including path loss and fast fading effects. N_0 is the variance of zero-mean additive white Gaussian noise (AWGN), and $\alpha \geq 0$ is the constant power price factor. The last term in (2.7) is used to penalize excessive transmission power. According to Section 2.1, let $g_t^{(k)}(\mathbf{i}_t) \in [0, 1]$ denote the achieved reward of player k at time t, as a function of joint action profile $\mathbf{i}_t$. We consider a game with noisy rewards where $g_t^{(k)}(\mathbf{i}) = f_t^{(k)}(\mathbf{i}) + \mathtt{C}^{(k)}(\mathbf{i})$, with $\mathtt{C}^{(k)}$ being some zero-mean random variable with bounded variance, independent of all other random variables. As it is well-known, in a non-cooperative game, the primary goal of each selfish player is to maximize its own accumulated reward. Formally, this can be written as

$$\underset{\left\{\left(i_t'^{(k)}, i_t''^{(k)}\right)\right\}_{t=1}^{n}}{\text{maximize}} \sum_{t=1}^{n} g_t^{(k)} \left(\mathbf{i}_t', \mathbf{i}_t''\right), \tag{2.8}$$

where $i_t'^{(k)} \in \mathcal{M}'$ and $i_t''^{(k)} \in \mathcal{M}''$. By Assumption (A1), however, it is clear that the objective function in (2.8) is not available. For this reason, we argue for a less ambitious goal, which is known as *regret minimization*. Formally, each player k attempts to achieve vanishing external regret in the sense that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} R_{\text{Ext}}^{(k)} \\ &= \lim_{n\to\infty} \frac{1}{n} \left(\max_{i=1,...,M} \sum_{t=1}^{n} g_t^{(k)} \left(i, \mathbf{i}_t^{(-k)}\right) - \sum_{t=1}^{n} \bar{g}_t^{(k)} \left(\mathbf{p}_t^{(k)}, \mathbf{i}_t^{(-k)}\right) \right) = 0. \end{aligned} \tag{2.9}$$

In addition to the individual strategy of each user aiming at satisfying (2.9), at the network level it is desired to achieve some steady state, i.e. equilibrium. Therefore, in the remainder of this chapter, we develop algorithmic solutions to the resource allocation problem with a twofold objective in mind: i) external regret of each user should vanish asymptotically according to (2.9) and ii) the actions of all players should converge to equilibrium.

By (2.4), the external regret of every user is upper-bounded by its internal regret. As a result, if all users select their actions according to some no-regret strategy, not only (2.9) is achieved by all of them (see also Remark 1), but also the corresponding game converges

to equilibrium in some sense, which immediately follows from Theorem 1. In Sections 2.3 and 2.4, we present two internal regret minimizing strategies that are shown to solve the game by achieving the two objectives mentioned above. Both algorithms can be applied in a *fully decentralized* manner by each player, since at each time, they only require the set of past rewards of the respective player.

Finally, it is worth noting that to our best knowledge, the set of correlated equilibria for the general time-varying discrete game defined by (2.7) cannot be characterized. However, in what follows, we characterize the set of equilibria for a relaxed version of this game. In doing so, we assume that for all $k \in \mathcal{K}$, the strategy set $\mathcal{M} = \mathcal{M}' \times \mathcal{M}''$ is a convex and compact subset of $\mathbb{R}^2$. With this assumption in mind, consider a game where each player has a time-invariant and bounded mean reward function such as

$$f^{(k)}(\mathbf{i}) = \log_2 \left(\frac{i''^{(k)} |h_{kk',i'^{(k)}}|^2}{\sum_{q \in \mathcal{Q}^{(k)}} i''^{(q)} |h_{qk',i'^{(k)}}|^2 + N_0} \right) - \alpha i''^{(k)}, \tag{2.10}$$

which implies that the average channel gains are time-invariant. By the following proposition, this game has a unique correlated equilibrium.

Proposition 1. *Consider a K-player game where the mean reward function of each player k is defined by (2.10). This game has a unique correlated equilibrium which places probability one on its unique pure strategy Nash equilibrium.*

Proof. See Section 1 of Appendix B. □

2.3. No-Regret Bandit Exponential-Based Weighted Average Strategy

The basic idea of an exponential-based weighted strategy is to assign each action, at every trial, some selection probability which is inversely proportional to the exponentially-weighted accumulated regret (or directly proportional to the exponentially-weighted accumulated reward) caused by that action in the past [HMc01]. Roughly speaking, if playing an action has resulted in large regret in the past, its future selection probability is small, and vice versa.

As described in Section 2.1.1, in a bandit formulation, players only observe the reward of the played action, and not those of others. Therefore the reward of each action i is estimated as [CBL06]

$$\tilde{g}_t^{(k)}(i) = \begin{cases} \frac{g_t^{(k)}(i)}{p_{i,t}^{(k)}} & i = i_t^{(k)} \\ 0 & \text{o.w.} \end{cases}, \tag{2.11}$$

which is an unbiased estimate of the true reward of action i; that is, $\mathrm{E}\left\{\tilde{g}^{(k)}(i)\right\} = g^{(k)}(i)$, where expectation is with respect to the distribution $\mathbf{p}_t$ of the random variable $i_t^{(k)}$. Estimated rewards are afterwards used to calculate regrets. For example, the regret of *not* playing action j instead of action i yields

$$\tilde{R}^{(k)}_{(i\to j),t-1} = \sum_{s=1}^{t-1} \tilde{r}^{(k)}_{(i\to j),s} = \sum_{s=1}^{t-1} p^{(k)}_{i,s} \left(\tilde{g}^{(k)}_s(j) - \tilde{g}^{(k)}_s(i)\right). \tag{2.12}$$

Despite exhibiting vanishing external regret, weighted average strategies yield in general large internal regret; as a result, even if all players play according to such strategies, the game does *not* converge to equilibrium. In the following, we utilize the bandit version of exponentially-weighted average strategy [CBL03], and convert it to an improved version that yields small internal regret, using the approach of Section 2.1.3. The strategy is called no-regret bandit exponentially-weighted average strategy (NR-BEWAS), and is described in Algorithm 1.

From Algorithm 1, NR-BEWAS has two parameters, namely γ_t and η_t. In the event that the game horizon, n, is known in advance, these two parameters are constant over time ($\eta_t = \eta$ and $\gamma_t = \gamma$), and the regret growth rate can be bounded precisely, mainly based on the results of [CBL06]. Otherwise, they vary with time. In this case, vanishing (sublinear in time) internal regret can be guaranteed; nevertheless, this bound might be loose. This discussion is formally summarized in the following propositions.

Proposition 2. *Let* $\eta_t = \eta = \left(\frac{\log(M)}{2Mn}\right)^{\frac{2}{3}}$ *and* $\gamma_t = \gamma = \left(\frac{M^2\log(M)}{4n}\right)^{\frac{1}{3}}$. *Then Algorithm 1 yields* $R^{(k)}_{\mathrm{Int}} \in O\left(\left(nM^2\right)^{\frac{2}{3}}\left(2\log\left(M\right)\right)^{\frac{1}{3}}\right)$, *hence vanishing internal regret.*

Proof. See Section 2 of Appendix B. □

Proposition 3. *Let* $\eta_t = \frac{\gamma_t^3}{M^2}$ *and* $\gamma_t = t^{-\frac{1}{3}}$. *Then Algorithm 1 (NR-BEWAS) yields vanishing internal regret; that is we have* $R^{(k)}_{\mathrm{Int}} \in o(n)$.

Proof. See Section 3 of Appendix B. □

The following corollaries follow from the above propositions and Theorem 1.

Corollary 1. *If all players play according to NR-BEWAS, then the empirical joint frequencies of the game converge to the set of correlated equilibria.*

Proof. The proof is a direct consequence of Theorem 1 and Proposition 2 or Proposition 3. □

Algorithm 1 No-Regret Bandit Exponential-Based Weighted Average Strategy (NR-BEWAS)

1: If the game horizon, n, is known, define γ_t and η_t as given in Proposition 2, otherwise as those given in Proposition 3.
2: Define $\Phi(\mathbf{u}) = \frac{1}{\eta_t} \log\left(\sum_{i=1}^{M} \exp(\eta_t u_i)\right)$, where $\mathbf{u} = (u_1, ..., u_M) \in \mathbb{R}^M$.
3: Let $\mathbf{p}_1^{(k)} = \left(\frac{1}{M}, ..., \frac{1}{M}\right)$ (uniform distribution).
4: Select an action using $\mathbf{p}_1^{(k)}$.
5: Play and observe the reward.
6: **for** $t = 2, ..., n$ **do**
7: Let $\mathbf{p}_{t-1}^{(k)} = \left(p_{1,t-1}^{(k)}, ..., p_{i,t-1}^{(k)}, ..., p_{j,t-1}^{(k)}, ..., p_{M,t-1}^{(k)}\right)$ be the mixed strategy at time $t-1$. Construct $\mathbf{p}_{(i\to j),t-1}^{(k)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{p}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{p}_{(i\to j),t-1}^{(k)} = \left(p_{1,t-1}^{(k)}, ..., 0, ..., p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}, ..., p_{M,t-1}^{(k)}\right)$.
8: Define [CBL06]

$$w_{(i\to j),t}^{(k)} = \frac{\exp\left(\eta_t \tilde{R}_{(i\to j),t-1}^{(k)}\right)}{\sum_{(m\to l):m\neq l} \exp\left(\eta_t \tilde{R}_{(m\to l),t-1}^{(k)}\right)}, \tag{2.13}$$

where $\tilde{R}_{(i\to j),t-1}^{(k)}$ is calculated by using (2.11) and (2.12).
9: Given $w_{(i\to j),t}^{(k)}$, solve the following fixed point equation to find $\mathbf{p}_t^{(k)}$:

$$\mathbf{p}_t^{(k)} = \sum_{(i\to j):i\neq j} \mathbf{p}_{(i\to j),t}^{(k)} w_{(i\to j),t}^{(k)}. \tag{2.14}$$

10: Final probability distribution yields

$$\mathbf{p}_t^{(k)} = (1-\gamma_t)\mathbf{p}_t^{(k)} + \frac{\gamma_t}{M}. \tag{2.15}$$

11: Using the final $\mathbf{p}_t^{(k)}$, given by (2.15), select an action.
12: Play and observe the reward.
13: **end for**

Corollary 2. *Let ϵ-correlated equilibrium approximate correlated equilibrium in the sense that $\bigcap_{\epsilon>0} \mathcal{C}_\epsilon = \mathcal{C}$. Assuming that the game horizon is known and all players play according to NR-BEWAS, then the minimum required number of trials to achieve ϵ-correlated equilibrium yields $\max_{k=1,...,K} \epsilon^{-\frac{3}{2}} O\left(MK\left(M^2 \log(M) + K^2 \log(K)\right)\right)$, which is proportional to $\epsilon^{-\frac{3}{2}}$ and increases polynomially in the number of actions and players.*

Proof. The proof follows from the bound of Proposition 2 and Remark 7.6 of [CBL06].[5] □

2.4. No-Regret Bandit Follow the Perturbed Leader Strategy

Similar to the weighted average strategy presented in the previous section, the strategy *follow the perturbed leader* is an approach to solve online decision making problems. In the basic version of this approach, called *follow the leader* [Han57], the action with the minimum regret in the past is selected at each trial. This rule is however deterministic and thus does not achieve vanishing regret against non-oblivious opponents. Therefore, in *follow the perturbed leader*, the player adds a random perturbation to the vector of accumulated regrets, and the action with the minimum perturbed regret in the past is selected [CBL06]. In [KE07], a bandit version of this algorithm is constructed, where unobserved rewards are estimated. The authors show that the developed algorithm exhibits vanishing external regret. Similar to NR-BEWAS, we here modify the algorithm of [KE07] to ensure vanishing internal regret, so that the convergence to equilibrium is guaranteed. The approach is called no-regret bandit follow the perturbed leader strategy (NR-BFPLS), and is summarized in Algorithm 2.

Algorithm 2 requires the knowledge of the probability assigned to each action by the follow the perturbed leader strategy at every trial. However, in contrast to NR-BEWAS, these probabilities are not assigned explicitly; therefore we explain how to calculate these values.

From (2.16), the selection probability of virtual action $(i \to j) \in \{1, ..., M(M-1)\}$ is the probability that $\tilde{R}_{(i\to j),t-1}$ plus perturbation $q_{(i\to j),t}$ is larger than those of other

[5] Details are omitted to avoid unnecessary restatement of existing analysis.

Algorithm 2 No-Regret Bandit Follow the Perturbed Leader Strategy (NR-BFPLS)

Define $\epsilon_t = \epsilon_n = \frac{\sqrt{\log(n)}}{3\sqrt{Mn}}$, and $\gamma_t = \min(1, M\epsilon_t)$. Note that unlike NR-BEWAS, here we know the game horizon, n, in advance.

Let $\mathbf{p}_1^{(k)} = \left(\frac{1}{M}, ..., \frac{1}{M}\right)$ (uniform distribution).

Select an action using $\mathbf{p}_1^{(k)}$.

Play and observe the reward.

for $t = 2, ..., n$ **do**

Let $\mathbf{p}_{t-1}^{(k)} = \left(p_{1,t-1}^{(k)}, ..., p_{i,t-1}^{(k)}, ..., p_{j,t-1}^{(k)}, ..., p_{M,t-1}^{(k)}\right)$ be the mixed strategy at time $t-1$. Construct $\mathbf{p}_{(i\to j),t-1}^{(k)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{p}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{p}_{(i\to j),t-1}^{(k)} = \left(p_{1,t-1}^{(k)}, ..., 0, ..., p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}, ..., p_{M,t-1}^{(k)}\right)$.

Calculate $\tilde{R}_{(i\to j),t-1}^{(k)}$ using (2.11) and (2.12).

Define $\sigma_{(i\to j),t-1} = \left(\sum_{s=1}^{t-1} \frac{1}{w_{(i\to j),s}^{(k)}}\right)^{\frac{1}{2}}$, which is the upper-bound of conditional variances of random variables $\tilde{R}_{(i\to j),t-1}^{(k)}$ [KE07].

Let $\tilde{R}_{(i\to j),t-1}^{(k)} = \tilde{R}_{(i\to j),t-1}^{(k)} - \sqrt{1+\sqrt{2/M}}\ \sigma_{(i\to j),t-1}\ \sqrt{\log(t)}$ [KE07].

Randomly select a perturbation vector $\mathbf{q}_t$ with $M(M-1)$ elements from two-sided exponential distribution with width ϵ_t.

Consider a selection rule which selects the action $(i \to j)$ given by

$$\text{argmax}\left\{\tilde{R}_{(i\to j),t-1}^{(k)} + q_{(i\to j),t}\right\},\ (i \to j) \in \{1, ..., M(M-1)\} \tag{2.16}$$

Note that in our setting $\tilde{R}_{(i\to j)}$ denotes the estimated regret of *not* playing action $(i \to j)$; hence we find the action with largest $\tilde{R}$.

By using (2.19), calculate the probability $w_{(i\to j),t}^{(k)}$ assigned to each pair $(i \to j)$.

Given $w_{(i\to j),t}^{(k)}$, solve the following fixed point equation to find $\mathbf{p}_t^{(k)}$.

$$\mathbf{p}_t^{(k)} = \sum_{(i\to j): i\neq j} \mathbf{p}_{(i\to j),t}^{(k)} w_{(i\to j),t}^{(k)}. \tag{2.17}$$

Final probability distribution yields

$$\mathbf{p}_t^{(k)} = (1-\gamma_t)\mathbf{p}_t^{(k)} + \frac{\gamma_t}{M}. \tag{2.18}$$

Using the final $\mathbf{p}_t^{(k)}$, given by (2.18), select an action.

Play and observe the reward.

end for

actions, i.e.

$$\begin{aligned}
&\Pr[i_t = (i \to j)] = w^{(k)}_{(i\to j),t-1} \\
&= \Pr[\tilde{R}^{(k)}_{(i\to j),t-1} + q_{(i\to j),t} \geq \tilde{R}^{(k)}_{(i'\to j'),t-1} + q_{(i'\to j'),t} \ \forall (i \to j) \neq (i' \to j')] \\
&= \int_{-\infty}^{\infty} \Pr[\tilde{R}^{(k)}_{(i\to j),t-1} + q_{(i\to j),t} = m \wedge \tilde{R}^{(k)}_{(i'\to j'),t-1} + q_{(i'\to j'),t} \leq m \ \forall (i \to j) \neq (i' \to j')] dm \\
&= \int_{-\infty}^{\infty} \Pr[\tilde{R}^{(k)}_{(i\to j),t-1} + q_{(i\to j),t} = m] \prod_{(i'\to j')\neq(i\to j)} \Pr[\tilde{R}^{(k)}_{(i'\to j'),t-1} + q_{(i'\to j'),t} \leq m] dm.
\end{aligned} \tag{2.19}$$

Since $\mathbf{q}_t$ is distributed according to a two-sided exponential distribution with width ϵ_n, the terms under integral can be calculated easily (see [HP05], for example).

Now we are in a position to show some properties of NR-BFPLS (Algorithm 2).

Proposition 4. *Let* $\epsilon_t = \epsilon = \frac{\sqrt{\log(n)}}{3\sqrt{Mn}}$ *and* $\gamma_t = \gamma = \min(1, M\epsilon_t)$. *Then Algorithm 2 yields vanishing internal regret with* $R^{(k)}_{\text{Int}} \in O\left(\left(2nM^2 \log(M)\right)^{\frac{1}{2}}\right)$.

Proof. By [KE07], we know that if the bandit follow the leader algorithm is applied to M actions, then $R^{(k)}_{\text{Ext}} \in O\left((nM \log(M))^{\frac{1}{2}}\right)$. Using this, the proof proceeds along similar lines as the proof of Proposition 2 and is therefore omitted. □

Corollary 3. *Assume that the game horizon is known and all players play according to NR-BFPLS. Then the minimum required number of trials to achieve* ϵ*-correlated equilibrium yields* $\max_{k=1,\ldots,K} \epsilon^{-2} O\left(MK\left(M^2 \log(M) + K^2 \log(K)\right)\right)$, *which is proportional to* ϵ^{-2} *and increases polynomially in the number of actions and players.*

Proof. The proof is a result of the bound of Proposition 4 and Remark 7.6 of [CBL06]. □

2.5. Bandit Experimental Regret-Testing Strategy

The basic idea behind exhaustive search algorithms for unknown games is that each player selects its action according to some predefined protocol, and observes its regret. According to such a protocol, each player switches between different (mixed) strategies until an efficient one is captured. Experimental regret-testing belongs to the large family of exhaustive search algorithms, and is comprehensively discussed in [GL07] and [CBL06] for bandit games. In this section, we briefly review this approach, and investigate its performance numerically later in Section 2.6.1.

First the time is divided into periods $j = 1, 2, ...$ of length T so that for each j we have $t \in [(j-1)T + 1, jT]$. At the beginning of period j, any player k randomly selects a mixed

strategy, denoted by $\mathbf{p}_j^{(k)}$. Moreover, some random variable $\mathtt{U}_{k,t}^{(j)} \in \{1, ..., M\}$ is defined as follows. For $t \in [(j-1)T+1, jT]$, and for each $i \in \mathcal{M}$, there are exactly s values of t such that $\mathtt{U}_{k,t}^{(j)} = i$, and $\mathtt{U}_{k,t}^{(j)} = 0$ for the remaining $t = T - sM$ trials. At time t, the action $i_t^{(k)}$ is selected to be [CBL03]

$$i_t^{(k)} : \begin{cases} \text{is distributed as } \mathbf{p}_j^{(k)} & \text{if } \mathtt{U}_{k,t}^{(j)} = 0 \\ \text{equals } i & \text{if } \mathtt{U}_{k,t}^{(j)} = i \end{cases}. \tag{2.20}$$

At the end of period j, player k calculates the experimental regret of playing each action i as [CBL03]

$$\hat{r}_{i,j}^{(k)} = \frac{1}{T - sM} \sum_{t=(j-1)T+1}^{jT} g_t^{(k)}\left(\mathbf{i}_t\right) \mathbf{1}_{\left\{\mathtt{U}_{k,t}^{(j)}=0\right\}} - \frac{1}{s} \sum_{t=(j-1)T+1}^{jT} g_t^{(k)}\left(i, \mathbf{i}_t^{(-k)}\right) \mathbf{1}_{\left\{\mathtt{U}_{k,t}^{(j)}=i\right\}}. \tag{2.21}$$

If the regret is smaller than an acceptable threshold ρ, the player continues to play its current mixed strategy. Otherwise, another mixed strategy is selected. The procedure is summarized in Algorithm 3. It is known that if the parameters of BERTS (e.g. T and ρ) are chosen appropriately, then, in a long run, the played mixed strategy profiles are in an approximate Nash equilibrium for almost all the time. Details can be found in [CBL06], and hence are omitted.

Algorithm 3 Bandit Experimental Regret Testing Strategy (BERTS) [CBL06]

1: Set T (period length), ρ (acceptable regret threshold), $\xi \ll 1$ (exploration parameter), $j = 1$ (period index). Notice that for each period $j = 1, ..., J$, we have $t \in [(j-1)T+1, jT]$.
2: Select a mixed strategy, $\mathbf{p}_j^{(k)}$ according to the uniform distribution, from the probability simplex with M dimensions.
3: For each $i \in \{1, ..., M\}$ select s exploring trials at random. Exploration trials that are dedicated to different actions should not overlap.
4: **for** $t = (j-1)T + y$, where $1 \leq y < T$ **do**
5: **if** t is an exploring trial dedicated to action i **then**
6: play action i and observe the reward.
7: **else**
8: select an action using $\mathbf{p}_j^{(k)}$. Play and observe the reward.
9: **end if**
10: **end for**
11: Calculate the experimental regret of period j, $\hat{r}_{i,j}^{(k)}$, using (2.21);
12: **if** $\max_{i=1,...,M} \hat{r}_{i,j}^{(k)} > \rho$, **then**
13: 1) set $j = j + 1$, 2) go to line 2.
14: **else**
15: • with probability ξ: 1) set $j = j + 1$, 2) go to line 2;
  • with probability $1 - \xi$: 1) let $\mathbf{p}_{j+1}^{(k)} = \mathbf{p}_j^{(k)}$, 2) set $j = j + 1$, 3) go to line 3.
16: **end if**

2.6. Numerical Analysis

Numerical analysis consists of two parts. In Section 2.6.1, we consider a simple network, and clarify the work flow of the developed algorithms. In Section 2.6.2, we consider a larger network, and study the performance of the proposed game model and algorithmic solutions in comparison with some other selection strategies.

2.6.1. Part One

Network model

The network consists of two transmitter-receiver pairs (users). There exist two orthogonal channels, C_1 and C_2, and two power-levels, P_1 and P_2. Hence, the action set of each user yields $\mathcal{M} = \{1 : (C_1, P_1), 2 : (C_1, P_2), 3 : (C_2, P_1), 4 : (C_2, P_2)\}$. The distribution of channel gains changes at each trial. We assume that the variance of the mean values of these distributions is relatively small, which corresponds to *low dynamicity*.[6] Channel matrices are $\mathbf{H}_1 = \begin{bmatrix} [0.50, 0.80] & [0.15, 0.20] \\ [0.01, 0.05] & [0.01, 0.09] \end{bmatrix}$ and $\mathbf{H}_2 = \begin{bmatrix} [0.02, 0.05] & [0.02, 0.06] \\ [0.05, 0.15] & [0.75, 0.95] \end{bmatrix}$, where $\mathbf{H}_x[u, v]$ ($x \in \{1, 2\}$), corresponds to the interval from which $|h_{uv,x,t}|$ is selected at each trial. Moreover, we assume $P_1 = 1$, $P_2 = 5$ and $\alpha = 10^{-3}$. Except for their instantaneous rewards, no other information is revealed to users. This information can be provided by the receiver feedback to transmitter. With these settings, it is easy to see that $((C_1, P_2), (C_2, P_2))$ is the unique pure strategy Nash equilibrium of this game.

Results and Discussion

We investigate the performance of three selection strategies, namely NR-BEWAS, NR-BFPLS and BERTS. The following strategies are also considered as benchmark:

- Centralized joint channel and power-level assignment using global statistical channel knowledge, so that the assignment corresponds to the most efficient pure strategy equilibrium in the sense of maximum aggregate average reward.
- Uniformly random selection.

Figure 2.1 compares the average rewards of NR-BEWAS and NR-BFPLS by those of random selection and centralized assignment. From the figure it can be concluded that despite being provided with no information, both NR-BFPLS and NR-BEWAS exhibit vanishing regret, in the sense that the achieved average reward converges to that of the

[6] Note that this assumption is made in order to simplify the implementation; as established theoretically, all proposed algorithms converge to equilibrium for general time-varying distributions.

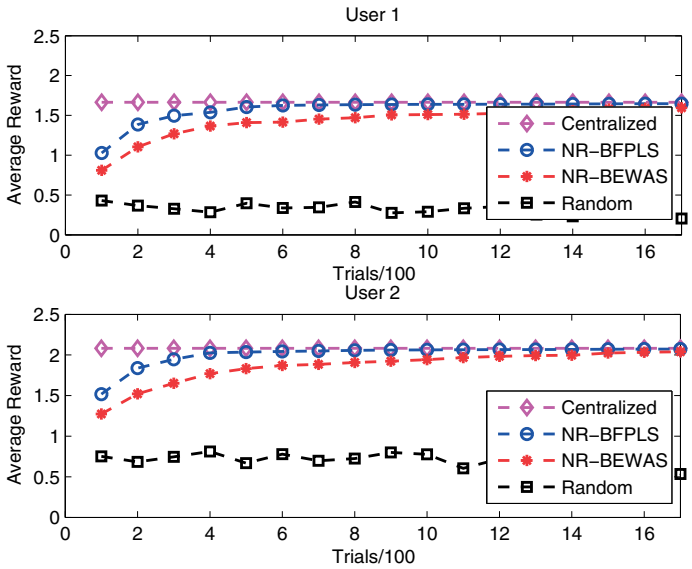


Figure 2.1.: Performance of four selection strategies. Both NR-BEWAS and NR-BFPLS exhibit vanishing regret; that is, their average rewards converge to that of centralized assignment.

centralized method. The convergence speed of NR-BFPLS is higher than that of NR-BEWAS.

Figures 2.2 and 2.3 illustrate the evolution of mixed strategies of the two users when NR-BEWAS is used. Figures 2.4 and 2.5, on the other hand, show the same variable when actions are selected by following NR-BFPLS. For both cases, the first and second users respectively converge to (C_1, P_2) and (C_2, P_2).

The performance of BERTS, however, is not an explicit function of the game duration. As described before, the procedure continues to search the space of mixed strategies until a suitable one, which yields a regret less than the selected threshold, is captured. This strategy is then played for the rest of the game. Theorem 7.8 of [CBL06] specifies the minimum game duration to guarantee the convergence of BERTS, which is relatively long even for small number of users and actions. Nevertheless, similar to other search-based algorithms, there is also the possibility of finding some acceptable strategy at early stages of the game. As a result, for relatively short games, the performance of BERTS is rather unpredictable. The other issue is the effect of regret threshold. On the one hand, larger threshold reduces the search time, since the set of acceptable strategies is large. On the other hand, large regret threshold might lead to performance loss, since there is the possibility that the user gets locked at some suboptimal strategy at early stages, thereby incurring large accumulated regret. It is worth noting that due to its

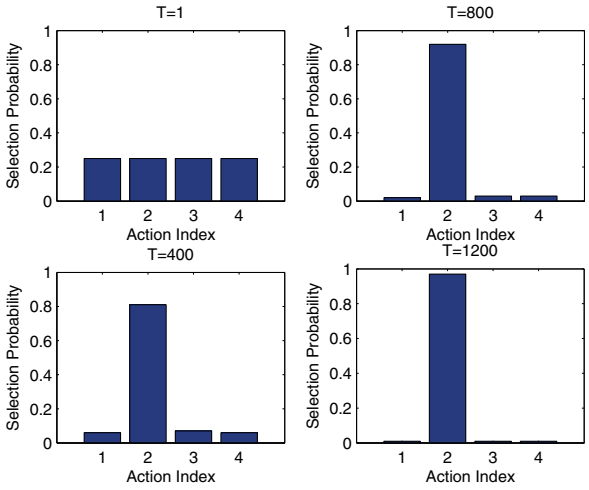


Figure 2.2.: Evolution of the mixed strategy of User 1, applying NR-BEWAS. The mixed strategy of User 1 converges to $(0, 1, 0, 0)$.

simplicity, and despite unpredictable performance, BERTS is an appealing approach in cases where computational effort should be minimized, and playing Nash equilibrium is desired. Figure 2.6 summarizes the results of few exemplary performances of BERTS. The parameters are selected as $T = 80$, $J = 1500$ and $\rho = 0.16$ (see Section 2.5). Simulation is performed for six *independent* rounds. The curve on the left side of Figure 2.6 depicts the period ($1 \leq j \leq 1500$) at which the algorithm finds an acceptable strategy. As expected, the results exhibit no specific pattern. The four subfigures on the right depict the mixed strategies selected by BERTS at rounds 1 and 2, together with average rewards. From this figure, at round 2, acceptable strategies are found earlier than round 1 by both users, leading to a better average performance. It is also worth noting that for User 2, the strategy of round 1 is in essence better than that of round 2; nevertheless, it is found too late. As a result, the average performance of round 2 is superior to that of round 1.

