CONTEXTUAL MULTI-ARMED BANDITS WITH STRUCTURED PAYOFFS

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Muhammad Anjum Qureshi
September 2020

Contextual Multi-Armed Bandits with Structured Payoffs
By Muhammad Anjum Qureshi
September 2020

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Cem Tekin(Advisor)

Orhan Arıkan

Savaş Dayanık

Elif Vural

Umut Orguner

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

CONTEXTUAL MULTI-ARMED BANDITS WITH STRUCTURED PAYOFFS

Muhammad Anjum Qureshi
Ph.D. in Electrical and Electronics Engineering
Advisor: Cem Tekin
September 2020

Multi-Armed Bandit (MAB) problems model sequential decision making under uncertainty. In traditional MAB, the learner selects an arm in each round, and then, observes a random reward from the arm's unknown reward distribution. In the end, the goal is to maximize the cumulative reward by learning to select optimal arms as much as possible. In the contextual MAB—an extension to MAB—the learner observes a context (side-information) in the beginning of each round, selects an arm, and then, observes a random reward whose distribution depends on both the arriving context and the chosen arm. Another MAB variant, called unimodal MAB, assumes that the expected reward exhibits a unimodal structure over the arms, and tries to locate the arm with the "peak" reward by learning the direction of increase of the expected reward. In this thesis, we consider an extension to unimodal MAB called contextual unimodal MAB, and demonstrate that it is a powerful tool for designing Artificial Intelligence (AI)-enabled radios by utilizing the special structure of the dependence of the reward to contexts and arms of the wireless environment.

While AI-enabled radios are expected to enhance the spectral efficiency of 5th generation (5G) millimeter wave (mmWave) networks by learning to optimize network resources, allocating resources over the mmWave band is extremely challenging due to rapidly-varying channel conditions. We consider several resource allocation problems in this thesis under various design possibilities for mmWave radio networks under unknown channel statistics and without any channel state information (CSI) feedback: i) dynamic rate selection for an energy harvesting transmitter, ii) dynamic power allocation for heterogeneous applications, and iii) distributed resource allocation in a multi-user network. All of these problems exhibit structured payoffs which are unimodal functions over partially ordered arms (transmission parameters) as well as unimodal or monotone functions over partially ordered contexts (side-information). Structure over arms helps in reducing the number of arms to be explored, while structure over contexts helps in

using past information from nearby contexts to make better selections.

We formalize dynamic adaptation of transmission parameters as a structured MAB, and propose frequentist and Bayesian online learning algorithms. We show that both approaches yield logarithmic in time regret. We also investigate dynamic rate and channel adaptation in a cognitive radio network serving heterogeneous applications under dynamically varying channel availability and rate constraints. We formalize the problem as a Bayesian learning problem, and propose a novel learning algorithm which considers each rate-channel pair as a two-dimensional action. The set of available actions varies dynamically over time due to variations in primary user activity and rate requirements of the applications served by the users. Additionally, we extend the work to cater to the scenario when the arms belong to a continuous interval as well as the contexts. Finally, we show via simulations that our algorithms significantly improve the performance in the aforementioned radio resource allocation problems.

Keywords: contextual MAB, unimodal MAB, Thompson sampling, volatile MAB, regret bounds, cognitive radio networks, AI-enabled radio, mmWave, resource allocation.

ÖZET

YAPILANDIRILMIŞ GETİRİLİ BAĞLAMSAL ÇOK KOLLU HAYDUTLAR

Muhammad Anjum Qureshi
Elektrik ve Elektronik Mühendisliği, Doktora
Tez Danışmanı: Cem Tekin
Eylül 2020

Çok kollu haydut (MAB) problemleri belirsizlik altında sıralı karar verme işlemini modellerler. Geleneksel MAB probleminde, öğrenici her turda bir kol seçer ve ardından seçilen kolun bilinmeyen ödül dağılımdan rastgele bir ödül gözlemler. Öğrenicinin amacı en yüksek ödülü veren kolları seçmeyi öğrenerek toplam ödülü ençoklamaktır. MAB'nin bir uzantısı olan bağlamsal MAB probleminde, öğrenici her turun başlangıcında bir bağlamı (yan bilgi) gözlemler, bir kol seçer ve dağılımı hem gelen bağlama hem de seçilen kola bağlı olan rastgele bir ödülü gözlemler. Başka bir MAB çeşidi olan tektepeli MAB'de ise beklenen ödülün kollar üzerinde tek tepeli bir yapı sergilediğini varsayar ve beklenen ödülün artış yönünü öğrenerek "zirve" ödüllü kolu tespit etmeye çalışır. Bu tezde, tek tepeli MAB probleminin bir uzantısı olan bağlamsal tek tepeli MAB problemi incelenmektedir ve bu modelin kollardaki ve bağlamlardaki yapılardan istifade ederek kablosuz iletişim alanında yapay zeka tabanlı radyo dizaynında güçlü bir araç olduğu gösterilmektedir.

Yapay zeka özellikli telsizlerin, ağ kaynaklarını optimize etmeyi öğrenerek 5. nesil (5G) milimetre dalga (mmWave) ağlarının spektral verimliliğini artırması beklenmektedir. Ancak, kaynakların mmWave bandı üzerinden tahsis edilmesi, hızla değişen kanal koşulları nedeniyle son derece zordur. Bu tezde, bilinmeyen kanal istatistikleri altında ve herhangi bir kanal durum bilgisi (CSI) geri bildirimi olmaksızın mmWave radyo ağları için çeşitli tasarım olanakları altında çeşitli kaynak tahsisi problemleri ele alınmıştır: i) bir enerji hasat vericisi için dinamik oran seçimi, ii) heterojen uygulamalar için dinamik güç tahsisi ve iii) çok kullanıcılı bir ağda dağıtılmış kaynak tahsisi. Tüm bu problemlerde beklenen ödül hem kısmen sıralı kollar (iletişim parametreleri) üzerinde tek tepeli hem de kısmen sıralı bağlamlar (yan-bilgi) üzerinde tek tepeli veya monoton bir yapı sergilemektedir. Kolların üzerindeki yapı, keşfedilecek kolların sayısını azaltmaya yardımcı olurken, bağlamlar üzerindeki yapı, daha iyi seçimler yapmak için yakın

bağlamlardan alınan geçmiş bilgileri kullanmaya yardımcı olur.

Tez kapsamında iletişim parametrelerinin dinamik uyarlanması problemi yukarıda belirtilen özelliklere sahip yapılandırılmış getirili MAB olarak modellenmiş ve sıklıkçı ve Bayes tabanlı yaklaşımlara dayalı çevrimiçi öğrenme algoritmaları önerilmiştir. Bu algoritmaların pişmanlıklarının zamanda logaritmik olduğu kanıtlanmıştır. Tez kapsamında bunlar haricinde, dinamik olarak değişen kanal kullanılabilirliği ve gönderim hızı kısıtlamaları altında heterojen uygulamalara hizmet veren bilişsel bir radyo ağında dinamik hız ve kanal adaptasyonu problemi da ele alınmıştır. Problem bir Bayes öğrenme problemi olarak modellenmiş ve her bir hız-kanal çiftini iki boyutlu bir eylem olarak ele alan yeni bir öğrenme algoritması önerilmiştir. Bu problemde kullanılabilir eylemler kümesi, birincil kullanıcı etkinliğindeki ve ağ uygulamalarının gönderim hızı gereksinimlerindeki değişimler nedeniyle zaman içinde dinamik olarak değişir. Bunlara ek olarak, seçilebilen kol kümesinin sürekli olduğu durumları içeren senaryolar da çalışılmıştır. Son olarak, geliştirilen algoritmaların yukarıda belirtilen kaynak tahsisi problemlerinde performansı önemli ölçüde iyileştirdiğini simülasyonlar yoluyla da gösterilmiştir.

Anahtar sözcükler: bağlamsal çok kollu haydutlar, tektepeli çok kollu haydutlar, Thompson örneklemesi, değişken çok kollu haydutlar, pişmanlık sınırları, bilişsel radyo ağları, yapay zeka özellikli radyo, mmWave, kaynak tahsisi.

Acknowledgement

I would like to thank All-mighty ALLAH for HIS blessings on me that I am able to reach such far in my life. I would like to express my gratitude to my supervisor Asst. Prof. Dr. Cem Tekin. I can easily distinguish my learning under his supervision from learning in the rest of my life before joining the Bilkent University. I have benefited a lot from his problem solving skills, technical writing and mathematical approach towards a problem, his continuous support and patience despite difficult times in my research has left a remarkable place in my life and was the reason for completion of this work. I would like to thank my Thesis committee members, Prof. Dr. Orhan Arikan and Prof. Dr. Savas Dayanik. They were kind enough to give their time for my regular thesis meetings, and provide useful comments those contributed towards improvement of the work and specially the last two chapters of the thesis. I want to mention external examiners Dr. Umut Orguner and Dr. Elif Vural to accept the invitation of serving as external examiners, and giving their precious time to read my thesis, and providing useful comments. I would to like to thank Prof. Dr. Sinan Gezici for being there whenever I needed departmental support, and guiding me throughout my stay.

I would like to thank Govt of Pakistan and my organization for supporting initial years of my study, and then TUBITAK for supporting me for the last year and so of my study under grants 116E229 and 215E342.

It is worth mentioning my family for their support in this journey. My parents, Saira Zulfiqar and Zulfiqar Ahmed Qureshi, are always there praying for me during all my period in Bilkent University. Whenever I wanted to start a task, I used to call my mother and ask her to pray for me, and then onwards I feel confident that I can do it. I would like to say thanks to love of my life, Faiza Anjum, for believing in me when I myself doubts to proceed. Her role in my career and life is far bigger than a wife, she is a friend, a mentor, an obedient partner. My kids Hassaan Ahmed Qureshi and Hannaan Ahmed Qureshi are always there praying for me with their little hands. Although, they see my strictness when they are studying or playing, but they know this is for their learning and improvement of skills. Almighty has given a gift of Hussaam Ahmed Qureshi during our stay in Turkey, he is gem of an addition to our family. I would like to mention my siblings,

Irum, Mehwish and Shehriyar for their constant love and support towards me.

My stay in Bilkent is beautiful because of my friends here. I would like to thank all of them for being there for me in difficult and good times. Notable mentions are Umar B Niazi, Ali Hassan Mirza, Mahzeb Fiaz, Wardah Sarmad, Hira Noor, Sadia Khaf, Mohsin Habib and Rabia Zafar Ali, with whom we had amazing trips and meals. This journey in Turkey may be very hard due to language, but I would like to mention Elif Dogan for her help and guidance everywhere to get us through smoothly, whether kids school or other city and language related issues, also thanks to her husband Salman Dar and their new family member Vera Melis Dar. I would like to thank King Saeed Ahmed and Mohammad Kazemi for providing guidance during my degree. I would like to thank Farhan khan and family, Asad Ali, Zakwan, Abdul Waheed and family, Salahuddin Zafar and family, Mr. Naveed and family, Humayun and family, Hamza and family, Umar bhai and family, Sabih and family, Bismillah Nasir and family, Amna Malik, Ali Sheraz, Samar Batool, Laila-tul-Qadr, Khushbakht Ali, Aamir and Mubashira Zaman for making my stay at Bilkent beautiful. I would like to thank my Turkish friends Kubilay Eksioglu, Eralp Turgay, Aras Yurtman, and Merve for being there for me whenever I needed them. I would like to thank my lab fellows, course fellows and CYBORG members, Cem Bulucu, Umitcan Sahin, Tolga Ergen, Safa, Nima Akbarzadeh, Muhammad Nabi, Alireza Javanmardi, Andi Nika, Alp and Kerem.

I would like to dedicate this thesis to my parents, my wife, and my kids.

Contents

List of Figures

List of Tables

List of Publications

This thesis includes content from the following publications:

1. M. A. Qureshi and C. Tekin, "Online Cross-layer Learning in Heterogeneous Cognitive Radio Networks without CSI," in Proc. 26th IEEE Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404793.

2. M. A. Qureshi and C. Tekin, "Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks," in IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 1, pp. 95-110, March 2020, doi: 10.1109/TCCN.2019.2953607.

3. M. A. Qureshi and C. Tekin, "Online Bayesian Learning for Rate Selection in Millimeter Wave Cognitive Radio Networks," in IEEE International Conference on Computer Communications (INFOCOM), 2020, pp. 1449-1458, doi: 10.1109/INFOCOM41043.2020.9155530.

4. M. A. Qureshi and C. Tekin "Rate and Channel Adaptation in Cognitive Radio Networks under Time-Varying Constraints," in IEEE Communications Letters, doi: 10.1109/LCOMM.2020.3015823.

Chapter 1

Introduction and Literature Survey

This section was published in [1] [1], [2] [2] , [3] [3], and [4] [4].

1.1 Introduction

Multi-Armed Bandit (MAB) problems model sequential decision making under uncertainty. In these problems, the learner has access to multiple arms, plays them one at a time and observes only the random reward of the played arms.

[1] © [2018] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Online cross-layer learning in heterogeneous cognitive radio networks without CSI", in Proc. 26th IEEE Signal Processing and Communications Applications Conference (SIU), May 2018.

[2] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks", IEEE Transactions on Cognitive Communications and Networking, March 2020.

[3] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Online Bayesian Learning for Rate Selection in Millimeter Wave Cognitive Radio Networks", in Proc. IEEE International Conference on Computer Communications (INFOCOM), July 2020.

[4] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Rate and Channel Adaptation in Cognitive Radio Networks under Time-Varying Constraints", in Communications Letters, 2020.

The goal is to come up with an arm selection strategy that maximizes the cumulative reward only based on the reward feedback. This requires balancing exploration (trying different arms to learn about them) and exploitation (playing the estimated optimal arm) in a judicious manner. MAB problems have been studied for many decades, since the introduction of Thompson sampling [16] and UCB-based index policies [17]. It is shown in [17] that for the MAB with independent arms the regret grows at least logarithmically in time. A policy that is able to achieve optimal asymptotic performance is also proposed in the same work.

Many variants of the classical MAB problem exist today. Two notable examples are the contextual MAB [18–20] and the unimodal MAB [21–23]. In the classical MAB, the reward is a random variable that depends on the chosen action, whereas in the contextual MAB (CMAB), the reward also depends on the context (side-information) that is revealed before action selection takes place. Thus, the regret in the classical MAB is defined with respect to the best fixed action, while the regret in the CMAB is defined with respect to the best sequence of actions given the contexts. Another variant of MAB, called unimodal MAB, assumes that the expected reward exhibits a unimodal structure over the arms, and tries to locate the arm with the "peak" reward by learning the direction of increase of the expected reward. In addition to these, volatile MAB is another important extension of MAB [24, 25], where the set of the arms vary over time.

In this thesis, we use MAB models to formalize adaptation of transmission parameters in next-generation communication systems. Explosion in the number of mobile devices and the proliferation of data-intensive wireless services considerably increased the demand for the frequency spectrum in the recent years, and rendered the commonly used sub-6 GHz portion of the spectrum overcrowded. To overcome this challenge, next-generation communication systems like 5th generation (5G) networks aim to utilize the millimeter wave (mmWave) band which spans the spectrum between 30 and 300 GHz. While this wide swath of spectrum provides unprecedented opportunities for new wireless technologies, communication over mmWave frequencies is heavily affected by various factors including

signal attenuation, atmospheric absorption, high path loss, penetration loss, mobility and other drastic variations in the environment [26]. This makes acquiring channel state information (CSI) costly and unreliable in mmWave networks, and thus, traditional communication protocols that rely on accurate CSI [27, 28] become futile in this adversarial environment.

On the one hand, statistical characteristics of mmWave communication motivates learning theory based solutions to perform resource allocation tasks [5, 29–33]. On the other hand, artificial intelligence (AI)-enabled cognitive radio networks (CRN) are conceived to further enhance the spectral efficiency of the mmWave band [34–36]. Moreover, energy harvesting based solutions are becoming ubiquitous for many self-sustainable real-world systems, where energy is continually proliferated from nature or man-made phenomena instead of conventional battery-powered generation [9, 37]. In these networks, the transmit power usually needs to be adjusted based on exogenous events. For instance, in the spectrum underlay paradigm [38], secondary users (SUs), are capable of sensing the spectrum and adapt their power so that the interference to primary users (PUs) remains below a threshold. The term interference temperature limit (ITL) sets a pre-defined threshold, which needs to be satisfied as long as the SU is using the specific frequency band. ITL is dependent on numerous factors including the location of the SU and the selected spectrum frequency [39]. As another example, in an energy harvesting CRN without any explicit battery or super-capacitor, the harvested energy is used by the system instantly without any storage. In the considered scenario, the transmit power of the SU is dependent on the current harvested energy.

Rate Adaptation (RA) is a fundamental mechanism that allows the transmitter to adapt the modulation and coding scheme to the channel conditions, where the target is to maximize the overall throughput which is defined as the product of the rate and the packet success probability over that rate [5, 30–32]. The packet transmission outcome is random and the packet success probabilities are not known a priori to the transmitter. These probabilities depend on the transmission power, and they need to be learned via interacting with the environment. We assume that the only feedback available after a transmission is the ACK/NACK

flag. The transmitter has to learn the best rate via utilizing this feedback and taking into account its input parameters, which motivates us to develop new online adaptive allocation strategies to learn faster.

In short, the highly dynamic and unpredictable nature of the mmWave band [40, 41] makes traditional wireless systems that rely on channel models and CSI impractical, and necessitates development of new AI-enabled wireless systems that learn to adapt to the evolving network conditions and user demands through repeated interaction with the mmWave environment. There exists a plethora of AI-based methods for adaptive resource optimization in wireless communications that learn from past experience to enhance the real-time performance. Examples include MABs used for dynamic rate and channel adaptation [5], artificial neural networks used for real-time characterization of the communication performance [42] and deep Q-learning used for selecting a proper modulation and/or coding scheme (MCS) for the primary transmission [43].

In order to highlight the importance of using the problem structure, we note that without any structure on the expected rewards, the best arm for each context can be learned only by exploring all context-arm pairs sufficiently enough, by running a separate instance of traditional MAB algorithms like UCB1 [44], which results in a regret that scales linearly in the number of context-arm pairs. Significant performance improvement can be achieved by exploiting the unimodal structure of the expected reward over the arms [5, 22, 45], which results in a regret that scales linearly in the number of contexts. Since mmWave channels have rapidly-varying characteristics, even this improvement may not be enough to learn the best arms fast enough.

Driven by the unique challenges of resource optimization in the mmWave band, in this thesis, we rigorously formulate the resource allocation under rapidly-varying wireless channels with unknown statistics as a contextual MAB, where in each round a decision maker observes a side-information known as the context [19] (e.g., data rate requirement, harvested energy, available channel), selects an arm (e.g., modulation and coding, transmit power), and then, observes a random reward (e.g., packet success indicator), whose distribution depends both on the

context and the chosen arm. Furthermore, in all of the resource allocation problems we consider in this thesis, including dynamic rate selection for an energy harvesting transmitter, dynamic power allocation for heterogeneous applications and distributed resource allocation in a multi-user network, the expected rewards under different contexts and arms are correlated and have a unimodal structure. For instance, in rate adaptation, we know that if the transmission succeeds (fails) at a certain rate, then it will also succeed (fail) at lower (higher) rates. However, we assume no structure on how contexts arrive over time, and aim to investigate how the unimodal structure and the context arrivals together affect the learning performance. Additionally, we provide extensions that consider how to handle time-varying resource constraints, and also propose other models where the context and arm sets are continuous.

1.2 Literature Survey

1.2.1 Related Work on Contextual Multi-armed Bandits

In the contextual MAB, before deciding on which arm to select in each round, the learner is provided with an additional information called the context. This allows the expected arm rewards to vary based on the context, and makes the contextual MAB a powerful model for real-world applications. Since the number of contexts can be large or even infinite, additional structure is required in order to learn efficiently. A common approach is to assume that the context-arm pairs lie in a similarity space with a predefined distance metric, and the expected reward is a Lipschitz continuous function of the distance between context-arm pairs. Using this structure, [19] proposes an algorithm that achieves $\tilde{O}(T^{1-1/(2+d_c)})$ regret, where d_c is the covering dimension of the similarity space. Furthermore, [20] proposes another algorithm with $\tilde{O}(T^{1-1/(2+d_z)})$ regret, where d_z is an optimistic covering dimension, also called the zooming dimension. Apart from these, in clustering of bandits [46, 47], similar contexts are grouped in clusters based on the Lipschitz assumption, and expected rewards are estimated for clusters of

contexts. Different from these works, we consider a unimodal structure over the contexts, which allows us to use confidence bounds of the neighboring contexts instead of their reward observations by completely avoiding approximation errors.

1.2.2 Related Work on Unimodal Bandits

Papers on the unimodal MAB assume that the expected reward exhibits a unimodal structure over the arms. Algorithms designed for the unimodal MAB tries to locate the arm with the "peak" reward by learning the direction of increase of the expected reward. In [22], an algorithm that exploits the unimodal structure based on Kullback-Leibler (KL)-UCB [45] indices is proposed. A similar approach is also used for dynamic rate adaptation in [5] and [29], where the expected reward is considered to be a graphical unimodal function of the arms. The regret of these algorithms is shown to be $O(|\mathcal{N}'(a^*)|\log(T))$, where a^* is the arm with the highest expected reward and $\mathcal{N}'(a^*)$ is the set of neighbors of arm a^*, which is defined based on the unimodality graph. In general, this set is much smaller than the set of all arms and does not grow as the set of arms increases, and thus, unlike standard MAB, the regret in the unimodal MAB is independent of the number of arms.

Earlier works on rate adaptation formalize the problem as a non-contextual MAB. For example, [5], [29] and [30] propose MAB algorithms based on Kullback-Leibler upper confidence bound (KL-UCB) indices that learn the optimal rate via utilizing the unimodal structure of the expected reward over the rates. On the other hand, [31] exploits rate unimodality by using a variant of Thompson sampling, called *modified Thompson sampling* (MTS), which achieves logarithmic regret by keeping independent priors for arms and by decoupling rate from success probability. However, the important structural property of rate unimodality is not exploited in MTS. It is shown in [23] that under a unimodal assumption on the expected reward function, it is also possible to achieve logarithmic regret. The proposed *unimodal Thompson sampling* (UTS) keeps independent priors and

exploits the arm unimodality like [22]. However, this algorithm is for general expected rewards and does not decouple the rate. A similar algorithm with detailed analysis for rank one bandits is proposed in [48], which explicitly calculates the constants in the regret. In [32], authors propose *constrained Thompson sampling* (CoTS) that exploits the structure more efficiently than MTS by assuming that the success probability is monotonic over the rate. It is shown in [31] and [23] via numerical experiments that Thompson sampling outperforms the frequentist approach based on KL-UCB indices.

In contrast to all these prior works, which do not consider contexts, our works exploit the unimodal structure over arms, the contextual information and the structure of the reward over the contexts. Exogeneity of the contexts makes regret analysis significantly different than non-contextual versions of KL-UCB and Thompson sampling that use the structure in rewards [5, 23, 30–32].

1.2.3 Related Work on mmWave Communication

Wireless communication over the mmWave band is envisioned to resolve spectrum scarcity and provide unmatched data rates for next-generation wireless technologies such as 5G [49, 50]. Meanwhile, communication over the mmWave band suffers from natural disadvantages such as blocking by dynamic obstacles, severe signal attenuation, high path loss and atmospheric absorption [51]. Numerous papers are devoted to investigate the propagation properties of mmWave channels [52]. Specifically, existing work on propagation models can be divided into indoor and outdoor channel models. In the indoor scenario, it is observed that the quality of the channel is severely influenced by dynamic activity (such as human activity) inside the building [53]. In the outdoor scenario, experiments demonstrate that penetration loss due to the geometry-induced blockage (building) is dependent on the building construction material and can also be significant.

Moreover, the dynamic blockage such as humans or cars introduces additional transient loss on the paths intercepting the moving object [54]. The take-away massage from these works is that the mmWave environment is highly dynamic

and unpredictable, and the channel dynamics are difficult to model. This inherent complexity of the mmWave environment is what justifies our learning theory based approach described in this thesis.

1.2.4 Related Work on AI-based Resource Allocation in Radio Networks

A large number of online learning algorithms are proposed for selecting the right transmission parameters under time-varying conditions in 802.11 and mmWave channels [5, 29, 31, 55]. In particular, these works study rate adaptation for throughput maximization. Among these, [5] and [29] propose an MAB model and upper confidence bound (UCB) policies that learn the optimal transmission rate by exploiting the unimodality of the expected reward over arms. Specifically, the method in [29] is shown to outperform the traditional SampleRate method [56], which sequentially selects transmission rates by estimating the throughput over a sliding window. Similarly, [31] proposes a Thompson sampling based algorithm for dynamic rate adaptation, and proves that it achieves logarithmic in time regret. The concept of unimodality is also used in beam alignment for mmWave communications [33].

None of these works investigate how contextual information about the wireless environment can be used for optimizing the transmission parameters, although this might be necessary for different applications. For instance, the transmit power constraint can be regarded as context as it may affect the packet success and throughput at a given rate [57]. There are a few exceptions, such as [58], which considers learning using contextual information for beam selection/alignment for mmWave communications. However, their proposed approach does not exploit unimodality of the expected reward over arms and contexts. In essence, utilizing contextual information in a structured way is what distinguishes our work from the prior art.

There also exist papers studying resource allocation using other AI-based techniques such as Q-learning, deep learning and neural networks [42, 43, 59–62]. Surveys on applying AI-based techniques in present and future communication systems can be found in [59] and [60]. Authors in [43] propose a method based on deep Q-learning for modulation and/or coding scheme selection. However, unimodality over the rates and the contextual information are not taken into account in this work. Similarly, [61] studies intelligent power control in cognitive communications by means of deep reinforcement learning but without exploiting the unimodal structure in power levels. In addition, a deep Q network (DQN) based algorithm for channel selection is proposed in [62]. While this algorithm originally requires an offline dataset for training, it is also extended to work under dynamic environments. Essentially, when a change in the system is detected, then the DQN based algorithm is retrained. Likewise, [42] addresses the problem of learning and adaptation in cognitive radios using neural networks, where backpropagation is used to train a multilayer feedforward neural network.

1.3 Problem Formulation

1.3.1 Description of the MAB Model

For $X, A \in \mathbb{Z}_+$, the set of contexts is given as $\mathcal{X} := \{x_1, x_2, \ldots, x_X\}$, where x_j represents the jth context, and the set of arms is given as $\mathcal{A} := \{a_1, a_2, \ldots, a_A\}$, where a_i represents the ith arm. For $x \in \mathcal{X}$ and $a \in \mathcal{A}$, j_x and i_a represent the indices of context x and arm a respectively, i.e., $x_{j_x} = x$ and $a_{i_a} = a$. Each context-arm pair (x, a) generates a random reward that comes from a fixed but unknown distribution bounded in $[0, 1]$ with expected value given as $\mu(x, a)$. The optimal arm for context x is denoted by $a_x^* := \operatorname{argmax}_{a \in \mathcal{A}} \mu(x, a)$ and its index is given as i_x^*, i.e., $a_x^* = a_{i_x^*}$. The context that gives the highest expected reward for arm a is denoted by $x_a^* := \operatorname{argmax}_{x \in \mathcal{X}} \mu(x, a)$ and its index is given as j_a^*, i.e., $x_a^* = x_{j_a^*}$. Without loss of generality we assume that a_x^* and x_a^* are unique. The suboptimality gap of arm a given context x is defined as $\Delta(x, a) := \mu(x, a_x^*) -$

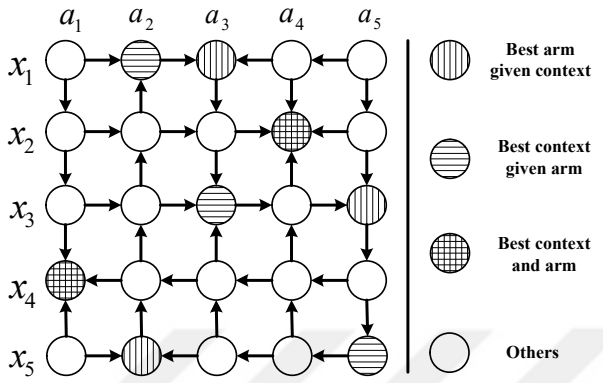


Figure 1.1: An example directed graph over which the expected reward function is jointly unimodal. Each node represents a context-arm pair and the arrows represent the direction of increase in the expected reward.

$\mu(x, a)$.

We assume that the elements of $\mathcal{X}$ and $\mathcal{A}$ are partially ordered, however, this partial order is not known to the learner a priori. The set of neighbors of context x_j (arm a_i) is given as $\mathcal{N}(x_j)$ ($\mathcal{N}(a_i)$). For $j \in \{2, \ldots, X-1\}$ ($i \in \{2, \ldots, A-1\}$), we have $\mathcal{N}(x_j) = \{x_{j-1}, x_{j+1}\}$ ($\mathcal{N}(a_i) = \{a_{i-1}, a_{i+1}\}$). We also have $\mathcal{N}(x_1) = \{x_2\}$ ($\mathcal{N}(a_1) = \{a_2\}$) and $\mathcal{N}(x_X) = \{x_{X-1}\}$ ($\mathcal{N}(a_A) = \{a_{A-1}\}$). We denote the lower indexed neighbor of context x (arm a) by x^- (a^-) and the upper indexed neighbor of context x (arm a) by x^+ (a^+), if they exist. The set of contexts (arms) that have indices lower than and higher than context x (arm a) are denoted by $[x]^-$ ($[a]^-$) and $[x]^+$ ($[a]^+$), respectively.

The system operates in a sequence of rounds indexed by $t \in \{1, 2, \ldots\}$. At the beginning of each round t, the learner observes a context $x(t)$ with index $j(t)$. After observing $x(t)$, the learner selects an arm $a(t)$ with index $i(t)$, and then, observes the random reward $r_{x(t),a(t)}(t)$ associated with the tuple $(x(t), a(t))$. The goal of the learner is to maximize its expected cumulative reward over rounds.

1.3.2 Joint Unimodality of the Expected Reward

We assume that the expected reward function $\mu(x, a)$ exhibits a unimodal structure over both the set of contexts and the set of arms. This structure can be explained via a graph whose vertices correspond to context-arm pairs.

Definition 1. *Let $G := (\mathcal{V}, E)$ be a directed graph over the set of vertices $\mathcal{V} := \{v_{x,a}, x \in \mathcal{X}, a \in \mathcal{A}\}$ connected via edges E (see Fig. 1.1 for an example). $\mu(x, a)$ is called* unimodal in the arms *if for any given context there exist a path from any non-optimal arm to the optimal arm along which the expected reward is strictly increasing. Similarly, $\mu(x, a)$ is called* unimodal in the contexts *if for any given arm there exist a path from any context to the context that gives the maximum expected reward for that particular arm along which the expected reward is strictly increasing. We say that $\mu(x, a)$ is* jointly unimodal *if it is unimodal both in the arms and the contexts.*

Based on Definition 1, joint unimodality implies the following.

(1) For all $x \in \mathcal{X}$:

- If $a_x^* \notin \{a_1, a_A\}$, then $\mu(x, a_1) < \ldots < \mu(x, a_x^*)$ and $\mu(x, a_x^*) > \ldots > \mu(x, a_A)$.
- If $a_x^* = a_1$, then $\mu(x, a_1) > \ldots > \mu(x, a_A)$.
- If $a_x^* = a_A$, then $\mu(x, a_1) < \ldots < \mu(x, a_A)$.

(2) For all $a \in \mathcal{A}$:

- If $x_a^* \notin \{x_1, x_X\}$, then $\mu(x_1, a) < \ldots < \mu(x_a^*, a)$ and $\mu(x_a^*, a) > \ldots > \mu(x_X, a)$.
- If $x_a^* = x_1$, then $\mu(x_1, a) > \ldots > \mu(x_X, a)$.
- If $x_a^* = x_X$, then $\mu(x_1, a) < \ldots < \mu(x_X, a)$.

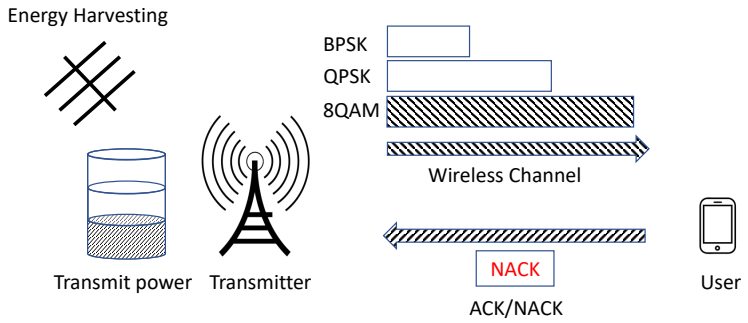


Figure 1.2: Rate selection under time-varying transmit power as a toy example.

As a side note, we emphasize that generalizing unimodal MAB [22] to handle joint unimodality over the set of context-arm pairs is non-trivial due to the fact that the learner does not know the context arrivals a priori and cannot control how they arrive over time. Furthermore, the context that gives the maximum expected reward for each arm and the arm that gives the maximum expected reward for each context can be different for each context and each arm. Since the goal of the learner is to maximize its cumulative reward, it needs to learn a separate optimal arm for each context by exploiting joint unimodality.

1.3.3 A Toy Example

We give power-aware rate selection problem in an energy harvesting wireless network as a toy example (see Fig. 1.2) to understand the structure. Here, arms correspond to different available rates and the context is the transmit power.

We consider a simple *harvest-then-transmit* model, where the transmitter solely relies on the harvesting source, and therefore, the power output from the energy source, is directly used by the load. The instantaneous harvested energy depends on the environmental conditions and varies with time. The expected reward is a unimodal function of the arms (see Fig. 1.3) as well as of the contexts (see Fig. 1.4), since for a given rate the higher value of transmit power (SNR) provides a higher transmission success probability.

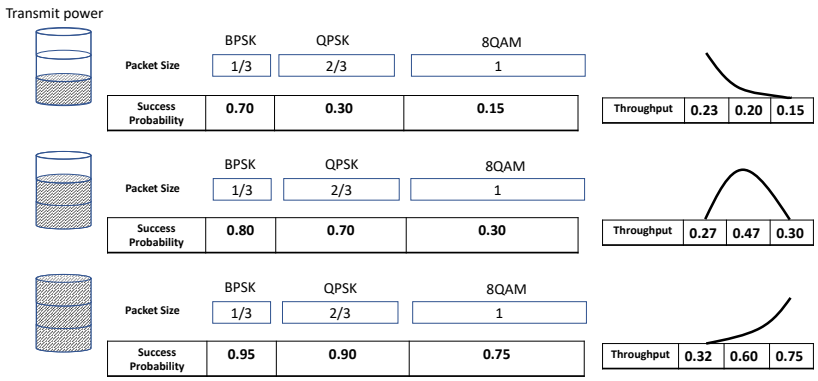


Figure 1.3: For a given transmit power, transmission success probability monotonically decreases with increase in the transmission rate, and therefore, throughput which is obtained by multiplying the transmission success probability with the corresponding transmission rate becomes unimodal function over transmission rates.

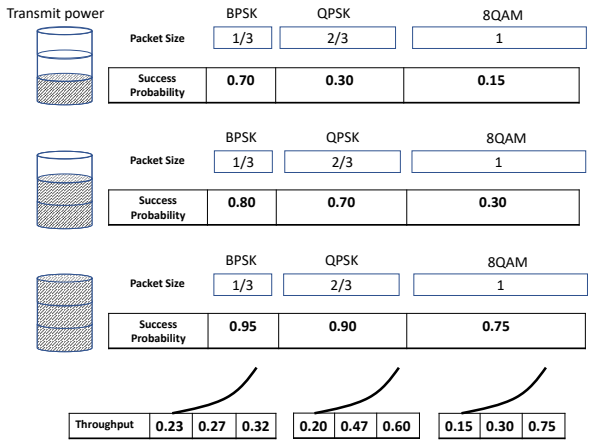


Figure 1.4: For a given transmission rate, transmission success probability monotonically increases with increase in the transmit power, and therefore, throughput which is obtained by multiplying the transmission success probability with the corresponding transmission rate also becomes unimodal function over transmit powers.

1.3.4 Definition of the Regret

Let $N_{x,a}(t)$ be the number of times arm a was selected for context x before round t by the learner and $N_x(t)$ be the number of times context x was observed before round t. The (pseudo) regret of the learner after the first T rounds is given as

$$\begin{aligned} R(T) &:= \sum_{t=1}^{T} \Big(\mu(x(t), a^*_{x(t)}) - \mu(x(t), a(t)) \Big) \\ &= \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \Delta(x, a) N_{x,a}(T+1). \end{aligned} \tag{1.1}$$

It is clear that maximizing the expected cumulative reward translates into minimizing the expected regret $\mathbb{E}[R(T)]$.