2.6.2. Part Two

In this section we consider a wireless network consisting of 5 users (transmitter-receiver pairs), that compete for access to three orthogonal channels at two power-levels (hence six actions). We compare NR-BFPLS and NR-BEWAS with the following selection approaches.[7]

[7] As mentioned before, observing the joint action profile and/or information exchange is not required for implementing NR-BEWAS, NR-BFPLS and BERTS. Therefore, they cannot be compared with strategies that include mutual observation and/or communications. A good example of such algorithms is the widely-used *best response dynamics*, where the strategy of each player is to play with the best

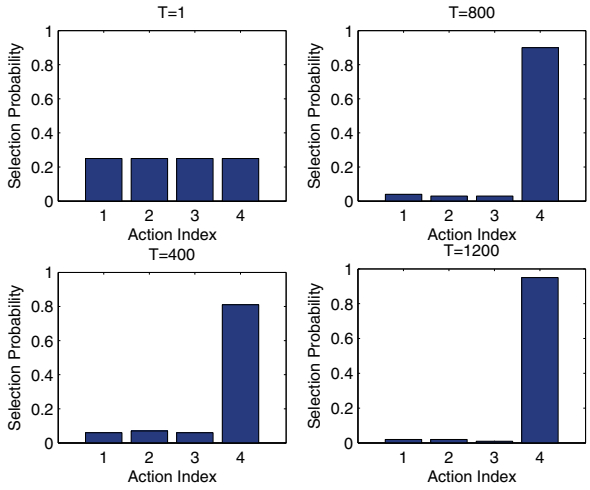


Figure 2.3.: Evolution of the mixed strategy of User 2, applying NR-BEWAS. The mixed strategy of User 2 converges to $(0, 0, 0, 1)$.

- Centralized action assignment as described in Section 2.6.1.
- Centralized no-collision action selection, where no reward is assigned to users that access the same channel. Thus, users are encouraged to avoid collisions. This curve can be considered as an upper-bound for the performance of learning algorithms that select actions based on collision avoidance, such as [KNJ12].
- ϵ-greedy algorithm, where at each trial, with probability ϵ (exploration parameter), an action is selected uniformly at random, while with probability $1 - \epsilon$ the best action so far is played. The average reward of selected action is updated after each play [NH99]. For stationary environments, ϵ is usually time-variant and converges to zero in the limit, while in adversarial cases, ϵ is preferred to remain fixed. Here we choose $\epsilon = 0.1$.
- Greedy approach, where at the beginning of the game, some trials are reserved for exploration, in which actions are selected at random (exploration period). The length of this period is a predefined fraction of the entire game duration. Based on the rewards of exploration period, the best action is selected, and is played for the rest of the game (exploitation period) [CBL06]. This approach is simple to implement; however, to our best knowledge, so far exists no analysis on the optimal length of the exploration period.

response to either the historical [CYC+11] or the predicted [5] joint action profile of opponents. Another example is the strategy suggested in [KNJ12], which is a combination of learning and auction algorithms and includes information exchange.

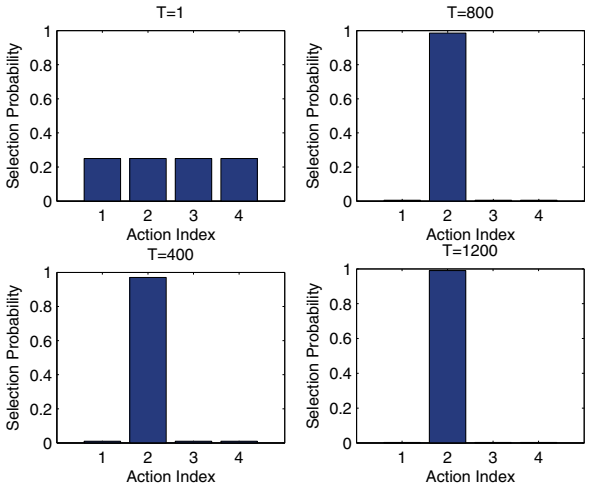


Figure 2.4.: Evolution of the mixed strategy of User 1, applying NR-BFPLS. The mixed strategy of User 1 converges to $(0, 1, 0, 0)$.

- Uniformly random selection.

The numerical results are depicted in Figure 2.7. From this figure, we can conclude the following.

- The performance of interference avoidance strategies is strongly influenced by channel matrices and tend to be poor specifically when the number of channels is less than that of users. The reason is that the sum reward of multiple interfering users with limited transmission powers might be larger than the maximum achievable reward of any single user.

- The rewards achieved by NR-BFPLS and NR-BEWAS converge to that of centralized approach. As expected, NR-BFPLS converges faster than NR-BEWAS and we point out that the convergence speed of both algorithms would be dramatically enhanced if some side-information were available to players, e.g. if users observed the actions of each other, or if information exchange were allowed among players. It is also worth noting that although NR-BFPLS converges faster than NR-BEWAS, calculating integral (2.19) might be computationally involved, especially for large number of actions [HP05].

- In general, ϵ-greedy and greedy approaches can be implemented easily with low computational cost; nevertheless, it can be seen that the greedy approaches are inferior to NR-BEWAS and NR-BFPLS in terms of asymptotic performance. Basically, these approaches are more suitable for static environments.

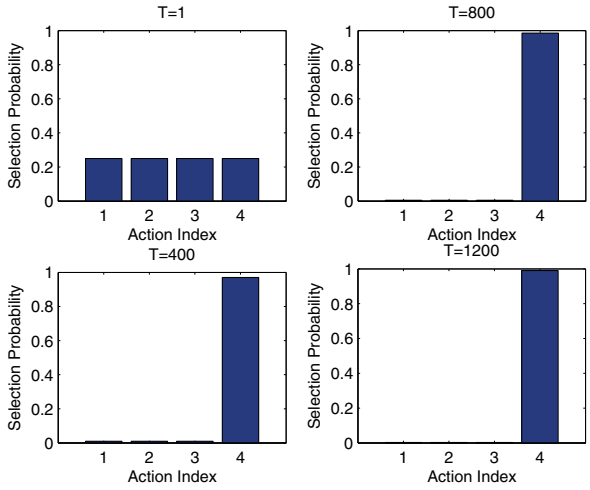


Figure 2.5.: Evolution of the mixed strategy of User 2, applying NR-BFPLS. The mixed strategy of User 2 converges to $(0, 0, 0, 1)$.

2.7. Conclusion and Remarks

In this chapter we studied the resource allocation problem in multi-user infrastructureless wireless networks, by formulating it as an adversarial multi-player multi-armed bandit game. In this model, given no prior- and/or side-information, players attempt to minimize some regret, expressed in terms of the loss of reward, by selecting appropriate actions on a given set of transmit power-levels and orthogonal frequency channels. Based on some recent mathematical results, we developed two joint channel selection and power control strategies, namely NR-BEWAS and NR-BFPLS. The analysis showed that the strategies not only provide vanishing regret for every player, but also guarantee that the empirical joint frequencies of the game converge to the set of correlated equilibria. The convergence rate of both strategies is polynomial in the number of actions and players, with NR-BFPLS converging faster than NR-BEWAS. The computational complexity of NR-BFPLS is however higher than that of NR-BEWAS, in particular for large number of actions. In addition to the theory, the proposed approaches were evaluated through extensive numerical analysis. In accordance to the theory, applying NR-BEWAS or NR-BFPLS in a resource allocation learning game results in convergence to equilibrium, provided that the game horizon is large enough. Moreover, given enough time, the average performance of both strategies is almost as well as that of the centralized strategy given global information. In addition, they exhibit superior performance compared to conventional learning strategies such as greedy and ϵ-greedy. Further, we numerically studied an annealed regret-testing strategy, namely BERTS. The results showed that the performance of BERTS is rather

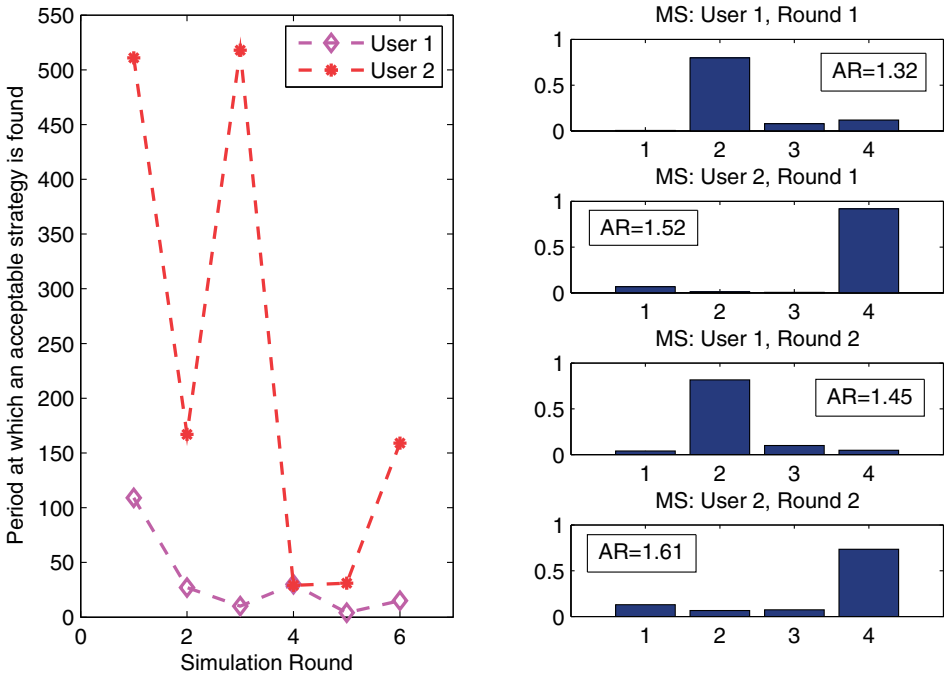


Figure 2.6.: Performance of BERTS. On the left, the two curves depict the period at which a suitable mixed strategy (MS) is found at each of the 6 rounds. On the right, these mixed strategies are shown for both users at rounds 1 and 2, together with average rewards (AR). The horizontal and vertical axes respectively depict the actions' indices and the selection probabilities.

unpredictable. More precisely, the required time for finding an acceptable mixed strategy by which the regret is less than a specific threshold cannot be estimated; nevertheless the algorithm is known to converge to Nash equilibrium, and the required convergence time can be upper-bounded.

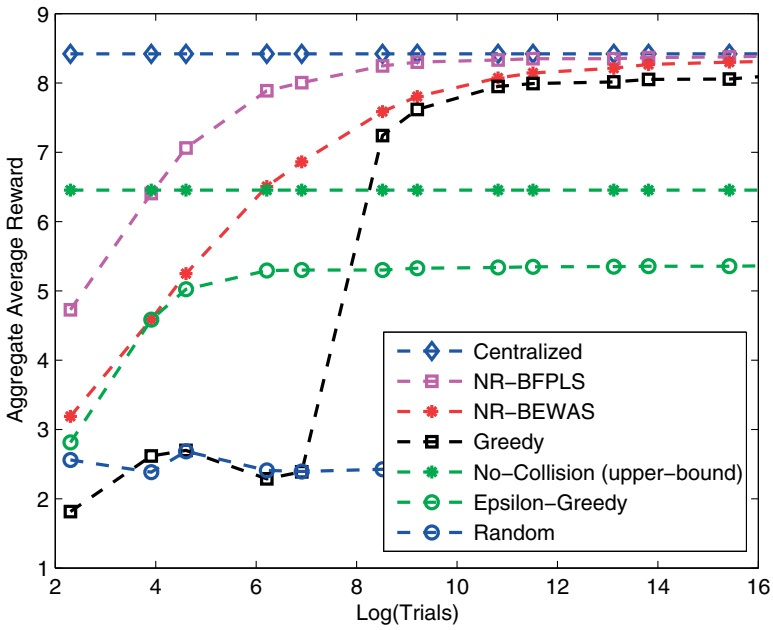


Figure 2.7.: Performance of NR-BFPLS and NR-BEWAS compared to some other selection strategies.

3. Distributed Resource Allocation in Stochastic Networks

In this chapter we study a multi-player adaptive decision making problem, where selfish players learn the optimal joint action profile from successive interactions with the environment, which, unlike the previous chapter, is assumed to be stochastic. This problem appears in many wireless networking scenarios, with a particular instance being the channel selection in a distributed D2D communications system integrated into a cellular network, as considered here. In Section 3.1 we present basic elements of stochastic MP-MAB games and define strong consistency. Section 3.2 briefly reviews calibrated forecasting [MS10], [KF08], [FV97], which is used in the proposed selection strategy. In Section 3.3 we model the channel selection problem as a stochastic bandit game among multiple learning agents that have no prior knowledge, but receive some side-information during the game. In Section 3.4 we propose a selection strategy that consists of two main blocks, namely calibrated forecasting and no-regret bandit learning [YZ02], [CBL06], [SJLS00], [CLRJ13]. Whereas calibrated forecasting is utilized to predict the joint action profile of selfish rational players, no-regret learning builds a reliable estimation of the reward generating functions of arms. We show that the proposed game model and selection strategy can be applied to both noise-limited (orthogonal channel access) and interference-limited (non-orthogonal channel access) transmissions. We prove that the gap between the average utility achieved by our approach and that of the best fixed strategy converges to zero as the game horizon tends to infinity. Moreover, by using our strategy, the empirical joint frequencies of the game converge to the set of correlated equilibria.

Our work generalizes previous studies in a variety of aspects, as listed briefly in the following.

- Some works such as [GLM04] and [FYX13] analyze the single-agent learning problem. In some others such as [KNJ12], although multi-agent problem is formulated, no convergence analysis is performed. We, however, propose an algorithmic solution for a multi-agent learning problem and show that by applying our approach the empirical joint frequencies of the game converge to the set of correlated equilibria. As any Nash

equilibrium belongs to the set of correlated equilibria, our solution is more general in comparison with the approaches that converge to a pure strategy Nash equilibrium, for example those proposed in [XWW+12], [XWS+13] and [XWW+13].

- In our model, we only use the natural assumption that the average reward of any action (channel) to a given player (user) depends on the nature (time-invariant average channel gains) and is a function of the users' joint action profile (interference). Although we mention the transmission rate as an example of such reward process, we do not restrict the reward functions to any specific parametric form in our analysis. As a result, the proposed selection algorithm is applicable to a wide range of problems. In contrast, in most previous works, including [XWW+12], [XWS+13] and [XWW+13], some *specific* utility function is defined, based on which the game is characterized as an exact potential game. The convergence analysis therefore holds only for the defined reward function.

- In our problem setting, both noise-limited and interference-limited transmission models are studied, and we do not impose any restriction on the interference pattern. This is in contrast with [KNJ12], [XWS+13] and [XWW+13], where the interference is either completely neglected or is limited to the neighboring users. This aspect of our model is important since in general channel allocation based on interference avoidance is suboptimal.

- In our work, every action pays different rewards to different users, i.e. the reward process of every action is user-specific. In a wireless network, this means that the variations in both channel availability and quality are taken into account. This stands in contrast to [XWW+12], where the reward of each specific channel is assumed to be equal for all users (common utility game), and only the channel availability is stochastic.

- Unlike [KNJ12] and [KNJ14], our algorithm does not require information exchange.

3.1. Stochastic Multi-Player Multi-Armed Bandit Games

As described in Chapter 2, multi-player multi-armed bandit game (MP-MAB) is a class of sequential decision making problems with limited information. At each trial t, any player $k \in \mathcal{K} = \{1, ..., K\}$ selects some action $i \in \mathcal{M} = \{1, ..., M\}$. The action of player k at time t is denoted by $i_t^{(k)}$. Upon being pulled at time t, any action $i \in \mathcal{M}$ generates some random reward $g_t^{(k)}\left(i, \mathbf{i}_t^{(-k)}\right) \coloneqq g_{i,t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right) \in \mathbb{R}^+$, which depends not only on the

action of player k itself, but also on the joint action profile of other players at time t, $\mathbf{i}_t^{(-k)}$. Let $i_t^{\circ(k)} \coloneqq \arg\max_{i\in\{1,\dots,M\}} g_{i,t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$ be the optimal action that yields a reward equal to $g_t^{\circ(k)}\left(\mathbf{i}_t^{(-k)}\right) \coloneqq g_{i_t^{\circ(k)},t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$. Ideally, at every time t, any player k selects the optimal action in the sense of maximum reward, thereby maximizing its accumulated utility. Therefore, its primary goal can be formulated as

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(g_{i_t^{(k)},t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)-g_t^{\circ(k)}\left(\mathbf{i}_t^{(-k)}\right)\right)=0, \tag{3.1}$$

with n being the game horizon. However, since players have no prior-information, solving (3.1) is impossible in general. Consequently, we argue in favor of another strategy where each player pursues a less ambitious goal: Minimize the asymptotic average regret, where regret is defined as the difference between the reward that could have been achieved by selecting the optimal channel in the sense of maximum average reward given the side-information, and that of the actual selected channel. We formulate this problem as follows.

For any action $i \in \{1,...,M\}$, let $f_i^{(k)}$ be the time-invariant mean reward process. Furthermore, assume that for each i and $\mathbf{i}^{(-k)}$, the achieved reward can be modeled as $g_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right)=f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)+\mathtt{C}_i^{(k)}\left(\mathbf{i}^{(-k)}\right)$, where $\mathtt{C}_i^{(k)}$ denotes a random error with zero mean and finite variance, independent from all other random variables. At trial t, let $i_t^{*(k)} \coloneqq \arg\max_{i\in\{1,\dots,M\}} f_i^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$ that results in $g_t^{*(k)}\left(\mathbf{i}_t^{(-k)}\right) \coloneqq g_{i_t^{*(k)},t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$. Then the goal of player k is to achieve

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(g_{i_t^{(k)},t}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)-g_t^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)\right)=0. \tag{3.2}$$

Now assume that $i_t^{*(k)}$ yields $f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right) \coloneqq \max_{i\in\{1,\dots,M\}} f_i^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$. In [YZ02], it is shown that (3.2) is equivalent to

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(f_{i_t^{(k)}}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)-f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)\right)=0, \tag{3.3}$$

provided that $g_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right)$ is bounded above and away from zero for $i \in \{1,..,M\}$. Therefore each player decides which action to take at successive rounds so that asymptotically the accumulated reward achieved by the played arms is not much less than that of the optimal arm. Obviously, this problem is an instance of the exploitation-exploration dilemma, in which a balance should be found between exploiting the arms that have exhibited good performance in the past (control), and exploring arms that might perform well in the future (learning). In this chapter we assume that $f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)$ and $f^{*(k)}\left(\mathbf{i}^{(-k)}\right)$ obey the

following assumption [YZ02].

Assumption (A2). $\forall\ k \in \mathcal{K},\ i \in \mathcal{M}$ *and* $\mathbf{i}^{(-k)} \in \bigotimes_{k=1}^{K-1} \{1, ..., M\}$,

a) $f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) \in [0, A]$ *for some* $A > 0$,

b) $B = \sup_i \sup_{\mathbf{i}^{(-k)}} \left(f^{*(k)}\left(\mathbf{i}^{(-k)}\right) - f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)\right) < \infty$,

c) $\mathrm{E}\left\{ f^{*(k)}\left(\mathbf{i}_1^{(-k)}\right)\right\} > 0$.

The last part of the assumption implies that the expected optimal reward is positive at least for the first round of the game, where the expectation is with respect to the mixed strategy. This assumption is later used to avoid any division by zero. We also assume that the achieved rewards of any particular player are revealed to that player only, whereas players' actions are observed by their opponents.

3.1.1. Strong Consistency and Bandit Problems

As described before, at the n-th play, the set of personal achieved rewards and observed actions up to time n are available to each player. The accumulated mean reward of player k up to time n is $\sum_{t=1}^{n} f_{i_t^{(k)}}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)$, while $\sum_{t=1}^{n} f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)$ is the optimal total reward of player k, which could have been achieved by pulling arm $i_t^{*(k)}$ for all trials up to n. Since the k-th player would attain the best performance if it selected the optimal arm at every trial, it is reasonable to evaluate any selection strategy κ used by player k based on the following performance metric [YZ02]

$$S_{\kappa,n} = \frac{\sum_{t=1}^{n} f_{i_t^{(k)}}^{(k)}\left(\mathbf{i}_t^{(-k)}\right)}{\sum_{t=1}^{n} f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)} \leq 1, \tag{3.4}$$

where $\sum_{t=1}^{n} f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right) > 0$ by Assumption (A2). Clearly, the closer $S_{\kappa,n}$ to 1, the better the selection strategy. Asymptotically as n tends to infinity, the most desired property is *strong consistency* [YZ02], defined below.

Definition 4 (Strong Consistency)**.** *A selection strategy* κ *is strongly consistent if* $S_{\kappa,n} \to 1$ *as* $n \to \infty$.

Remark 2 ([YZ02])**.** *If* $\frac{1}{n}\sum_{t=1}^{n} f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)$ *is bounded above and away from* 0 *with probability* 1*, then* $S_{\kappa,n} \to 1$ *almost surely implies*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(f_{i_t^{(k)}}^{(k)}\left(\mathbf{i}_t^{(-k)}\right) - f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)\right) = 0. \tag{3.5}$$

Referring to "$f^{(k)}_{i_t^{(k)}}\left(\mathbf{i}_t^{(-k)}\right) - f^{*(k)}\left(\mathbf{i}_t^{(-k)}\right)$" as the "regret" at time t, this implies that strong consistency is equivalent to achieving per-round vanishing (zero-average) regret.

From the game-theoretical point of view, for each player $k \in \{1, ..., K\}$, an MP-MAB is a game with two agents: the first agent is player k itself, and the second agent is the set of all other $K-1$ players, whose joint action profile affects the rewards of player k. Since the reward of any player k depends on the decisions of other players, the key idea of the proposed approach is to enable each user to *forecast* the future actions of its opponents based on public knowledge and to proceed by best responding to the predicted joint action profile using some bandit strategy. In Section 3.2, we discuss how reliable forecasting can be performed. Later in Section 3.4 we describe how players should proceed using this side-information.

3.2. Calibration and Construction of a Calibrated Forecaster

In this section, we briefly describe basic elements of calibrated forecasting. We then explain how calibrated forecasting is related to the concept of equilibria in games. Later in Section 3.4 we use these materials in order to develop a convergent channel selection strategy.

3.2.1. Calibration

Following [MS10], consider a random experiment with a finite set of outcomes $\mathcal{D}$ of cardinality D, and let δ_{d_t} stand for the Dirac probability distribution on some outcome d at time t. The set of probability distributions over $\mathcal{D}$ is denoted by $\mathcal{P} \subseteq \mathbb{R}^D$. Equip $\mathcal{P}$ with some norm $\|\cdot\|$. At time t, the forecaster outputs a probability distribution $\mathbf{p}_t$ over the set of outcomes.

Definition 5 (Calibrated Forecaster [MS10])**.** *A forecaster is said to be calibrated if* $\forall\ \epsilon > 0$ *and* $\forall\ \mathbf{p} \in \mathcal{P}$*, almost surely,*

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\{\|\mathbf{p}_t - \mathbf{p}\| \leq \epsilon\}} \left(\mathbf{p}_t - \delta_{d_t}\right) \right\| = 0. \tag{3.6}$$

A relaxed notion of calibration is ϵ-calibration. Given $\epsilon > 0$, an ϵ-calibrated forecaster considers some finite covering of $\mathcal{P}$ by N_ϵ balls of radius ϵ. Denoting the centers of these balls by $\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N_\epsilon)}$, the forecaster selects only forecasts $\mathbf{p}_t \in \left\{\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N_\epsilon)}\right\}$. Using this, ϵ-calibration is defined as follows [MS10].

Definition 6 (ϵ-Calibrated Forecaster)**.** *Define Q_t to be the index in $\{1, ..., N_\epsilon\}$ such that $\mathbf{p}_t = \mathbf{p}^{(Q_t)}$. A forecaster is said to be ϵ-calibrated if almost surely,*

$$\limsup_{n\to\infty} \sum_{q=1}^{N_\epsilon} \left\| \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\{Q_t=q\}} \left(\mathbf{p}^{(q)} - \delta_{d_t} \right) \right\| \leq \epsilon. \tag{3.7}$$

Note that none of the two definitions makes any assumption on the nature of the random experiment whose outcome is being predicted. The following result can be found in [FV97], [KF08], and [CBL06].

Theorem 2. *Consider a game with K players provided with M actions. Let $\mathcal{C}$ stand for the set of correlated equilibria, and define the joint empirical frequencies of play as (2.5). For a player k, let $\mathcal{D} = \mathcal{I}^-$ be the set of joint action profiles of opponents. Assume that each player plays by best responding to a calibrated forecast of the opponents joint action profile in a sequence of plays; that is, for each player k we have*

$$i_t^{(k)} = \underset{i\in\{1,...,M\}}{\arg\max} \sum_{d=1}^{D} p_{d,t}^{(k)} f_i^{(k)}(d), \tag{3.8}$$

where $\mathbf{p}_t^{(k)} = \left(p_{1,t}^{(k)}, ..., p_{D,t}^{(k)} \right)$ stands for the output of the forecaster, which is a probability distribution over $D = M^{K-1}$ possible joint action profiles of the opponents. Accordingly, each d represents a realization of the joint action profile of opponents of player k, i.e. $\mathbf{i}^{(-k)}$. Then the distance $\inf_{\pi\in\mathcal{C}} \sum_{\mathbf{i}} |\hat{\pi}_n(\mathbf{i}) - \pi(\mathbf{i})|$ between the empirical joint distribution of plays and the set of correlated equilibria converges to 0 almost surely as $n \to \infty$.

3.2.2. Construction of a Calibrated Forecaster

For constructing a calibrated forecaster, an approach is to use *doubling-trick* [MS10]. In the first step, an ϵ-calibrated forecaster is constructed for some $\epsilon > 0$. Then, the time is divided into periods of increasing length, and the procedure of ϵ-calibration is repeated as a subroutine over periods, where ϵ decreases gradually to zero (that is, N_ϵ-grid becomes finer at each period). In Algorithm 4, we review this procedure. The proof of calibration follows from the *Blackwell's approachability* theorem. See [MS10] for details and the proof of calibration.

Theorem 3 ([MS10])**.** *The forecasting procedure (Algorithm 4) is calibrated. That is, it satisfies (3.6).*

Algorithm 4 A Calibrated Forecaster [MS10]

Define $T_j = 2^j$ where T_j is the length of period $j = 1, 2, ...$

For any period j, define a two-player game, where the first player is the forecaster with the action set $\mathcal{Z} = \left\{1, ..., N_{\epsilon_j}\right\}$ and the second player is the nature with action set $\mathcal{D}$. With respect to our model, the first player is some agent k, while the second player is the set of all other $K-1$ agents; hence $\mathcal{D} = \mathcal{I}^-$ and $D = M^{K-1}$. Accordingly, any outcome d is the realization of a joint action profile of $K-1$ players, that is $\mathbf{i}^{(-k)}$.

For each period j, let $\epsilon_j = 2^{-j/(D+1)}$.

Define the vector-valued regret of the first player as $\mathbf{u}\{q, d\} = \left(\mathbf{0}, ..., \mathbf{0}, \mathbf{p}^{(q)} - \delta_d, \mathbf{0}, ..., \mathbf{0}\right)$ for each $q \in \left\{1, ..., N_{\epsilon_j}\right\}$, $d \in \mathcal{D}$.

Define the target set $\mathcal{F}$ as follows:

- Write (DN_{ϵ_j})-dimensional vectors of $\mathbb{R}^{DN_{\epsilon_j}}$ as N_{ϵ_j}-dimensional vectors with components in $\mathbb{R}^D$, i.e. $\mathbf{x} = \left(\mathbf{x}_1, ..., \mathbf{x}_{N_{\epsilon_j}}\right)$, where $\mathbf{x}_l \in \mathbb{R}^D$ for all $l \in \left\{1, ..., N_{\epsilon_j}\right\}$.
- $\mathcal{F}$ is a subset of the ϵ_j-ball around $(\mathbf{0}, ..., \mathbf{0})$ for the calibration norm $\|\cdot\|$, which is a closed convex set.

Define the sequence of the vector-valued regrets up to time T $(1 \leq T \leq T_j)$ as

$$\mathbf{u}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}\{Q_t, d_t\} = \frac{1}{T}\left(\sum_{t=1}^{T} \mathbf{1}_{\{Q_t=1\}}\left(\mathbf{p}^{(1)} - \delta_t\right), ..., \sum_{t=1}^{T} \mathbf{1}_{\{Q_t=N_{\epsilon_j}\}}\left(\mathbf{p}^{(N_{\epsilon_j})} - \delta_t\right)\right). \tag{3.9}$$

Now, (3.7) (condition of ϵ_j-calibration) can be restated as the convergence of $\mathbf{u}_{T_j}$ to the set $\mathcal{F}$ almost surely. In the following, $\mathbf{u}^{(j)} \coloneqq \mathbf{u}_{T_j}$ denotes the final regret of period j.

repeat

  for $t = 1 \to T_j$ **do**

    if $(j = 1 \wedge t = 1)$ **then**

      Select an action Q_t from $\mathcal{Z}$ according to the uniform distribution over the action set, i.e. let $\psi_1 = \left(\frac{1}{N_{\epsilon_j}}, ..., \frac{1}{N_{\epsilon_j}}\right)$. Note that ψ_t is the mixed strategy at time t, while $\psi^{(j)} \coloneqq \psi_{T_j}$ denotes the final mixed strategy of period j.

    else if $(j > 1 \wedge t = 1)$ **then**

      Select an action Q_t from $\mathcal{Z}$, according to a probability distribution in a small neighborhood of $\psi^{(j-1)}$ (localization of search).

    else

      Select an action Q_t from $\mathcal{Z}$ at random according to a distribution $\psi_t = \left(\psi_{t,1}, ..., \psi_{t,N_{\epsilon_j}}\right)$ on $\left\{1, ..., N_{\epsilon_j}\right\}$ such that $\forall d \in \mathcal{D}$,

$$\left(\mathbf{u}_{t-1} - \Pi_{\mathcal{F}}\left(\mathbf{u}_{t-1}\right)\right) \cdot \left(\mathbf{u}\{\psi_t, d\} - \Pi_{\mathcal{F}}\left(\mathbf{u}_{t-1}\right)\right) \leq 0, \tag{3.10}$$

      where $\Pi_{\mathcal{F}}$ denotes the projection in l_2-norm onto $\mathcal{F}$ and $\cdot$ denotes the inner product in $\mathbb{R}^{DN_{\epsilon_j}}$. See [MS10] for details.

    end if

  end for

  Calculate the final regret of the current period, $\mathbf{u}^{(j)} = \mathbf{u}_{T_j}$. Also, let $\psi^{(j)} = \psi_{T_j}$.

  if $\mathbf{u}^{(j)} > \epsilon_j$, **then**

  - Let $j = 1$ and $t = 1$.

  else

  - Let $j = j + 1$ and $t = 1$.

  end if

until convergence (j is large enough so that $\epsilon_j \approx 0$)

3.3. Bandit-Theoretical Model of D2D Channel Selection Problem

3.3.1. System Model

We study a distributed D2D communications system as an underlay to a single-cell wireless network with a set $\mathcal{M}$ of M licensed orthogonal channels. The D2D system consists of a set $\mathcal{K}$ of K suitably-selected single-antenna transmitter-receiver pairs. Each pair is referred to as a D2D user, and is denoted either by just k or by the pair (k, k'). In order to eliminate the adverse effects of D2D transmission on cellular users, any channel is available to D2D users only if it is not occupied by some cellular user.[1] As the D2D data is not forwarded via the BS, conventional pilot signals cannot be used, and therefore the BS is not in possession of D2D channel knowledge. D2D users have neither channel (quality and availability) nor network knowledge. We assume that the BS observes the transmission channels of all D2D and cellular users. D2D users do not exchange information. However, there exists a channel through which the BS broadcasts some signals referred to as side-information. Based on the physical characteristics of the radio propagation medium, the broadcast signal is assumed to be heard by all D2D users. Note that the broadcast channel is occupied only until convergence, and therefore the overhead remains low. Throughout the chapter, $|h_{uv,x}|^2 \in \mathbb{R}^+$ is used to denote the *average* gain of some channel x between u and v, including path loss and fast fading effects. Note that unlike Chapter 2, the distributions of all channels are assumed to have time-invariant expected values. The variance of zero-mean AWGN is denoted by N_0. We assume that all users transmit with fixed average power P. This is also in contrast to Chapter 2, where users choose a transmission power from a set of power-levels.

3.3.2. Bandit-Theoretical Problem Formulation

At the beginning of each transmission round, every D2D user selects a channel to sense and informs the BS about its choice. For simplicity, we assume that sensing is perfect. If the selected channel is free, then the transmission begins. The primarily duration of transmission is denoted by T_r. At the end of transmission, the BS broadcasts the D2D indices along with the indices of their selected channels. Consequently, every D2D user knows which channels are selected by other users. This side-information is then used by each D2D user to learn the environment as well as the behavior of other users.

Since all D2D users are allowed to select among M (probably available) channels, collision might occur. We consider the following multiple access protocols.

[1] As we see later, this assumption can be simply relaxed.

- Orthogonal multiple access (noise-limited case): In case of collision, carrier sense multiple access (CSMA) is implemented [ZTSC07]. Since interference is avoided, transmissions are corrupted only by AWGN. The mean reward of some D2D user k transmitting at some channel $i \in \mathcal{M}$ is defined as

$$f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) = \frac{\tau_i^{(k)}\left(\mathbf{i}^{(-k)}\right)}{T_r} \log_2\left(1 + \frac{P\left|h_{kk',i}\right|^2}{N_0}\right)\theta_i, \tag{3.11}$$

  where $\mathbf{i}^{(-k)}$ denotes the set of channels selected by all D2D users except for k. Moreover, $\tau_i^{(k)}$ is the expected value of the random useful transmission time of user k through channel i as a function of $\mathbf{i}^{(-k)}$, which depends on the exact applied CSMA scheme. θ_i is the expected value of a Bernoulli random variable that indicates whether channel i is occupied by some cellular user or not.

- Non-orthogonal multiple access (interference-limited case): In case of collision, colliding users transmit simultaneously, giving rise to the interference. The mean reward is given by

$$f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) = \log_2\left(1 + \frac{P\left|h_{kk',i}\right|^2}{P\sum_{q\in\mathcal{Q}^{(k)}}\left|h_{qk',i}\right|^2 + N_0}\right)\theta_i, \tag{3.12}$$

  where $\mathcal{Q}^{(k)}$ denotes the set of D2D users that share channel i with user k.

Finally, we emphasize that the reward functions can be freely defined as long as the mean reward remains some time-invariant function of average channel gains and users' joint action profile. For instance, consider the case where D2D user k is allowed to transmit simultaneously in the same frequency band with cellular user c. In this case, the mean reward function of user k should be modified so that it incurs some cost due to disturbing the cellular user. For example the mean reward function can be defined as

$$f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) = \log_2\left(1 + \frac{P\left|h_{kk',i}\right|^2}{P\sum_{q\in\mathcal{Q}^{(k)}}\left|h_{qk',i}\right|^2 + P_c\left|h_{ck',i}\right|^2 + N_0}\right) - \alpha\left|h_{kb,i}\right|^2 P, \tag{3.13}$$

where α is a cost factor determined by the BS, denoted by b, and P_c is the transmission power of cellular user c. Another example is a network where the cellular users require some quality of service (QoS) guarantee. In this case the reward can be defined to be some small constant when the QoS is violated, and the D2D transmission rate otherwise. As we see later, as a result of the learning process, D2D users abandon the joint action profiles that result in low reward, in order to avoid large regret. Hence the QoS of cellular users will be satisfied.

Finally it should be mentioned that the proposed model can be easily generalized to include the power allocation, as briefly described in the following. Instead of transmitting with fixed average power P, assume that each D2D user selects one of the M'' quantized power-levels in order to transmit in one of the M' orthogonal channels. Thus, every D2D user is provided with $M' \times M''$ actions, where each action i is understood as a pair (i', i''), i.e. it consists of one channel and one power-level. Correspondingly the mean reward function is defined as

$$f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) = \log_2\left(\frac{i''|h_{kk',i'}|^2}{\sum_{q\in\mathcal{Q}^{(k)}} i''^{(q)}|h_{qk',i'}|^2 + N_0}\right)\theta_i - \alpha i'', \tag{3.14}$$

where $i''^{(q)}$ is the transmission power of interference q. The proposed approach can be also applied to this problem, provided that every user has the incentive to announce its transmission power-level to the network or to the BS, at each transmission round until convergence.

By comparing our system model (Section 3.3.1) with MP-MAB (Section 3.1), we observe that the distributed channel selection problem is in great harmony with MP-MAB settings. Therefore, we model this problem as an MP-MAB game in which every D2D user is modeled as a player, while frequency channels are regarded as arms, implying that choosing a channel corresponds to pulling an arm. Clearly, the reward achieved by any player[2] depends on the selected channel of the user itself and also on those of other users, according to (3.11) or (3.12). According to this model the goal of each D2D user k is to satisfy (3.3). By Remark 2, (3.3) is equivalent to strong consistency. Therefore in the following section we develop a strongly consistent channel selection strategy.

3.4. Calibrated Bandit Strategy

As it is clear from (3.11) and (3.12), the performance of each D2D user depends on two factors: 1) channel quality and availability, and 2) the joint channel selection profile of all D2D users. Given no initial information, the impacts of these factors on the average reward are learned over time, and the average reward function is estimated by means of a regression process. Here we make the following assumption.

Assumption (A3). *The regression process is strongly consistent in L_∞ norm for each $f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)$; that is, $\|\hat{f}_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right) - f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)\|_\infty \to 0$, for all $i \in \{1, ..., M\}$, $k \in \{1, ..., K\}$*

[2]Note that we here mentioned only some exemplary reward functions, which can be substituted by any other time-invariant utility or cost function. As we see later, the proposed approach relies on non-parametric regression and hence offers high flexibility with respect to the reward function.

and $\mathbf{i}^{(-k)} \in \bigotimes_{k=1}^{K-1} \{1, ..., M\}$*, almost surely as* $t \to \infty$*, where* $\hat{f}_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right)$ *denotes the regression estimate of* $f_i^{(k)}\left(\mathbf{i}^{(-k)}\right)$ *at the* t*-th trial.*

In Section 3.3.2, we modeled the channel selection as a bandit game. In what follows, we describe our proposed strategy to solve this game and investigate its convergence characteristics.

3.4.1. Selection Strategy

The game horizon is first divided into periods $j = 1, 2, ...$ of increasing length T'_j. We also define a sequence Z_j for $j = 1, 2, ...$, so that T'_j and Z_j satisfy the following assumption.

Assumption (A4). *We assume that* $\left\{T'_j\right\}_{j=1,2,...}$ *and* $\{Z_j\}_{j=1,2,...}$ *are selected so that*

a) $\left\{\left\lceil T'_j Z_j \right\rceil\right\}_{j=1,2,...}$ *is an increasing sequence of integers,*

b) $\lim_{J\to\infty} \sum_{j=1}^{J} \left\lceil T'_j Z_j \right\rceil \to \infty$*,*

c) $\lim_{J\to\infty} \frac{\sum_{j=1}^{J} \lceil T'_j Z_j \rceil}{\sum_{j=1}^{J} T'_j} = 0.$

At each period j, $\left\lceil T'_j Z_j \right\rceil$ randomly-selected trials are reserved for exploration, and the rest of the trials are used for exploitation in the following manner.

- Exploitation: In an exploitation trial, every player k first receives a probability distribution over all possible joint action profiles of other $K-1$ players, which is the output of its forecasting procedure. Based on this information, and by using the estimated mean reward functions, it selects the action with the highest estimated mean reward; that is, it acts with the best response to the predicted joint action profile of its opponents.

- Exploration: In an exploration trial, with probability $\gamma \ll 1$ the best response is played, whereas with probability $1-\gamma$, an action is selected uniformly at random.

In all trials, after selection, the player's estimation of the mean reward process of the selected action is improved based on the achieved reward. Moreover, actions of other players are observed (here by hearing the broadcast message). This observation is used by the forecaster, as described in Algorithm 4. The procedure is summarized in Algorithm 5.

Note that for larger period indices (large j), the fraction of time reserved for exploration is smaller, as depicted in Figure 3.1. Therefore, the strategy belongs to the class of algorithms that follow the *greedy in the limit with infinite exploration* (GLIE) principal [SJLS00]. Intuitively, this concept states that in a (near-) static environment, the

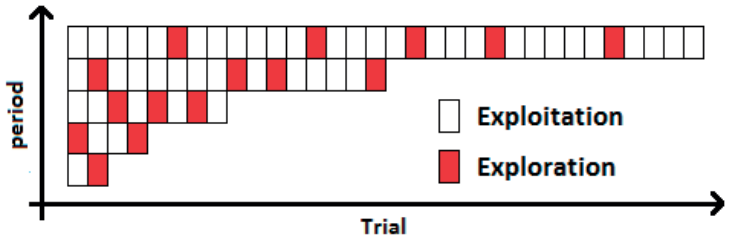


Figure 3.1.: Exemplary exploration-exploitation trade-off for few successive periods.

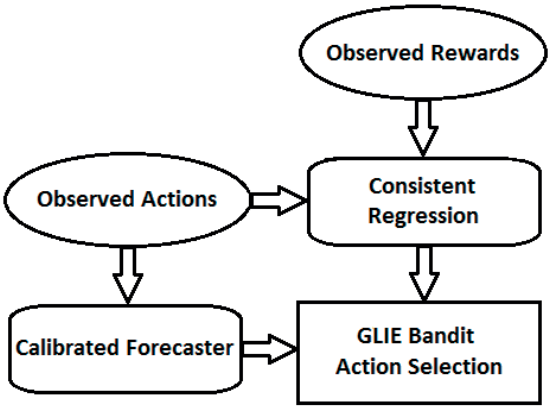


Figure 3.2.: Flowchart diagram of calibrated bandit selection strategy (Algorithm 5)

estimation of arms' reward functions becomes more reliable with time, and therefore less exploration is required. Also note that the regression and the forecasting perform two different tasks; while the former deals with estimating the reward functions, the latter predicts the joint action profile of opponents, as visualized by Figure 3.2.

3.4.2. Strong Consistency and Convergence

The following results ensure the consistency and declare the convergence characteristics of the proposed selection strategy.

Lemma 1. $\left\{T'_j\right\}_{j=1,2,\ldots} = 2^j$ *and* $\{Z_j\}_{j=1,2,\ldots} = \frac{j}{2^j}$ *satisfy Assumption (A4).*

Proof. The lemma can be easily verified by direct calculation using theorems concerning limits of infinite sequences. □

Lemma 2. *Consider a selection strategy* κ *so that each player* k *plays with actions based on* δ_{d_t}*, i.e.* $i_t^{(k)} = \kappa\left(\delta_{d_t}\right)$*, where* δ_{d_t} *is the Dirac probability distribution on the true joint action profile of opponents at time* t*. Let* κ' *be some strategy identical to* κ*, except that*

Algorithm 5 Calibrated Bandit Selection Strategy (CBS)

Define an increasing sequence of integers, $T'_j = 2^j$ for $j = 1, 2, ...$ Each member T'_j of this sequence is the length of period j, i.e. the number of trials included in it.
Define a decreasing sequence of numbers, $Z_j = \frac{j}{2^j}$ for $j = 1, 2, ...$
Set the period $j = 1$ and select the exploration parameter $\gamma \ll 1$.
repeat
  Select $\lceil T'_j Z_j \rceil$ exploration trials belonging to $[1 + \sum_j T'_{j-1}, \sum_j T'_j]$ uniformly at random.
  for $t = s + \sum_j T'_{j-1}$, $1 \leq s < T'_j$, **do**
    if t is an exploring trial, **then**
      with probability $1 - \gamma$, select an arm equally at random;
      with probability γ,
      1. receive the output of the forecaster (Algorithm 4),
      2. using this information, select the arm with the highest estimated mean reward.
    else
      Receive the input from the forecaster.
      Using this information, select the arm with the highest estimated mean reward.
    end if
    Play the selected arm and observe the reward.
    Observe the actions of other players and inform the forecaster (forecaster's input).
    Improve the estimation of the mean reward function of the played arm.
  end for
  $j = j + 1$.
until convergence (j is sufficiently large)

$\mathbf{p}_t$ is used in the place of δ_{d_t}, i.e. $i_t^{(k)} = \kappa(\mathbf{p}_t)$, where $\mathbf{p}_t$ is a probability distribution over all possible joint action profiles of opponents, produced by a calibrated forecaster. Then, $\lim_{n\to\infty} S_{\kappa,n} = 1$ implies that $\lim_{n\to\infty} S_{\kappa',n} = 1$, where $S_{\kappa,n}$ and $S_{\kappa',n}$ are defined by (3.4).

Proof. See Section 1 of Appendix C. □

Lemma 2 simply states that if a strategy is strongly consistent given true joint action profiles, then its consistency is preserved by using the calibrated forecast of the joint action profiles.

Lemma 3. *Asymptotically, the bandit selection strategy (CBS) samples each action $i \in \{1, ..., M\}$ and also each joint action profile $\mathbf{i} = \left(i^{(1)}, ..., i^{(K)}\right) \in \bigotimes_{k=1}^{K} \{1, ..., M\}$ infinitely often.*

Proof. See Section 2 of Appendix C. □

Theorem 4. *Under Assumptions (A2), (A3) and (A4), the proposed selection strategy (CBS), is strongly consistent.*

Proof. See Section 3 of Appendix C. □

Theorem 5. *Consider a K-player MAB game where each player is provided with M actions. Let $\mathcal{C}$ denote the set of correlated equilibria and $\mathbf{i} = \left(i^{(1)}, ..., i^{(K)}\right) \in \bigotimes_{k=1}^{K} \{1, .., M\}$. Define the empirical joint frequencies of play as (2.5). If all players play according to the CBS, then the distance $\inf_{\pi \in \mathcal{C}} \sum_{\mathbf{i}} |\hat{\pi}_n(\mathbf{i}) - \pi(\mathbf{i})|$ between the empirical joint distribution of plays and the set of correlated equilibria converges to 0 almost surely as $n \to \infty$.*

Proof. See Section 4 of Appendix C. □

Remark 3. *In Section 3.3.2, we mentioned that every player is interested in optimizing its performance in the sense of regret minimization, and no player intends to ruin the performance of others. Therefore players are rational and not malicious. By Remark 2 and Theorem 4, CBS yields vanishing regret; Thus the assumption that all players use this strategy is justified.*

3.4.3. Some Notes on Convergence Rate

As it is clear from Algorithm 5 (see also Figure 3.2), for final convergence, the forecasting and regression procedures must converge to the true joint action profile and the true reward functions, respectively. In what follows, we discuss the impact of some variables, including number of actions (M) and users (K), as well as the exploration parameter (γ), on the convergence rate of these procedures.

Theorem 6 ([MS10])**.** *For the calibrated forecaster given in Algorithm 4 we have*

$$\limsup_{n\to\infty} \frac{n^{\frac{1}{D+1}}}{\sqrt{\log(n)}} \sup_{\mathcal{B}\in\mathfrak{B}} \left\| \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}_{\mathbf{p}_t \in \mathcal{B}} \left(\mathbf{p}_t - \delta_{d_t}\right) \right\|_1 \leq \Gamma_D, \tag{3.15}$$

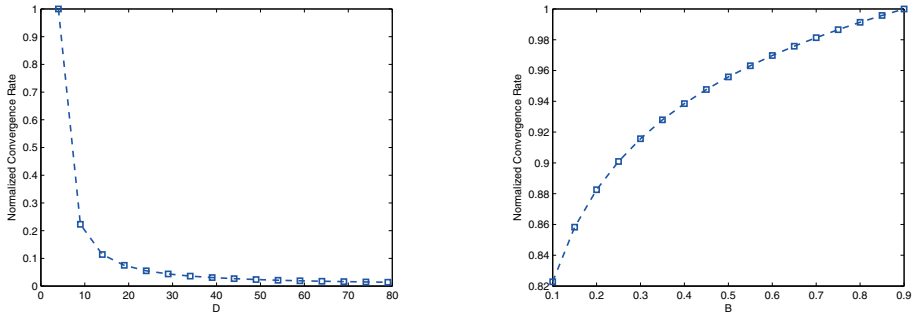
where $\mathfrak{B}$ is the Borel sigma-algebra of $\mathcal{P}$ and the constant Γ_D depends only on D.

From the algorithm we know that $D = M^{(K-1)} > 1$. Figure 3.3(a) shows how the convergence rate scales with D for $\Gamma_D = D$. As expected, convergence speed decreases for larger number of users K and/or actions M, thereby larger D. Note that the effect of increasing K on D is more than that of M.

Now, consider the regression process, which is assumed to be non-parametric. Then the following holds.

Theorem 7 ([Sto82])**.** *Consider a p-times differentiable unknown regression function f and a d-dimensional measurement variable. Let $\hat{f}$ denote an estimator of f based on a training sample of size n, and let $\left\|\hat{f}_n - f\right\|_\infty$ be the L_∞ norm of $\hat{f}_n - f$. Under appropriate regularity condition, the optimal rate of convergence for $\left\|\hat{f}_n - f\right\|_\infty$ to zero is $\left(\frac{\log(n)}{n}\right)^\eta$ where $\eta = \frac{p}{2p+d}$.*

Based on Theorem 7, Algorithm 5 impacts the convergence rate through changing the sampling rate. For any given action, in the worst-case, samples are gathered only at exploration trials. Let R be the number of periods (game horizon). By the algorithm, each joint action profile is expected to be played $\frac{1-\gamma}{M^K}\sum_{j=1}^{J} j = \frac{1-\gamma}{M^K} \cdot \frac{J(J+1)}{2}$ times during J periods (see also the proof of Lemma 3). Moreover, suppose that some fixed number of samples is required to estimate the reward of each joint action profile with sufficient precision. Therefore, it is clear that increasing M and/or K, as well as increasing γ, degrades the sampling rate and with it the convergence speed, since larger game horizon is required for sufficient sampling. Let $B = \frac{1-\gamma}{M^K} < 1$. Figure 3.3(b) shows the convergence speed of the regression process as a function of B.



(a) Calibrated forecaster (b) Regression

Figure 3.3.: Scaling convergence rate with variables and parameters (M, K and γ), $D = M^{(K-1)} > 1$ and $B = \frac{1-\gamma}{M^K} < 1$.

3.4.4. Some Notes on Complexity

As suggested by Algorithm 5 (see also Figure 3.2), the computational burden of the proposed algorithm is due to the calibrated forecasting and regression process. In [MS10], it is shown that at each period j, the $2\epsilon_j$ calibrated forecaster has a complexity in the order of $\epsilon_j^{-(D+1)}$ per step. As the most complex part of the forecaster is to calculate ψ_t, the complexity can be dramatically reduced by approximating ψ_t instead of exactly calculating it. This can be done for instance by using the *adaptive multiplicative weight* algorithm [FS99]. The complexity of the regression process depends on the exact procedure being used; for example, for *Gaussian regression* and *support vector machine*, the complexity is cubical in the number of sample points [RW05], [BCDW07]. Note that in our algorithm, there exist at most $t-1$ samples for each action at each step t.

3.5. Numerical Results

This section consists of two parts. First, the algorithm's work flow is described in a small network. Afterwards we consider a large network, in which the performance of the proposed strategy (CBS) is compared with some other resource allocation methods. Through this section and according to the system model, transmitter-receiver pairs are predefined. Moreover, for each pair, the average channel gains are time-invariant, i.e. they remain fixed during the simulation.

3.5.1. Part One

Network model

We consider an underlay D2D network consisting of two transmitter-receiver pairs (two D2D users), i.e. $K = 2$. There exist two orthogonal channels ($M = 2$), whose availability follows Bernoulli distribution with parameter $\frac{1}{2}$. D2D users are only allowed to use these channels upon availability, i.e. in case they are not occupied by any cellular user. We implement the following selection strategies.

- Statistical centralized strategy: Given global *statistical* channel knowledge and by exhaustive search, a central controller assigns each D2D user some transmission channel so that the assignment corresponds to the most efficient pure strategy equilibrium in the sense of maximum aggregate average reward.

- Calibrated bandit strategy (CBS): D2D users utilize the proposed selection strategy (Algorithm 5).

Since $M = 2$, for each D2D user $k \in \{1, 2\}$, we have $p_1^{(k)} + p_2^{(k)} = 1$, where $p_i^{(k)}$ is the likelihood of D2D user k to take action $i \in \{1, 2\}$, by following the mixed strategy $\left(p_1^{(k)}, p_2^{(k)}\right)$. This implies that for each player k, the probability distribution over all joint action profiles of opponents reduces to the mixed strategy of the other player. We assume $N_\epsilon = 40$.[3] Therefore, the ϵ-grid defines 40 possible mixed strategies (quantized vectors).[4] The primal output of the forecaster of each player is a vector of weights including 40 elements, where every element denotes the likelihood of one of the *quantized* mixed strategies to be played by the other player. The final output of the forecaster is then a mixed strategy extracted from the set of quantized mixed strategies according to this distribution, as described in Algorithm 4.

[3] In general, smaller N_ϵ can be used at early periods to reduce the computational burden. We however consider fixed N_ϵ in order to highlight the evolution of forecasts.

[4] Quantized vectors are indexed as $i = 1, ..., 40$. While $(p_1, p_2) = (0, 1)$ for $i = 1$, we have $(p_1, p_2) = (1, 0)$ for $i = 40$. That is, p_1 increases with the index of quantized vector, whereas p_2 decreases.

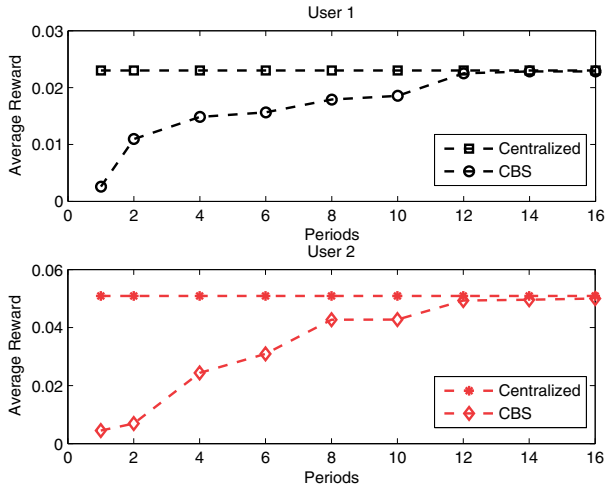


Figure 3.4.: Average reward of CBS versus centralized strategy; orthogonal access.

Orthogonal Multiple Access

D2D users follow here an orthogonal transmission scheme. The average joint rewards of players under all possible joint action profiles are summarized in Table 3.1. From this table, the channel selection game has a pure strategy Nash equilibrium[5] that yields the maximum aggregate reward of the two D2D users, and is achieved when first and second D2D users transmit in channels 1 and 2, respectively.

Table 3.1.: Reward Matrix for Orthogonal Access (u_i:user i, c_j:channel j; $i, j \in \{1, 2\}$)

u_1 \ u_2	c_1	c_2
c_1	0.012,0.000	0.023,0.054
c_2	0.016,0.000	0.008,0.027

The average rewards of players are depicted in Figure 3.4. It can be concluded that for sufficiently large game horizon (large number of periods), the average achieved reward of our strategy converges to that of equilibrium. Players' actions are shown in Figure 3.5, at both early and final stages of the game (before and after the convergence), for 10 randomly-selected trials. By comparing this figure with Table 3.1, it follows that the game converges to equilibrium, which is the joint action $(1, 2)$. The initial forecasters' outputs (ψ, see Algorithm 4) are shown in Figure 3.6(a), for some randomly-selected trial before

[5]Note that Nash equilibrium is a special case of correlated equilibrium.