1.4 Our Contributions

- Our key contributions in the work using frequentist approach are summarized as follows:
  - We formulate resource allocation problems for rapidly-varying mmWave channels such as dynamic rate selection for an energy harvesting transmitter, dynamic power allocation for heterogeneous applications and distributed resource allocation in a multi-user network as a new structured reinforcement learning problem called contextual unimodal MAB.
  - We propose a learning algorithm called CUL for the contextual unimodal MAB and prove that it achieves improved regret bounds compared to previously known state-of-the-art algorithms, where the expected regret scales logarithmically in time and sublinearly in the number of arms and contexts for a wide range of context arrivals. Our algorithm does not depend on channel model parameters such as indoor/outdoor, line-of-sight (LOS)/non-line-of-sight (NLOS) etc., and hence, can be deployed in any mmWave environment.

  - We show via experiments that CUL provides significant performance improvement compared to the state-of-the-art by using the unimodality of the expected reward jointly in the arms and the contexts.

- Our key contributions in the work using Bayesian approach are summarized as follows:

  - We consider the problem of rate selection under time-varying transmit power over an mmWave channel and formalize the problem as a contextual MAB.
  - We propose a Bayesian learning algorithm called DRS-TS which exploits the structure among rates as well as among transmit powers. We prove that the regret of DRS-TS scales logarithmically in time and the leading term in the regret is independent of the number of rates. To the best of our knowledge, this is the first work that analyses Thompson sampling in a contextual unimodal MAB.
  - We compare DRS-TS with other state-of-the-art learning algorithms and show that it significantly outperforms its competitors via numerical experiments.

- Our key contributions in the work under time-varying constraints are listed below:

  - We cast rate-channel pair selection problem as an online learning problem, and solve it by designing a learning algorithm called Volatile Constrained Thompson Sampling (V-CoTS), which not only cater to the volatility of resource set but also exploits the structure in the success probability over the available transmission rates.
  - We consider learning in an unknown environment, where the SU is initially unaware of the channel characteristics. The available channels need neither to be identically distributed nor to have similar fading characteristics.
  - We provide experimental results on cognitive communications over dynamically varying resource sets, and demonstrate that the proposed algorithm achieves considerably higher performance when compared to

other state-of-the-art methods. We further provide numerical results to demonstrate significant performance gains in the following simplified scenarios: exploiting the monotone structure in the transmission rates for non-volatile channels (see Fig. 4.7), and when the action set is one-dimensional (see Fig. 4.8).

- Our key contributions in the work for continuous arm/contexts are summarized as follows:
  - We consider the contextual multiarmed bandit problem in a continuous and structured stochastic setting where the expected reward is unimodal function over the partially ordered arms and optimal arms are monotone over the contexts.
  - We propose an online learning algorithm, called Structured Contextual Unimodal Bandits (S-CUB), which exploits the structure among arms as well as among contexts. The proposed algorithm uniformly partitions the context space into intervals, and then, learn the best arm for the interval to which arriving context belongs. To the best of our knowledge, this is the first work that consider unimodal and monotone structure in MAB when both the arms and the contexts belong to a continuous interval.
  - We numerically compare S-CUB with earlier works and demonstrate that it is able to significantly outperform its competitors.

1.5 Organization of the Thesis

The rest of this thesis is organized as follows:

In chapter 2, we propose a new methodology for dynamic resource allocation in rapidly-varying mmWave wireless channels that enables extremely fast learning by exploiting the structure among arms (transmission parameters) and contexts (side-information). We discuss some resource allocation problem in wireless networks in Section 2.1. The learning algorithm is proposed in Section 2.2, and its

regret is analyzed in Section 2.3. Experimental results for the proposed resource allocation problems are provided in Section 2.4. The proof of lemmas used in the main theorem are given in Section 2.5 and 2.6.

In chapter 3, we consider the problem of rate selection under time-varying transmit power over an mmWave channel, and propose a Bayesian learning algorithm, called DRS-TS, that exploits the structure of the throughput in rates as well as in transmit powers efficiently. Problem formulation for dynamic rate selection under time varying transmit power is provided in Section 3.1. The learning algorithm is given in Section 3.2. The detailed regret analysis is provided in Section 3.3. Illustrative results which demonstrate the superiority of the proposed algorithm are discussed in Section 3.4.

In Chapter 4, we propose a Bayesian learning algorithm for dynamic rate and channel selection under unknown channel conditions and with time-varying PU user activity and application rate requirements. We discuss the dynamic rate and channel adaptation problem in Section 4.1, and comparison with related works is provided is Section 4.2. Problem formulation for the considered problem is given in Section 4.3, followed by the proposed algorithm in Section 4.4. Experimental results are provided in Section 4.5.

In Chapter 5, we consider the contextual unimodal bandits where both the context and arms are form continuous interval and are structured, and propose a learning algorithm, called S-CUB, that exploits the structure of the expected reward in arms as well as the strucuture of optimal arms in contexts efficiently. We discuss problem formulation for continuous arms and contexts in Section 5.1. The learning algorithm is proposed in Section 5.2 followed by illustrative results in Section 5.3. The summary and conclusions are discussed in Chapter 6.

Chapter 2

Exploiting Structure in Contextual Multi-Armed Bandits via Frequentist Approach

This section was published in [2].[1]

In this chapter, we fuse contextual MAB with unimodal MAB and investigate how learning can be made faster by exploiting unimodality over both arms and contexts. Our proposed algorithm achieves a regret that is $O(\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}\log(T))$, where $\mathcal{X}$ is the finite set of contexts, a_x^* is the arm with the highest expected reward for context x, $\mathcal{N}'(x,a_x^*)$ is its neighbor set and $\gamma_{x,a}\in[0,1]$ is a constant that depends on the context arrival process and the arm selections of the learning algorithm. When the context arrivals are favorable (see Section 2.4 for an example), $\gamma_{x,a}$ is close to zero, and hence, the regret becomes small in the number of contexts. Table 2.1 compares our work with prior works on the contextual MAB and the unimodal MAB.

In this chapter, we propose an AI-based algorithm called Contextual Unimodal

[1] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks", IEEE Transactions on Cognitive Communications and Networking, March 2020.

Table 2.1: Comparison of CUL with state-of-the-art algorithms

Algorithm	Arm unimodality	Contextual	Context unimodality	Regret
KL-UCB-U [5]	✓	×	×	$O(\lvert\mathcal{N}'(a^*)\rvert \log(T))$
Contextual Zooming [20]	×	✓	×	$\tilde{O}(T^{1-1/(2+d_z)})$
CUCB [44]	×	✓	×	$O(XA\log(T))$
CUL (this chapter)	✓	✓	✓	$O(\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}\log(T))$, $\gamma_{x,a}\in[0,1]$

Learning (CUL), which is able to learn very fast by exploiting unimodality jointly over the contexts and the arms. Essentially, CUL exploits unimodality over arms to reduce the number of arms to be explored and unimodality over contexts to select good arms for the current context by using past information from nearby contexts. This results in a regret that increases logarithmically in time and sublinearly both in the number of arms and contexts under a wide range of context arrivals. Exploiting unimodality over contexts is significantly different from exploiting unimodality over arms, since the context arrivals are exogenous, and thus, proving regret bounds for our algorithm requires substantial innovations in technical analysis. Specifically, unimodality over contexts is exploited via comparing the upper and lower confidence bounds of neighboring contexts. Based on this comparison, a modified neighborhood set that contains the contexts that have higher rewards (e.g., throughput) than the current context with high probability is obtained, and the generated set is then used to refine the reward estimates for the current context. This method of reducing explorations enables fast learning.

Most importantly, this new way of learning significantly improves the performance in a variety of resource allocation problems related to mmWave networks compared to the state-of-the-art, and our findings emphasize that instead of working with black-box reinforcement learning models, AI-enabled radios should be designed by considering the structure of the environment.

2.1 Resource Allocation Problems in mmWave Wireless Channels

In this subsection, we detail three resource allocation problems in mmWave wireless channels. We consider a very general mmWave channel model and assume that neither the channel statistics nor the CSI is available. However, we assume that the channel distribution does not change over time. In practice, this assumption can be relaxed to allow abruptly changing or slowly evolving non-stationary channels by designing sliding-window or discounted variants of the proposed algorithm [63,64]. Rather than dealing with this additional complication, we focus on the more fundamental problem of how joint unimodality can be used to achieve fast learning. Our results for the stationary channels indirectly imply that similar gains in performance will be observed by exploiting joint unimodality in non-stationary environments.

In the settings we consider here, the only feedback that the transmitter receives after the transmission of a data packet is ACK/NACK. We assume that there is perfect CRC-based error detection at the receiver and ACK/NACK packets are transmitted over an error-free channel. The *signal-to-noise ratio* (SNR) represents the quality of the channel.

2.1.1 Dynamic Rate Selection for an Energy Harvesting Transmitter [1–9]

It is well known that dynamic rate selection over rapidly varying wireless channels can be modeled as an MAB problem [5, 29, 33]. In the MAB equivalent of the aforementioned problem, in each round the learner selects a modulation scheme, transmits a packet with the rate imposed by the selected modulation scheme, receives as feedback ACK/NACK for the transmitted packet, and collects the expected reward as the rate multiplied by the transmission success probability. It is shown in [5] that this formulation is asymptotically equivalent to maximizing

the number of packets successfully transmitted over a given time horizon.

Consider a power-aware rate selection problem in an energy harvesting mmWave radio network. Here, arms correspond to different available rates and the context is the harvested energy available for transmission. We consider a simple *harvest-then-transmit* model, where the transmitter solely relies on the harvesting source, e.g., a solar cell, a wind turbine or an RF energy source. Therefore, the power output from the energy source, which is denoted by $p(t)$ at time t,[2] is directly used by the load [9]. The instantaneous harvested energy depends on the environmental conditions and varies with time. In practice, transmit power is assigned from a discrete set [65], and hence, the best power management strategy is to match the transmit power to the available harvested energy. Since the optimal rate may be different for each transmit power, traditional dynamic rate selection [5] results in a non-optimal solution. The expected reward is a unimodal function of the arms as discussed in [5] as well as of the contexts, since for a given rate the higher value of transmit power (SNR) provides a higher transmission success probability.

At each time t, a transmit power $p(t) \in \{p_1, \ldots, p_L\}$ is presented to the user, and the user chooses a rate from the set $\mathcal{A} := \{a_1, \ldots, a_A\}$, where L and A are the number of available transmit powers and rates, respectively. The context and arm sets are ordered, i.e., $p_1 < p_2 < \ldots < p_L$ and $a_1 < a_2 < \ldots < a_A$. Let $X_{p,a}(t)$ be a Bernoulli random variable, which represents the success ($X_{p,a}(t) = 1$) or failure ($X_{p,a}(t) = 0$) of the packet transmission for a given power-rate pair (p, a). The (random) reward for power-rate pair (p, a) is given as

$$r_{p,a}(t) = \begin{cases} a/a_A & X_{p,a}(t) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Division with the maximum rate ensures that the rewards lie in the unit interval. The expected reward is given as

$$\mu(p, a) = \mathbb{E}[r_{p,a}] = \left(\frac{a}{a_A}\right) F_{p,a} \tag{2.2}$$

where $F_{p,a} := \mathbb{P}(X_{p,a} = 1)$ is the transmission success probability for power-rate pair (p, a).

[2]Here, t represents the index for data packet transmission.

2.1.2 Dynamic Power Allocation for Heterogeneous Applications [10–12]

Radio networks usually serve heterogeneous users/applications with different QoS requirements [66]. In this section, we consider a setting where the context represents the rate constraint for the current application and the goal is to select the transmission power that maximizes the performance-to-power ratio. At each time t, the transmitter observes the target rate $a \in \{a_1, \ldots, a_A\}$, and chooses a power level from the discrete ordered set $\mathcal{P} := \{p_1, \ldots, p_L\}$, where A and L are the number of available rates and power levels, respectively. The normalized (random) reward of rate-power pair (a, p) is given as

$$r_{a,p}(t) = \begin{cases} p_1/p & X_{p,a}(t) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

and the expected reward of rate-power pair (a, p) is given as

$$\mu(a, p) = \mathbb{E}[r_{a,p}] = \left(\frac{p_1}{p}\right) F_{p,a} \ . \tag{2.4}$$

Here, $\mu(a, p)$ represents the packet success probability to power ratio.

Note that for a fixed transmit power p, transmission success probability monotonically decreases with the rate, i.e., $F_{p,a_1} > F_{p,a_2} > \ldots > F_{p,a_A}$. This implies that given a fixed transmit power p, the expected reward is monotone (hence unimodal) in the contexts, i.e., $(p_1/p)F_{p,a_1} > (p_1/p)F_{p,a_2} > \ldots > (p_1/p)F_{p,a_A}$. In addition, for a given context (rate a), the transmission success probability increases as a function of the transmit power (SNR) [67], i.e., $F_{p_1,a} < F_{p_2,a} < \ldots < F_{p_L,a}$. Hence, when multiplied with (p_1/p) the expected reward is a unimodal function of the transmit power in general (a case in which this holds is given in our numerical experiments), i.e., $(p_1/p_1)F_{p_1,a} < \ldots < (p_1/p_k)F_{p_k,a} > \ldots > (p_1/p_L)F_{p_L,a}$.[3]

[3] A similar discussion for the unimodality of throughput in the transmission rate is given in [5]. In that case, the success probability decreases with r and the throughput, defined as $r\theta_r$, which is the product of an increasing and a decreasing function in r, becomes unimodal.

2.1.3 Distributed Resource Allocation in a Multi-user Network [13–15]

Consider a cooperative multi-player multi-channel setting in which the users select the channels in round robin manner to ensure fairness [68]. Let M be the number of users and $N \geq M$ be the number of channels. Assume that the channels are ranked based on their quality (SNR). Throughput of the users can be maximized by dividing learning into two phases. In the ranking phase, the channel ranks will be estimated, and in the exploitation phase, orthogonal channels will be selected in a round robin manner while the optimal transmission rate is learned for each channel. In this section, we focus on the exploitation phase over the orthogonal channels, and thus assume the channel ranking is known by the users.[4]

The learning problem of a user can be stated as follows. At each time t, based on the round robin schedule, a channel c from a finite set $\mathcal{C} := \{c_1, \ldots, c_N\}$ is provided to the user, and the user chooses a rate from the finite set $\mathcal{A} := \{a_1, \ldots, a_A\}$ for that channel, where A is the number of available rates. Let $X_{c,a}(t)$ be a Bernoulli random variable, which represents the success or failure of the transmission for a given channel-rate pair (c, a). The (random) reward of a user for channel-rate pair (c, a) is given as

$$r_{c,a}(t) = \begin{cases} a/a_A & X_{c,a}(t) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

and the expected reward of channel-rate pair (c, a) is given as

$$\mu(c, a) = \mathbb{E}[r_{c,a}(t)] = \Big(\frac{a}{a_A}\Big) F_{c,a} \tag{2.6}$$

where $F_{c,a} := \mathbb{P}(X_{c,a} = 1)$ is the transmission success probability on channel c at rate a.

[4]This can be achieved by using simple algorithms as in [15].

2.2 The Learning Algorithm

We propose Contextual Unimodal Learning (CUL), an algorithm based on a variant of KL-UCB that takes into account joint unimodality of $\mu(x, a)$ [5, 22, 45] to minimize the expected regret (pseudocode is given in Algorithm 1, graphical demonstration in Fig. 2.1). CUL exploits unimodality of $\mu(x, a)$ in arms in a way similar to KL-UCB-U in [5]. Its main novelty comes from exploiting the contextual information as well as the unimodaliy in contexts, which is substantially different from exploiting the unimodality in arms, since the learner does not have any control over how the contexts arrive.

For each context-arm pair (x, a), CUL keeps the sample mean estimate of the rewards obtained from rounds in which context was x and arm a was selected prior to the current round, denoted by $\hat{\mu}_{x,a}$, and the number of times arm a was selected when the context was x prior to the current round, denoted by $N_{x,a}$. Values of these parameters at the beginning of round t are denoted by $\hat{\mu}_{x,a}(t)$ and $N_{x,a}(t)$, respectively.

The *leader* for context $x \in \mathcal{X}$ in round t is defined as the arm with the highest sample mean reward, i.e.,

$$L_x(t) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, \hat{\mu}_{x,a}(t) \text{ (ties are broken arbitrarily).}$$

Letting $\mathbf{1}(\cdot)$ denote the indicator function, we define

$$b_{x,a}(t) = \sum_{t'=1}^{t-1} \mathbf{1}(x(t') = x, a = L_x(t'))$$

as the number of times arm a was a leader when the context was x up to (before) round t. After observing $x(t)$ in round t, CUL identifies the leader $L_{x(t)}(t)$ and calculates $b_{x(t),L_{x(t)}(t)}(t)$. If

$$\frac{b_{x(t),L_{x(t)}(t)}(t) - 1}{3} \in \mathbb{N}$$

CUL selects the leader (exploitation). Similar to KL-UCB-U [5], this ensures that the number of times an arm has been the leader bounds the number of

Algorithm 1 CUL

Input: X, A
Initialize: $j_a^+(0) = 1, j_a^-(0) = X,\ \forall a \in \mathcal{A}, t = 1$
Counters: $N_{x,a}(1) = 0, \hat{\mu}_{x,a}(1) = 0, b_{x,a}(1) = 0, \forall a \in \mathcal{A}, \forall x \in \mathcal{X}$
while $t \geq 1$ **do**
    Observe context $x(t)$
    $L_{x(t)}(t) = \text{argmax}_{a \in \mathcal{A}}\, \hat{\mu}_{x(t),a}(t)$
    if $\frac{b_{x(t),L_{x(t)}(t)}(t)-1}{3} \in \mathbb{N}$
        $a(t) = L_{x(t)}(t)$
    else
        $\mathcal{T} = \{L_{x(t)}(t)\} \cup \mathcal{N}(L_{x(t)}(t))$
        for $a \in \mathcal{T}$
            Calculate $u_{x_j,a}(t)$ (2.7), $l_{x_j,a}(t)$ (2.9), $\forall x_j \in \mathcal{X}$
            $I^+_{x_j,a}(t) = \mathbf{1}\{l_{x_j,a}(t) \geq u_{x_j^-,a}(t)\},\ \forall x_j \in \mathcal{X}$
            $I^-_{x_j,a}(t) = \mathbf{1}\{l_{x_j,a}(t) \geq u_{x_j^+,a}(t)\},\ \forall x_j \in \mathcal{X}$
            $j_a^+(t) = \max\{j : I^+_{x_j,a}(t) = 1\}$
            $j_a^-(t) = \min\{j : I^-_{x_j,a}(t) = 1\}$
            Find $\mathcal{U}_{x(t),a}(t)$ using (2.10)
            $\underline{u}_{x(t),a}(t) = \min_{x' \in \{\mathcal{U}_{x(t),a}(t)\ \cup\ x(t)\}} u_{x',a}(t)$
        end for
        $a(t) = \text{argmax}_{a \in \mathcal{T}}\, \underline{u}_{x(t),a}(t)$
    end if
    Observe reward $r_{x(t),a(t)}(t)$
    Update parameters for $(x(t), a(t))$ (other parameters retain their values in round t):
    $\hat{\mu}_{x(t),a(t)}(t+1) = \frac{\hat{\mu}_{x(t),a(t)}(t) N_{x(t),a(t)} + r_{x(t),a(t)}(t)}{N_{x(t),a(t)}(t)+1}$
    $N_{x(t),a(t)}(t+1) = N_{x(t),a(t)}(t) + 1$
    $b_{x(t),L_{x(t)}(t)}(t+1) = b_{x(t),L_{x(t)}(t)}(t) + 1$
    $t = t + 1$
end while

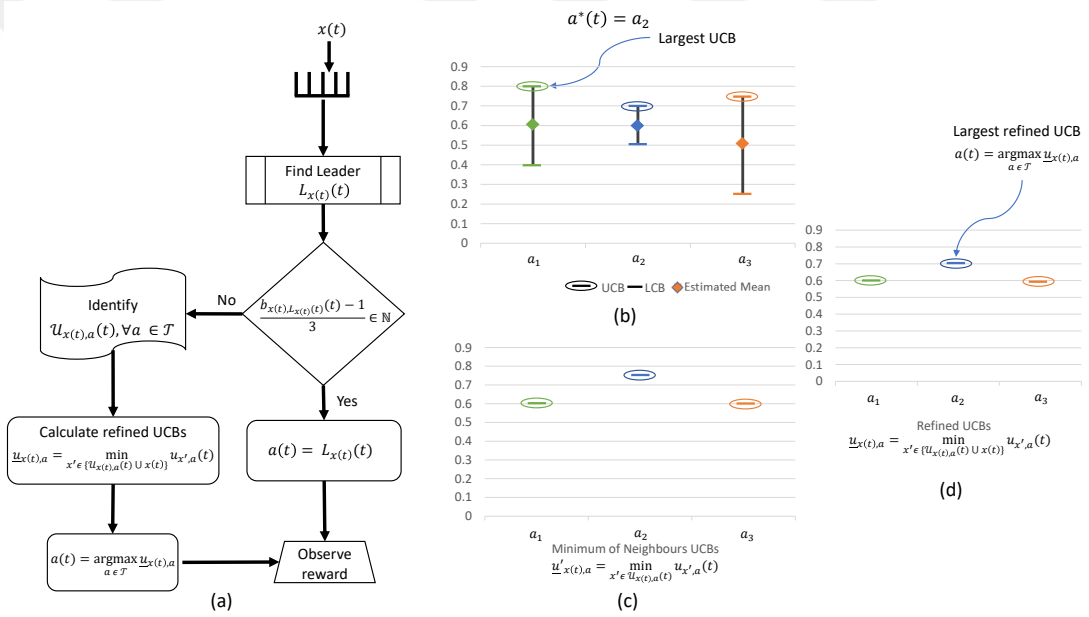


Figure 2.1: Graphical demonstration of CUL algorithm for an example arm set: (a) Flow chart of CUL (b) In a given time slot t, a_2 is the optimal arm for a given context $x(t)$, a_2 is the leader and we have $\mathcal{T} = \{a_1, a_2, a_3\}$. If contextual unimodality is ignored, a_1 is selected based on calculated UCBs to cater to the exploration. (c) As an intermediate step, minimum of the neigbours UCBs in $\mathcal{U}_{x(t),a}(t)$ for arms in $\mathcal{T}$ are shown. (d) Refined UCBs are calculated by using $\mathcal{U}_{x(t),a}(t)$ and $x(t)$ for arms in $\mathcal{T}$, and therefore, utilizing the contextual unimodality results in selection of the optimal arm a_2 in the current time slot t.

times that arm has been selected. Otherwise, CUL judiciously tries to balance exploration and exploitation by using joint unimodality. Essentially, it restricts the exploration only to the current leader and arms which lie in the neighborhood of the current leader in round t, given by $\mathcal{T} = \{L_{x(t)}(t)\} \cup \mathcal{N}(L_{x(t)}(t))$. This restricted exploration strategy works due to the fact that there are no local optima due to unimodality, and hence, CUL always finds the direction towards the global optimum.

In order to select an arm from $\mathcal{T}$, the Bernoulli KL-UCB index for all (x, a) such that $a \in \mathcal{T}$ is calculated as

$$u_{x,a}(t) = \max \left\{ f \in [0, K_a] : N_{x,a}(t) d \left(\frac{\hat{\mu}_{x,a}(t)}{K_a}, \frac{f}{K_a} \right) \leq \log(t) + 3 \log(\log(t)) \right\} \tag{2.7}$$

where $K_a \in [0, 1]$ is the normalization constant set to a/a_A for rate selection and p_1/p for power allocation applications given in Section 2.1. Here, $d(p, q)$ represents the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q, and is given as

$$d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \tag{2.8}$$

with $0 \log 0/0 = 0$ and $x \log x/0 = +\infty$ for $x > 0$. Likewise, the Bernoulli KL-LCB (lower confidence bound) index is calculated as

$$l_{x,a}(t) = \min \left\{ f \in [0, K_a] : N_{x,a}(t) d \left(\frac{\hat{\mu}_{x,a}(t)}{K_a}, \frac{f}{K_a} \right) \leq \log(t) + 3 \log(\log(t)) \right\}. \tag{2.9}$$

For a context x for which $N_x(t)$ is small, $u_{x,a}(t)$ can be much higher than $\mu(x, a)$. In order to utilize joint unimodality to learn faster, the learner should refine its UCB for $(x(t), a)$ using rewards collected from arm a in other contexts. For instance, if the learner knows with a high probability that $\mu(x', a) > \mu(x(t), a)$ for all x' in some subset $\mathcal{U}_{x(t),a}(t)$ of $\mathcal{X}$, then it can simply set its refined UCB as

$$\underline{u}_{x(t),a}(t) = \min_{x' \in \{\mathcal{U}_{x(t),a}(t) \ \cup \ x(t)\}} u_{x',a}(t).$$

After this, it will select the arm in $\mathcal{T}$ with the highest refined UCB, i.e., $a(t) = \text{argmax}_{a \in \mathcal{T}} \underline{u}_{x(t),a}(t)$.

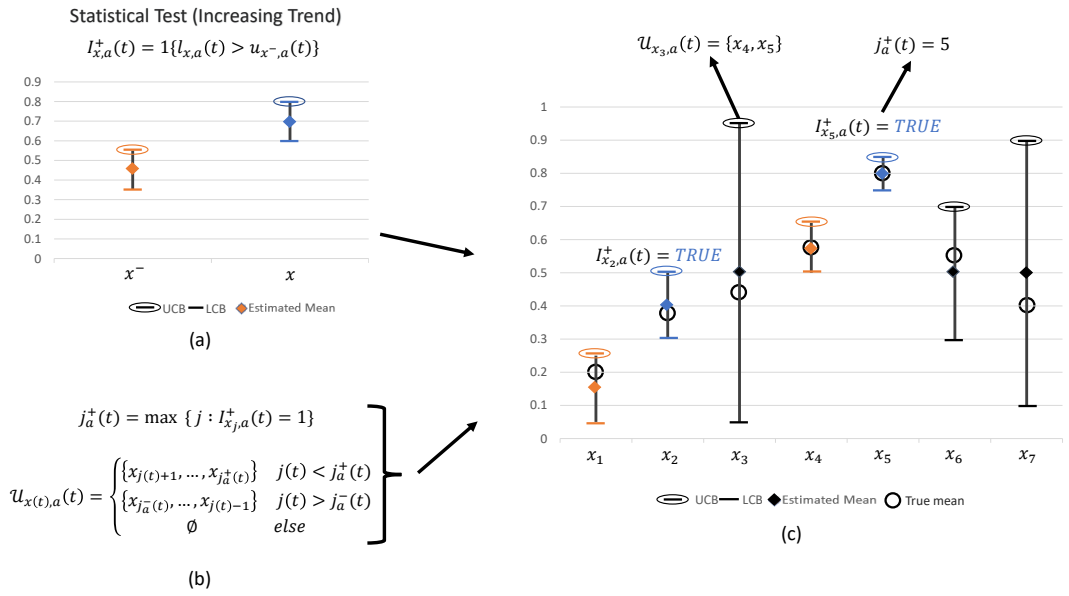


Figure 2.2: Graphical demonstration of finding the increasing trend in contexts: (a) Statistical test for increasing trend over contexts for a given arm a in time slot t. (b) $j^+_a(t)$ represents the index of the context with the highest expected reward with high probability, and $\mathcal{U}_{x(t),a}(t)$ for a given context contains the contexts with higher values of expected rewards with high probability. (c) An example with 7 contexts for a given arm a is shown, statistical tests for x_2 and x_5 are TRUE, and therefore, $j^+_a(t)$ is 5 i.e., the highest index for which statistical test is true, and $\mathcal{U}_{x_3,a}(t)$ for context x_3 contains x_4 and x_5 which have higher expected rewards than x_3.

Next, we explain how such a subset $\mathcal{U}_{x,a}(t)$ is constructed by CUL. As a first step, the learner needs to infer the direction of increase of the expected reward of arm a as a function of the context. For this, the *increasing trend in contexts indicator* $I^+_{x,a}(t)$ of context-arm pair (x, a) in round t is defined as

$$I^+_{x,a}(t) = \mathbf{1}\{l_{x,a}(t) > u_{x^-,a}(t)\}.$$

$I^+_{x,a}(t) = 1$ implies that (x^-, a) has a lower expected reward than (x, a) with a high probability, which implies due to unimodality that (x', a) has a lower expected reward than (x, a) for all $x' \in [x]^-$ with a high probability. For each $a \in \mathcal{A}$, let $j^+_a(0) = 1$ and $j^+_a(t) = \max\{j : I^+_{x_j,a}(t) = 1\}$. If $j^+_a(t)$ is empty, then $j^+_a(t) = j^+_a(0)$ (see Fig. 2.2 for demonstration).

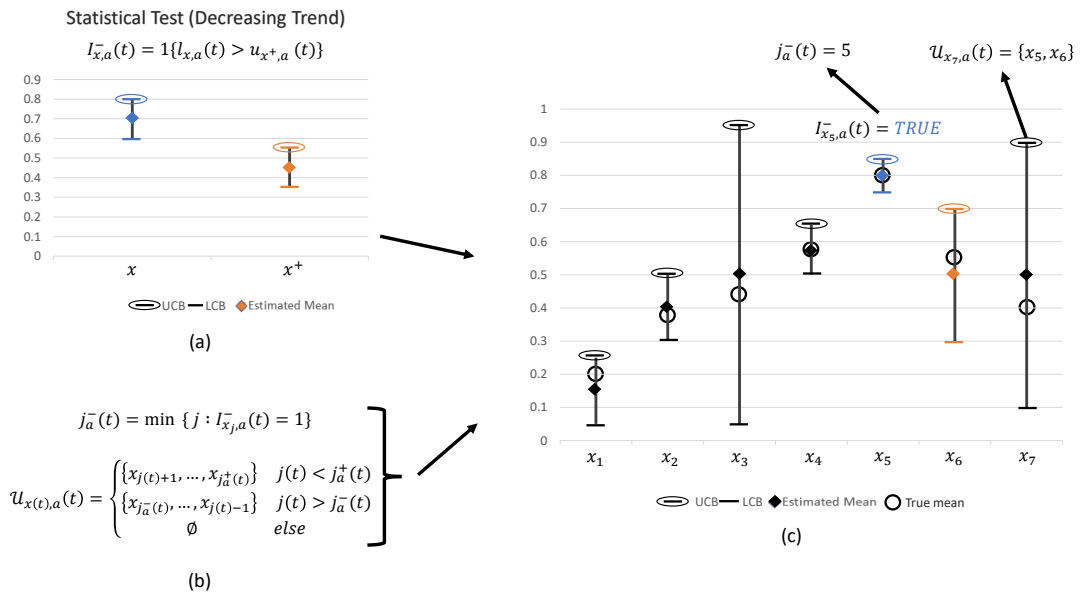


Figure 2.3: Graphical demonstration of finding the decreasing trend in contexts: (a) Statistical test for decreasing trend over contexts for a given arm a in time slot t. (b) $j^-_a(t)$ represents the index of the context with the highest expected reward with high probability, and $\mathcal{U}_{x(t),a}(t)$ for a given context contains the contexts with higher values of expected rewards with high probability. (c) An example with 7 contexts for a given arm a is shown, statistical tests for x_5 is TRUE, and therefore, $j^-_a(t)$ is 5 i.e., the lowest index for which statistical test is true, and $\mathcal{U}_{x_7,a}(t)$ for context x_7 contains x_5 and x_6 which have higher expected rewards than x_7.

Similarly, the *decreasing trend in contexts indicator* $I^-_{x,a}(t)$ of context-arm pair (x, a) in round t is defined as

$$I^-_{x,a}(t) = \mathbf{1}\{l_{x,a}(t) > u_{x^+,a}(t)\}.$$

$I^-_{x,a}(t) = 1$ implies that (x^+, a) has a lower expected reward than (x, a) with a high probability, which again implies due to unimodality that (x', a) has a lower expected reward than (x, a) for all $x' \in [x]^+$ with a high probability. For each $a \in \mathcal{A}$, let $j^-_a(0) = X$ and $j^-_a(t) = \min\{j : I^-_{x_j,a}(t) = 1\}$. If $j^-_a(t)$ is empty, then $j^-_a(t) = j^-_a(0)$ (see Fig. 2.3 for demonstration).

Based on the definitions above, the following occurs with a high probability in round t:

- If $j(t) < j_a^+(t)$, then $\mu(x(t), a) < \mu(x_j, a)$, $\forall j \in \{j(t)+1, \ldots, j_a^+(t)\}$.
- If $j(t) > j_a^-(t)$, then $\mu(x(t), a) < \mu(x_j, a)$, $\forall j \in \{j_a^-(t), \ldots, j(t)-1\}$.

Based on this, when $j_a^+(t) \leq j_a^-(t)$ or $j_a^+(t) > j_a^-(t)$ and $j(t) \notin (j_a^-(t), j_a^+(t))$ happen, we let

$$\mathcal{U}_{x(t),a}(t) = \begin{cases} \{x_{j(t)+1}, \ldots, x_{j_a^+(t)}\} & \text{if } j(t) < j_a^+(t) \\ \{x_{j_a^-(t)}, \ldots, x_{j(t)-1}\} & \text{if } j(t) > j_a^-(t) \\ \emptyset & \text{otherwise.} \end{cases} \tag{2.10}$$

Since, KL-UCB index is an upper bound on true mean and KL-LCB index is a lower bound on true mean with high probability, the event $\{j_a^+(t) \leq j_a^-(t)\}$ happens with high probability, which in turn ensures the fact that $\mu(x', a) > \mu(x, a), \forall x' \in \mathcal{U}_{x(t),a}(t)$ due to the unimodality assumption. On the other hand, if KL-UCB of at least one context in $\mathcal{X}$ underestimates its expected value or its KL-LCB over-estimates the expected value for arm a, the event $\{j_a^+(t) > j_a^-(t)\}$ may happen. In this case, if $\{j_a^+(t) > j(t) > j_a^-(t)\}$, then $\mathcal{U}_{x(t),a}(t)$ will be the union of $\{x_{j(t)+1}, \ldots, x_{j_a^+(t)}\}$ and $\{x_{j_a^-(t)}, \ldots, x_{j(t)-1}\}$, which in turn allows $\mu(x', a) < \mu(x, a)$ for a context $x' \in \mathcal{U}_{x(t),a}(t)$. The bound on probability of the event $\mu(x', a) < \mu(x, a)$ for some context $x' \in \mathcal{U}_{x(t),a}(t)$ is given in Lemma 2, and is used in (2.14).

Intuitively, given a particular arm, when the neighbors of the current context are sufficiently learned, and the expected reward at a neighbor is greater than the expected reward at the current context, then the UCB of the neighbor will be higher than the true mean reward at the current context, and thus, it can be used as a UCB for the current context. Note that exploiting the unimodality over contexts depends on how the contexts arrive. We illustrate how context arrivals affect the regret of CUL in Section 2.4.7.

2.3 Regret Analysis of CUL

In this section, we bound the expected regret of CUL.

2.3.1 Preliminaries

The expected regret can be rewritten as

$$\mathbb{E}[R(T)] = \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \Delta(x,a)\mathbb{E}[N_{x,a}(T+1)]. \tag{2.11}$$

Let $\tau_x(t)$ denote the round in which context x arrives for the tth time. Let $\tilde{N}_{x,a}(t) := N_{x,a}(\tau_x(t))$, $\tilde{\mu}_{x,a}(t) := \hat{\mu}_{x,a}(\tau_x(t))$, $\tilde{a}_x(t) := a(\tau_x(t))$, $\tilde{b}_{x,a}(t) := b_{x,a}(\tau_x(t))$, $\tilde{L}_x(t) := L_x(\tau_x(t))$, $\tilde{u}_{x,a}(t) := u_{x,a}(\tau_x(t))$, $\tilde{l}_{x,a}(t) := l_{x,a}(\tau_x(t))$, and $\tilde{\underline{u}}_{x,a}(t) := \underline{u}_{x,a}(\tau_x(t))$.