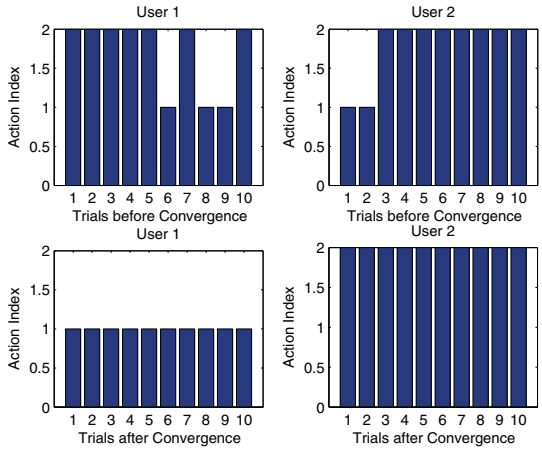
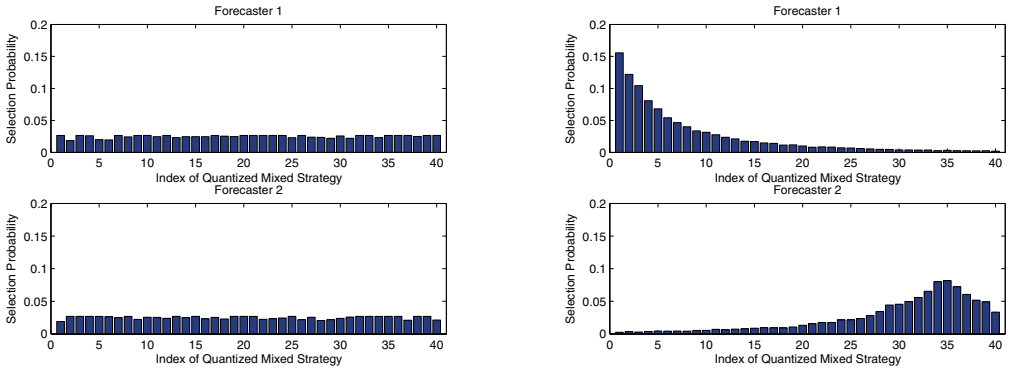


Figure 3.5.: Selected actions of CBS before and after convergence, at 10 random trials; orthogonal access.

convergence. It is clear that outputs are almost uniformly distributed, i.e. all quantized mixed strategies are almost equally likely to occur. This conclusion is in agreement with Figure 3.5, where selected channels before convergence do not follow any specific pattern. Forecasters' outputs at some randomly-selected trial after the convergence are depicted in Figure 3.6(b). In this figure, Forecaster 1 assigns higher weights to the quantized mixed strategies with $p_2 > p_1$, while Forecaster 2 emphasizes the strategies with $p_1 > p_2$. This means that the first and second players are expected by their opponents to select channels 1 and 2, respectively. The predictions are approved by Figure 3.5, where the first and second D2D users finally settle at the first and second channels, respectively.



(a) Before convergence.

(b) After convergence.

Figure 3.6.: Forecasters' outputs at some random trial; orthogonal access.

Non-orthogonal Multiple Access

In case of non-orthogonal multiple access, conflicting D2D users decide whether to transmit in different channels or in a common channel. In order to clarify this, we perform two experiments. For the first and second experiments, the average joint reward matrices are given in Tables 3.2(a) and 3.2(b). From these tables it can be concluded that the most

Table 3.2.: Reward Matrices for Non-Orthogonal Access (u_i:user i, c_j:channel j; $i, j \in \{1, 2\}$)

(a) Case 1

u_1 \ u_2	c_1	c_2
c_1	0.024,0.040	0.024,0.021
c_2	0.075,0.042	0.063,0.021

(b) Case 2

u_1 \ u_2	c_1	c_2
c_1	0.024,0.001	0.024,0.021
c_2	0.075,0.000	0.063,0.021

efficient pure strategy equilibrium points for the first and second games are joint actions $(2, 1)$ and $(2, 2)$, respectively. This means that in the first case, it is beneficial for players to transmit in different channels, whereas in the second case, D2D users achieve higher gains if they both use the second channel.

The average rewards of users are shown in Figures 3.7(a) and 3.7(b), correspondingly, for the two experiments. Moreover, Figures 3.8(a) and 3.8(b) show the users' actions for the first and second experiments. Figures 3.9(a) and 3.9(b) show the forecasters' outputs, at a random trial after convergence (the outputs before convergence are similar to Figure 3.6(a)). Descriptions are similar to the orthogonal case, and hence are omitted.

3.5.2. Part Two

We consider a network with ten D2D users (transmitter-receiver pairs, $K = 10$), and five primary channels ($M = 5$). The channels are available according to a Bernoulli random process with parameter $\frac{1}{2}$, and the performance metric is the aggregate average reward of D2D users. We compare the following approaches.

- Statistical centralized strategy: As described in Section 3.5.1, this approach requires global statistical channel knowledge and a central controller. Stability is here therefore not relevant, as no competition takes place.
- Calibrated bandit strategy (CBS): This approach is proposed in this paper. As explained before, CBS requires no prior knowledge and also no pairwise information exchange. Convergence to the set of correlated equilibria is proved for the general reward generating process, provided that the mean reward is time-invariant and

depends only on the nature (here average channel gains) and the players' joint action profile. A broadcast channel is required only until convergence.

- Game-theoretical pricing [1]: In this strategy, D2D users are modeled as buyers, whereas the BS is the seller. Any given channel is sold to only one D2D user, thereby orthogonal channel access. Information exchange is required among buyers and the seller. Pure strategy Nash equilibrium can be achieved upon existence.

- Bandit with fixed rewards [XWW+12]: In this model, any given channel offers a fixed transmission rate that is equally shared in case of collision. For comparison, the fixed reward of any channel is here selected by averaging the achievable rates of all D2D users through that channel. No prior information or information exchange is required. Convergence to Nash equilibrium is proved in [XWW+12] and [XWS+13], for two specific utility functions.

- No collision bandit strategy [KNJ12]: This is a bandit strategy where upon collision, no reward is paid to the colliding users. Information exchange is required. Stability is not discussed.

- ϵ-greedy Q-learning strategy [BGN11]: At each trial, every player selects the best action so far with probability $1-\epsilon$, and some random action with probability ϵ. No forecasting and/or best response dynamics is performed. No information exchange or prior knowledge is required. Stability is not guaranteed in general, but might exist for potential games.

- Bernoulli bandit strategy [LZK10]: In this model the learning process does not include channel qualities. More precisely, only the availability and the number of users willing to transmit through each channel are learned. No prior knowledge or information exchange is required. Stability is not guaranteed.

- Uniformly random strategy: At each trial, an action is selected uniformly at random.

Results are depicted in Figure 3.10, and some important notes are shortly discussed in the following.

- CBS requires some time to converge to the centralized approach in terms of aggregate average reward of all users. It can be however implemented in a distributed manner with low overhead.

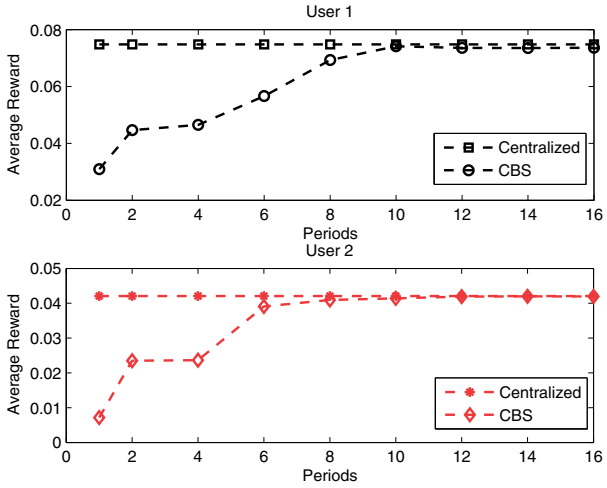
- Greedy Q-learning algorithm exhibits inferior performance in comparison with the CBS, which is mainly due to the absence of forecasting and best response dynam-

ics. It should be also noted that for the general utility functions, convergence to equilibrium is not guaranteed by this algorithm.

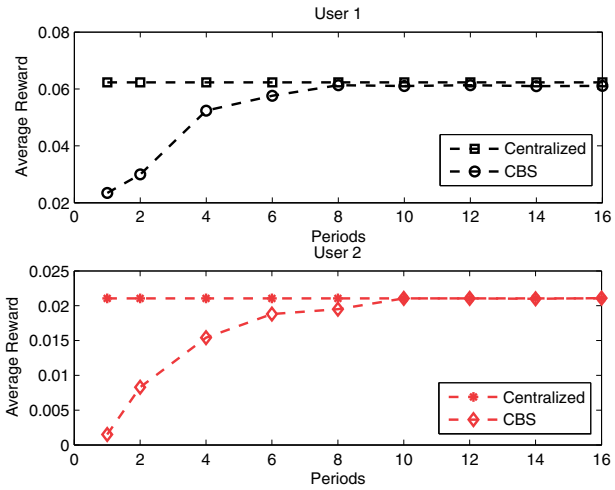
- As expected, orthogonal multiple access is in general suboptimal. In essence, the performances of selection strategies that are based on orthogonal access are inferior to CBS that allows interference.
- The performance of no-collision bandit scheme is poor since transmission through a common channel yields zero reward to all colliding users, which is clearly suboptimal for selfish users and when $M < K$.
- If the transmission channels are different only with respect to the availability, then the performance of the Bernoulli bandit strategy is acceptable; nonetheless, in case the channel quality is also taken into account (as in our model), Bernoulli bandit strategy performs poor.

3.6. Conclusion and Remarks

We studied a channel selection problem in an underlay distributed D2D communications system, where spectrum vacancies of the cellular network are utilized by selfish D2D users. We showed that the channel selection problem boils down to a multi-player multi-armed stochastic bandit game with side-information, and proposed a selection strategy, called CBS, based on no-regret learning and calibrated forecasting. Analysis established that CBS is strongly consistent; that is, for each D2D user, the average accumulated reward in the long run is equal to that of the optimal strategy. Moreover, we proved that if applied by all players, CBS guarantees that the game converges to equilibrium in some sense. In addition, we discussed the convergence rate and complexity issues. As expected intuitively, the convergence speed reduces with increasing number of players and actions, while complexity increases. It was also concluded that the convergence rate and complexity can be controlled by changing the exploration parameter, sampling rate and the applied regression process. The first part of numerical studies confirmed the analytical results with respect to the convergence to an efficient equilibrium. In the second part the proposed approach was compared with some other selection strategies. It was concluded that CBS performs better than greedy Q-learning, as it applies the forecasting procedure. Moreover, the approaches that are based on orthogonal channel access or assign zero reward to the colliding users exhibit inferior performance compared to the CBS.

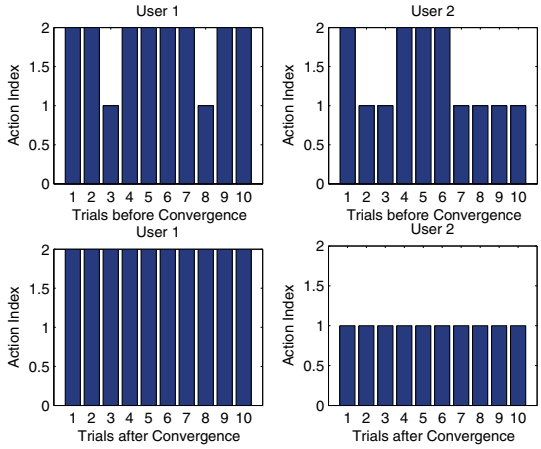


(a) Case 1

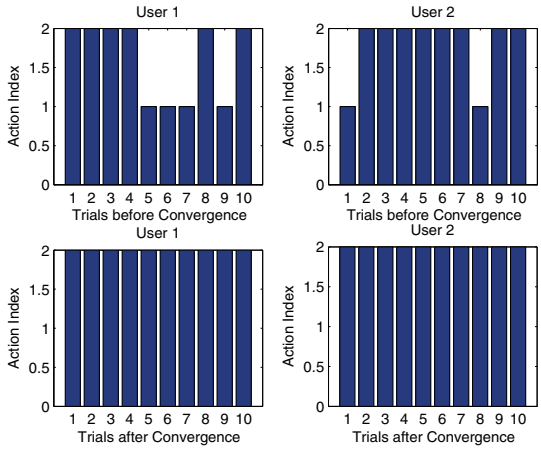


(b) Case 2

Figure 3.7.: Average reward of CBS versus centralized strategy; non-orthogonal access.

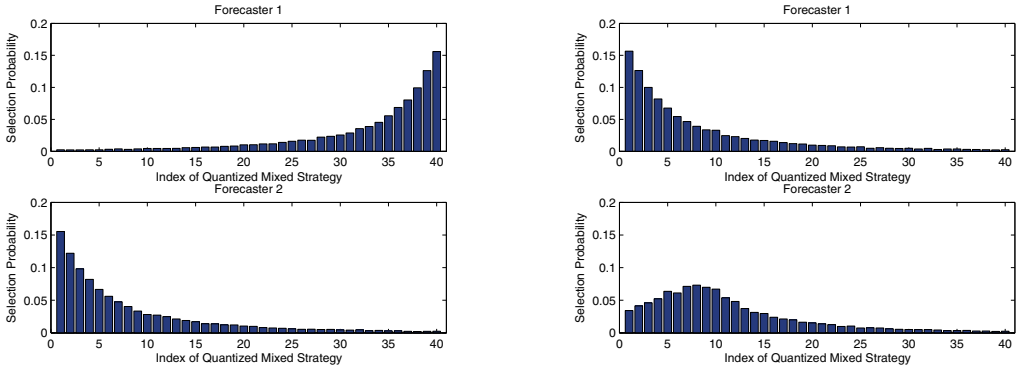


(a) Case 1



(b) Case 2

Figure 3.8.: Selected actions of CBS before and after convergence, at 10 random trials; non-orthogonal access.



(a) Case 1 (b) Case 2

Figure 3.9.: Forecasters' outputs at a random trial after convergence; non-orthogonal access.

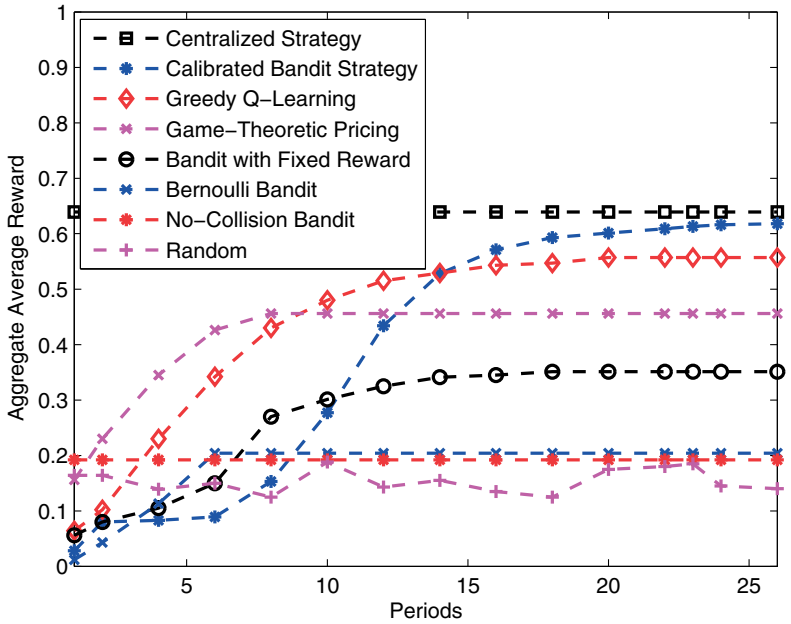


Figure 3.10.: Performance of calibrated bandit strategy (CBS) compared to some other selection strategies.

4. Hybrid Centralized-Distributed Resource Allocation

In this chapter we focus on a decision making problem in a network with two types of agents, namely primary and secondary agents. We assume the existence of an authority that regularizes the network in favor of primary agents by assigning restricted resources to secondary agents. The secondary agents therefore compete with each other in order to access the limited assigned resources. In this context, this chapter studies a resource allocation problem in a single-cell wireless network with multiple D2D users, sharing the available radio frequency channels with cellular users. As described in Chapter 1, in underlay D2D networks, there might be also the possibility of simultaneous D2D and cellular transmissions, provided that the adverse effects of D2D transmissions (secondary agents) on cellular users (primary agents) is minimized, and cellular users are given the priority in using wireless resources. In Section 4.1.1, we introduce the network model and formulate a joint channel allocation and power control problem. The system model considered in this work generalizes the state-of-the-art in the following crucial aspects:

- In a great majority of previous works, including [YDRT11] and [XSH+12a], the network model consists of a specific and limited number of cellular and/or D2D users. In contrast, in our model, arbitrary number of cellular and D2D users coexist in the network.

- Some works such as [BFA11], [FLYW+13] and [JKR+09] study a system with multiple D2D users; however, in every time slot, only one D2D user is allowed to transmit in any frequency channel, primarily allocated to some cellular user. Contrary to these works, we allow multiple D2D users to share a given channel with some cellular user.

- A large body of existing resource allocation schemes are centralized, i.e., the resource management is performed by a base station of the cellular infrastructure that is assumed to be provided with the global channel and network knowledge. To this category belong [PHK13], [ATN+14], [WCC+11] and [HYXY12], for instance. The centralized solution, however, yields large overhead, as well as heavy computational cost to the cellular network. Unlike these studies, in our model we assume that

the BS is only aware of the statistical channel knowledge of cellular users and the geographical locations of D2D users. This information can be simply acquired by using pilot signals for cellular users and the GPS (Global Positioning System) data of D2D users, yielding considerably lower overhead compared to the full information case. Moreover, by using our proposed strategy, the resource allocation problem is solved by the BS only partially, thereby reducing the computational cost.

- While in many studies including [WSH+13], [XSH+13], [SNHH14] and [LJYH14], it is assumed that users are provided with some prior knowledge, in our model D2D and cellular users do not have any information.

In Section 4.1.2, we prove a lower-bound on the aggregate utility of cellular users. Based on this lower-bound, and by taking the higher priority of cellular users into account, we decompose the resource allocation problem into two cascaded problems related to channel allocation and D2D power control. The former problem, which is investigated in Section 4.2, is a multi-objective combinatorial optimization problem that is very costly to solve with respect to the time and computational complexity. Therefore we propose a suboptimal, but efficient, graph-theoretical heuristic solution that involves *maximum-weighted bipartite matching* [Kuh55] and *minimum-weighted graph partitioning* [Bar81]. The problem can be then solved in a centralized manner by the BS, since the solution relies only on strictly limited information. The approach also offers high flexibility with respect to the performance criteria, since quality of service (QoS) or fairness can be taken into account. The latter problem, in turn, deals with maximizing the aggregate utility of D2D users by means of power control, desirably in a distributed manner. In Section 4.3, we model the power control problem as a game with incomplete information, which, in contrast to most previous studies, is defined on a discrete strategy set. We show that this game is an exact potential game and characterize the set of Nash equilibria. Furthermore, we use the *Q-learning fictitious play* strategy [CLRJ13] in order to converge to Nash equilibrium. Finally, extensive numerical analysis is performed to evaluate the performance of the proposed approach in practical cases.

4.1. System Model and Problem Formulation

4.1.1. System Model

We consider the downlink of a single-cell network with one BS denoted by b and a set $\mathcal{L}$ consisting of L cellular users, each denoted by l. The cell is provided with a set $\mathcal{Q}$ of $Q = L$ orthogonal frequency channels. There exists also a set $\mathcal{K}$ of K predefined D2D users, where each D2D user consists of a transmitter-receiver pair, and is represented

either by k or by the pair (k, k'). The BS is able to communicate with multiple cellular users simultaneously, possibly by using multiple antennas. The data stream intended to any given cellular user is transmitted with fixed average power P_c. Each D2D user selects a power-level from the set $\mathcal{M} = \{P_1, P_2, ..., P_M\}$ where $1 < P_1 < P_2 < \cdots < P_M$. We assume that $P_M \ll P_c$, since in general the BS has access to larger energy resources in comparison with user devices. Primarily, each channel $q \in \mathcal{Q}$ is used i) by the BS in order to transmit to some set $\mathcal{L}_q \subseteq \mathcal{L}$ of L_q cellular users, and ii) by a set $\mathcal{K}_q \subseteq \mathcal{K}$ of K_q D2D users for direct communications. We assume that $L_q = 1 \; \forall \; q \in \mathcal{Q}$; that is, each channel is assigned exactly one cellular user and therefore no vacant channel exists. This assumption is made in order to protect the cellular users from an excessive interference due to a high BS power. We use $\mathbf{i}_q = \left(i^{(1)}, ..., i^{(K_q)}\right)$ to denote the vector of transmission powers of the D2D users that transmit through channel q. Similar to the previous chapters, $|h_{uv,q}|^2 > 0$ is the *average* gain of channel q from transmitter u to the receiver v. We assume that $|h_{uv,q}|^2 = |h'_{uv}|^2 \left|h''_{uv,q}\right|^2$, where $0 < |h'_{uv}|^2 \leq 1$ and $0 < \left|h''_{uv,q}\right|^2 \leq 1$ stand for the path loss and fast fading components, respectively. We assume that the channel gains of any given link are drawn from some distribution with time-invariant mean value. Moreover, due to the channel reciprocity, we have $|h_{uv,q}|^2 = |h_{vu,q}|^2$. Signal-to-interference ratio (SIR) is denoted by γ. We consider a high SIR regime where $1 < \gamma$, so that $\log(1 + \gamma) \approx \log(\gamma)$. When treating interference as noise, $\log(\gamma)$ represents the achievable transmission rate of interference-limited point to point transmission.

The average utility of some cellular user $l \in \mathcal{L}_q$ that occupies channel q is defined as

$$f^{(l)}(q, \mathbf{i}_q) = \log\left(\frac{P_c \left|h_{bl,q}\right|^2}{1 + \sum_{k \in \mathcal{K}_q} i^{(k)} \left|h_{kl,q}\right|^2}\right), \tag{4.1}$$

which corresponds to the achievable transmission rate, as described before.

Since D2D users are also involved in power control in addition to channel allocation, the average utility of any D2D user $k \in \mathcal{K}_q$ that shares channel q with some cellular user $l \in \mathcal{L}_q$ is defined as

$$f^{(k)}(q, \mathbf{i}_q) = \log\left(\frac{i^{(k)} \left|h_{kk',q}\right|^2}{1 + \sum_{j \in \mathcal{K}_q, j \neq k} i^{(j)} \left|h_{jk',q}\right|^2 + P_c \left|h_{bk',q}\right|^2}\right) - \alpha i^{(k)}, \tag{4.2}$$

where α is a fixed power price factor to penalize excessive power usage. Therefore, by definition, the utility of a D2D user corresponds to its transmission rate (see above) minus a cost that is paid to the cellular user in order to reimburse the adverse effects of spectrum sharing. The price factor can either be equal for all D2D users (as in (4.2)), or be user specific; for instance, proportional to the channel gain (or distance) between a D2D user

and the cellular user transmitting in the same channel [1]. The analysis in this chapter holds for both cases.

We consider a model with strictly limited information, as described in the following assumption.

Assumption (A5). *Each of the following is assumed throughout this chapter.*

a) The BS has the knowledge of i) geographical locations of cellular and D2D users as well as the path loss exponent, thus $|h'_{lk}|^2 \ \forall\ l \in \mathcal{L}, k \in \mathcal{K}$*, and ii) the average gain of all channels from every cellular user to the BS, i.e.* $|h_{bl,q}|^2 \ \forall\ l \in \mathcal{L}, q \in \mathcal{Q}$*.*

b) The BS has no information about the fast fading component of cellular to cellular or D2D to D2D links.

c) Cellular and D2D users have no channel knowledge.

4.1.2. Problem Formulation

Network aggregate utility is conventionally regarded as a measure for evaluating the performance of resource management protocols [CNT08], [RW08], [FP14], [CMRWS12]. Based on this criterion, the problem is to allocate channels and power-levels to the cellular and D2D users so as to maximize the network aggregate utility. With (4.1) and (4.2) in hand, this problem can be stated formally as

$$\underset{\mathcal{L}_q, \mathcal{K}_q, \mathbf{i}_q}{\text{maximize}} \ \sum_{q=1}^{Q} \left(\sum_{l \in \mathcal{L}_q} f^{(l)}\left(q, \mathbf{i}_q\right) + \sum_{k \in \mathcal{K}_q} f^{(k)}\left(q, \mathbf{i}_q\right) \right), \tag{4.3}$$

where $\mathcal{L}_q \subseteq \mathcal{L}$, $\mathcal{K}_q \subseteq \mathcal{K}$ and $\mathbf{i}_q \in \bigotimes_{k=1}^{K_q} \{P_1, ..., P_M\}$. Note that unlike some previous works such as [XWW+12] and [XWS+13], the utility functions defined here are user-specific, i.e. any given channel pays different rewards to different users. As a result, the *set* of D2D and cellular users allocated to each channel is required to be determined, and not the *number* of users.

Such formulation however does not comply with the underlay D2D concept, and suffers from the following drawbacks that make it difficult or even impossible to deal with: i) The objective in (4.3) is not available at the BS due to the lack of information (see Assumption (A5)), ii) The higher priority of cellular users is not taken into account, and iii) The objective function depends on both channel and power allocations, which are mutually dependent. Therefore a solution to (4.3) is difficult to obtain and is expected to be not amenable to distributed implementation. Our goal is therefore to develop a sophisticated heuristic approach. To this end, we first prove a lower-bound on the aggregate utility

of cellular users that enables us to decouple the channel allocation and power control problems.

Proposition 5. *For any* $\mathbf{i}_q, P_c$ *and channel gains, we have*

$$\sum_{q=1}^{Q}\sum_{l\in\mathcal{L}_q} f^{(l)}\left(q,\mathbf{i}_q\right) > \sum_{q=1}^{Q}\sum_{l\in\mathcal{L}_q}\log\left(P_c\left|h_{bl,q}\right|^2\right) - \sum_{q=1}^{Q}\sum_{l\in\mathcal{L}_q}\sum_{k\in\mathcal{K}_q} P_M\left|h'_{kl}\right|^2. \tag{4.4}$$

Proof. See Section 1 of Appendix D. □

In words, the lower-bound in (4.4) corresponds to the worst-case scenario where all D2D users transmit with the maximum available power, and the fast fading components of all D2D to cellular links equal one, yielding maximum interference. In essence, the bound does not depend on D2D power allocation and relies only on the available information at the BS; hence it may serve as a basis for resource management.

Since the cellular users are assumed to have a higher priority and therefore should be served first, we propose a two-step resource allocation strategy. In the first step, the objective is to maximize the lower-bound of the aggregate utility of cellular users, given by (4.4). More precisely, given P_M, P_c and imperfect channel knowledge, we aim at assigning channels to cellular and D2D users so as

$$\underset{\mathcal{L}_q,\mathcal{K}_q}{\text{maximize}} \sum_{q=1}^{Q}\sum_{l\in\mathcal{L}_q}\log\left(P_c\left|h_{bl,q}\right|^2\right) - \sum_{q=1}^{Q}\sum_{l\in\mathcal{L}_q}\sum_{k\in\mathcal{K}_q} P_M\left|h'_{kl}\right|^2, \tag{4.5}$$

subject to

$$L_q = 1,\ \forall\ q\in\mathcal{Q}. \tag{4.6}$$

This problem is investigated in Section 4.2.

Once the channels are allocated, in the second step we address the power control problem for D2D users, with the goal of maximizing the aggregate utility of D2D users as formalized below.

$$\underset{\mathbf{i}_q\in\bigotimes_{k=1}^{K_q}\{P_1,\ldots,P_M\}}{\text{maximize}} \sum_{q=1}^{Q}\sum_{k\in\mathcal{K}_q} f^{(k)}\left(q,\mathbf{i}_q\right). \tag{4.7}$$

Section 4.3 is devoted to this problem.

Summarizing, the resource allocation problem is decomposed to a channel allocation problem for all users followed by a power control problem for D2D users. As we see later, the first problem is solved at the BS using a centralized method, whereas the second problem is solved by D2D users in a distributed manner. Using such a two-stage scheme, not only a higher priority of cellular users is taken into account, but also D2D users utilize

the assigned channels efficiently. Moreover, the limited available information is exploited with low computational effort.

4.2. Channel Allocation

This section deals with the first step of resource management, i.e. the channel assignment, with the goal of optimizing the performance of cellular users in terms of (4.5).

4.2.1. The Channel Allocation Scheme

We notice that the first and second terms in (4.5) are respectively *proportional* to the sum of desired signals and interferences, over all cellular users. Moreover, while the first term depends only on cellular users, the second term depends also on D2D users. Roughly speaking, the problem in (4.5) can be rephrased as $\text{maximize}_{x,y}\ f(x) - g(x,y)$, where x and y respectively denote the cellular and D2D channel assignments. This problem is a multi-objective combinatorial optimization problem that is NP-hard and hence notoriously difficult to solve. Therefore we propose the following suboptimal, but simple and efficient, heuristic approach: At the beginning maximize the first term of (4.5) (weighted signal sum), so that the sets $\mathcal{L}_q$, $q \in \mathcal{Q}$, are defined. Afterwards, given $\mathcal{L}_q$, allocate D2D users to the frequency channels in a way that the second term (interference sum) is minimized. Formally,

$$\underset{\mathcal{L}_q}{\text{maximize}} \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log\left(P_c \left|h_{bl,q}\right|^2\right), \tag{4.8}$$

subject to (4.6) and

$$\underset{\mathcal{K}_q}{\text{minimize}} \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \sum_{k \in \mathcal{K}_q} P_M \left|h'_{kl}\right|^2. \tag{4.9}$$

We call (4.8) and (4.9) as *assignment* and *clustering* problems, respectively. In what follows, we show that these problems boil down to two classic graph-theoretical problems on the induced network graph, namely *maximum-weighted bipartite graph matching* and *minimum-weighted graph partitioning*.

Assignment Problem

In the following, we show that the problem in (4.8) can be formulated as a weighted bipartite matching problem, defined below.

Definition 7 (Weighted Bipartite Matching)**.** *Let $G = (\mathcal{V}, \mathcal{E})$ be a weighted bipartite graph where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$, $\mathcal{V}_1 \cap \mathcal{V}_2 = \varnothing$ and $\mathcal{E} \subseteq \mathcal{V}_1 \times \mathcal{V}_2$. Each edge $e \in \mathcal{E}$ connecting any two*

vertices $x \in \mathcal{V}_1$ and $y \in \mathcal{V}_2$ is associated with some weight w_{xy}. The weights are gathered in the $V_1 \times V_2$ graph matrix denoted by $\boldsymbol{W} = [w_{xy}]$.
Matching: A matching is a subset $\mathcal{M} \subseteq \mathcal{E}$ such that $\forall v \in \mathcal{V}$ at most one edge in $\mathcal{M}$ is incident upon v.
Maximum Matching: A matching $\mathcal{M}$ such that every other matching $\mathcal{M}'$ satisfies $W_{\mathcal{M}'} \leq W_{\mathcal{M}}$, where $W_{\mathcal{U}}$ denotes the total weight of the selected edges for some matching $\mathcal{U}$.
Minimum Matching: A matching $\mathcal{M}$ such that every other matching $\mathcal{M}'$ satisfies $W_{\mathcal{M}} \leq W_{\mathcal{M}'}$.

Based on Definition 7, consider the bipartite graph $G_L(\mathcal{V}, \mathcal{E})$, with $\mathcal{V}_1 = \mathcal{L}$ (the set of cellular users) and $\mathcal{V}_2 = \mathcal{Q}$ (the set of channels). The weight of the edge connecting $l \in \mathcal{L}$ and $q \in \mathcal{Q}$, w_{lq}, is defined as the weighted average gain of channel q between the cellular user l and the BS, i.e. $\log\left(P_c \left|h_{bl,q}\right|^2\right)$. The problem is to assign each cellular user a channel so that (4.6) and (4.8) are satisfied. Let the assignment be presented by an $L \times Q$ assignment matrix $\mathbf{A} = [a_{lq}]$, where

$$a_{lq} = \begin{cases} 1 & \text{if } l \in \mathcal{L}_q \\ 0 & \text{otherwise} \end{cases}. \tag{4.10}$$

Therefore $\mathbf{A}$ satisfies the following constraints

$$\sum_{l=1}^{L} a_{lq} \leq 1 \quad , \; q \in \{1, 2, ..., Q\}, \tag{4.11}$$

$$\sum_{q=1}^{Q} a_{lq} = 1 \quad , \; l \in \{1, 2, ..., L\}, \tag{4.12}$$

$$a_{lq} \in \{0, 1\} \quad , \; \forall \; l, q. \tag{4.13}$$

While (4.11) implies that each channel serves at most one cellular user, (4.12) means that each cellular user is served by exactly one channel. Note that equality holds in (4.11) since we assume $Q = L$ (see Section 4.1.1). The sum of edges' weights hence yields

$$\sum_{q=1}^{Q} \sum_{l \in \mathcal{L}} w_{lq} a_{lq} = \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} w_{lq}. \tag{4.14}$$

Thus the problem in (4.8) subject to (4.6) is equivalent to maximizing (4.14), subject to (4.11), (4.12), and (4.13); that is, it corresponds to the maximum matching of G_L.

Clustering Problem

This step consists of allocating channels to D2D users, with the goal of minimizing the total interference at cellular users, over all channels. In order to address this problem, we

first define the network graph.

Definition 8 (Network Graph)**.** *The network graph for any channel $q \in \mathcal{Q}$ is an undirected graph $G_N = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$, where $\mathcal{V}_1$ and $\mathcal{V}_2$ represent the set of K D2D transmitters and L cellular receivers, respectively. The weight of an edge between any pair of graph vertices (x, y) is denoted by w_{xy}, where w_{xy} is equal to the average gain of channel q between x and y.*

However, by Assumption (A5), only limited channel state information is available at the BS; therefore the network graph cannot be constructed. As a result, we define the *estimated* network graph, which can be reproduced by the BS using the available information.

Definition 9 (Estimated Network Graph)**.** *Estimated network graph is an undirected graph $G_E = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$, where $\mathcal{V}_1$ and $\mathcal{V}_2$ represent the set of K D2D transmitters and L cellular receivers, respectively. The weight of an edge between any D2D transmitter k and cellular receiver l is defined as $w_{kl} = P_M |h'_{kl}|^2$. The weight of the edge between any two cellular users and any two D2D users are respectively equal to some constant $C > KP_M$ and zero.*[1]

Next we show that problem (4.9) can be rephrased as Q-way minimum-weighted graph partitioning of the estimated network graph G_E.

Definition 10 (Q-way Weighted Partitioning)**.** *Let $G = (\mathcal{V}, \mathcal{E})$ be a weighted graph where each edge $e \in \mathcal{E}$ connecting any two vertices x and y is associated with some weight w_{xy}. The weights are gathered in a $V \times V$ matrix denoted by $\mathbf{W} = [w_{xy}]$. The minimum-weighted Q-way partitioning problem divides the set of vertices into Q disjoint subsets, in a way that the sum weights of edges whose incident vertices fall into the same subset is minimized.*

Now consider the estimated network graph, G_E. Then solving (4.9) is equivalent to finding some $(L + K) \times Q$ assignment matrix $\mathbf{B} = [b_{jq}]$, where

$$b_{jq} = \begin{cases} 1 & \text{if } j \in \mathcal{L}_q \cup \mathcal{K}_q \\ 0 & \text{otherwise} \end{cases}. \tag{4.15}$$

Thus each column in $\mathbf{B}$, e.g. $\mathbf{B}_q = \left[b_{1q}, b_{2q}, ..., b_{(L+K)q}\right]^T, q \in \{1, 2, ..., Q\}$, is an indicator describing cluster q. Therefore b_{jq} satisfies the following constraints

$$\textstyle\sum_{j=1}^{L+K} b_{jq} = L_q + K_q, \quad q \in \{1, 2, ..., Q\}, \tag{4.16}$$

[1] Later we see that this definition results in some form of clustering that reduces the cellular to cellular and also the D2D to cellular interferences. D2D to D2D interference is however neglected, implying that in the absence of full and precise channel knowledge, the priority is to protect cellular users.

$$\sum_{q=1}^{Q} b_{jq} = 1, \quad j \in \{1, 2, ..., L+K\}, \tag{4.17}$$

and

$$b_{jq} \in \{0, 1\}, \quad \forall\, j, q. \tag{4.18}$$

The sum of edges' weights connecting the users in cluster q hence follows as

$$\frac{1}{2} \sum_{j \in \mathcal{L} \cup \mathcal{K}} \sum_{j' \in \mathcal{L} \cup \mathcal{K}} w_{jj'} b_{jq} b_{j'q} = \frac{1}{2} \mathbf{B}_q^T \mathbf{W}_E \mathbf{B}_q, \tag{4.19}$$

where $\mathbf{W}_E$ is the weight matrix of G_E. As a result, the sum of weights of the edges that are not cut by the Q-way partitioning of G_E yields

$$\begin{aligned} &\frac{1}{2} \sum_{q=1}^{Q} \sum_{j \in \mathcal{L} \cup \mathcal{K}} \sum_{j' \in \mathcal{L} \cup \mathcal{K}} w_{jj'} b_{jq} b_{j'q} \\ &\quad = \frac{1}{2} \sum_{q=1}^{Q} \sum_{j \in \mathcal{K}} \sum_{j' \in \mathcal{K}} w_{jj'} b_{jq} b_{j'q} + \frac{1}{2} \sum_{q=1}^{Q} \sum_{j \in \mathcal{L}} \sum_{j' \in \mathcal{L}} w_{jj'} b_{jq} b_{j'q} \\ &\qquad + 2 \times \frac{1}{2} \sum_{q=1}^{Q} \sum_{j \in \mathcal{L}} \sum_{j' \in \mathcal{K}} w_{jj'} b_{jq} b_{j'q}. \end{aligned} \tag{4.20}$$

The first term on the right-hand side of (4.20) is zero by the definition of G_E. Also, by the following proposition, the second term equals zero as well, since every minimum-weighted partitioning assigns exactly one cellular user to each cluster.

Proposition 6. *Any minimum-weighted Q-way partitioning of the estimated network graph G_E assigns exactly one cellular user to each cluster, that is $L_q = 1\ \forall q \in \mathcal{Q}$.*

Proof. See Section 2 of Appendix D. □

By Proposition 6 and comparing (4.19) with (4.20), we have

$$\begin{aligned} \frac{1}{2} \sum_{q=1}^{Q} \mathbf{B}_q^T \mathbf{W}_E \mathbf{B}_q &= \sum_{q=1}^{Q} \sum_{j \in \mathcal{L}} \sum_{j' \in \mathcal{K}} w_{jj'} b_{jq} b_{j'q} \\ &= \sum_{q=1}^{Q} \sum_{j \in \mathcal{L}_q} \sum_{j' \in \mathcal{K}_q} w_{jj'}. \end{aligned} \tag{4.21}$$

By comparing (4.21) with (4.9), and by using the definition of G_E, it can be concluded that (4.9) is equivalent to the minimum-weighted Q-way partitioning of G_E.

4.2.2. Time and Computational Complexity

In principal, the proposed channel allocation scheme solves two problems, maximum-weighted matching and minimum-weighted partitioning. The latter problem, however, can be itself reformulated as a minimum-weighted matching, due to the special characteristics of the defined estimated network graph. This is described formally in the following proposition.

Proposition 7. *Define a bipartite graph $G'(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}_1 = \mathcal{K}$ and $\mathcal{V}_2$ is produced by K times replicating $\mathcal{L}$, i.e. $\mathcal{V}_2 = \underbrace{\mathcal{L} \cup \mathcal{L} ... \cup \mathcal{L}}_{\times K}$. The weight of any edge connecting some D2D user $k \in \mathcal{V}_1$ to each copy $l' \in \mathcal{V}_2$ of some cellular user $l \in \mathcal{L}$ is $w_{lk} = \mathbf{W}_E[k, l] = P_M |h'_{kl}|^2$, i.e. equal to the weight of the edge connecting k and l in the estimated network graph, G_E. Then the minimum-weighted Q-way partitioning of G_E is equivalent to a minimum-weighted bipartite matching of G'.*

Proof. See Section 3 of Appendix D. □

Based on Proposition 7, it can be concluded that the channel allocation algorithm solves two (parallel) weighted matching problems. Weighted matching is a classic graph-theoretical problem for which numerous efficient algorithmic solutions exist. A well-known solution is the Hungarian algorithm [Kuh55]. For a bipartite graph $G(\mathcal{V}, \mathcal{E})$, the space complexity of the Hungarian algorithm is $O(V^2 E)$ with $V = \max\{V_1, V_2\}$,[2] that is polynomial in the number of vertices and also in the number of edges. The running time is $O(V^3)$, which is therefore polynomial in the number of vertices. In our model, for the first matching we have $V = L$ and $E = L^2$ by the definition of G_L.[3] For the second matching, on the other hand, we have $V = KL$ and $E = (KL)^2$, by the definition of G_E and Proposition 7. More algorithmic solutions can be found in [Gal86] and [MV80] for instance.

4.2.3. QoS Guarantee and Fairness

Despite being suboptimal, the decoupling approach described in Section 4.2.1 enables us to solve the channel allocation problem efficiently under a variety of constraints, thereby offering high flexibility and applicability. Two examples are given below.

- **QoS requirement for cellular users**: By problem (4.5), the goal of channel allocation is to provide *every* D2D user with some transmission channel, in a way that the aggregate utility of cellular users is maximized; as a result, the *individual*

[2] In case $V_1 \neq V_2$, dummy vertices are added. See [Kuh55] for details.

[3] The number of edges corresponds to the worst-case, where the bipartite graph is complete, i.e. there exists an edge between any pair $x \in \mathcal{V}_1$ and $y \in \mathcal{V}_2$.

performances of cellular users are ignored. In many networks, however, cellular users require some specific QoS that restricts the amount of tolerable interference. Let each cellular user l require some minimum utility, $f_{\min}^{(l)}$, by which its QoS is guaranteed. After solving problem (4.8), each cellular user is assigned a channel. Assume that it is feasible to satisfy the required QoS of any cellular user in the assigned channel. As the nominator of (4.1) is known, the maximum tolerable interference of each cellular user l, say $I_{\max}^{(l)}$, can be calculated given $f_{\min}^{(l)}$. We construct a bipartite graph with $\mathcal{V}_1 = \mathcal{K}$ and $\mathcal{V}_2 = \mathcal{L}$. The problem is then to assign *as many as possible* D2D users to cellular users (thus to channels), so that no interference experienced by any cellular user exceeds the maximum tolerable value. Formally, the problem is to find an $K \times L$ assignment matrix $\mathbf{X} = [x_{kl}]$ so that

$$\text{maximize} \sum_{l=1}^{L} \sum_{k=1}^{K} x_{kl}, \tag{4.22}$$

subject to the following constraints

$$\sum_{k \in \mathcal{K}} w_{kl}\ x_{kl} \leq I_{\max}^{(l)},\ \forall l \in \mathcal{L}, \tag{4.23}$$

$$\sum_{l=1}^{L} x_{kl} \leq 1,\ \forall k \in \mathcal{K}, \tag{4.24}$$

and

$$x_{kl} \in \{0, 1\},\ \forall l, k. \tag{4.25}$$

Note that by the definition of the estimated network graph, $w_{kl} = P_M \left|h'_{kl}\right|^2$, i.e. it is an upper-bound of the interference experienced by cellular user l due to D2D user k. This problem is known as the *generalized assignment problem* which is NP-hard; nonetheless, efficient approximate solutions exist. See [CKR06] for an example.

- **Fairness requirement**: The problem is here similar to the partitioning problem described in Section 4.2.1, with the additional requirement that the resulted clusters are balanced, in the sense that the interference experienced by cellular users due to D2D users are *almost* equal. Formally, desired is to solve (4.9), subject to (4.16), (4.17) and (4.18), so that $\sum_{l \in \mathcal{L}_1} \sum_{k \in \mathcal{K}_1} w_{kl} \approx \sum_{l \in \mathcal{L}_2} \sum_{k \in \mathcal{K}_2} w_{kl} \approx ... \approx \sum_{l \in \mathcal{L}_Q} \sum_{k \in \mathcal{K}_Q} w_{kl}$. It should be emphasized that in this context, the burden of D2D communications is divided (almost) equally among cellular users; this constraint however does not necessarily mean that all cellular users achieve equal utilities.

4.3. Power Control

This section deals with the second step of resource assignment, i.e. D2D power control, with the aim of optimizing the performance of D2D users in terms of (4.7).

4.3.1. Power Control Game

As described in the foregoing section, while performing the channel assignment, the BS ignores the potential interferences that might arise among D2D users, due to the lack of information and also their lower priority. Precisely, D2D users are partitioned into clusters and each cluster is assigned a single channel. Given no information, each D2D user therefore craves to maximize its own utility, thereby causing interference to the users with whom it shares a channel. By means of power control, however, interference can be managed so that the channel assigned to each cluster is utilized efficiently. We model the power control problem as a game with incomplete information, defined on a discrete strategy set. We show that the game is potential and characterize the set of Nash equilibria. Potential games and Nash equilibrium are defined in Section 1 of Appendix A.

As clusters are assigned orthogonal channels, the actions of D2D users inside any given cluster do not affect the utilities of the users outside that cluster. Therefore the power allocation problem in any cluster $q \in \{1, ..., Q\}$ can be formulated as a game among K_q D2D users, described formally in the following.

Definition 11 (Cluster Power Allocation Game)**.** *The power allocation game of cluster $q \in \{1, ..., Q\}$ is a strategic game defined as $\mathfrak{G}_q = \left\{\mathcal{K}_q, \mathcal{I}, \left\{f^{(k)}\right\}_{k\in\mathcal{K}_q}\right\}$, where $\mathcal{K}_q$ is the set of D2D users assigned to channel q, $\mathcal{I} = \bigotimes_{k=1}^{K_q} \{P_1, P_2, ..., P_M\}$ is the set of joint actions with realizations $\mathbf{i}_q = \left(i^{(1)}, ..., i^{(K_q)}\right)$, and $f^{(k)} : \mathcal{I} \to \mathbb{R}^+$ is the payoff function of player $k \in \{1, ..., K_q\}$ defined by (4.2).*

A crucial difference between the cluster power allocation game and the standard power control games investigated in other studies including [SBP06] is that the players' strategy is here selected from a *discrete set*, while in the previous contributions the strategy set is continuous. Consequently, most of the existing results do not hold, and hence we proceed to the following theorem.

Theorem 8. *a) The cluster power allocation game (Definition 11) is an exact potential game with potential*

$$v(\mathbf{i}_q) = \sum_{k\in\mathcal{K}_q} \log\left(i^{(k)}\right) - \sum_{k\in\mathcal{K}_q} \alpha i^{(k)}. \tag{4.26}$$

b) Denote the set of potential maximizers by $\mathcal{V}_{\max}$. Then, a joint action profile $\mathbf{i}_q$ is a Nash equilibrium if and only if $\mathbf{i}_q \in \mathcal{V}_{\max}$.

Proof. See Section 4 of Appendix D. □

4.3.2. Q-Learning Better-Reply Dynamics

According to the system model, in the cluster power allocation game (Definition 11), the utility functions are not known by the players (D2D users) in advance. Therefore they require interacting with the environment in order to learn the optimal joint action profile in the sense of aggregate utility maximization, and to achieve equilibrium. We consider the cluster power allocation game to be a game with noisy payoffs. In such games, for each joint action profile $\mathbf{i}_q \in \mathcal{I}$ of K_q players, the utility achieved by player k at each interaction can be written as $g_t^{(k)}(\mathbf{i}_q) = f^{(k)}(\mathbf{i}_q) + \mathtt{C}^{(k)}(\mathbf{i}_q)$, where $\mathtt{C}^{(k)}$ is a random fluctuation with zero-mean and bounded variance, independent from all other random variables. During the learning process, each player faces a trade-off between gathering information (learning) on the one side and using the information to achieve higher utility in future (control) on the other side. This trade-off is known as the exploration-exploitation dilemma. In order to deal with this dilemma and also to achieve an efficient equilibrium in a distributed manner, we use *Q-learning better-reply dynamics* [CLRJ13]. This strategy consists of three main steps that are performed recursively: 1) Observe the personal reward and also the actions of opponents.[4] 2) Update the Q-values of the played joint action profile, 3) With a small probability $\epsilon \ll 1$, select an action uniformly at random, while with a large probability, $1-\epsilon$, play according to the better-reply dynamics that is described in the following definition.

Definition 12 (Better-Reply Dynamics [CLRJ13])**.** *Assume that at some trial $t-1$, a player k plays with action $p_{k,t-1}$. Then, at trial t, with probability ζ_k, the player selects the same action as in the previous trial, $t-1$, i.e. $i_t^{(k)} = i_{t-1}^{(k)}$. With probability $1-\zeta_k$, however, the player selects an action according to a distribution that puts positive probabilities only on actions that are better replies to its (finite) memory than $i_{t-1}^{(k)}$. For instance, it selects an action according a uniform distribution over all better-replies.*

For readers' convenience, the detailed strategy is described in Algorithm 6 for some player $k \in \mathcal{K}_q$.

[4] When using multi-agent Q-learning algorithms, conventionally it is assumed that every agent observes the state of the environment and/or the actions of its opponents [VH04b]. In our model, players are therefore required to announce their transmission powers, for example by broadcasting in a specific time period, borrowed from the total transmission time. This overhead, however, is much less than that of the frequent and pairwise information exchange, for which usually a control channel is required [LLK12]. The reason is that after convergence, which is achieved relatively fast, the transmission powers of players remain fixed. Therefore no more broadcasting is required, and the borrowed time period is again available for useful data transmission.

Algorithm 6 Q-Learning Better-Reply Dynamics [CLRJ13]

1: Select arbitrary positive constants c_λ and c_ε.
2: Select learning parameters $\rho_\lambda \in \left[\frac{1}{2}, 1\right]$.
3: Let $\mathbf{p}_t^{(k)} = \left(p_{1,t}^{(k)}, ..., p_{M,t}^{(k)}\right)$ be the mixed strategy of player k at time t. Let $\mathbf{p}_1^{(k)}$ be the uniform distribution over all actions (power levels).
4: Select an action, $i_1^{(k)}$, using $\mathbf{p}_1^{(k)}$. Play and observe the reward.
5: **for** $t = 2, ..., T$ **do**
6: Let

$$\varepsilon_t = c_\varepsilon t^{\frac{-1}{K_q}}. \tag{4.27}$$

7:
- With probability ε_t, let $\mathbf{p}_t^{(k)}$ be the uniform distribution over all actions.
- With probability $1 - \varepsilon_t$, perform the following (better-reply dynamics):
  - With probability ζ_k, let $\mathbf{p}_t^{(k)}$ be the Dirac probability distribution on $i_{t-1}^{(k)}$.
  - With probability $1 - \zeta_k$, let $\mathbf{p}_t^{(k)}$ be the uniform distribution over all actions that are better replies to the full (finite) memory than $i_{t-1}^{(k)}$.

8: Using $\mathbf{p}_t^{(k)}$, select the action of time t, $i_t^{(k)}$, and play.
9: Announce the selected action. Moreover, observe the played joint action profile of other players, $\mathbf{i}_t^{(-k)}$, and also the achieved reward, $g_t^{(k)}(\mathbf{i}_{q,t})$, $\mathbf{i}_{q,t} = \left(i_t^{(k)}, \mathbf{i}_{q,t}^{(-k)}\right) = \left(i_t^{(1)}, ..., i_t^{(k)}, ..., i_t^{(K_q)}\right)$.
10: Update the Q-value of the played joint action profile as

$$Q_{t+1}^{(k)}(\mathbf{i}_{q,t}) = Q_t^{(k)}(\mathbf{i}_{q,t}) + \lambda_t \left(g_t^{(k)}(\mathbf{i}_{q,t}) - Q_t^{(k)}(\mathbf{i}_{q,t})\right) \mathbf{1}_{\mathbf{i}_{q,t}}, \tag{4.28}$$

with

$$\lambda_t = \left(c_\lambda + \#^t[\mathbf{i}_{q,t}]\right)^{-\rho_\lambda}, \tag{4.29}$$

where $\#^t[\mathbf{i}_{q,t}]$ denotes the number of trials in which $\mathbf{i}_{q,t}$ is played, and $\mathbf{1}_{\mathbf{i}_{q,t}}$ is the indicator function.
11: **end for**

Theorem 9 ([CLRJ13])**.** *The Q-learning better-reply dynamics (Algorithm 6) with ε_t, λ_t given by (4.27) and (4.29) respectively, converges to a pure Nash equilibrium in games with noisy unknown rewards that are generic and admit a potential function.*

Corollary 4. *By using Q-learning better-reply dynamics, the cluster power allocation game (Definition 11) converges to a Nash equilibrium that maximizes the potential function.*

Proof. The proof directly follows from Theorem 8 and Theorem 9. □

Remark 4 (Price of Stability)**.** *Note that the equilibrium achieved by Q-learning better-reply dynamics does not necessarily maximizes the sum utilities of all players (social welfare), although such solution is desired by (4.7). The inefficiency of equilibrium is in fact the price of the absence of an authority to mandate agents to use a specific transmission power, and is formally referred to as the 'price of stability' [GC12].*

4.4. Numerical Analysis

We consider an underlay D2D communications system, consisting of twelve D2D users ($K = 12$) and five cellular users ($L = 5$), as depicted in Figure 4.1. Note that only the transmitter sides of D2D users are shown in the figure, as receivers do not cause any interference to the cellular users and therefore do not impact the channel allocation (see also the definition of the estimated network graph in Section 4.2.1). The (cellular and D2D) users' locations and also the channel gains are selected randomly. According to the system model (Section 4.1.1), there exist five orthogonal channels ($Q = 5$). Each D2D user $k \in \mathcal{K}$ selects a transmit power from the set of power-levels, $\mathcal{M} = \{2, 4\}$. Moreover, the transmit power of the BS to the cellular users is selected to be $P_c = 7$.

4.4.1. Channel Allocation

Table 4.1 includes $|h_{bl,q}|^2$ (cellular-BS average channel gains) for $l, q \in \{1, ..., 5\}$, which is assumed to be known by the BS, together with the network topology (Figure 4.1), according to Assumption (A5) (Section 4.1). Based on this information, and by using the graph-theoretical channel allocation scheme described in Section 4.2, the BS assigns each (cellular and D2D) user a channel, as summarized in Table 4.2. Based on Table 4.1 and Figure 4.1, it can be concluded that by the channel allocation given in Table 4.2, both (4.8) and (4.9) are satisfied.

As discussed in Section 4.2.3, it is also possible to change the criterion of channel allocation from maximizing the social welfare to address the QoS guarantee or fairness

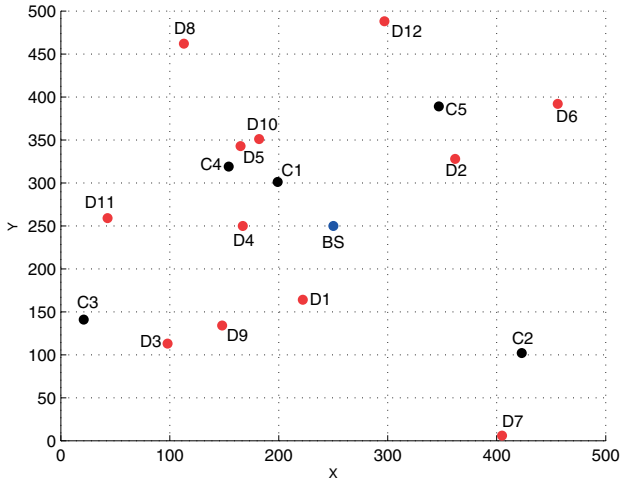


Figure 4.1.: Network model consisting of D2D transmitters (Di, $i \in \{1, ..., 12\}$) and cellular receivers (Ci, $i \in \{1, ..., 5\}$).