Let $\tilde{y}_{x,a}(t) := \text{argmin}_{\{x' \in \mathcal{U}_{x,a}(\tau_x(t))\}} u_{x',a}(\tau_x(t))$ denote the target context for a context-arm pair (x,a) in round $\tau_x(t)$ (if it exists). Next, we introduce two variables: $\tilde{N}^x_{x',a'}(t)$ and $\tilde{\mu}^x_{x',a'}(t)$. The first one is the number of times arm a' has been selected when the context was x' up to round $\tau_x(t)$, and the second one is the sample mean reward of context-arm pair (x',a') at the beginning of round $\tau_x(t)$. Similarly, $\tilde{u}^x_{x',a'}(t)$ and $\tilde{l}^x_{x',a'}(t)$ are the KL-UCB index and KL-LCB index of context-arm pair (x',a') at the beginning of round $\tau_x(t)$, respectively. For ease of notation, we denote these parameters for target context $\tilde{y}_{x,a}(t)$ of context-arm pair (x,a) as $\tilde{M}_{x,a}(t) := \tilde{N}^x_{\tilde{y}_{x,a}(t),a}(t)$, $\tilde{\eta}_{x,a}(t) := \tilde{\mu}^x_{\tilde{y}_{x,a}(t),a}(t)$, $\tilde{w}_{x,a}(t) := \tilde{u}^x_{\tilde{y}_{x,a}(t),a}(t)$, and $\tilde{o}_{x,a}(t) := \tilde{l}^x_{\tilde{y}_{x,a}(t),a}(t)$.

Similarly, let $\tau_{x,a}(s)$ denote the round in which context is x and arm a is selected for the sth time. Let $\tilde{y}^s_{x,a}$ denote the target context in round $\tau_{x,a}(s)$, $\tilde{N}^{x,a,s}_{x',a'}$ denote the number of times arm a' has been selected when the context was x' up to round $\tau_{x,a}(s)$, and $\tilde{\mu}^{x,a,s}_{x',a'}$ denote the sample mean reward of context-arm pair (x',a') at the beginning of round $\tau_{x,a}(s)$. When $(x',a') = (x,a)$, we simply write $\tilde{N}^s_{x,a} := \tilde{N}^{x,a,s}_{x',a'}$ and $\tilde{\mu}^s_{x,a} := \tilde{\mu}^{x,a,s}_{x',a'}$. Thus, we have $\tilde{N}^s_{x,a} = (s-1)$ and $\tilde{\mu}^s_{x,a} = \frac{1}{(s-1)} \sum_{k=1}^{(s-1)} Y^k_{x,a}$, where $Y^k_{x,a}$ is the reward of arm a when it is selected for kth time when the context is x, with convention $\tilde{\mu}^s_{x,a} = 0$ for $s = 1$. When $s = \tilde{N}_{x,a}(t) + 1$, we have $\tilde{\mu}_{x,a}(t) = \tilde{\mu}^s_{x,a}$. For ease of notation, we represent the target context $(\tilde{y}^s_{x,a})$ dependent quantities $\tilde{N}^{x,a,s}_{\tilde{y}^s_{x,a},a}$ as $\tilde{M}^s_{x,a}$ and $\tilde{\mu}^{x,a,s}_{\tilde{y}^s_{x,a},a}$ as $\tilde{\eta}^s_{x,a}$. If the refined neighborhood is an empty set, there is no target context, and quantities related to it are zero, i.e., $\tilde{M}^s_{x,a} = 0$ and $\tilde{\eta}^s_{x,a} = 0$.

It is important to note that if there exists an arm $a' \in \mathcal{N}(a_x^*)$ for context x such that $K_{a'} < \mu(x, a_x^*)$, then for $\mu(x, a_x^*) \leq \underline{\tilde{u}}_{x,a_x^*}(t)$, a' can never be selected. Since by definition in (2.7) we have $\tilde{u}_{x,a'}(t) \in [0, K_{a'}]$, and hence, $\underline{\tilde{u}}_{x,a'}(t) \leq \tilde{u}_{x,a'}(t) < \mu(x, a_x^*) < \underline{\tilde{u}}_{x,a_x^*}(t), \forall t$ i.e., the refined index of arm a' is always less than refined index of a_x^*. For such an arm a', if it exists, we define $\mathcal{N}'(x, a_x^*) := \mathcal{N}(a_x^*) \backslash a'$, otherwise $\mathcal{N}'(x, a_x^*) := \mathcal{N}(a_x^*)$.

For $\frac{x}{K_a}, \frac{y}{K_a} \in [0, 1]$, let $d_a(x, y) := d(\frac{x}{K_a}, \frac{y}{K_a})$, $d_a^+(x, y) := d_a(x, y)\mathbf{1}\{x < y\}$, and $f(t) := \log(t) + 3\log(\log(t))$. For context x and $a \in \mathcal{N}'(x, a_x^*)$, let

$$K_{x,a}^T := \left\lfloor \frac{1+\epsilon}{d_a^+(\mu(x,a), \mu(x,a_x^*))} f(T) \right\rfloor$$

for some $\epsilon > 0$.

2.3.2 Main Result

Theorem 1. *For all $\epsilon > 0$, there exist constants $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that the expected regret of CUL satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \Delta(x,a) \Big((1+\epsilon) \frac{\gamma_{x,a} \log(T)}{d(\frac{\mu(x,a)}{K_a}, \frac{\mu(x,a_x^*)}{K_a})} + \gamma_{x,a} + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}} \Big) + O\big(\log(\log(T))\big)$$

where $\gamma_{x,a} := \frac{\sum_{s=1}^{K_{x,a}^T+1} \mathbb{P}\left(\tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) < f(T)\right)}{K_{x,a}^T+1} \in [0,1]$.

The term $\gamma_{x,a}$ depends on how well the target context is learned for a context-arm pair (x, a). When there are no arrivals to the target context, $\gamma_{x,a}$ is 1. If the number of samples from the target contexts of context-arm pair (x, a) is small, then the value of $\gamma_{x,a}$ is close to 1, and decreases as the target context becomes more and more confident about arm a. If we extend the analysis in [5] to the contextual case without contextual unimodality, the upper bound contains $\sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \log(T)$. The term $\gamma_{x,a} \in [0, 1]$ ensures the fact that $\sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \gamma_{x,a} \log(T) \leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \log(T)$. Its exact value depends on the context arrivals to the target context and number of selections

of arm a at the target context. In short, as the confidence of the target context about any arm a increases, the regret of CUL decreases. The effect of context arrivals on $\gamma_{x,a}$ and the regret is discussed in Section 2.4 via numerical experiments.

2.3.3 Proof of Theorem 1

First, we state two lemmas that will be used in the proof. Their proofs are given in Sections 2.5 and 2.6.

Lemma 1. *For $a \in \mathcal{N}'(x, a_x^*)$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \mu(x, a_x^*) \leq \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\}\right]$$
$$\leq \sum_{s=1}^{T} \mathbb{P}((s-1)d_a^+(\tilde{\mu}_{x,a}^s, \mu(x, a_x^*)) < f(T), \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x, a_x^*)) < f(T)) \ .$$

Lemma 2. *For all (x, a) and $\delta = \log(\tau_x(t)) + 3\log(\log(\tau_x(t)))$, we have*

$$\mathbb{P}(\mu(x, a) > \underline{\tilde{u}}_{x,a}(t)) \leq 2(X+1)e^{\lceil \delta \log(\tau_x(t)) \rceil} \exp(-\delta).$$

Next, we proceed with the proof. For x such that $N_x(T+1) > 0$ and $a \neq a_x^*$, the expectation in (2.11) is decomposed into two terms as in Theorem 4.2(a) [22]:

$$\mathbb{E}[N_{x,a}(T+1)] = \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{a}_x(t) = a\}\right]$$
$$= \mathbb{E}\Bigg[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\} + \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\}\Bigg].$$

We say that an arm a is a suboptimal arm for a given context x if $\Delta(x, a) > 0$. The first term inside the expectation corresponds to the number of times a_x^* is not the leader and the suboptimal arm a is selected, whereas the second term indicates the number of times a_x^* is the leader and the suboptimal arm a is selected. Since only the leader and its neighbors are explored, when the leader is a_x^* we only

select from arms that lie in $\{\mathcal{N}(a_x^*) \cup a_x^*\}$. Therefore, the expected regret for $a \neq a_x^*$ can be rewritten as

$$\begin{aligned}
&\mathbb{E}[R(T)] \\
&= \sum_{x=x_1}^{x_X} \Bigg(\sum_{a \neq a_x^*} \Delta(x,a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\} \right] \\
&+ \sum_{a \in \mathcal{N}(a_x^*)} \Delta(x,a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\} \right] \Bigg). \qquad (2.12)
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \Delta(x,a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\} \right] \\
&\leq \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\} \right] \\
&\leq \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a\} \right] \\
&\leq \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}[b_{x,a}(T+1)].
\end{aligned}$$

Similar to Theorem C.1 in [22], a suboptimal arm a can be the leader for a given context x only for a small number of times (in expectation). Thus, we have $\mathbb{E}[b_{x,a}(T+1)] = O(\log(\log(T)))$.

For $a \in \mathcal{N}(a_x^*)$, we decompose the expectation in the second term in (2.12) as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\} \right] \\
&\leq \mathbb{E}\Big[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\mu(x,a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\} \Big] \\
&+ \mathbb{E}\Big[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \mu(x,a_x^*) \leq \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\} \Big]. \qquad (2.13)
\end{aligned}$$

The first term on the right hand side (r.h.s.) of the inequality corresponds to the event that the refined index of the optimal arm a_x^* underestimates its expected

value, and the second term is the event that the refined index of the optimal arm a_x^* is an upper bound on its expected value and the suboptimal arm a is selected. We bound the first term in (2.13) by using the concentration inequality in Lemma 2 as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\mu(x,a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\}\right] \\
&\quad= \sum_{t=1}^{N_x(T+1)} \mathbb{P}\{\mu(x,a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\} \\
&\quad\le \sum_{t=1}^{T} \left(2(X+1)e^{\lceil \log(t)(\log(t)+3\log(\log(t)))\rceil} \exp(-\log(t)-3\log(\log(t)))\right) \\
&\quad\le \sum_{t=1}^{T} \frac{2(X+1)e^{\lceil \log(t)^2+3\log(t)\log(\log(t))\rceil}}{t\log(t)^3} \\
&\quad\le C_1 \log(\log(T)) \qquad (2.14)
\end{aligned}$$

for a positive constant C_1 such that $C_1 \le 14X + 14$.

Next, we bound the second term in (2.13) by using Lemma 1 and the facts that when $\mu(x,a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t)$ and $\tilde{L}_x(t) = a_x^*$, the only suboptimal arms that can be selected are in $\mathcal{N}'(x,a_x^*)$, and for any two events A and B, $\mathbb{P}(A,B) \le \mathbb{P}(A)$. Thus, we have for $a \in \mathcal{N}'(x,a_x^*)$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \mu(x,a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\}\right] \\
&\le \mathbb{E}\left[\sum_{s=1}^{T} \mathbf{1}\big((s-1)d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) < f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) < f(T)\big)\right] \\
&\le \sum_{s=1}^{K_{x,a}^T+1} \mathbb{P}\big((s-1)d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) < f(T), \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) < f(T)\big) \\
&\quad+ \sum_{s=K_{x,a}^T+2}^{T} \mathbb{P}\big((s-1)d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) < f(T), \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) < f(T)\big) \\
&\le \sum_{s=1}^{K_{x,a}^T+1} \mathbb{P}\big(\tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) < f(T)\big)
\end{aligned}$$

$$+ \sum_{s=K^T_{x,a}+2}^{T} \mathbb{P}\big((s-1)d_a^+(\tilde{\mu}^s_{x,a}, \mu(x,a^*_x)) < f(T)\big). \tag{2.15}$$

We represent the first term in (2.15) as

$$\sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}\big(\tilde{M}^s_{x,a} d_a^+(\tilde{\eta}^s_{x,a}, \mu(x,a^*_x)) < f(T)\big) = \gamma_{x,a}(K^T_{x,a}+1)$$

where $\gamma_{x,a} = \frac{\sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}\left(\tilde{M}^s_{x,a} d_a^+(\tilde{\eta}^s_{x,a}, \mu(x,a^*_x)) < f(T)\right)}{K^T_{x,a}+1}$ and lies in $[0,1]$.

We bound the second term in (2.15) by using Lemma 8 in [45] as,

$$\begin{aligned}
&\sum_{s=K^T_{x,a}+2}^{T} \mathbb{P}((s-1)d_a^+(\tilde{\mu}^s_{x,a}, \mu(x,a^*_x)) < f(T)) \\
&\quad \leq \sum_{s=K^T_{x,a}+2}^{\infty} \mathbb{P}((K^T_{x,a}+1)d_a^+(\tilde{\mu}^s_{x,a}, \mu(x,a^*_x)) < f(T)) \\
&\quad \leq \sum_{s=K^T_{x,a}+2}^{\infty} \mathbb{P}(d_a^+(\tilde{\mu}^s_{x,a}, \mu(x,a^*_x)) < \frac{d_a(\mu(x,a), \mu(x,a^*_x))}{1+\epsilon}) = \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}
\end{aligned}$$

where $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ are the constants from Lemma 8 in [45]. Using above bounds, we obtain

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{x=x_1}^{x_X} \sum_{a \neq a^*_x} \Delta(x,a)\mathbb{E}[N_{x,a}(T+1)] \\
&\leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a^*_x)} \Delta(x,a)\Big(\gamma_{x,a}(K^T_{x,a}+1) + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}\Big) + O\big(\log(\log(T))\big) \\
&\leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a^*_x)} \Big((1+\epsilon)\gamma_{x,a}\log(T)\frac{\Delta(x,a)}{d(\frac{\mu(x,a)}{K_a}, \frac{\mu(x,a^*_x)}{K_a})} \\
&\quad + \gamma_{x,a}\Delta(x,a) + \Delta(x,a)\frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}\Big) + O(\log(\log(T))).
\end{aligned}$$

CUL achieves improved regret bounds compared with the state-of-the-art under a wide range of context arrivals. We note that the time averaged regret, given as $\mathbb{E}[R(T)]/T$ gives the rate of convergence of the cumulative reward of the algorithm to that of the optimal oracle strategy. The regret bound given in Theorem 1 not only proves that $\lim_{T\to\infty} \mathbb{E}[R(T)]/T = 0$ but also implies that the cumulative

performance of CUL is closer to that of the optimal oracle strategy compared to the state-of-the-art under a wide range of context arrivals. We further highlighted this in Table 2.1, which compares CUL with other-state-of-the-art learning algorithms. We also note that our regret bound given in Theorem 1 holds for any sequence of context arrivals and does not require contexts generated by the environment to have a stochastic pattern. This allows CUL to work efficiently in many different environments ranging from contexts arriving uniformly at random to periodically over time as shown in Section 2.4.

2.4 Experiments

In order to evaluate the performance of CUL, we perform multiple experiments for the applications given in Section 2.1, and compare the performance with the following algorithms.

KL-UCB-U [5]: The MAB algorithm that uses KL-UCB indices to exploit unimodality in arms only. This algorithm neglects the contexts.

Sliding Window Graphical Optimal Rate Sampling (SW-G-ORS) [29]: A unimodal MAB algorithm designed to work in non-stationary environments by using a sliding window of length τ. This algorithm also neglects the contexts.

Contextual UCB: Runs a different instance of UCB1 [44] for each context.

Since the actions in the considered applications are ordered, the unimodal graph given as input to KL-UCB-U and SW-G-ORS is a straight line. For SW-G-ORS, the length of sliding window is chosen as $\tau = T/50$.

2.4.1 5G Channel Model and Traces

For experiments, we create traces via performing simulations using communication and 5G toolbox in MATLAB. For multiple transmit power and rate pairs, we send packets through a TDL channel with Rayleigh fading to estimate packet success probabilities. We note that for the Rayleigh fading channel, probability density function of the instantaneous received SNR is given as

$$p_\gamma(\gamma) = \frac{1}{\overline{\gamma}} \exp(-\frac{\gamma}{\overline{\gamma}}) \tag{2.16}$$

where $\overline{\gamma}$ represents the average channel SNR. Then, Bernoulli random numbers are generated by using the obtained probabilities, and random rewards in each experiment are generated by using definitions in (2.1), (2.3) and (2.5). The presented results are averaged over 50 experiment runs.

2.4.2 Performance Metrics

We define the throughput error at time t as the difference of the achieved throughput by the algorithm and the oracle's throughput (optimal throughput) at that time, the *throughput error* in an experiment as the average over all t, and the *averaged throughput error* (averaged performance-to-power error in case of dynamic power allocation) as an average over all repetitions of the experiment. Furthermore, accuracy is defined as the number of times the optimal resource is selected by the algorithm and is calculated in percentage, and the *averaged accuracy* as the average over all repetitions of the experiment. The comparison of averaged throughput error and averaged accuracy of CUL with state-of-the-art algorithms for applications discussed in Section 2.1 are shown in Tables 2.2, 2.3 and 2.4.

2.4.3 Experiment 1: Dynamic Rate Selection for an Energy Harvesting Transmitter (Fig. 2.4-2.6, Table 2.2)

We consider the problem in Section 2.1.1. In this experiment, arms are the transmission rates and contexts are the transmit powers imposed based on the harvested energy. The modulation scheme is selected from a discrete set $\mathcal{A} :=$ {QPSK, 16QAM, 64QAM, 256QAM}, which corresponds to rates of $\{2, 4, 6, 8\}$ *bits per symbol* (bps), respectively. We consider 9 power levels, which correspond to the average SNRs $\bar{\gamma}$ of {4.75, 4.80, 4.85, 4.90, 11.45, 11.50, 17.25, 17.30, 17.35} dBs, and set $T = 5 \times 10^4$.

Fig. 2.4 shows the throughput (expected reward) as a function of transmit powers (contexts) and rates (arms).[5] It is easy to see that the throughput is unimodal in rates (transmit powers) given a fixed transmit power (rate), and thus, satisfies joint unimodality. In order to simulate power arrivals, we use the typical harvested solar energy pattern from Fig. 3 in [69]. This figure shows that the harvested solar energy (context) increases from sunrise to noon, and then drops down after noon. Thus, the context arrival itself exhibits a unimodal structure. Based on this, we create a sequence of transmit powers that matches the harvested energy pattern. We assume that in the evening and night, there is some backup energy which slowly diminishes, and that low power levels correspond to this backup energy based transmit powers.

Throughput and resource selection of KL-UCB-U, SW-G-ORS, CUCB and CUL are compared in Fig. 2.5 and Fig. 2.6. As the optimal rates for different transmit powers are different, the achieved optimal throughput (oracle's throughput) varies as transmit power varies along the time t. As the transmit power increases, the optimal rate increases and then decreases with decrease in allowable transmit power. Fig. 2.5 provides the comparison in terms of moving average throughput, where the value at time t represents the average throughput of previous 200 packets and each curve is averaged over 50 repetitions of the

[5] In all figures, the optimal arm for each context is represented in dark blue.

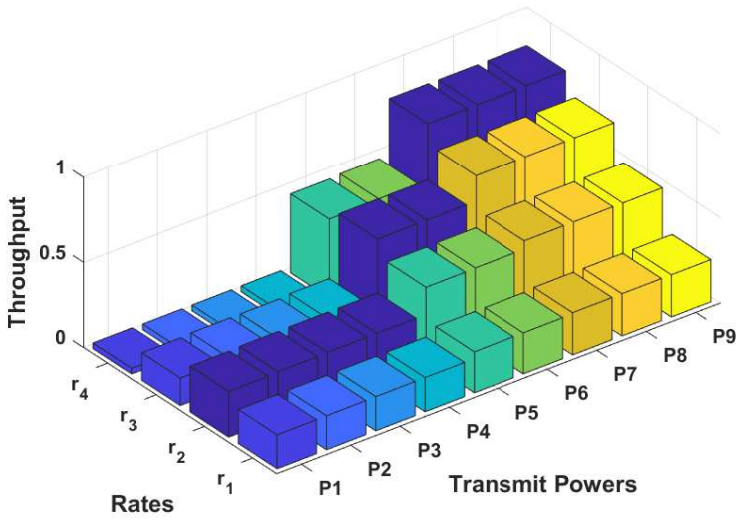


Figure 2.4: Throughput vs. power-rate pairs in the dynamic rate selection experiment.

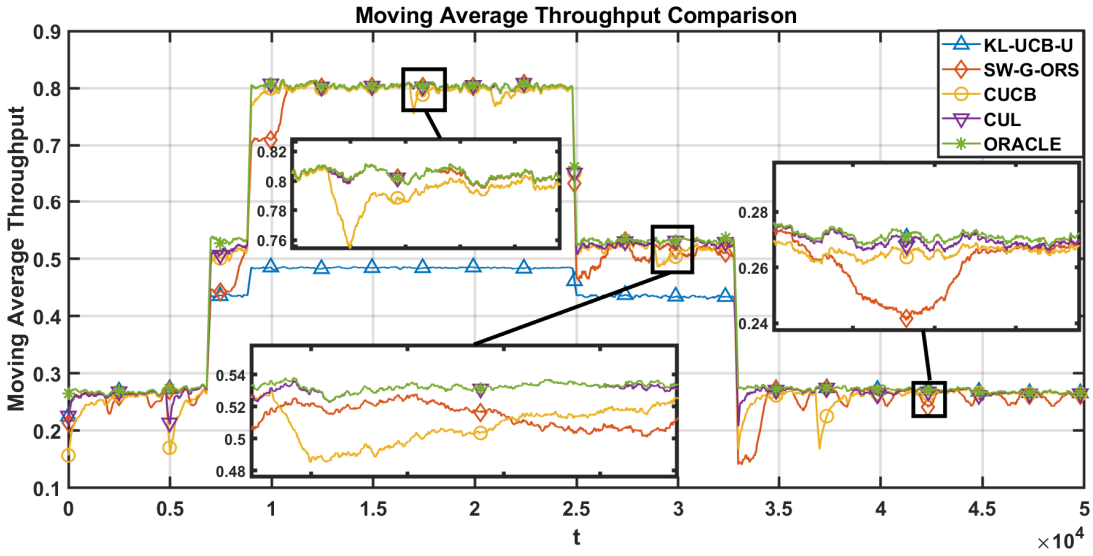


Figure 2.5: Throughput in the dynamic rate selection experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

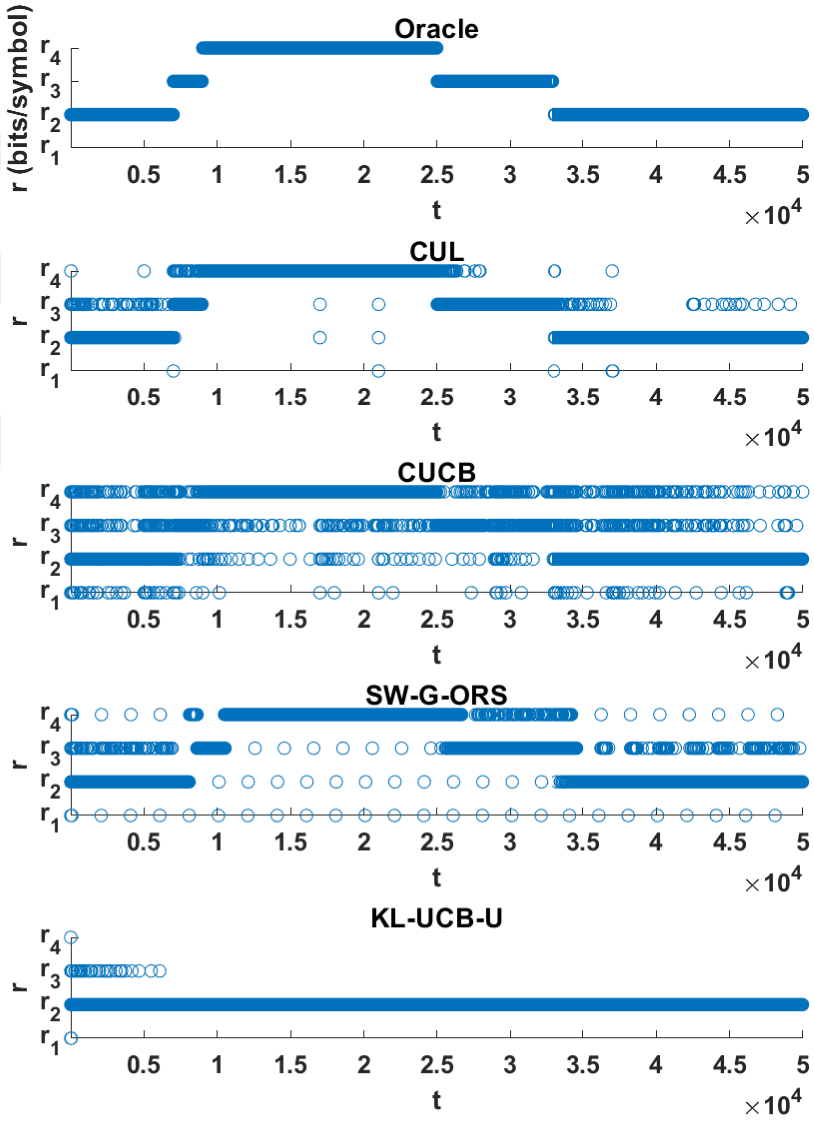


Figure 2.6: Resource selection over time in the dynamic rate selection experiment.

Table 2.2: Averaged throughput error and averaged accuracy in the dynamic rate selection experiment.

Algorithm	Averaged Throughput Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	122.3	47.5
SW-G-ORS	16.9	82.4
CUCB	7.7	88.4
CUL	2.3	95.6

experiment. It is evident that initially algorithms tend to explore the available rates, and as more contexts arrive the decisions improve. However, different algorithms have different convergence rates, and the throughput curve of CUL is the closest one to the oracle. Fig. 2.6 shows a snapshot of the resources selected by the algorithms over time. Clearly, CUL is able to track the oracle much better than the other algorithms. A comparison in terms of averaged throughput error and averaged accuracy is shown in Table 2.2.

KL-UCB-U tries to find a global optimal rate for all transmit powers, and hence, achieves worse performance. Although the energy arrival pattern looks favorable for SW-G-ORS, it also achieves worse performance possibly due to reinitialization of the parameters in each sliding window, which may not perfectly match with the energy arrivals. On the other hand, CUCB finds the optimal rate for each transmit power but does not exploit the unimodality, and hence, learns slowly. Finally, CUL achieves considerably better performance than all of the other algorithms since it utilizes both the contextual information and joint unimodality to learn fast.

2.4.4 Experiment 2: Dynamic Power Allocation for Heterogeneous Applications (Fig. 2.7-2.9, Table 2.3)

We consider the problem in Section 2.1.2. In this experiment, in each round the learner selects a transmit power level (arm) in order to serve an application with a rate constraint (context). The modulation scheme is selected from a discrete

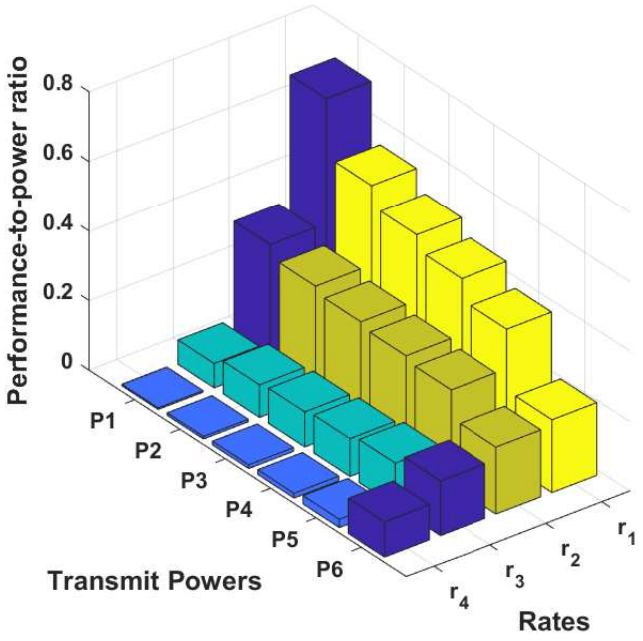


Figure 2.7: Performance-to-power ratio vs. power-rate pairs in the dynamic power allocation experiment.

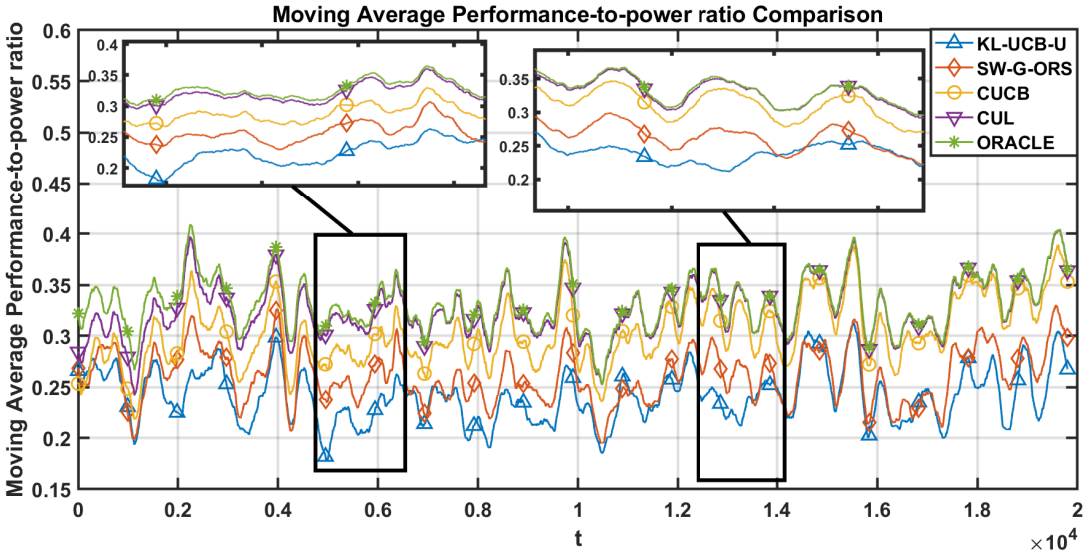


Figure 2.8: Performance-to-power ratio in the dynamic power allocation experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

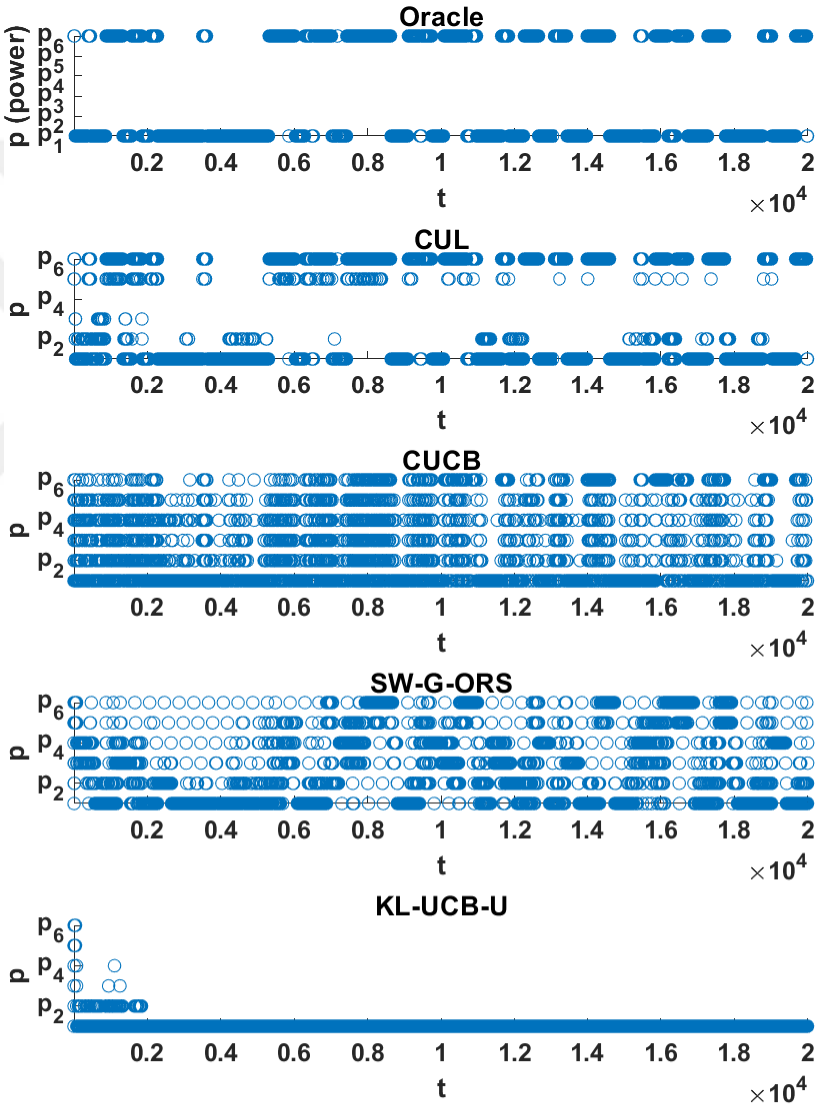


Figure 2.9: Resource selection over time in the dynamic power allocation experiment.

Table 2.3: Averaged performance-to-power error and averaged accuracy in the dynamic power allocation experiment.

Algorithm	Averaged Performance-to-power Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	91.0	38.8
SW-G-ORS	71.4	34.5
CUCB	29.0	60.2
CUL	6.4	86.9

set $\mathcal{X} := \{$QPSK, 16QAM, 64QAM, 256QAM$\}$, which corresponds to rates of $\{2, 4, 6, 8\}$ bps, respectively. The context is selected uniformly at random from $\mathcal{X}$, and duration in terms of time slots for which the application remains active is drawn from the uniform distribution in $[0, T/50]$, for $T = 2 \times 10^4$. We consider 6 power levels, which correspond to the average SNRs $\bar{\gamma}$ of $\{$2.5, 3.5, 4.0, 4.5, 5.5, 11.5$\}$ dBs.

Fig. 2.7 shows that the performance-to-power ratio given in (2.4) is jointly unimodal in rates and transmit powers. Performances of KL-UCB-U, SW-G-ORS, CUCB and CUL are compared in Fig. 2.8 and 2.9. In contrast to Experiment 1, the context arrivals are uniformly distributed, and thus, the optimal performance-to-power ratio is rapidly varying as shown in Fig. 2.8. The corresponding optimal rate, which maximizes the performance-to-power ratio, also switches frequently as shown in Fig. 2.9. It is shown that CUL consistently outperforms the other algorithms as the significant enhancement in the performance can already be achieved by using contextual information and transmit power's unimodality. In addition, average performance-to-power error and average accuracy metrics reported in Table 2.3 show the superiority of CUL.

2.4.5 Experiment 3: Distributed Resource Allocation in a Multi-user Network (Fig. 2.10-2.12, Table 2.4)

We consider the problem in Section 2.1.3. There are $N = 24$ channels indexed by the set $\mathcal{C} := \{c_1, \ldots, c_{24}\}$, which are ordered based on their average SNRs

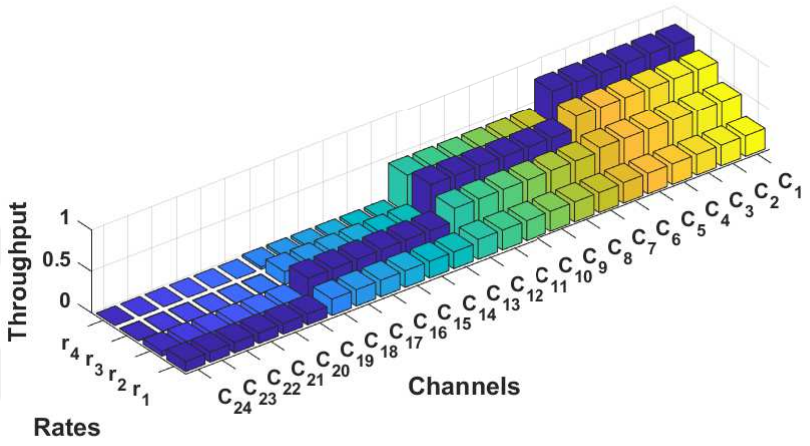


Figure 2.10: Throughput vs. channel-rate pairs in the distributed resource allocation experiment.