Table 4.1.: BS to Cellular Average Channel Gains

User \ Channel	1	2	3	4	5
C1	0.04	0.01	0.27	0.12	0.04
C2	0.29	0.06	0.15	0.18	0.26
C3	0.31	0.46	0.24	0.19	0.06
C4	0.12	0.06	0.29	0.34	0.16
C5	0.24	0.08	0.23	0.41	0.07

issues of cellular users. Assume that the required QoS of any cellular user $l \in \mathcal{L}$ is satisfied if it achieves some minimum utility, say $f_{\min}^{(l)} = 3.5$.[5] Given Table 4.1, the maximum tolerable interference of each cellular user can be simply calculated. Using the proposed channel allocation scheme, a channel allocation that guarantees the QoS satisfaction of all cellular users is summarized in Table 4.3. Moreover, the result of channel assignment based on fairness among cellular users is given in Table 4.4.[6]

The achieved average rewards of cellular users under all three criteria are shown in Figure 4.2. It can be seen that if the allocation criterion is to achieve the highest utility sum, then some cellular users might experience no interference, whereas some others might be strongly disturbed. In case of QoS guarantee, however, users with higher channel

[5]Note that the QoS requirements of cellular users are not necessarily similar.
[6]Note that the solutions are approximately-optimal and also might not be unique.

Table 4.2.: Channel Allocation, Maximum Aggregate Utility for Cellular Users

Channel	User
1	C5,D1,D3,D9
2	C3,D2,D6,D7,D12
3	C1
4	C4
5	C2,D4,D5,D8,D10,D11

Table 4.3.: Channel Allocation, QoS Guarantee for Cellular Users

Channel	User
1	C5,D3,D9
2	C3,D1,D2,D11,D12
3	C1,D8,D10
4	C4,D4
5	C2,D4

gains experience more interference and vice versa, so that at the end all cellular users are satisfied, upon feasibility. Moreover, by Table 4.3, in the current setting, all D2D users can be served without violating the QoS requirement of cellular users; however, it is not necessarily always the case. In the last criterion, all cellular users experience almost equal amounts of interference, regardless of their achieved utilities.

For our primary channel allocation criterion, i.e. maximizing the aggregate utility of cellular users, it is of interest to investigate the performance loss of cellular users, caused by sharing the spectrum with D2D users. This performance degradation is visualized in Figure 4.3, where the achievable utilities of cellular users without any interference (no channel sharing) are shown in comparison with the case where *all* D2D users are assigned some channel. It can be seen than by less than 15% performance loss, all D2D users can be served.

4.4.2. Power Control

From Table 4.2, it can be observed that the minimum-weighted partitioning divides the D2D and cellular users into five clusters, each allocated a frequency channel. In this section, we investigate the power control game of the first cluster, i.e. the cluster that includes three D2D users (D1, D3 and D9), and is assigned channel one. The games of

Table 4.4.: Channel Allocation, Fairness Among Cellular Users

Channel	User
1	C5,D3,D9,D11
2	C3,D2,D12
3	C1,D1,D8
4	C4,D6,D7
5	C2,D4,D5,D10

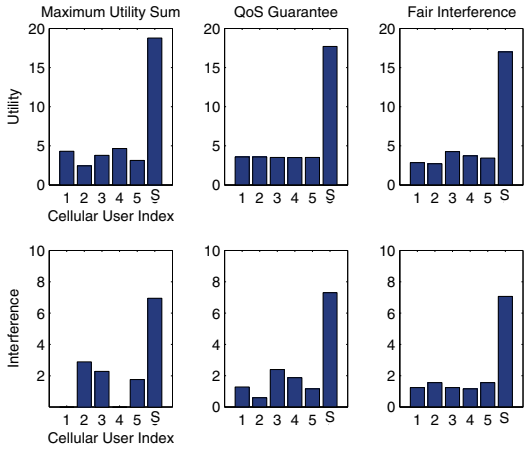


Figure 4.2.: Average utility and interference experienced by cellular users under three criteria (S:Sum).

other clusters are similar. The game horizon is $n = 2 \times 10^3$. Moreover, we assume $\alpha = 0.1$. The joint rewards of D2D users as a result of different joint action profiles are calculated by using (4.2), as given in Table 4.5. From this table, the action profile $(2, 2, 2)$, or in other words, (P_2, P_2, P_2), is the Nash equilibrium, which also maximizes the potential function. Hence the game converges theoretically to this point. Figure 4.4 describes the frequency in which any given action is played by each D2D user. It can be seen that the equilibrium point is played almost all the time. Figure 4.5 depicts the average utility of D2D users versus the equilibrium reward, confirming that in a short time, the average reward of every player converges to that of equilibrium. This result can be also concluded from Figure 4.4, which shows that the equilibrium is played almost all the time.

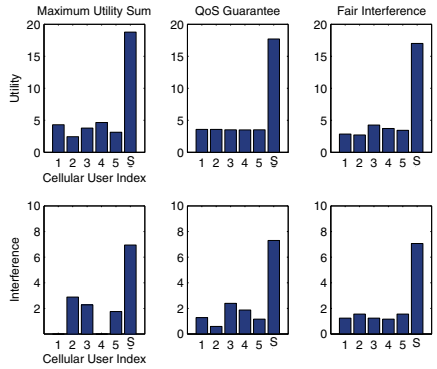


Figure 4.3.: Performance loss of cellular users due to sharing channels with D2D users, with the allocation criterion being the maximization of the cellular utility sum.

Table 4.5.: Joint Reward Table

Joint Action	Joint Reward	Joint Action	Joint Reward
$(1,1,1)$	$(1.30, 1.18, 1.05)$	$(2,2,1)$	$(1.40, 1.27, 0.15)$
$(1,2,1)$	$(0.90, 1.68, 0.65)$	$(1,2,2)$	$(0.61, 1.27, 1.14)$
$(2,1,1)$	$(1.79, 0.78, 0.64)$	$(2,1,2)$	$(1.40, 0.49, 1.14)$
$(1,1,2)$	$(0.90, 0.78, 1.54)$	$(2,2,2)$	$(1.10, 0.99, 0.85)$

4.4.3. Overall Performance

In order to evaluate the overall performance of the proposed hybrid resource allocation strategy (HRAS), we compare it with three other strategies that are described below.

- Centralized approach that is based on the exhaustive search, given global information. In accordance with the concept of underlay D2D networks, the priority is here given to the cellular users. Formally, the selected joint channel and power allocation vector maximizes $\sum_{l=1}^{L} f^{(l)}$, and ties are broken in favor of the allocation vector that yields a higher aggregate D2D utility, i.e. larger $\sum_{k=1}^{K} f^{(k)}$.

- Centralized approach that is based on the exhaustive search given global information, but *without* considering the priority for cellular users. Formally, the algorithm searches for the joint channel and power allocation vector that maximizes $\sum_{l=1}^{L} f^{(l)} + \sum_{k=1}^{K} f^{(k)}$.

- Random resource allocation, where the channel and power-levels are all assigned using uniform distribution.

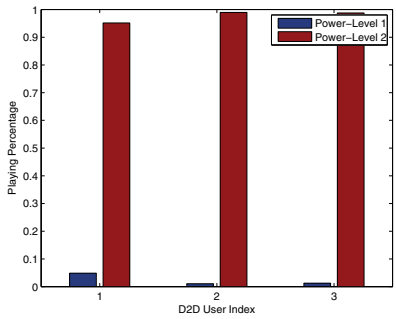


Figure 4.4.: Fraction of trials in which any given action is played by D2D users.

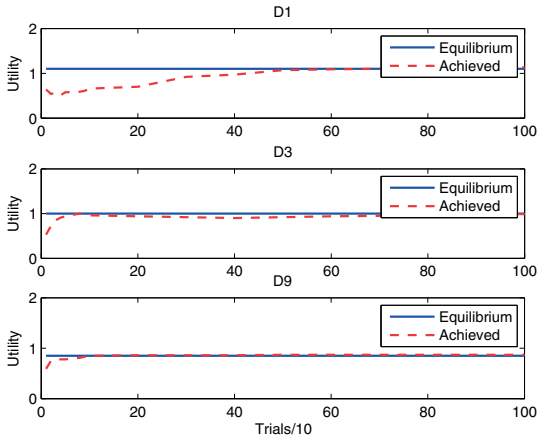


Figure 4.5.: The utilities achieved by D2D users versus the equilibrium reward.

As applying the exhaustive search to the large network investigated before (Figure 4.1) yields high computational and time complexity ($5^{16} \times 2^{12}$ cases should be searched), we turn to a smaller network with $L = Q = M = 2$ and $K = 6$. Ten experiments are performed while channel gains as well as users' locations are changed randomly. Results are depicted in Figure 4.6. From this figure, it can be concluded that the utility achieved by our proposed resource allocation scheme is almost equal to the highest possible aggregate network utility, when taking the priority of cellular users into account. It is clear that larger network utility sum can be achieved by neglecting the cellular priority; nevertheless, such setting does not comply with the concept of underlay D2D communications, since cellular users might be extremely disturbed. It is also worth mentioning that the number of possible channel and power allocation vectors grows exponentially in the number of users (D2D and cellular) and polynomially in the number of actions (channels and power-levels). As a result, for large networks, centralized resource allocation based on exhaustive

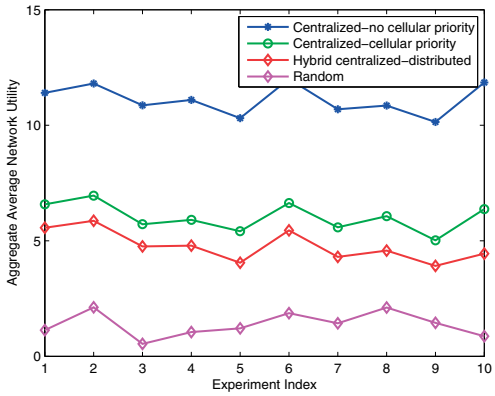


Figure 4.6.: Overall performance of the hybrid resource allocation strategy (HRAS) compared to some other approaches.

search yields excessive cost in terms of time and computational complexity, and hence cannot be practiced. Our approach, in contrast, offers polynomial complexity, and hence is specifically suitable for large networks.

4.5. Conclusion and Remarks

We studied an underlay D2D communications system, where device pairs are primarily allowed to transmit simultaneously with cellular users through a common channel. A two-stage hybrid centralized-distributed resource allocation strategy was proposed that takes the priority of cellular users into account, and relies on strictly limited information. In the first stage, centralized channel allocation is performed by using a graph-theoretical method. The method was shown to offer high flexibility in selecting the allocation criteria, for instance aggregate utility, fairness or QoS guarantee. It was also concluded that both time and computational complexities are polynomial in the number of users. In the second stage, power control problem is modeled as a game with incomplete information. We proved that the game is an exact potential game defined on a discrete strategy set, and therefore Q-learning fictitious play can be used by the players in order to achieve an efficient Nash equilibrium in a distributed manner. The set of Nash equilibria was shown to be equivalent to the set of potential maximizers. Extensive numerical analysis demonstrated the applicability of the proposed approach, specifically in the context of large-scale networks. Moreover, the results showed that the number of D2D users that can be served by cellular resources depends on the QoS requirement of cellular users. If no QoS requirement exists, serving all D2D users causes a degradation of the cellular

aggregate utility, which depends on the channel qualities as well as the number of D2D users. In addition, it was concluded that using Q-learning fictitious play strategy results in a fast convergence to the most efficient equilibrium point.

A. Some Auxiliary Definitions and Results

1. Game Theory

Throughout this part, we consider a game $\mathfrak{G}$ consisting of a set $\mathcal{K}$ of K players. The strategy set is denoted by $\mathcal{M}$ with a generic element $i^{(k)} = \left(i_1^{(k)}, ..., i_L^{(k)}\right)$, for player $k \in \mathcal{K}$, which implies that the action has L components. The set of joint strategy profiles of players is denoted by $\mathcal{I}$ with a generic element $\mathbf{i} = \left(i^{(1)}, ..., i^{(K)}\right) \in \bigotimes_{k=1}^{K} \{1, ..., M\}$. Accordingly, $\mathbf{i}^{(-k)} \in \bigotimes_{k=1}^{K-1} \{1, ..., M\}$ stands for the joint action profile of all players except for player k. Moreover, $f^{(k)}(\mathbf{i})$ is the mean reward function of player k.

Definition 13 (Pure Strategy Nash Equilibrium)**.** *A joint action* $\mathbf{i} = \left(i^{(1)}, ..., i^{(k)}, ..., i^{(K)}\right)$ *is called a pure strategy Nash equilibrium if for all* $k \in \mathcal{K}$ *and all joint action profiles* $\mathbf{i}' = \left(i^{(1)}, ..., i'^{(k)}, ..., i^{(K)}\right)$,

$$f^{(k)}(\mathbf{i}) \geq f^{(k)}\left(\mathbf{i}'\right). \tag{A.1}$$

Definition 14 (Mixed Strategy Nash Equilibrium)**.** *Let* $\mathbf{p}^{(k)} = \left(p_1^{(k)}, ..., p_M^{(k)}\right)$ *be a mixed strategy of player* k*. Moreover, let* $\pi = \mathbf{p}^{(1)} \times ... \times \mathbf{p}^{(k)} \times ... \times \mathbf{p}^{(K)}$ *denote the joint mixed strategy profile. The probability of each joint action profile* $\mathbf{i} = \left(i^{(1)}, ..., i^{(k)}, ..., i^{(K)}\right)$ *yields* $\prod_{k \in \mathcal{K}} p_{i^{(k)}}^{(k)}$*. For a player* $k \in \mathcal{K}$ *and for each* $\mathbf{i}$*, define* $\bar{f}^{(k)}(\pi) = \sum_{\mathbf{i} \in \mathcal{I}} \left(\prod_{j \in \mathcal{K}} p_{i^{(j)}}^{(j)}\right) f^{(k)}(\mathbf{i})$*. Then* π *is called a mixed strategy Nash equilibrium when for all* $k \in \mathcal{K}$ *and all mixed strategies* $\mathbf{p}'^{(k)}$*, if* $\pi' = \mathbf{p}^{(1)} \times ... \times \mathbf{p}'^{(k)} \times ... \times \mathbf{p}^{(K)}$*, then*

$$\bar{f}^{(k)}(\pi) \geq \bar{f}^{(k)}(\pi'). \tag{A.2}$$

In words, Nash equilibrium refers to a steady state in which no player can achieve higher reward by changing its strategy profile unilaterally.

Definition 15 (Correlated Equilibrium)**.** *Let* π *be a probability distribution over the set* $\bigotimes_{k=1}^{K} \{1, ..., M\}$ *of all possible* K*-tuples of actions. Also, let* $\pi(\mathbf{i}) = \pi\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ *denote the probability vector of the joint action profile* $\mathbf{i} = \left(i^{(k)}, \mathbf{i}^{(-k)}\right)$*. Then* π *is called a correlated equilibrium if for all* $k \in \mathcal{K}$ *and all* $\mathbf{i}' = \left(i'^{(k)}, \mathbf{i}^{(-k)}\right)$,

$$\sum_{\mathbf{i}^{(-k)}} f^{(k)}\left(i^{(k)}, \mathbf{i}^{(-k)}\right) \pi\left(i^{(k)}, \mathbf{i}^{(-k)}\right) \geq \sum_{\mathbf{i}^{(-k)}} f^{(k)}\left(i'^{(k)}, \mathbf{i}^{(-k)}\right) \pi\left(i^{(k)}, \mathbf{i}^{(-k)}\right). \tag{A.3}$$

Correlated equilibrium can be interpreted as if all players are provided with a private instruction from a trusted third-party, and $K-1$ players follow this instruction, then no player k can improve its expected reward by deviating from the recommendation of the third-party.

Definition 16 (Smooth Game)**.** *A game $\mathfrak{G}$ is smooth if, for each $k \in \mathcal{K}$, $f^{(k)}(\mathbf{i})$ has continuous partial derivatives with respect to the components of $i^{(k)}$.*

Definition 17 (Strictly Monotone Payoff Gradient)**.** *Let $\nabla f^{(k)} = \left(\frac{\partial f^{(k)}}{\partial i_1^{(k)}}, \cdots, \frac{\partial f^{(k)}}{\partial i_L^{(k)}} \right)$, and call $\left(\nabla f^{(k)} \right)_{k \in \mathcal{K}}$ the payoff gradient of a smooth game $\mathfrak{G}$. We say that the payoff gradient is strictly monotone if*

$$\sum_{k=1}^{K} \left(\nabla f^{(k)}(\mathbf{i}) - \nabla f^{(k)}(\mathbf{j}) \right)^T \left(i^{(k)} - j^{(k)} \right) < 0 \tag{A.4}$$

holds $\forall\ \mathbf{i}, \mathbf{j} \in \mathcal{I}$ with $\mathbf{i} \neq \mathbf{j}$.

Theorem 10 ([Ui08a])**.** *Consider a smooth game $\mathfrak{G}$ with compact strategy sets. If the payoff gradient of $\mathfrak{G}$ is strictly monotone then it has a unique correlated equilibrium, which places probability one on a unique pure strategy Nash equilibrium.*

Definition 18 (Exact Potential Game)**.** *A game $\mathfrak{G}$ is (exact) potential if there exists a function $v : \mathcal{I} \to \mathbb{R}$ such that*

$$f^{(k)}\left(i^{(k)}, \mathbf{i}^{(-k)} \right) - f^{(k)}\left(j^{(k)}, \mathbf{i}^{(-k)} \right) = v\left(i^{(k)}, \mathbf{i}^{(-k)} \right) - v\left(j^{(k)}, \mathbf{i}^{(-k)} \right), \tag{A.5}$$

$\forall\ i^{(k)}, j^{(k)} \in \mathcal{M}$ and $k \in \mathcal{K}$. Then v is called a potential of the game $\mathfrak{G}$.

Now consider some set $\mathcal{Q}$ with cardinality Q. In the following, v stands for a function defined on a discrete set $\mathcal{X} \subseteq \mathbb{Z}^Q$, where $\mathcal{X} = \prod_{q=1}^{Q} \mathbf{x}_q$ and $\mathbf{x}_q = \left\{ x_q \in \mathbb{Z}, \underline{x}_q \leq x_q \leq \overline{x}_q \right\}$. Moreover $\|\mathbf{x}\| = \sum_q x_q$ denotes the l_1-norm of a vector $\mathbf{x} \subseteq \mathbb{Z}^Q$.

Definition 19 (Larger Midpoint Property (LMP))**.** *We say that a function $v : \mathcal{X} \to \mathbb{R}$ satisfies the larger midpoint property if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ with $\|\mathbf{x} - \mathbf{y}\| = 2$,*

$$\underset{\mathbf{z} \in \mathcal{X}: \|\mathbf{x}-\mathbf{z}\| = \|\mathbf{y}-\mathbf{z}\| = 1}{\text{maximum}} v(\mathbf{z}) \geq t v(\mathbf{x}) + (1-t) v(\mathbf{x}) \ \ (\exists\ t \in (0,1)), \tag{A.6}$$

or

$$\underset{\mathbf{z} \in \mathcal{X}: \|\mathbf{x}-\mathbf{z}\| = \|\mathbf{y}-\mathbf{z}\| = 1}{\text{maximum}} v(\mathbf{z}) \begin{cases} > \min\{v(\mathbf{x}), v(\mathbf{y})\} & \text{if } v(\mathbf{x}) \neq v(\mathbf{y}) \\ \geq v(\mathbf{x}) = v(\mathbf{y}) & \text{o.w.} \end{cases}. \tag{A.7}$$

Definition 20 (Separable Concave Function)**.** *A function $v : \mathcal{X} \to \mathbb{R}$ is said to be separable concave if it can be written as $v(\mathbf{x}) = \sum_{q\in\mathcal{Q}} v_q(x_q)$, where $v_q(x_q) \geq \frac{v_q(x_q-1)+v_q(x_q+1)}{2}$ for all $x_q \neq \underline{x}_q, \overline{x}_q$.*

Lemma 4 ([Ui08b])**.** *If $v : \mathcal{X} \to \mathbb{R}$ is a separable concave function, then (A.6) holds, and therefore v satisfies the larger midpoint property.*

Proposition 8 ([Ui08b])**.** *Let $\mathfrak{G}$ be an exact potential game with a potential function v that satisfies the LMP property. Then $\mathbf{i} \in \mathcal{I}$ maximizes v if and only if it is a Nash equilibrium.*

2. Multi-Armed Bandits

Lemma 5. *Let R_n and R_{Ext} be given by (2.1) and (2.2), respectively. Then, for any $\delta \in (0, \frac{1}{2}]$, we have*[1]

$$\Pr\left(|R_n - R_{\text{Ext}}| \leq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta, \tag{A.8}$$

from which it follows that if $R_n \in o(n)$, then we have $R_{\text{Ext}} \in o(n)$, with arbitrarily high probability.[2]

Proof. By comparing (2.1) and (2.2), it suffices to show that

$$\Pr\left(\left|\sum_{t=1}^{n} g_t(i_t) - \sum_{t=1}^{n} \bar{g}_t(\mathbf{p}_t)\right| \leq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \tag{A.9}$$

To this end, define $S \coloneqq \sum_{t=1}^{n} g_t(i_t)$, where $g_t(i_t) \in [0,1]$, $1 \leq t \leq n$, are independent random variables (see also Section 2.1.1). Further note that $\bar{S} = \mathrm{E}\{S\} = \sum_{t=1}^{n} \bar{g}_t(\mathbf{p}_t)$. Therefore, by Hoeffding's inequality [CBL06],

$$\begin{aligned}\Pr\left(|R_n - R_{\text{Ext}}| \geq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) =& \Pr\left(\left|S - \bar{S}\right| \geq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) \\ \leq& 2\exp\left(-\frac{2\frac{n}{2}\log\left(\frac{1}{\delta}\right)}{n}\right) = 2\delta.\end{aligned} \tag{A.10}$$

[1] Throughout this section and in order to simplify the notation, the player index (k) is omitted unless ambiguity arises.

[2] Here and hereafter, the statement "$X(n) \in o(n)$ with arbitrarily high probability" for some nonnegative random sequence $X(n) \in \mathbb{R}$ means that the probability of $X(n) \notin o(n)$ can be made arbitrarily small, provided that some parameter is chosen sufficiently small.

Hence with $\Pr\left(|R_n - R_{\text{Ext}}| \leq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) = 1 - \Pr\left(|R_n - R_{\text{Ext}}| \geq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right)$ the Lemma follows. □

Lemma 6. *Let R_{Ext} be given by (2.2). Define $\tilde{R}_n = \max_{i=1,\ldots,M} \sum_{t=1}^n g_t(i) - \sum_{t=1}^n \bar{\tilde{g}}_t(\mathbf{p}_t)$, where $\bar{\tilde{g}}_t(\mathbf{p}_t) = \sum_{i=1}^M p_{i,t}\tilde{g}_t(i)$ and $\tilde{g}_t(i)$ is given by (2.11). Then we have*

$$\Pr\left(\left|\tilde{R}_n - R_{\text{Ext}}\right| \leq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \tag{A.11}$$

Hence, for sufficiently small $\delta > 0$, $R_{\text{Ext}} \in o(n)$ implies that $\tilde{R}_n \in o(n)$, with arbitrarily high probability.

Proof. Similar to the proof of Lemma 5, it follows from (2.2) and the definition of $\tilde{R}_n$ that it is sufficient to show that for $\delta \in (0, \frac{1}{2}]$,

$$\Pr\left(\left|\sum_{t=1}^n \bar{\tilde{g}}_t(\mathbf{p}_t) - \sum_{t=1}^n \bar{g}_t(\mathbf{p}_t)\right| \leq \sqrt{\frac{n}{2}\log\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \tag{A.12}$$

To this end, note that $\bar{\tilde{g}}_t(\mathbf{p}_t) \in [0,1]$, $1 \leq t \leq n$, are independent random variables. Moreover, since $\tilde{g}_t(i)$ is an unbiased estimate of $g_t(i)$, we have $\mathrm{E}\left\{\sum_{t=1}^n \bar{\tilde{g}}_t(\mathbf{p}_t)\right\} = \sum_{t=1}^n \bar{g}_t(\mathbf{p}_t)$. Hence, defining $S \coloneqq \bar{\tilde{g}}_t(\mathbf{p}_t) - \bar{g}_t(\mathbf{p}_t)$ and proceeding as in the proof of Lemma 5 with the Hoeffding's inequality in hand proves the lemma. □

Proposition 9. *Let R_n be given by (2.1) and $\tilde{R}_n$ be defined as in Lemma 6. Then $R_n \in o(n)$ implies that $\tilde{R}_n \in o(n)$.*

Proof. Lemma 5 implies that $R_n \in o(n) \Rightarrow R_{\text{Ext}} \in o(n)$ with arbitrarily high probability, while by Lemma 6, we have $R_{\text{Ext}} \in o(n) \Rightarrow \tilde{R} \in o(n)$. Therefore, if $R_n \in o(n)$, then $\tilde{R}_n \in o(n)$ with arbitrarily high probability. □

Theorem 11 ([CBL06])**.** *Let $\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^M \phi(u_i)\right)$. Consider a selection strategy that at time t selects action i_t according to distribution $\mathbf{p}_t$, whose elements $p_{i,t}$ are defined as*

$$p_{i,t} = (1 - \gamma_t)\frac{\phi'\left(R_{i,t-1}\right)}{\sum_{j=1}^M \phi'\left(R_{j,t-1}\right)} + \frac{\gamma_t}{M}, \tag{A.13}$$

where $R_{i,t-1} = \sum_{s=1}^{t-1}\left(g_s(i) - g_s(i_s)\right)$. Assume the followings:
A1. $\sum_{t=1}^n \frac{1}{\gamma_t^2} = o\left(\frac{n^2}{\log(n)}\right)$.

A2. For all vectors $\mathbf{v}_t = (v_{1,t}, ..., v_{n,t})$ *with* $|v_{i,t}| \leq \frac{M}{\gamma_t}$*, we have*

$$\lim_{n\to\infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^{n} C(\mathbf{v}_t) = 0, \tag{A.14}$$

where $C(\mathbf{v}_t) = \sup_{\mathbf{u}\in\mathbb{R}^M} \psi' \left(\sum_{i=1}^{M} \phi(u_i)\right) \sum_{i=1}^{M} \phi''(u_i) v_{i,t}^2$.
A3. For all vectors $\mathbf{u}_t = (u_{1,t}, ..., u_{n,t})$*, with* $u_{i,t} \leq t$*,*

$$\lim_{n\to\infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^{n} \gamma_t \sum_{i=1}^{M} \nabla_i \Phi(\mathbf{u}_t) = 0. \tag{A.15}$$

A4. For all vectors $\mathbf{u}_t = (u_{1,t}, ..., u_{n,t})$*, with* $u_{i,t} \leq t$*,*

$$\lim_{n\to\infty} \frac{\log(n)}{\psi(\phi(n))} \sqrt{\sum_{t=1}^{n} \frac{1}{\gamma_t^2} \left(\sum_{i=1}^{M} \nabla_i \Phi(\mathbf{u}_t)\right)^2} = 0. \tag{A.16}$$

Then the selection strategy satisfies

$$\lim_{n\to\infty} \frac{1}{n} \left(\max_{i=1,...,M} \sum_{t=1}^{n} g_t(i) - \sum_{t=1}^{n} g_t(i_t) \right) = 0, \tag{A.17}$$

or equivalently, $R_n \in o(n)$*, where* R_n *is given by (2.1).*

B. Additional Proofs of Chapter 2

1. Proof of Proposition 1

In order to prove Proposition 1, we use Theorem 10. As the strategy set is compact, in order to use this theorem, we show that i) the game is smooth, and ii) the payoff gradient is strictly monotone.

According to our system model, by changing the channel index, $i'^{(k)}$, the channel gain and interference change. Therefore we define $i_1^{(k)} := \frac{|h_{kk',i'^{(k)}}|^2}{\sum_{q\in\mathcal{Q}^{(k)}} i''^{(q)}|h_{qk',i'^{(k)}}|^2+N_0}$ and $i_2^{(k)} := i''^{(k)}$, from which we have $f^{(k)}(\mathbf{i}) = \log_2\left(i_1^{(k)} i_2^{(k)}\right) - \alpha i_2^{(k)}$. Thus

$$\frac{\partial f^{(k)}}{\partial i_1^{(k)}} = \frac{1}{\mu i_1^{(k)}}, \tag{B.1}$$

and

$$\frac{\partial f^{(k)}}{\partial i_2^{(k)}} = \frac{1}{\mu i_2^{(k)}} - \alpha, \tag{B.2}$$

with $\mu = \log(2)$. Hence by Definition 16, the game is smooth. On the other hand,

$$\begin{aligned}
&\left(\nabla f^{(k)}(\mathbf{i}) - \nabla f^{(k)}(\mathbf{j})\right)^T \left(i^{(k)} - j^{(k)}\right) \\
&\quad= \frac{1}{\mu}\begin{bmatrix} \frac{1}{i_1^{(k)}} - \frac{1}{j_1^{(k)}} & \frac{1}{i_2^{(k)}} - \frac{1}{j_2^{(k)}} \end{bmatrix} \begin{bmatrix} i_1^{(k)} - j_1^{(k)} \\ i_2^{(k)} - j_2^{(k)} \end{bmatrix} \\
&\quad= \frac{1}{\mu}\left(\frac{1}{i_1^{(k)}} - \frac{1}{j_1^{(k)}}\right)\left(i_1^{(k)} - j_1^{(k)}\right) + \frac{1}{\mu}\left(\frac{1}{i_2^{(k)}} - \frac{1}{j_2^{(k)}}\right)\left(i_2^{(k)} - j_2^{(k)}\right),
\end{aligned} \tag{B.3}$$

which is always negative as for any $x, y > 0$ with $x \neq y$, $x - y > 0$ yields $\frac{1}{x} - \frac{1}{y} < 0$ and vice versa. Thus

$$\sum_{k=1}^{K} \nabla f^{(k)} < 0, \tag{B.4}$$

i.e. the payoff gradient is strictly monotone by Definition 17.

As a result, by Theorem 10, the game has a unique correlated equilibrium that places probability one on the unique Nash equilibrium.

2. Proof of Proposition 2

Lemma 7. *Consider a bandit exponential-based weighted average strategy (BEWAS) that uses* $\mathbf{p}_t = (p_{1,t}, ..., p_{M,t})$ *to select an action among* M *possible choices, where* $p_{i,t}$ *is calculated as*[1]

$$p_{i,t} = (1-\gamma)\frac{\exp\left(\eta\tilde{R}_{i,t-1}\right)}{\sum_{j=1,...,M}\exp\left(\eta\tilde{R}_{j,t-1}\right)} + \frac{\gamma}{M}, \tag{B.5}$$

and $\tilde{R}_{i,t-1}$ *denotes the estimated accumulated regret of not playing action* i.[2] *Then selecting* γ *and* η *as given by Proposition 2 yields* $R_{\text{Ext}} \in O\left((nM)^{\frac{2}{3}}(\log(M))^{\frac{1}{3}}\right)$.

Proof. The proof is a direct corollary of Theorem 6.6 of [CBL06]. □

Given Lemma 7, we follow the approach of [SL05] for the rest of the proof.

Recall that by Section 2.1.3 and Algorithm 1, the mixed strategy of each player is defined as

$$\mathbf{p}_t = \sum_{(i\to j):i\neq j} \mathbf{p}_{(i\to j),t}\delta_{(i\to j),t}. \tag{B.6}$$

Hence,

$$\bar{g}_t(\mathbf{p}_t) = \sum_{(i\to j):i\neq j} \bar{g}_t(\mathbf{p}_{(i\to j),t})\delta_{(i\to j),t}. \tag{B.7}$$

Lemma 7 specifies the growth rate of the external regret of BEWAS. On the other hand, as described in Section 2.1.3, NR-BEWAS applies the BEWAS algorithm for $M(M-1) \leq M^2$ actions. Therefore, (B.7) together with Lemma 7 yields

$$\max\sum_{t=1}^{n}\bar{g}_t(\mathbf{p}_{(i\to j),t}) - \sum_{t=1}^{n}\bar{g}_t(\mathbf{p}_t) \in O\left(\left(M^2n\right)^{\frac{2}{3}}(2\log(M))^{\frac{1}{3}}\right), \tag{B.8}$$

and by the definition of internal regret it holds $\max_{i\neq j} R_{(i\to j),n} \in O\left(\left(M^2n\right)^{\frac{2}{3}}(2\log(M))^{\frac{1}{3}}\right)$, which concludes the proof. Details can be found in [SL05], and hence are omitted.

[1] Throughout this section and in order to simplify the notation, the player index (k) is omitted unless ambiguity arises.

[2] This definition should not be mistaken for the general regret defined in Section 2.1.1.

3. Proof of Proposition 3

Lemma 8. *Consider a BEWAS that uses* $\mathbf{p}_t = (p_{1,t}, ..., p_{M,t})$ *to select an action among* M *possible choices, where* $p_{i,t}$ *is calculated as*

$$p_{i,t} = (1-\gamma_t)\frac{\exp\left(\eta_t \tilde{R}_{i,t-1}\right)}{\sum_{j=1,...,M}\exp\left(\eta_t \tilde{R}_{j,t-1}\right)} + \frac{\gamma_t}{M}, \tag{B.9}$$

and $\tilde{R}_{i,t-1}$ *denotes the estimated accumulated regret of not playing action* i*. Then, for* γ_t *and* η_t *as given by Proposition 3, this strategy yields vanishing external regret, i.e.* $R_{\text{Ext}} \in o(n)$*.*

Proof. By Proposition 9, if (A.17) is satisfied for a selection strategy (that is, if $R_n \in o(n)$), then the growth rate of the external regret caused by the bandit version of that strategy (which uses the estimated rewards instead of true ones) grows sublinearly in n, i.e. $\tilde{R}_n \in o(n)$. Therefore, in order to prove the proposition, we show that our selected parameters $\gamma_t = t^{-\frac{1}{3}}$ and $\eta_t = \frac{\gamma_t^3}{M^2}$ satisfy axioms A1-A4 of Theorem 11.[3] Note that in our strategy we have $\Phi(\mathbf{u}) = \frac{1}{\eta_t}\log\left(\sum_{i=1}^{M}\exp(\eta_t u_i)\right)$.

A1. For $\gamma_t = t^{-\frac{1}{3}}$, we have

$$\sum_{t=1}^{n}\frac{1}{\gamma_t^2} = \sum_{t=1}^{n} t^{\frac{2}{3}} = \text{Harmonic Number}[n, -\frac{2}{3}] \coloneqq \mathrm{H}_n[\frac{-2}{3}]. \tag{B.10}$$

Then,

$$\lim_{n\to\infty}\frac{\log(n)}{n^2}\sum_{t=1}^{n}\gamma_t^2 = \lim_{n\to\infty}\frac{\log(n)}{n^2}\mathrm{H}_n[\frac{-2}{3}] = 0. \tag{B.11}$$

A2. For $\psi(x) = \frac{1}{\eta_t}\log(x)$ and $\phi(x) = \exp(\eta_t x)$, we obtain

$$C(\mathbf{v}_t) = \sup\left(\eta_t\sum_{i=1}^{M} v_{i,t}^2\right) = \frac{\eta_t M^3}{\gamma_t^2}. \tag{B.12}$$

Thus,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{\psi(\phi(n))}\sum_{t=1}^{n}C(\mathbf{v}_t) &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}t^{\frac{-1}{3}} \\ &= \lim_{n\to\infty}\frac{1}{n}\mathrm{H}_n[\frac{1}{3}] = 0.\end{aligned} \tag{B.13}$$

[3] Simple calculus steps are omitted.

A3. For $\Phi(\mathbf{u}) = \frac{1}{\eta_t} \log\left(\sum_{i=1}^{M} \exp(\eta_t u_i)\right)$, $\nabla_i \Phi(\mathbf{u}_t)$ yields

$$\nabla_i \Phi(\mathbf{u}_t) = \frac{\exp(\eta_t u_i)}{\sum_{i=1}^{M} \exp(\eta_t u_i)}. \tag{B.14}$$

Therefore,

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^{n} \gamma_t \sum_{i=1}^{M} \nabla_i \Phi(\mathbf{u}_t) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} t^{\frac{-1}{3}} \sum_{i=1}^{M} \frac{\exp(\eta_t u_i)}{\sum_{i=1}^{M} \exp(\eta_t u_i)} \\ &= \lim_{n\to\infty} \frac{1}{n} \, \mathrm{H}_n[\frac{1}{3}] = 0. \end{aligned} \tag{B.15}$$

A4. *A*4 follows simply by substituting (B.14) in (A.16).

Hence, all axioms A1-A4 are satisfied, and therefore (A.17) holds, which, together with Proposition 9, completes the proof. □

By Lemma 8, the external regret of the BEWAS described before grows sublinearly in n. Therefore, similar to the proof of Proposition 2, (B.7) yields

$$\max \sum_{t=1}^{n} \bar{g}_t(\mathbf{p}_{(i\to j),t}) - \sum_{t=1}^{n} \bar{g}_t(\mathbf{p}_t) \in o(n), \tag{B.16}$$

and the definition of internal regret ensures that $\max_{i\neq j} R_{(i\to j),n} \in o(n)$, which concludes the proof.

C. Additional Proofs of Chapter 3

1. Proof of Lemma 2

We follow a root suggested in [CLRJ13]. Suppose that $\lim_{n\to\infty} S_{\kappa,n} = 1$ holds. In order to prove $\lim_{n\to\infty} S_{\kappa',n} = 1$, it is sufficient to show that after some finite time, the actions of the player that are selected based on $\mathbf{p}_t$ are equal to those based on δ_{d_t}. By (3.6), we know that with probability 1, there exists an $\nu > 0$ so that after a time point $\theta < \infty$, $|\mathbf{p}_t - \delta_{d_t}| < \nu$ holds for all $t > \theta$ (see also Theorem 3). At the same time, according to our system model and by Assumption (A2), the reward functions are bounded, and the action space and memory are finite. This implies that if $|\mathbf{p}_t - \delta_{d_t}| < \nu$ holds, then the actions of the player evolve as if it were aware of the true joint action profile of its opponents. Hence the lemma follows.

2. Proof of Lemma 3

From Algorithm 5, at each period j, $\left\lceil T'_j Z_j \right\rceil$ trials are selected for exploration by each player. At each one of these trials, with probability $1-\gamma$, an arm i is selected equally at random. Since these processes are independent, the probability that arm i is pulled at some exploration trial yields $\frac{1-\gamma}{M}$. Now, let $\left\{\mathtt{W}_t^{(i)}\right\}_{t=1}^{n}$ be a sequence of random variables, where $\mathtt{W}_t^{(i)} = 1$ if arm i is played at time t, and $\mathtt{W}_t^{(i)} = 0$ otherwise, and the outcomes $\mathtt{W}_t^{(i)}$ are independent over time. In the worst-case, arm i never becomes the best response, and hence its chance of being played is limited to the exploration trials. As a result, $\Pr\left(\mathtt{W}_t^{(i)} = 1\right) = \frac{1-\gamma}{M}$, and the sum of probabilities for the event $\mathtt{W}_t^{(i)} = 1$ yields $\sum_{t=1}^{n} \Pr\left(\mathtt{W}_t = 1\right) = \frac{1-\gamma}{M} \sum_{j=1}^{J} \left\lceil T'_j \cdot Z_j \right\rceil$. By using Assumption (A4) and Lemma 1, we conclude that $\lim_{J\to\infty} \sum_{j=1}^{J} \Pr\left(\mathtt{W}_t = 1\right) \to \infty$. Thus, by the second Borel-Cantelli lemma ([CLRJ13], [Fel68]), it follows that the probability of arm i being pulled infinitely often equals 1. On the other hand, players select their actions independently. Therefore, the probability of playing each joint action profile is $\frac{1-\gamma}{M^K}$. By the same argument, each joint action profile is also played infinitely often. Hence, the Lemma is proved.

3. Proof of Theorem 4

Since the proof is identical for all players, we omit the player index, k, in order to simplify the notation. For example, the selected action and the joint action profile of opponents are shown by i and $\mathbf{i}^-$, respectively. The proof is inspired by [YZ02], where the authors showed the consistency of an allocation rule for single-player contextual bandit games, which, similar to our algorithm, is based on the GLIE concept. For brevity, we refer to the calibrated bandit strategy (CBS) as strategy χ.

In the following, we consider a selection strategy χ', which is identical to χ, except that at each time t and before taking any action, the player is informed about the *true* joint action profile of other $K-1$ players, that is $\mathbf{i}^-$. We prove that χ' is strongly consistent. Therefore, from Lemma 2, it follows that χ is strongly consistent as well.

Let i_t denote the selected arm at time t, while $\hat{i}_t$ stands for the arm with the highest *estimated* mean reward at time t. That is, at time t we have $\hat{f}_{\hat{i}_t}(\mathbf{i}_t^-) \coloneqq \max_{i\in\{1,\dots,M\}} \hat{f}_i(\mathbf{i}_t^-)$. Moreover, i_t^* denotes the arm with the highest *true* mean reward at time t so that $f_{i_t^*}(\mathbf{i}_t^-) \coloneqq \max_{i\in\{1,\dots,M\}} f_i(\mathbf{i}_t^-)$. Ties are broken using some deterministic rule.

From Definition 4, $S_{\chi',n}$ is upper-bounded by 1. Therefore, it is sufficient to prove a lower-bound on $S_{\chi',n}$ that converges to 1 as $n\to\infty$. To this end, we rewrite $S_{\chi',n}$ as [YZ02]

$$S_{\chi',n} = \frac{\sum_{t=1}^n f_{\hat{i}_t}\left(\mathbf{i}_t^-\right)}{\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)} + \frac{\sum_{t=1}^n \left(f_{i_t}\left(\mathbf{i}_t^-\right) - f_{\hat{i}_t}\left(\mathbf{i}_t^-\right)\right)}{\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)} \leq 1. \tag{C.1}$$

By Assumption (A2), it follows from (C.1) that

$$S_{\chi',n} \geq \frac{\sum_{t=1}^n f_{\hat{i}_t}\left(\mathbf{i}_t^-\right)}{\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)} - \frac{\frac{1}{n}\sum_{t=1}^n B\mathbf{1}_{\left\{i_t\neq\hat{i}_t\right\}}}{\frac{1}{n}\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)}. \tag{C.2}$$

The remainder of the proof consists of two parts. In the first part we show that

$$\frac{\frac{1}{n}\sum_{t=1}^n B\mathbf{1}_{\left\{i_t\neq\hat{i}_t\right\}}}{\frac{1}{n}\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)} \xrightarrow{\text{a.s.}} 0, \text{ as } n\to\infty, \tag{C.3}$$

where *a.s.* stands for almost surely. In the second part we establish that

$$\frac{\sum_{t=1}^n f_{\hat{i}_t}\left(\mathbf{i}_t^-\right)}{\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)} \xrightarrow{\text{a.s.}} 1, \text{ as } n\to\infty. \tag{C.4}$$

Combining (C.3) and (C.4) with (C.2) and $S_{\chi',n}\leq 1$ proves the strong consistency.
(i) By Assumption (A2), $\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)$ is positive. As a result, $\frac{1}{n}\sum_{t=1}^n f_{i_t^*}\left(\mathbf{i}_t^-\right)$ converges

to $\mathrm{E}\left\{f_{i_t^*}\left(\mathbf{i}_t^-\right)\right\} > 0$ almost surely. Hence, it suffices to show that $\frac{1}{n}\sum_{t=1}^{n} B\mathbf{1}_{\left\{i_t \neq \hat{i}_t\right\}} \to 0$, almost surely. To show this, we consider the worst-case; that is, we assume that in all exploration trials, inferior arms are selected (i.e. the best response is never selected by chance). Therefore

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} B\mathbf{1}_{\left\{i_t \neq \hat{i}_t\right\}} = \lim_{J\to\infty}\frac{\sum_{j=1}^{J}\left\lceil T_j' Z_j\right\rceil}{\sum_{j=1}^{J} T_j'} = 0, \tag{C.5}$$

where the second equality follows from Assumption (A4) and Lemma 1. This proves (C.3).
(ii) First, we note that (C.4) is equivalent to [YZ02][1]

$$\frac{\sum_{t=1}^{n}\left(f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right)\right)}{\sum_{t=1}^{n} f_{i_t^*}\left(\mathbf{i}_t^-\right)} \overset{\text{a.s.}}{\to} 0, \text{ as } n \to \infty. \tag{C.6}$$

Moreover, by (C.1), (C.2) and (C.3), we conclude that

$$\frac{\sum_{t=1}^{n}\left(f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right)\right)}{\sum_{t=1}^{n} f_{i_t^*}\left(\mathbf{i}_t^-\right)} \leq 0. \tag{C.7}$$

Clearly,

$$\begin{aligned} f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right) = {} & f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - \hat{f}_{\hat{i}_t,t-1}\left(\mathbf{i}_t^-\right) + \hat{f}_{\hat{i}_t,t-1}\left(\mathbf{i}_t^-\right) \\ & - \hat{f}_{i_t^*,t-1}\left(\mathbf{i}_t^-\right) + \hat{f}_{i_t^*,t-1}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right). \end{aligned} \tag{C.8}$$

On the other hand, for every trial t, $\hat{f}_{\hat{i}_t,t-1}\left(\mathbf{i}_t^-\right) \geq \hat{f}_{i_t^*,t-1}\left(\mathbf{i}_t^-\right)$ holds. Hence we can write [YZ02]

$$\begin{aligned} & f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right) \\ & \geq f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - \hat{f}_{\hat{i}_t,t-1}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right) + \hat{f}_{i_t^*,t-1}\left(\mathbf{i}_t^-\right) \\ & \geq -2 \sup_{i\in\{1,\ldots,M\}} \left\| \hat{f}_{i,t-1}\left(\mathbf{i}_t^-\right) - f_i\left(\mathbf{i}_t^-\right)\right\|_\infty. \end{aligned} \tag{C.9}$$

This yields

$$\begin{aligned} & \frac{\sum_{t=1}^{n}\left(f_{\hat{i}_t}\left(\mathbf{i}_t^-\right) - f_{i_t^*}\left(\mathbf{i}_t^-\right)\right)}{\sum_{t=1}^{n} f_{i_t^*}\left(\mathbf{i}_t^-\right)} \\ & \geq \frac{\frac{-2}{n}\sum_{t=1}^{n}\sup_{1\leq i\leq M}\left\|\hat{f}_{i,t-1}\left(\mathbf{i}_t^-\right) - f_i\left(\mathbf{i}_t^-\right)\right\|_\infty}{\frac{1}{n}\sum_{t=1}^{n} f_{i_t^*}\left(\mathbf{i}_t^-\right)}. \end{aligned} \tag{C.10}$$

For brevity, let us rewrite (C.10) in a shorter form as $a \geq b$. By Assumption (A3),

[1] This part of the proof is almost identical to [YZ02]; the difference is that here we use the fact that each action and also each joint action profile is played infinitely often (Lemma 3) in order to complete the proof.

$\|\hat{f}_{i,n}(\mathbf{i}^-) - f_i(\mathbf{i}^-)\|_\infty \to 0$ as $n \to \infty$. However, in order to use this assumption, we need to ensure that not only each arm, but also each joint action profile is played infinitely many times, as $n \to \infty$. This is established in Lemma 3. Therefore, the right-hand side of (C.10) converges to zero, i.e. $b \to 0$ and hence $a \geq 0$. On the other hand, by (C.7), the left-hand side is upper-bounded by zero, that is $a \leq 0$. As a result, (C.6) follows, which completes the second part of the proof.

4. Proof of Theorem 5

Consider a K-player MAB game, as described in Section 3.1. By Theorem 1, if each player plays by best responding to a calibrated forecast of the joint action profile of opponents, then

$$\inf_{\pi \in \mathcal{C}} \sum_{\mathbf{i}} |\hat{\pi}_n(\mathbf{i}) - \pi(\mathbf{i})| \to 0, \tag{C.11}$$

as $n \to \infty$. We refer to this selection strategy as χ'. In order to prove the Theorem, we show that CBS (here referred to as χ), in which the true mean rewards of joint action profiles are not known and are gradually learned by exploration, exhibits the same convergence characteristics as χ'.

First, we rearrange the K-player MAB game to a two-agent game where the first agent is any player k and the second agent is the set of its opponents, i.e. the set of $K-1$ players. For this game, any joint action profile of the two agents can be written as $\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$, where $i \in \{1, ..., M\}$ and $\mathbf{i}^{(-k)} \in \bigotimes_{k=1}^{K-1} \{1, ..., M\}$. Let $\hat{\pi}_n\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ denote the fraction of time until n in which some joint action $\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ is played. According to selection strategy χ, $\hat{\pi}_n\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ can be written as

$$\hat{\pi}_n\left(i^{(k)}, \mathbf{i}^{(-k)}\right) = \hat{\pi}_{n,m}\left(i^{(k)}, \mathbf{i}^{(-k)}\right) + \hat{\pi}_{n,l}\left(i^{(k)}, \mathbf{i}^{(-k)}\right), \tag{C.12}$$

where $\hat{\pi}_{n,m}\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ and $\hat{\pi}_{n,l}\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ denote the fractions of time in which $\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ is played by exploration (i.e. by chance), and by exploitation (i.e. according to the best response rule given by (3.8)), respectively. According to Algorithm 5, the total number of exploration trials is given by $\sum_{j=1}^{J} \left\lceil T'_j Z_j \right\rceil$, and by Assumption (A4) we know

$$\lim_{J \to \infty} \frac{\sum_{j=1}^{J} \left\lceil T'_j Z_j \right\rceil}{\sum_{j=1}^{J} T'_j} = 0. \tag{C.13}$$

This implies that $\hat{\pi}_{n,m}\left(i^{(k)}, \mathbf{i}^{(-k)}\right) = 0$ holds for $n \to \infty$. Therefore, in the limit,

$\hat{\pi}_{n,m}\left(i^{(k)}, \mathbf{i}^{(-k)}\right)$ can be neglected when calculating the empirical frequencies of plays, and

$$\hat{\pi}_n\left(i^{(k)}, \mathbf{i}^{(-k)}\right) = \hat{\pi}_{n,l}\left(i^{(k)}, \mathbf{i}^{(-k)}\right) \tag{C.14}$$

holds asymptotically.

In order to complete the proof, it is sufficient to show that after some finite time, the player's actions that are based on $\hat{f}_{i,t}^{(k)}$ are equal to those based on $f_i^{(k)}$. By Assumption (A2), with probability 1, there exists an $\nu > 0$ so that for every $i \in \{1, ..., M\}$ and after a time point $\theta < \infty$, $\| \hat{f}_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right) - f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) \| < \nu$ holds for all $t > \theta$. At the same time, according to our system model and by Assumption (A2), the reward functions are bounded, and the action space and memory are finite. This implies that if $\| \hat{f}_{i,t}^{(k)}\left(\mathbf{i}^{(-k)}\right) - f_i^{(k)}\left(\mathbf{i}^{(-k)}\right) \| < \nu$ holds, then the player's actions evolve as if it were aware of the true mean reward of each joint action profile, which completes the proof.

D. Additional Proofs of Chapter 4

1. Proof of Proposition 5

According to our system model, $i^{(k)} \leq P_M \ \forall \ k \in \mathcal{K}$. Besides, $|h_{uv,q}|^2 = |h'_{uv}|^2 \left|h''_{uv,q}\right|^2$, with $0 < |h'_{uv}|^2 \leq 1$ and $0 < \left|h''_{uv,q}\right|^2 \leq 1$. Hence,

$$\sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log \left(\frac{P_c \left|h_{bl,q}\right|^2}{1 + \sum_{k \in \mathcal{K}_q} i^{(k)} \left|h_{kl,q}\right|^2} \right) \geq \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log \left(\frac{P_c \left|h_{bl,q}\right|^2}{1 + \sum_{k \in \mathcal{K}_q} P_M \left|h'_{kl}\right|^2} \right). \tag{D.1}$$

By basic properties of the logarithmic function, the right-hand side of (D.1) can be written as

$$\begin{aligned} &\sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log \left(P_c \left|h_{bl,q}\right|^2 \right) - \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log \left(1 + \sum_{k \in \mathcal{K}_q} P_M \left|h'_{kl}\right|^2 \right) \\ > &\sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \log \left(P_c \left|h_{bl,q}\right|^2 \right) - \sum_{q=1}^{Q} \sum_{l \in \mathcal{L}_q} \sum_{k \in \mathcal{K}_q} P_M \left|h'_{kl}\right|^2, \end{aligned} \tag{D.2}$$

where the inequality follows from the standard logarithm inequality, $\frac{a}{1+a} \leq \log(1+a) \leq a, \ \forall a > -1$ [Lov80]. Therefore the result follows.

2. Proof of Proposition 6

We proceed by contraposition, i.e. we show that if $\{q \in \mathcal{Q} | L_q \neq 1\} \neq \varnothing$, then the partitioning is suboptimal.

Let $\mathcal{H}$ be the set of all possible Q-way partitionings of $L+K$ vertices of G_E. Assume that there exists some partitioning $c \in \mathcal{H}$, by which the graph is partitioned into Q_a clusters with $L_q > 1$, $q \in \{1, ..., Q_a\}$. As $L = Q$ (see Section 4.1.1), there remain $Q_b = Q - Q_a$ clusters with $L_q = 0$, $q \in \{Q_a + 1, ..., Q\}$. In what follows, we show that partitioning c is suboptimal, by constructing another partitioning whose cost is less than that of c.

Index Q_a and Q_b clusters of partitioning c by $1, ..., Q_a$ and $Q_a + 1, ..., Q$, respectively. Moreover, let T_a and T_b correspondingly denote the sum of edges' weights inside all clusters

with and without cellular users. Thus we have

$$T_a = \sum_{q=1}^{Q_a} \sum_{l \in \mathcal{L}_q} \left(\sum_{j \in \mathcal{L}_q} w_{jl} + \sum_{k \in \mathcal{K}_q} w_{kl} \right), \tag{D.3}$$

and $T_b = 0$ by Definition 9. Let T_c denote the total cost of partitioning c. In order to establish that partitioning c is suboptimal, we show that

$$T_c = T_a + T_b > \min_{\mathcal{H}} \sum_{q=1}^{Q_a} \sum_{l \in \mathcal{L}_q} \left(\sum_{j \in \mathcal{L}_q} w_{jl} + \sum_{k \in \mathcal{K}_q} w_{kl} \right). \tag{D.4}$$

To this end, we construct some partitioning c', with $T_{c'} < T_c$. Assume that we change only one cluster of c, say cluster $r \in \{1, ..., Q_a\}$ with $L_r > 1$, by removing a cellular user $J \in \mathcal{L}_r$. Since all vertices must be included in the partitioning, J is added in some cluster $r' \in \{1, ..., Q\} - \{r\}$. Therefore, one of the following holds:

- $r' \in \{1, ..., Q_a\} - \{r\}$, or
- $r' \in \{Q_a + 1, ..., Q\}$.

It is clear that the first case results in the original problem. Hence we assume that the cellular user J is included in $r' \in \{Q_a + 1, ..., Q\}$, and refer to the new partitioning by c'. Then we have

$$T_{c'} = T_c - \sum_{j \in \mathcal{L}_r} w_{jJ} - \sum_{k \in \mathcal{K}_r} w_{kJ} + \sum_{k \in \mathcal{K}_{r'}} w_{kJ}. \tag{D.5}$$

Since $0 \leq w_{kJ} \leq P_M$, we have $0 \leq \sum_{k \in \mathcal{K}_x} w_{kJ} \leq KP_M$, for any clusters x. Moreover, since $\mathcal{L}_r > 1$ and $w_{jJ} = C$ for $j, J \in \mathcal{L}$, $j \neq J$, we have $\sum_{j \in \mathcal{L}_r} w_{jJ} \geq C$ (see also Definition 9). Hence the worst-case occurs when: i) $\sum_{k \in \mathcal{K}_r} w_{kJ} = 0$, which means that in cluster r, no D2D user causes interference to the cellular user J, ii) $\sum_{k \in \mathcal{K}_{r'}} w_{kJ} = KP_M$, that is cluster r' includes all D2D users and they cause the maximum interference to the cellular user J, and iii) $\sum_{j \in \mathcal{L}_r} w_{jJ} = C$, i.e. $L_r = 2$. As a result,

$$T_{c'} \leq T_c - C + KP_M < T_c, \tag{D.6}$$

as we assume $C > KP_M$ by Definition 9. Therefore by (D.6) partitioning c is suboptimal, which is the contraposition and hence the proof is complete.