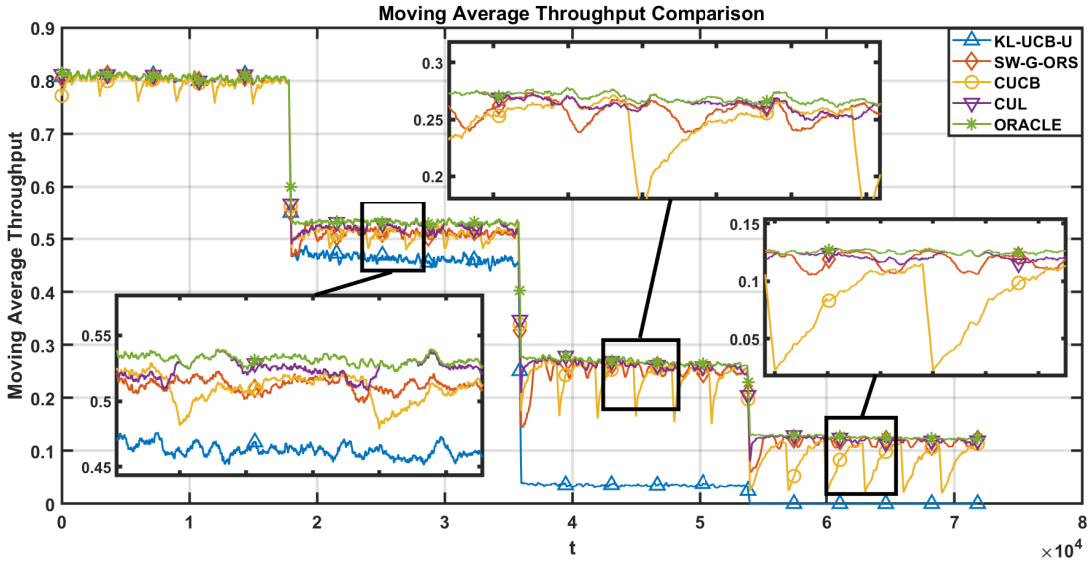


Figure 2.11: Throughput in the distributed resource allocation experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

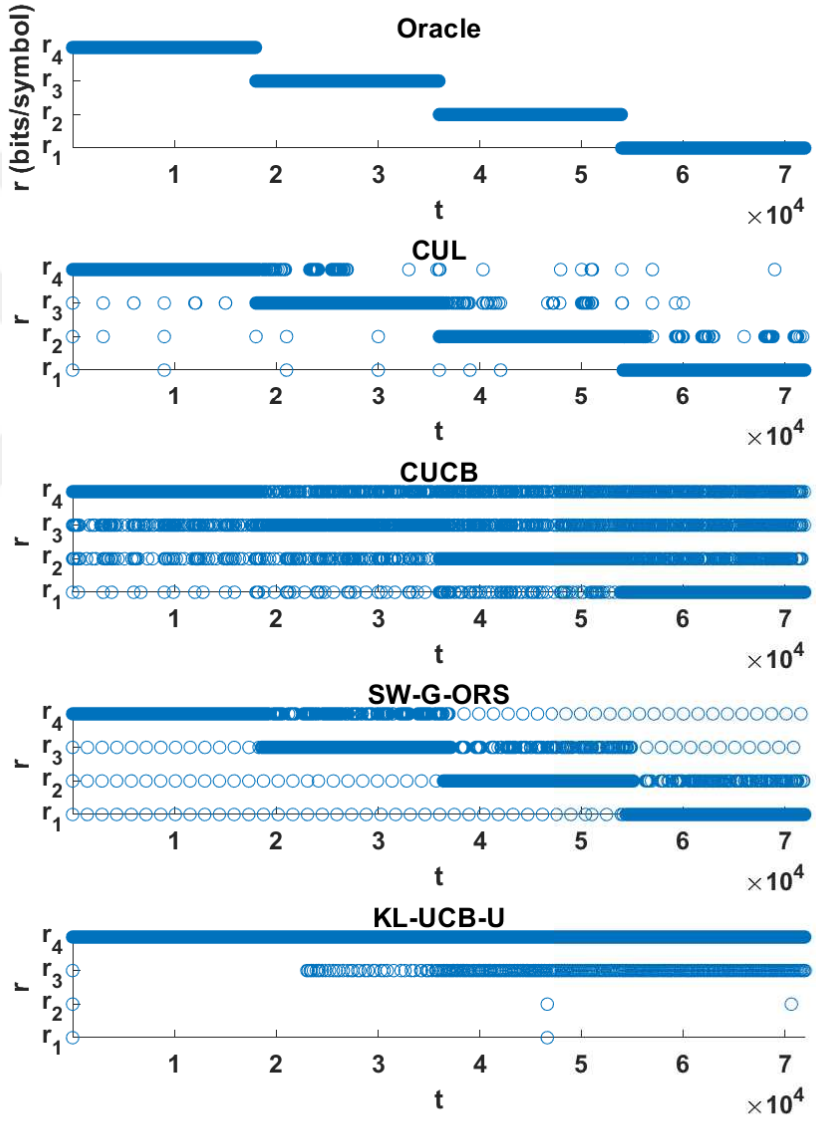


Figure 2.12: Resource selection over time in the distributed resource allocation experiment.

Table 2.4: Averaged throughput error and averaged accuracy in the distributed resource allocation experiment.

Algorithm	Averaged Throughput Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	107.0	25.2
SW-G-ORS	12.2	83.9
CUCB	24.9	76.2
CUL	4.9	93.7

given in 4 different sets of 6 channels each. Channels in each set are separated by 0.05 dB and lie in $\{17.25, \ldots, 17.50\}$, $\{11.25, \ldots, 11.50\}$, $\{4.65, \ldots, 4.90\}$ and $\{-0.50, \ldots, -0.25\}$ dBs, respectively. The number of users is $M = 24$. The modulation scheme is selected from a discrete set $\mathcal{A} := \{$QPSK, 16QAM, 64QAM, 256QAM$\}$, which corresponds to rates of $\{2, 4, 6, 8\}$ bps, respectively. Users select these channels in a round-robin fashion in blocks of $\tau' = T/24$ rounds, for $T = 7.2 \times 10^4$. For instance, user 1 selects channel c_1 in the first τ' rounds, and then moves to channel c_2 and selects it for the next τ' rounds, and so on. Similar to Experiment 1, Fig. 2.10 provides the throughput (of a particular user) as a function of channels (contexts) and rates (arms). Fig. 2.11 and Fig. 2.12 compare KL-UCB-U, SW-G-ORS, CUCB and CUL. The contexts arrive in a sequential manner and the corresponding optimal throughput varies with assigned channels as shown in Fig. 2.11. Similarly, the corresponding optimal rate for channels varies in a sequential manner as shown in Fig. 2.12. Average throughput error and average accuracy of the algorithms are compared in Table 2.4. This shows that CUL consistently outperforms the other algorithms when the context arrivals are periodic due to the round-robin channel selection.

2.4.6 Complexity Analysis

This section investigates computation and storage overhead complexities of CUL and compares them with the other algorithms.

- KL-UCB-U calculates and updates the UCB index only for the current leader and its neighbors, so the computational complexity is $O(1)$. The computational complexity of SW-G-ORS is similar to that of KL-UCB because it essentially calculates and updates the indexes for the current sliding window. In CUCB, an index is calculated for all A arms given an arriving context, so the computational complexity is $O(A)$. In CUL, indexes are calculated for the current leader and its neighbors not only for the current context but also for the other contexts to exploit joint unimodality, so the computational complexity is $O(X)$.

- *Storage overhead complexity* equals to the number of variables that an algorithm needs to store. KL-UCB-U keeps index for each arm and also counters for estimated mean of an arm, number of times an arm is selected and number of times an arm is the leader, so the storage complexity is $\approx 4A$. In addition to variables used in KL-UCB-U, SW-G-ORS keeps the arm selections and instantaneous rewards for all of the instances of sliding window, so the storage complexity is $\approx (4 + 2\tau)A$. CUCB keeps index for each context-arm pair and also counters for estimated reward of an arm and number of times an arm is selected for each context, so the storage complexity is $\approx 3XA$. CUL keeps indices (UCB and LCB) for each context-arm pair and also counters for estimated reward of an arm, number of times an arm is selected and number of times an arm is the leader for each context, so the storage complexity is $\approx 5XA$.

We note that KL-UCB-U has the lowest computational complexity since it is agnostic to the contexts. However, the performance achieved by this algorithm is very low. On the other hand, SW-G-ORS achieves relatively good performance, but with a higher storage overhead complexity, since it stores arms selections and instantaneous rewards for a sliding window of length τ. Similarly, CUCB achieves better performance, but with a higher computational cost, since it keeps a different set of parameters for each context. CUL, which outperforms the other algorithms in terms of the performance, has the highest computational complexity since it needs to perform a larger number of operations than CUCB to exploit the joint unimodality. Nevertheless, the computational complexity of CUL is linear

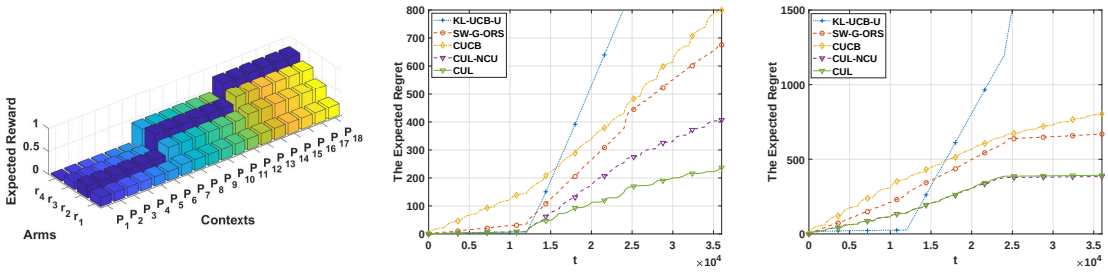
in the number of arms and contexts, thus it can easily handle a large number of arms and contexts and run on devices with limited memory and processing capability.

2.4.7 Experiment 4: Effect of Context Arrivals (Fig. 2.13)

This experiment analyzes the performance of CUL under different sequences of context arrivals for the dynamic rate selection for an energy harvested transmitter application. There are 4 arms and 18 contexts, time horizon is set as $T = 3.6\times10^4$, and the expected rewards are shown in Fig. 2.13(i). We introduce *CUL with no Contextual Unimodality* (CUL-NCU) to study the effect of context arrivals. CUL-NCU is a variant of CUL that exploits the contextual information and unimodality over arms but not unimodality over the contexts. It runs a separate instance of a unimodal (in arms) MAB algorithm for each context. We included this variant as a competitor algorithm in order to observe the effect of exploiting contextual unimodality on the regret.

Firstly, we analyze the performance for a favorable sequence of contexts arrivals. We assume that contexts with higher expected rewards arrive first, then followed by contexts with lower expected rewards. The gap between the regrets of CUL-NCU and CUL given in Fig. 2.13(ii) shows that the performance gain is achieved solely by exploiting contextual unimodality. In this case, it is possible to use a much tighter UCB that comes from the target context in the modified neighborhood. If the unimodality over arms and contexts is not exploited, the regret's upper bound contains a term $\sum_{x=x_1}^{x_X}\sum_{a\neq a_x^*}\log(T) = 54\log(T)$. On the other hand for CUL, the regret contains the term $\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}$, which reduces to 6.91 when calculated for this setting. CUL-NCU beats all other algorithms except CUL as it does not use contextual unimodality. It is shown again that CUL consistently outperforms the other algorithms.

Secondly, we analyze the performance for an unfavorable sequence of contexts arrivals, where the contexts with lower expected rewards arrive first. In this scenario, exploring the modified neighborhood and the target context is difficult.



(i) (ii) (iii)

Figure 2.13: Comparison of CUL with CUL-NCU, CUCB, KL-UCB-U and SW-G-ORS for rate selection given power (i) Expected Rewards (ii) Favorable context arrivals (iii) Non-favorable context arrivals.

For most of the time, the modified neighborhood is empty, which makes the benefit of using unimodality in contexts insignificant. As seen in Fig. 2.13(iii), CUL behaves similar to CUL-NCU in this case. However, CUL and CUL-NCU still significantly outperform the other algorithms thanks to exploiting contextual information and arm unimodality.

2.5 Proof of Lemma 1

Note that $\tilde{L}_x(t) = a_x^*$, $\tilde{a}_x(t) = a$ and $\mu(x, a_x^*) \leq \underline{\tilde{u}}_{x,a_x^*}(t)$ together imply that $\underline{\tilde{u}}_{x,a}(t) \geq \underline{\tilde{u}}_{x,a_x^*}(t) \geq \mu(x, a_x^*)$. Thus, we have

$$\begin{aligned}\{\underline{\tilde{u}}_{x,a}(t) \geq \mu(x, a_x^*)\} &= \{\min\{\tilde{u}_{x,a}(t), \tilde{w}_{x,a}(t)\} \geq \mu(x, a_x^*)\} \\ &= \{\tilde{u}_{x,a}(t) \geq \mu(x, a_x^*), \tilde{w}_{x,a}(t) \geq \mu(x, a_x^*)\}.\end{aligned}$$

Condition $\tilde{u}_{x,a}(t) \geq \mu(x, a_x^*)$ implies that

$$\begin{aligned}\tilde{N}_{x,a}(t) d_a^+(\tilde{\mu}_{x,a}(t), \mu(x, a_x^*)) &\leq \tilde{N}_{x,a}(t) d_a(\tilde{\mu}_{x,a}(t), \tilde{u}_{x,a}(t)) \\ &\leq \log(T) + 3\log(\log(T)) = f(T).\end{aligned}$$

Similarly, condition $\tilde{w}_{x,a}(t) \geq \mu(x, a_x^*)$ implies that

$$\tilde{M}_{x,a}(t) d_a^+(\tilde{\eta}_{x,a}(t), \mu(x, a_x^*)) \leq \tilde{M}_{x,a}(t) d_a(\tilde{\eta}_{x,a}(t), \tilde{w}_{x,a}(t)) \leq f(T).$$

Using the inequalities above, we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t)=a_x^*, \mu(x,a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t)=a\}\right] \\
&\le \mathbb{E}\Bigg[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\big(\tilde{N}_{x,a}(t) d_a^+(\tilde{\mu}_{x,a}(t), \mu(x,a_x^*)) \le f(T), \\
&\qquad \tilde{M}_{x,a}(t) d_a^+(\tilde{\eta}_{x,a}(t), \mu(x,a_x^*)) \le f(T), \tilde{a}_x(t)=a\big)\Bigg] \\
&\le \mathbb{E}\Bigg[\sum_{t=1}^{N_x(T+1)} \sum_{s=1}^{t} \mathbf{1}\big(\tilde{N}_{x,a}(t)=(s-1), \tilde{a}_x(t)=a, \\
&\qquad\qquad (s-1) d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) \ \le f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) \le f(T)\big)\Bigg] \\
&\le \mathbb{E}\Bigg[\sum_{t=1}^{T} \sum_{s=1}^{t} \mathbf{1}\big(\tilde{N}_{x,a}(t)=(s-1), \tilde{a}_x(t)=a\big) \\
&\qquad\qquad \mathbf{1}\big((s-1) d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) \le f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) \le f(T)\big)\Bigg].
\end{aligned}
$$

By the change of variables, the sum above can be rewritten as

$$
\begin{aligned}
&\le \mathbb{E}\Bigg[\sum_{s=1}^{T} \sum_{t=s}^{T} \mathbf{1}\big(\tilde{N}_{x,a}(t)=(s-1), \tilde{a}_x(t)=a\big) \\
&\qquad\qquad \mathbf{1}\big((s-1) d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) \le f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) \le f(T)\big)\Bigg] \\
&\le \mathbb{E}\Bigg[\Big(\sum_{s=1}^{T} \mathbf{1}\big((s-1) d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) \ \le f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) \le f(T)\big) \\
&\qquad \sum_{t=s}^{T} \mathbf{1}\big(\tilde{N}_{x,a}(t)=(s-1), \tilde{a}_x(t)=a\big)\Big)\Bigg] \\
&\le \sum_{s=1}^{T} \mathbb{P}\big((s-1) d_a^+(\tilde{\mu}_{x,a}^s, \mu(x,a_x^*)) \le f(T),\ \tilde{M}_{x,a}^s d_a^+(\tilde{\eta}_{x,a}^s, \mu(x,a_x^*)) \le f(T)\big)
\end{aligned}
$$

where the last inequality holds since $s \in \{1,\dots,T\}$, $\sum_{t=s}^{T} \mathbf{1}\big(\tilde{N}_{x,a}(t) = (s-1), \tilde{a}_x(t)=a\big) \le 1$.

2.6 Proof of Lemma 2

Let $A = \{\mu(x,a) \leq \mu(\tilde{y}_{x,a}(t), a)\}$. For $\delta > 0$, we have

$$\begin{aligned}
\mathbb{P}(\mu(x,a) > \underline{\tilde{u}}_{x,a}(t)) &= \mathbb{P}(\mu(x,a) > \min\{\tilde{u}_{x,a}(t), \tilde{w}_{x,a}(t)\}) \\
&\leq \mathbb{P}(\mu(x,a) > \tilde{u}_{x,a}(t)) + \mathbb{P}(\mu(x,a) > \tilde{w}_{x,a}(t)) \\
&= \mathbb{P}(\mu(x,a) > \tilde{u}_{x,a}(t)) \\
&+ \mathbb{P}(\mu(x,a) > \tilde{w}_{x,a}(t) \mid A)\ \mathbb{P}(A) + \mathbb{P}(\mu(x,a) > \tilde{w}_{x,a}(t) \mid A^c)\ \mathbb{P}(A^c) \\
&\leq \mathbb{P}(\mu(x,a) > \tilde{u}_{x,a}(t)) + \mathbb{P}(\mu(\tilde{y}_{x,a}(t), a) > \tilde{w}_{x,a}(t)) + \mathbb{P}(A^c) \\
&\leq 2e\lceil \delta \log(\tau_x(t)) \rceil \exp(-\delta) + \mathbb{P}(A^c)
\end{aligned}$$

where the last inequality follows from Theorem 10 in [45]. Next, we bound the probability of event A^c as

$$\mathbb{P}(A^c) \leq 2Xe\lceil \delta \log(\tau_x(t)) \rceil \exp(-\delta).$$

Since the KL-UCB index is an upper bound on true mean with high probability and the KL-LCB index is a lower bound on true mean with high probability, A^c implies that there exists at least one context for arm a whose KL-UCB index under-estimates its mean value or KL-LCB index over-estimates its mean value, i.e., given

$$\begin{aligned}
A_1^c &= \cup_{x' \in \mathcal{X}} \{\tilde{u}^x_{x',a}(t) < \mu(x',a)\}, \\
A_2^c &= \cup_{x' \in \mathcal{X}} \{\tilde{l}^x_{x',a}(t) > \mu(x',a)\}
\end{aligned}$$

we have $A^c \subset A_1^c \cup A_2^c$.