3. Proof of Proposition 7

By Proposition 6, any optimal partitioning of the estimated network graph G_E includes exactly one cellular user in each cluster, that is $w_{ij} = 0$, $\forall i, j \in \mathcal{L}_q, q \in \mathcal{Q}$. Moreover, by Definition 9, $w_{ij} = 0$ $\forall i, j \in \mathcal{K}$. Therefore we define a complete bipartite graph G with $\mathcal{V}_1 = \mathcal{K}$ and $\mathcal{V}_2 = \mathcal{L}$. The weight of the edge connecting $k \in \mathcal{K}$ and $l \in \mathcal{L}$ is equal to the weight of the corresponding edge in G_E, i.e. $w_{kl} = \mathbf{W}_E[k, l]$. We then augment $\mathcal{V}_2$, by K times replicating each node $l \in \mathcal{L}$, resulting in a set $\mathcal{L}' = \underbrace{\mathcal{L} \cup \mathcal{L} ... \cup \mathcal{L}}_{\times K}$. Using this set, a bipartite graph G' is constructed, where $\mathcal{V}_1 = \mathcal{K}$ and $\mathcal{L}_2 = \mathcal{L}'$. The weight of any edge connecting any $k \in \mathcal{K}$ to every copy $l' \in \mathcal{L}'$ of some l is $w_{kl'} = w_{kl}$. On graph G', a bipartite minimum-weighted matching results in an $K \times (K \times L)$ assignment matrix $\mathbf{B} = [b_{kl'}]$, so that the sum

$$\sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}'} w_{kl'} b_{kl'} \tag{D.7}$$

is minimized. For each l, let the set of its copies be denoted by $\mathcal{U}_l$. Moreover, the set of all users $k \in \mathcal{K}$ that are assigned to any copy of l is denoted by $\mathcal{A}_l$. Therefore (D.7) can be reformulated as

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{U}_l} \sum_{j' \in \mathcal{A}_l} b_{jl} w_{jj'} b_{j'l}, \tag{D.8}$$

which is identical to (4.21). Hence the proposition follows.

4. Proof of Theorem 8

The proof consists of two parts. First we show that the power allocation game defined in Definition 11 is an exact potential game, by deriving a potential function. This will prove the first part of Theorem 8. Afterwards we establish that the potential function satisfies the LMP property, and we characterize the set of Nash equilibria using Proposition 8. This will prove the second part of the theorem.

Part One

By Definition 18, we need to find a function $v : \mathcal{I} \to \mathbb{R}^+$ that satisfies (A.5). With $f^{(k)}(\mathbf{i})$ given by (4.2) we have

$$f^{(k)}\left(i^{(k)}, \mathbf{i}_q^{(-k)}\right) - f^{(k)}\left(i'^{(k)}, \mathbf{i}_q^{(-k)}\right) = \log\left(\frac{i^{(k)}}{i'^{(k)}}\right) - \alpha\left(i^{(k)} - i'^{(k)}\right). \tag{D.9}$$

Define

$$v(\mathbf{i}_q) = \sum_{k \in \mathcal{K}_q} \log\left(i^{(k)}\right) - \sum_{k \in \mathcal{K}_q} \alpha i^{(k)}. \tag{D.10}$$

Then by simple calculus it follows that

$$v^{(k)}\left(i^{(k)}, \mathbf{i}_q^{(-k)}\right) - v^{(k)}\left(i'^{(k)}, \mathbf{i}_q^{(-k)}\right) = \log\left(\frac{i^{(k)}}{i'^{(k)}}\right) - \alpha\left(i^{(k)} - i'^{(k)}\right). \tag{D.11}$$

Therefore, according to Definition 18 and by comparing (D.11) with (D.9), the power allocation game is an exact potential game with a potential function defined in (D.10).

Part Two

Lemma 9. *The potential function given by (D.10) is separable concave.*

Proof. Clearly, the potential function can be written as $v\left(\mathbf{i}_q\right) = \sum_{k \in \mathcal{K}_q} v^{(k)}\left(i^{(k)}\right)$ with

$$v^{(k)}\left(i^{(k)}\right) = \log\left(i^{(k)}\right) - \alpha i^{(k)}. \tag{D.12}$$

Thus, by the assumption $i^{(k)} > 1$ (see Section 4.1.1), we have

$$\begin{aligned} \frac{v^{(k)}\left(i^{(k)}+1\right) + v^{(k)}\left(i^{(k)}-1\right)}{2} &= \frac{\log\left(\left(i^{(k)}\right)^2 - 1\right) - 2\alpha i^{(k)}}{2} \\ \leq \frac{\log\left(\left(i^{(k)}\right)^2\right) - 2\alpha i^{(k)}}{2} &= \log\left(i^{(k)}\right) - \alpha i^{(k)}. \end{aligned} \tag{D.13}$$

Therefore, by Definition 20, the function is separable concave. □

Lemma 10. *The potential function given by (D.10) satisfies the larger midpoint property.*

Proof. The proof directly follows from Lemma 4 and Lemma 9. □

Therefore, since the potential function satisfies the LMP property, the second part of Theorem 8 follows directly from Proposition 8.

Publication List

[1] S. Maghsudi and S. Stanczak. A hybrid centralized-decentralized resource allocation scheme for two-hop transmission. In *International Symposium on Wireless Communication Systems (ISWCS)*, pages 96–100, Aachen-Germany, Nov 2011.

[2] S. Maghsudi and S. Stanczak. A delay-constrained rateless coded incremental relaying protocol for two-hop transmission. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 168–172, Paris-France, April 2012.

[3] S. Maghsudi and S. Stanczak. On network-coded rateless transmission: Protocol design, clustering and cooperator assignment. In *International Symposium on Wireless Communication Systems (ISWCS)*, pages 306–310, Paris-France, Aug 2012.

[4] S. Maghsudi and S. Stanczak. Joint power allocation and relay selection for network-coded two-way relaying. In *Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, New Jersey-USA, March 2012.

[5] S. Maghsudi and S. Stanczak. Relay selection with no side information: An adversarial bandit approach. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 715–720, Shanghai-China, April 2013.

[6] S. Maghsudi and S. Stanczak. Dynamic bandit with covariates: Strategic solutions with application to wireless resource allocation. In *IEEE International Conference on Communications (ICC)*, pages 5898–5902, Budapest-Hungary, June 2013.

[7] S. Maghsudi and S. Stanczak. Relay selection problem in wireless networks: A solution concept based on stochastic bandits and calibrated forecasters. In *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 385–389, Darmstadt-Germany, June 2013.

[8] S. Maghsudi and S. Stanczak. Joint channel selection and power control in infrastructureless wireless networks: A multi-player multi-armed bandit framework. *IEEE Transactions on Vehicular Technology (TVT)*, 2014. *accepted*.

[9] S. Maghsudi and S. Stanczak. Transmission mode selection for network-assisted device to device communication: A levy-bandit approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013, Florence-Italy, May 2014.

[10] S. Maghsudi and S. Stanczak. Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. *IEEE Transactions on Wireless Communications (TWC)*, 14(3):1309–1322, March 2015.

[11] S. Maghsudi and S. Stanczak. Hybrid centralized-distributed resource allocation for device-to-device communications underlaying cellular networks. *IEEE Transactions on vehicular Technology (TVT)*, 2015. *accepted*.

[12] S. Maghsudi and S. Stanczak. Joint channel allocation and power control for underlay D2D transmission. In *IEEE International Conference on Communications (ICC)*, London-UK, June 2015.

[13] S. Maghsudi and S. Stanczak. On channel selection for energy-constrained rateless-coded D2D transmission. 2015. *submitted as invited paper*.

[14] S. Maghsudi and S. Stanczak. Distributed channel selection for underlay device-to-device communications: A game-theoretical learning framework. In *ser. Signals and Communication Technology*. Springer, 2015. to appear.

Bibliography

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002.

[ACBFS03] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1), 2003.

[AO10] P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[ATN+14] A. Aijaz, M. Tshangini, M.R. Nakhai, X. Chu, and A.-H. Aghvami. Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees. *IEEE Transactions on Communications*, 62(7):2353–2365, July 2014.

[Bar81] E.R. Barnes. An algorithm for partitioning the nodes of a graph. In *IEEE Conference on Decision and Control*, volume 20, pages 303–304, Dec 1981.

[BCB12] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[BCDW07] L. Bottou, O. Chapelle, D. Decoste, and J. Weston. *Large-scale Kernel Machines*. The MIT Press, 2007.

[BDT08] R. Branzei, D. Dimitrov, and S. Tijs. *Models in Cooperative Game Theory*. Springer, Dordrecht, 2008.

[Bel56] R. Bellman. A problem in the sequential design of expriments. *Sankhya*, 16, 1956.

[BF11] D.R. Brown and F. Fazel. A game theoretic study of energy efficient cooperative wireless networks. *Journal of Communications and Networks*, 13(3):266–276, June 2011.

[BFA11] M. Belleschi, G. Fodor, and A. Abrardo. Performance analysis of a distributed resource allocation scheme for D2D communications. In *IEEE Global Communication Workshops*, pages 358–362, Dec 2011.

[BGN11] M. Bennis, S. Guruacharya, and D. Niyato. Distributed learning strategies for interference mitigation in femtocell networks. In *IEEE Global Telecommunications Conference*, pages 1–5, Dec 2011.

[BM07] A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, Dec 2007.

[BPSF13] Z. Bnaya, R. Puzis, R. Stern, and A. Felner. Bandit algorithms for social network queries. In *International Conference on Social Computing*, pages 148–153, Sept 2013.

[CBFH+97] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, May 1997.

[CBL03] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Journal of Machine Learning*, 51(3):239–261, 2003.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CJL11] D. Chen, H. Ji, and V.C.M. Leung. Distributed optimal relay selection for improving TCP throughput over cognitive radio networks: A cross-layer design approach. In *IEEE International Conference on Communications*, June 2011.

[CKR06] R. Cohen, L. Katzir, and D. Raz. An efficient approximation for the generalized assignment problem. *Information Processing Letters*, 100(4):162–166, Nov 2006.

[CKRU08] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal. Mortal multi-armed bandits. In *Neural Information Processing Systems Conference*, pages 273–280, 2008.

[Cla89] M.K. Clayton. Covarite models for Bernoulli bandits. *Journal of Sequential Analysis*, 8:405–426, 1989.

[CLRJ13] A.C. Chapman, D.S. Leslie, A. Rogers, and N.R. Jennings. Convergent learning algorithms for unknown reward games. *SIAM Journal on Control and Optimization*, 51(4):3154–3180, 2013.

[CMRWS12] M.H. Cheung, H. Mohsenian-Rad, V.W.S Wong, and R. Schober. Utility-optimal random access for wireless multimedia networks. *IEEE Wireless Communications Letters*, 1(4):340–343, Aug 2012.

[CNT08] C. Curescu and S. Nadjm-Tehrani. A bidding algorithm for optimized utility-based resource allocation in ad hoc networks. *IEEE Transactions on Mobile Computing*, 7(12):1397–1414, Dec 2008.

[CWWL10] Y. Chen, Y. Wu, B. Wang, and K.J.R. Liu. Spectrum auction games for multimedia streaming over cognitive radio networks. *IEEE Transactions on Communications*, 58(8):2381–2390, August 2010.

[CYC+11] J. Chen, Q. Yu, P. Cheng, Y. Sun, Y. Fan, and X. Shen. Game theoretical approach for channel allocation in wireless sensor and actuator networks. *IEEE Transactions on Automatic Control*, 56(10):2332–2344, 2011.

[CZKD06a] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin. Transmission scheduling for sensor network lifetime maximization: A shortest path bandit formulation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–IV, May 2006.

[CZKD06b] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin. Transmission scheduling for sensor network lifetime maximization: A shortest path bandit formulation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, page 4, May 2006.

[CZKD07] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin. Transmission scheduling for optimizing sensor network lifetime: A stochastic shortest path approach. *IEEE Transactions on Signal Processing*, 55(5):2294–2309, May 2007.

[CZZ13] X. Chen, Z. Zhao, and H. Zhang. Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks. *IEEE Transactions on Mobile Computing*, 12(11):2155–2166, Nov 2013.

[DL13] S. Dong and J. Lee. Greedy confidence bound techniques for restless multi-armed bandit based cognitive radio. In *Annual Conference on Information Sciences and Systems*, pages 1–4, March 2013.

[DRW+09] K. Doppler, M. Rinne, C. Wijting, C.-B. Ribeiro, and K. Hugl. Device-to-device communication as an underlay to LTE-advanced networks. *IEEE Communications Magazine*, 47(12):42–49, Dec 2009.

[FDM+12] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi. Design aspects of network assisted device-to-device communications. *IEEE Communications Magazine*, 50(3):170–177, March 2012.

[Fel68] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1968.

[FL96] D. Fudenberg and D.K. Levine. *The Theory of Learning in Games.* MIT Press, 1996.

[FLYW+13] D. Feng, L. Lu, Y. Yuan-Wu, G.Y. Li, G. Feng, and S. Li. Device-to-device communications underlaying cellular networks. *IEEE Transactions on Communications*, 61(8):3541–3551, Aug 2013.

[FP14] A. Ferragut and F. Paganini. Network resource allocation for users with multiple connections: Fairness and stability. *IEEE/ACM Transactions on Networking*, 22(2):349–362, April 2014.

[FS99] Y. Freund and R.E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

[FV97] D.P. Foster and R. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

[FY03] D.P. Foster and H. P. Young. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior*, 45:73–96, 2003.

[FYX13] X. Fang, D. Yang, and G. Xue. Taming wheel of fortune in the air: An algorithmic framework for channel selection strategy in cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 62(2):783–796, Feb 2013.

[Gal86] Z. Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys*, 18(1):23–28, March 1986.

[GC12] A. Sidiropoulosc G. Christodouloua, V.-S. Mirroknib. Convergence and approximation in potential games. *Theoretical Computer Science*, 438:13–27, 2012.

[GD14] N. Gulati and K.R. Dandekar. Learning state selection for reconfigurable antennas: A multi-armed bandit approach. *Antennas and Propagation, IEEE Transactions on*, 62(3):1027–1038, March 2014.

[GGW11] J. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices*. Wiley, 2 edition, 2011.

[Git79] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society*, 41(9):148–177, 1979.

[GKJ10] Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *IEEE Symposium on New Frontiers in Dynamic Spectrum*, pages 1–9, April 2010.

[GL07] F. Germano and G. Lugosi. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior*, 60(1):154, July 2007.

[GLM04] M. Guo, Y. Liu, and J. Malec. A new Q-learning algorithm based on the metropolis criterion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(5):2140–2143, Oct 2004.

[GMS10] S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *Journal of ACM*, 58(1):3, 2010.

[GVJ10] J. Gao, S.A. Vorobyov, and H. Jiang. Cooperative resource allocation games under spectral mask and total power constraints. *IEEE Transactions on Signal Processing*, 58(8):4379–4395, Aug 2010.

[GXW11] L. Gao, Y. Xu, and X. Wang. Map: Multiauctioneer progressive auction for dynamic spectrum access. *IEEE Transactions on Mobile Computing*, 10(8):1144–1161, Aug 2011.

[Han57] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3(39):97–139, 1957.

[HGF13] J. Huang, X. Gan, and X. Feng. Multi-armed bandit based opportunistic channel access: A consideration of switch cost. In *IEEE International Conference on Communications*, pages 1651–1655, June 2013.

[HHCP08] J. Huang, Z. Han, M. Chiang, and H.V. Poor. Auction-based resource allocation for cooperative communications. *IEEE Journal on Selected Areas in Communications*, 26(7):1226–1237, September 2008.

[HJL07] Z. Han, Z. Ji, and K.J.R. Liu. Non-cooperative resource competition game by virtual referee in multi-cell ofdma networks. *IEEE Journal on Selected Areas in Communications*, 25(6):1079–1090, August 2007.

[HLF11] S.-K. Hsu, J.-S. Lin, and K.-T. Feng. Stochastic multiple channel sensing protocol for cognitive radio networks. In *IEEE Wireless Communications and Networking Conference*, pages 227–232, March 2011.

[HMc01] S. Hart and A. Mas-colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, May 2001.

[HP05] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, Dec 2005.

[HYXY12] T. Han, R. Yin, Y. Xu, and G. Yu. Uplink channel reusing selection optimization for device-to-device communication underlaying cellular networks. In *IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pages 559–564, Sept 2012.

[JKR+09] P. Janis, V. Koivunen, C.-B. Ribeiro, K. Doppler, and K. Hugl. Interference-avoiding MIMO schemes for device-to-device radio underlaying cellular networks. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 2385–2389, Sept 2009.

[JMMM13] K. Jagannathan, S. Mannor, I. Menache, and E. Modiano. A state action frequency approach to throughput maximization over uncertain wireless channels. *Internet Mathematics*, 9(2-3):136–160, Jun 2013.

[JR11] G.A. Jehle and P.J. Reny. *Advanced Microeconomic Theory*. Financial Times-Prentice Hall, 2011.

[KA08] B. Kaufman and B. Aazhang. Cellular networks with an overlaid device to device network. In *Asilomar Conference on Signals, Systems and Computers*, pages 1537–1541, Oct 2008.

[KD07] V. Krishnamurthy and D.V. Djonin. Structured threshold policies for dynamic sensor scheduling-a partially observed Markov decision process approach. *IEEE Transactions on Signal Processing*, 55(10), Oct 2007.

[KE05] J. Kujala and T. Elomaa. On following the perturbed leader in the bandit setting. In *Algorithmic Learning Theory*, pages 371–385, Oct 2005.

[KE07] J. Kujala and T. Elomaa. Following the perturbed leader to gamble at multi-armed bandits. In *Algorithmic Learning Theory*, volume 4754, pages 166–180, 2007.

[KF08] S.M. Kakade and D.P. Foster. Deterministic calibration and Nash equilibrium. *Elsevier Journal of Computer and System Sciences*, 74:115–130, 2008.

[KJ87] M.N. Katehakis and A.F. Veinott Jr. The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[KNJ12] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multi-player multi-armed bandits. In *IEEE Annual Conference on Decision and Control*, pages 3960–3965, Dec 2012.

[KNJ14] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, April 2014.

[Kuh55] H.-W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2(1):83–97, 1955.

[Li09] H. Li. Multi-agent q-learning of channel selection in multi-user cognitive radio systems: A two by two case. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1893–1898, Oct 2009.

[LJP08] L. Lai, H. Jiang, and H.V. Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *Asilomar Conference on Signals, Systems and Computers*, pages 98–102, Oct 2008.

[LJYH14] Y. Li, D. Jin, J. Yuan, and Z. Han. Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks. *IEEE Transactions on Wireless Communications*, 13(7):3965–3977, July 2014.

[LLK12] S. Liu, L. Lazos, and M. Krunz. Cluster-based control channel allocation in opportunistic cognitive radio networks. *IEEE Transactions on Mobile Computing*, 11(10):1436–1449, Oct 2012.

[LMY+13] C. Luo, G. Min, F.R. Yu, M. Chen, L.T. Yang, and V.C.M. Leung. Energy-efficient distributed relay and power control in cognitive radio cooperative communications. *IEEE Journal on Selected Areas in Communications*, 31(11):2442–2452, November 2013.

[Lov80] E.R. Love. Some logarithm inequalities. *The Mathematical Gazette*, 64(427):55–57, 1980.

[LPT13] M. Lelarge, A. Proutiere, and M.S. Talebi. Spectrum bandit optimization. In *IEEE Information Theory Workshop*, pages 1–5, Sept 2013.

[LR85] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[LTH+07] J.-W. Lee, A. Tang, J. Huang, M. Chiang, and A.R. Calderbank. Reverse-engineering mac: A non-cooperative game model. *IEEE Journal on Selected Areas in Communications*, 25(6):1135–1147, August 2007.

[LYW+11] B. Li, P. Yang, J. Wang, Q. Wu, and N. Xia. Non-bayesian learning of channel sensing order for dynamic spectrum access networks. In *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 529–534, Oct 2011.

[LZ09] K. Liu and Q. Zhao. On the myopic policy for a class of restless bandit problems with applications in dynamic multichannel access. In *IEEE Conference on Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference*, pages 3592–3597, Dec 2009.

[LZ10a] K. Liu and Q. Zhao. Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 3010–3013, March 2010.

[LZ10b] 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 2010.

[LZK10] K. Liu, Q. Zhao, and B. Krishnamachari. Decentralized multi-armed bandit with imperfect observations. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1669–1674, Sep 2010.

[MKZ09] M. Maskery, V. Krishnamurthy, and Q. Zhao. Decentralized dynamic spectrum access for cognitive radios: cooperative design of a non-cooperative game. *IEEE Transactions on Communications*, 57(2):459–469, February 2009.

[MS10] S. Mannor and G. Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010.

[MV80] S. Micali and V. Vazirani. An algoithm for finding maximum matching in general graphs. In *Annual Symposium on Foundations of Computer Science*, pages 17–27, Oct 1980.

[NH99] J. Nie and S. Haykin. A q-learning-based dynamic channel assignment technique for mobile communication systems. *IEEE Transactions on Vehicular Technology*, 48(5):1676–1687, Sep 1999.

[NH07] D. Niyato and E. Hossain. Radio resource management games in wireless networks: an approach to bandwidth allocation and admission control for polling service in ieee 802.16 [radio resource management and protocol engineering for ieee 802.16]. *IEEE Wireless Communications*, 14(1):27–35, Feb 2007.

[NM44] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[NM01] J. Nino-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98, 2001.

[NRTV07] N. Nisan, T. Roughgarden, E. Tardos, and V.V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.

[OKP12] J. Oksanen, V. Koivunen, and H.V. Poor. A sensing policy based on confidence bounds and a restless multi-armed bandit model. In *Asilomar Conference on Signals, Systems and Computers*, pages 318–323, Nov 2012.

[OMES11] W. Ouyang, S. Murugesan, A. Eryilmaz, and N.B. Shroff. Exploiting channel memory for joint estimation and scheduling in downlink networks. In *IEEE Conference on Computer Communications*, pages 3056–3064, April 2011.

[PCA07] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. In *Proceedings of the International Conference on Machine Learning*, pages 721–728, 2007.

[PHK13] P. Phunchongharn, E. Hossain, and D.-I. Kim. Resource allocation for device-to-device communications underlaying LTE-advanced networks. *IEEE Wireless Communications*, 20(4):91–100, Aug 2013.

[PO13] F.H. Panahi and T. Ohtsuki. Optimal channel-sensing policy based on fuzzy q-learning process over cognitive radio systems. In *IEEE International Conference on Communications*, pages 2677–2682, June 2013.

[Put94] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, NY, USA, 1st edition, 1994.

[Rob52] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[RW05] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[RW08] AH.M. Rad and V.W.S. Wong. Cross-layer fair bandwidth sharing for multichannel wireless mesh networks. *IEEE Transactions on Wireless Communications*, 7(9):3436–3445, Sept 2008.

[SBP06] G. Scutari, S. Barbarossa, and D.P. Palomar. Potential games: A framework for vector power control problems with coupled constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, page 4, May 2006.

[SHD+09] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar. Coalitional game theory for communication networks. *IEEE Signal Processing Magazine*, 26(5):77–97, September 2009.

[SJLS00] S. Singh, T. Jaakkola, M.L. Littman, and C. Szepesvari. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.

[SL05] G. Stoltz and G. Lugosi. Internal regret in on-line portfolio selection. *Journal of Machine Learning*, 59(1):125–159, 2005.

[SNHH14] L. Song, D. Niyato, Z. Han, and E. Hossain. Game-theoretic resource allocation methods for device-to-device communication. *IEEE Wireless Communications*, 21(3):136–144, June 2014.

[Sto82] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[SYJL08] P. Si, F.R. Yu, H. Ji, and V.C.M. Leung. Distributed sender scheduling for multimedia transmission in wireless peer-to-peer networks. In *IEEE Global Telecommunications Conference*, pages 1–5, Nov 2008.

[SYJL10] P. Si, F.R. Yu, H. Ji, and V.C.M Leung. Distributed multisource transmission in wireless mobile peer-to-peer networks: A restless-bandit approach. *IEEE Transactions on Vehicular Technology*, 59(1):420–430, Jan 2010.

[Ui08a] T. Ui. Correlated equilibrium and concave games. *International Journal of Game Theory*, 37(1):1–13, April 2008.

[Ui08b] T. Ui. Discrete concavity for potential games. *International Game Theory Review*, 10(1):137, 2008.

[VH04a] J.G. Vlachogiannis and N.D. Hatziargyriou. Reinforcement learning for reactive power control. *IEEE Transactions on Power Systems*, 19(3):1317–1325, Aug 2004.

[VH04b] J.G. Vlachogiannis and N.D. Hatziargyriou. Reinforcement learning for reactive power control. *IEEE Transactions on Power Systems*, 19(3):1317–1325, Aug 2004.

[WAM08] Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, Dec 2008.

[WCC+11] B. Wang, L. Chen, X. Chen, X. Zhang, and D. Yang. Resource allocation optimization for device-to-device communication underlaying cellular networks. In *IEEE Vehicular Technology Conference*, pages 1–6, 2011.

[Web92] R. Weber. On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.

[Whi80] P. Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society*, 42(2):143–149, 1980.

[WHL09] B. Wang, Z. Han, and K.J.R. Liu. Distributed relay selection and power control for multiuser cooperative communication networks using stackelberg game. *IEEE Transactions on Mobile Computing*, 8(7):975–990, July 2009.

[WKP05] C.C. Wang, S.R. Kulkarni, and H.V. Poor. Advances in applied mathematics. *Arbitrary Side observations in bandit problems*, 34(4):903–938, May 2005.

[WSH+13] F. Wang, L. Song, Z. Han, Q. Zhao, and X. Wang. Joint scheduling and resource allocation for device-to-device underlay communication. In *IEEE Wireless Communications and Networking Conference*, pages 134–139, April 2013.

[WWJ+14] Q. Wang, W. Wang, S. Jin, H. Zhu, and N.T. Zhang. Quality-optimized joint source selection and power control for wireless multimedia d2d communication using stackelberg game, 2014.

[XSH+12a] C. Xu, L. Song, Z. Han, D. Li, and B. Jiao. Resource allocation using a reverse iterative combinatorial auction for device-to-device underlay cellular networks. In *IEEE Global Communications Conference*, pages 4542–4547, Dec 2012.

[XSH+12b] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, and B. Jiao. Interference-aware resource allocation for device-to-device communications as an underlay using sequential second price auction. In *IEEE International Conference on Communications*, pages 445–449, June 2012.

[XSH+13] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao. Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach. *IEEE Journal on Selected Areas in Communications*, 31(9):348–358, Sept 2013.

[XWS+13] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan. Opportunistic spectrum access with spatial reuse: Graphical game and uncoupled learning solutions. *IEEE Transactions on Wireless Communications*, 12(10):4814–4826, October 2013.

[XWW+12] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.D. Yao. Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution. *IEEE Transactions on Wireless Communications*, 11(4):1380–1391, April 2012.

[XWW+13] Y. Xu, Q. Wu, J. Wang, L. Shen, and A. Anpalagan. Opportunistic spectrum access using partially overlapping channels: Graphical game and uncoupled learning. *IEEE Transactions on Communications*, 61(9):3906–3918, September 2013.

[YDRT11] C.H. Yu, K. Doppler, C.B. Ribeiro, and O. Tirkkonen. Resource sharing optimization for device-to-device communication underlaying cellular networks. *IEEE Transactions on Wireless Communications*, 10(8):2752–2763, Aug 2011.

[YFX12] D. Yang, X. Fang, and G. Xue. Game theory in cooperative communications. *IEEE Wireless Communications*, 19(2):44–49, April 2012.

[YZ02] Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

[ZLNW13] Y. Zhang, C. Lee, D. Niyato, and P. Wang. Auction approaches for resource allocation in wireless systems: A survey. *IEEE Communications Surveys Tutorials*, 15(3):1020–1041, Third 2013.

[ZTL+11] Y. Zhai, P. Tehrani, L. Li, J. Zhao, and Q. Zhao. Dynamic pricing under binary demand uncertainty: A multi-armed bandit with correlated arms. In *Asilomar Conference on Signals, Systems and Computers*, pages 1597–1601, Nov 2011.

[ZTSC07] Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework. *IEEE Journal on Selected Areas in Communications*, 25(3):589–600, 2007.

[ZYC12] G. Zhang, K. Yang, and H.-H. Chen. Resource allocation for wireless cooperative networks: a unified cooperative bargaining game theoretic framework. *IEEE Wireless Communications*, 19(2):38–43, April 2012.