Since by the definition of KL-UCB in (2.7), $q \mapsto d(p,q)$ is increasing function for $q \geq p$, if $\tilde{u}^x_{x',a}(t) < \mu(x',a)$, then

$$\tilde{N}^x_{x',a}(t) d\left(\frac{\tilde{\mu}^x_{x',a}(t)}{K_a}, \frac{\mu(x',a)}{K_a}\right) > \delta.$$

Similarly by the definition of KL-LCB in (2.9), $q \mapsto d(p,q)$ is decreasing function for $q \leq p$, if $\tilde{l}^x_{x',a}(t) > \mu(x',a)$ then:

$$\tilde{N}^x_{x',a}(t) d\left(\frac{\tilde{\mu}^x_{x',a}(t)}{K_a}, \frac{\mu(x',a)}{K_a}\right) > \delta.$$

Therefore,

$$A_1^c \cup A_2^c \subseteq \bigcup_{x' \in \mathcal{X}} \left\{ \tilde{N}^x_{x',a}(t) d\left(\frac{\tilde{\mu}^x_{x',a}(t)}{K_a}, \frac{\mu(x',a)}{K_a} \right) > \delta \right\}.$$

Again by using Theorem 10 in [45], we get

$$\mathbb{P}(A^c) \leq 2Xe^{\lceil \delta \log(\tau_x(t)) \rceil} \exp(-\delta)$$

and hence,

$$\mathbb{P}(\mu(x,a) > \underline{\tilde{u}}_{x,a}(t)) \leq 2(X+1)e^{\lceil \delta \log(\tau_x(t)) \rceil} \exp(-\delta).$$

Chapter 3

Exploiting Structure in Contextual Multi-Armed Bandits via Bayesian Approach

This section was published in [3].[1]

In this chapter, we rigorously formulate the rate adaptation over a wireless communication system as a contextual Bayesian unimodal MAB. We refer to each MCS as an arm and to the transmit power as a side information (context). The user (learner) selects an arm after observing the context in each round. The expected reward is defined as the throughput and consistent with real-world observations [30, 67] it is assumed to exhibit unimodal structure under different MCSs and a monotone structure under different transmit powers.

[1] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Online Bayesian Learning for Rate Selection in Millimeter Wave Cognitive Radio Networks", in Proc. IEEE International Conference on Computer Communications (INFOCOM), July 2020.

3.1 Problem Formulation

We consider a single transmitter and receiver based wireless link, where the transmitter (i) is powered from an energy harvested source in case of the energy harvesting CRN or (ii) obeys an ITL in case of the underlay CRN when selecting its transmit power. The transmitter can transmit by using one of the K available transmission rates by choosing the corresponding MCS and one of the P possible transmit powers.[2] Let $\mathcal{K} = [K]$ denote the set of MCS and r_i, $i \in \mathcal{K}$, denote the rate that corresponds to the ith MCS. The set of rates $\mathcal{R} = \{r_1, \ldots, r_K\}$ and the set of transmit powers $\mathcal{P} = \{p_1, \ldots, p_P\}$ are ordered such that $r_1 < \ldots < r_K$ and $p_1 < \ldots < p_P$. For $p \in \mathcal{P}$ and $r \in \mathcal{R}$, j_p and i_r represent the indices of transmit power p and rate r, i.e., $p_{j_p} = p$ and $r_{i_r} = r$.

In the MAB formulation, the SU makes decisions sequentially over rounds indexed by $t \in [T]$,[3] where T represents the time horizon. At the beginning of round t, the SU receives the transmit power it should use in that round, denoted by $p(t)$ with index $j(t)$, i.e., $p(t) = p_{j(t)}$. In case of the energy harvesting CRN, this comes from the power control algorithm of the energy harvesting device [9], while in case of the underlay CRN, this comes from interference temperature measurements [39]. Then, the SU chooses a modulation and coding scheme from $\mathcal{K}$ with corresponding rate $r(t)$ and transmits with power $p(t)$ and rate $r(t)$. At the end of the round it receives ACK/NACK feedback $x_{p(t),r(t)}(t)$ indicating whether the transmission was successful or not, and collects as throughput the (normalized) reward $g_{p(t),r(t)}(t) = (r(t)/r_K)x_{p(t),r(t)}(t)$. Here, $x_{p,r}(t)$ is a Bernoulli random variable with expected value $\psi(p, r)$ that takes value 1 if the transmission given power-rate pair (p, r) in round t is successful and value 0 otherwise. We call $\psi(p, r)$ the transmission success probability and $\mu(p, r) = (r/r_K) \cdot \psi(p, r)$ the normalized throughput associated with power-rate pair (p, r). We also call transmit powers *contexts* and rates *arms* of the MAB.

In case of the energy harvesting CRN, we consider no embedded energy supply

[2] Similar to [70] and [65], we adopt a discrete power set.

[3] For an integer T, we let $[T] := \{1, \ldots, T\}$.

so that the CRN solely relies on the harvested energy. The dynamic power output from an energy harvesting source (e.g., RF energy source, solar cell, wind turbine) is directly used by the load [9]. In case of the underlay CRN, the SU uses its cognitive capabilities to determine the transmit power such that the dynamic ITL is satisfied [71]. For a single dynamic SU, ITL is dependent on the current location of the SU [72], and assuming known primary location, the SU adjusts its transmit power to satisfy the current ITL. For cooperative multi-player homogeneous CRN in which the SUs are assigned channels in round robin manner, ITL which is dependent on the channel frequency, is calculated beforehand for each frequency channel [73], and the SU adjusts its transmit power to satisfy ITL for the currently assigned frequency channel.

The optimal rate given a transmit power p is denoted by $r_p^* = \text{argmax}_{r\in\mathcal{R}}\, \mu(p,r)$ and its index is given as i_p^*, i.e., $r_p^* = r_{i_p^*}$. Without loss of generality we assume that r_p^* is unique. The suboptimality gap of rate r given transmit power p is defined as $\Delta(p,r) = \mu(p,r_p^*) - \mu(p,r)$. The set of neighbors of rate r_i is given as $\mathcal{N}(r_i)$. We have $\mathcal{N}(r_i) = \{r_{i-1}, r_{i+1}\}$ for $i \in \{2,\ldots,K-1\}$, $\mathcal{N}(r_1) = \{r_2\}$ and $\mathcal{N}(r_K) = \{r_{K-1}\}$. We denote the lower and upper indexed neighbors of rate r by r^- and r^+ given that they exist. Moreover, the set of rates lower than and higher than rate r are denoted by $[r]^-$ and $[r]^+$.

The expected reward function $\mu(p,r)$ exhibits a unimodal structure over the set of rates [5, 29]. For a given transmit power, this structure can be explained via a line graph whose vertices correspond to the rates. More specifically, $\mu(p,r)$ is called *unimodal in the rates* if for any given p there exist a path from any suboptimal rate to the optimal rate along which the expected reward is strictly increasing, i.e., $\forall p \in \mathcal{P}$, $\mu(p,r_1) < \ldots < \mu(p,r_p^*)$ and $\mu(p,r_p^*) > \ldots > \mu(p,r_K)$. Furthermore, for any given r, $\mu(p,r)$ exhibits a monotone structure over the set of transmit powers, i.e., $\forall r \in \mathcal{R}$, $\mu(p_1,r) \leq \ldots \leq \mu(p_P,r)$.

Let $N_{p,r}(t)$ be the number of times rate r was selected by the SU given transmit power p before round t. For a given sequence of transmit powers $p(1),\ldots,p(T)$,

the (pseudo) regret after the first T rounds is defined as

$$R(T) = \sum_{t=1}^{T} \Big(\mu(p(t), r^*_{p(t)}) - \mu(p(t), r(t)) \Big)$$
$$= \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}} \Delta(p, r) N_{p,r}(T+1). \tag{3.1}$$

Our goal is to design a learning algorithm for the SU that minimizes the growth rate of the expected regret.

3.2 The Learning Algorithm

We propose Dynamic Rate Selection via Thompson Sampling (DRS-TS), which is a learning algorithm that takes into account unimodality and monotonicity of $\mu(p, r)$ to minimize the expected regret (pseudocode is given in Algorithm 2, graphical demonstration in Fig. 3.1). DRS-TS exploits unimodality of $\mu(p, r)$ in rates somewhat similar to UTS [23] and MTS [31]. The main novelty of DRS-TS comes from introduction of transmit power (contextual) information and exploiting the monotonicity of $\mu(p, r)$ in transmit powers. It is important to note that, the learner does not have any control over the transmit power arrivals and the exogenous nature of the arrivals makes efficient learning much more challenging than the non-contextual prior works.

For each power-rate pair (p, r), DRS-TS keeps the counters $N_{p,r}(t)$ and $S_{p,r}(t)$, where $S_{p,r}(t)$ counts the number of successful transmissions in rounds in which transmit power was p and rate r was selected before round t. It also keeps sample mean estimate of the rewards $\hat{\mu}_{p,r}(t)$ that are obtained from rounds in which transmit power was p and rate r was selected prior to the round t.[4]

The *rate leader* for transmit power $p \in \mathcal{P}$ in round t is defined as the rate with the highest sample mean reward, i.e., $L_p(t) = \text{argmax}_{r \in \mathcal{R}} \hat{\mu}_{p,r}(t)$ (ties are broken arbitrarily). Letting $\mathbf{1}(\cdot)$ denote the indicator

[4]When the current round is clear from the context, we suppress the round index.

Algorithm 2 DRS-TS

Input: P, K
Initialize: $t = 1$
Counters: $N_{p,r} = 0, \hat{\mu}_{p,r} = 0, S_{p,r} = 0, b_{p,r} = 0, \forall r \in \mathcal{R}, \forall p \in \mathcal{P}$
while $t \geq 1$ **do**
    Observe transmit power $p(t)$
    $L_{p(t)} = \text{argmax}_{r \in \mathcal{R}} \hat{\mu}_{p(t),r}$
    if $\frac{b_{p(t),L_{p(t)}} - 1}{3} \in \mathbb{N}$
        $r(t) = L_{p(t)}$
    else
        $\overline{\mathcal{R}} \leftarrow \mathcal{N}(L_{p(t)}) \cup \{L_{p(t)}\}$
        for $r \in \overline{\mathcal{R}}$
            Draw $\phi_{p_j,r}$ from $\pi_{p_j,r}$ in (3.2), $\forall p_j \in \mathcal{P}$
            $\theta_{p_j,r} = \frac{r}{r_K} \phi_{p_j,r}$, $\forall p_j \in \mathcal{P}$
            Find $\overline{\mathcal{M}}_{p(t),r}(t)$ using (3.3)
            $\underline{\theta}_{p(t),r} = \min_{p' \in \overline{\mathcal{M}}_{p(t),r}(t) \cup \{p(t)\}} \theta_{p',r}$
        end for
        $r(t) = \text{argmax}_{r \in \overline{\mathcal{R}}} \ \underline{\theta}_{p(t),r}$
    end if
    Observe feedback $x_{p(t),r(t)}(t)$ and reward $g_{p(t),r(t)}(t)$
    $b_{p(t),L_{p(t)}} = b_{p(t),L_{p(t)}} + 1$
    $N_{p(t),r(t)} = N_{p(t),r(t)} + 1$
    $\hat{\mu}_{p(t),r(t)} = \frac{\hat{\mu}_{p(t),r(t)}(N_{p(t),r(t)}-1)+g_{p(t),r(t)}(t)}{N_{p(t),r(t)}}$
    $S_{p(t),r(t)} = S_{p(t),r(t)} + x_{p(t),r(t)}(t)$
    $t = t + 1$
end while

function, we define

$$b_{p,r}(t) = \sum_{t'=1}^{t-1} \mathbf{1}(p(t') = p, r = L_p(t'))$$

as the number of times rate r was a rate leader when the transmit power was p up to round t.

After observing $p(t)$ in round t, DRS-TS identifies the rate leader $L_{p(t)}(t)$ and calculates $b_{p(t),L_{p(t)}(t)}(t)$. If $(b_{p(t),L_{p(t)}(t)}(t) - 1)/3 \in \mathbb{N}$, then DRS-TS exploits the rate leader to ensure that the current rate leader is selected more often. Similar to [23] and [22], this significantly simplifies the regret analysis. Otherwise, DRS-TS tries to learn the optimal rate for the given transmit power by utilizing

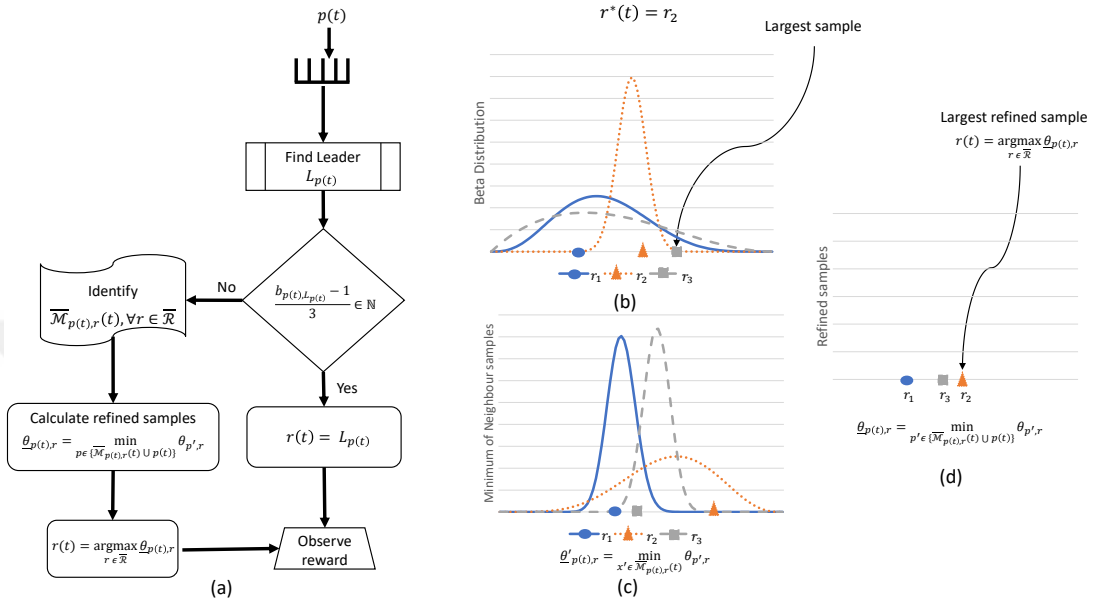


Figure 3.1: Graphical demonstration of DRS-TS algorithm for an example rate set: (a) Flow chart of DRS-TS (b) In a given time slot t, r_2 is the optimal rate for a given transmit power $p(t)$, r_2 is the leader and we have $\overline{\mathcal{R}} = \{r_1, r_2, r_3\}$. If transmit power monotonicity is ignored, r_3 is selected based on calculated samples to cater to the exploration. (c) As an intermediate step, minimum of the neigbours samples and their corresponding distributions in $\overline{\mathcal{M}}_{p(t),r}(t)$ for rates in $\overline{\mathcal{R}}$ are shown. (d) Refined samples are calculated by using $\overline{\mathcal{M}}_{p(t),r}(t)$ and $p(t)$ for rates in $\overline{\mathcal{R}}$, and therefore, utilizing the transmit power monotonicity results in selection of the optimal rate r_2 in the current time slot t.

unimodality in rates and monotonicity in transmit powers. As a first step, it calculates $\overline{\mathcal{R}}(t) = \mathcal{N}(L_{p(t)}(t)) \cup \{L_{p(t)}(t)\}$. Thanks to unimodality, exploring over $\overline{\mathcal{R}}(t)$ is sufficient for DRS-TS to identify the optimal rate.

Selecting from $\overline{\mathcal{R}}(t)$ is performed by using Thompson sampling. The posterior distribution of transmission success probability for all (p, r) in round t is calculated as

$$\pi_{p,r}(t) = Beta\big(1 + S_{p,r}(t),\ 1 + N_{p,r}(t) - S_{p,r}(t)\big) \tag{3.2}$$

where $Beta(\alpha, \beta)$ is the beta distribution with parameters α and β. A sample drawn from $\pi_{p,r}(t)$ is denoted by $\phi_{p,r}(t)$, and the *throughput sample* is obtained via $\theta_{p,r}(t) = \frac{r}{r_K}\ \phi_{p,r}(t)$. These samples are then used to transfer the knowledge

obtained from other transmit powers to the current transmit power. As the first step, we define the monotone neighborhood of $p(t)$ with index $j(t)$, which contains all the contexts greater than the current context as

$$\mathcal{M}_{p(t),r}(t) = \begin{cases} \{p_{j(t)+1}, \ldots, p_P\} & \text{if } j(t) < P \\ \emptyset & \text{otherwise} \end{cases}$$

Monotonicity of $\mu(p, r)$ in transmit powers implies that for a given rate r, it is certain that (p', r) has higher expected reward than (p, r) for all $p' \in \mathcal{M}_{p,r}(t)$. Since the number of observations from a given (p', r) such that $p' \in \mathcal{M}_{p(t),r}(t)$ may be small, to help learning for the current context $\mathcal{M}_{p(t),r}(t)$ is further sorted and only the pairs (p', r) that are observed more than $(p(t), r)$ are selected. Thus, we define the refined monotone neighborhood of $p(t)$ as

$$\overline{\mathcal{M}}_{p(t),r}(t) = \left\{ p' \in \mathcal{M}_{p(t),r}(t) : N_{p',r}(t) > N_{p(t),r} \right\}. \tag{3.3}$$

Using the above facts, DRS-TS simply sets its refined throughput sample as

$$\underline{\theta}_{p(t),r}(t) = \min_{p' \in \overline{\mathcal{M}}_{p(t),r}(t) \,\cup\, \{p(t)\}} \theta_{p',r}(t)$$

and then selects the arm with the highest refined sample, i.e., $r(t) = \text{argmax}_{r \in \overline{\mathcal{R}}(t)} \; \underline{\theta}_{p(t),r}(t)$. Note that randomness of the posterior samples ensures sufficient exploration. Finally, at the end of round t, the ACK/NACK feedback $x_{p(t),r(t)}(t)$ is observed, the reward $g_{p(t),r(t)}(t)$ is collected and the sample mean estimates and counters that correspond to the pair $(p(t), r(t))$ are updated.

3.3 Regret Analysis of DRS-TS

In this section, we analyze expected value of the regret of DRS-TS given in (3.1). We build on the non-contextual analysis in [74] to show that DRS-TS achieves logarithmic regret in the contextual setting.

3.3.1 Preliminaries

The expected regret can be written as

$$\mathbb{E}[R(T)] = \sum_{p=p_1}^{p_P} \sum_{r \neq r_p^*} \Delta(p,r)\mathbb{E}[N_{p,r}(T+1)]. \tag{3.4}$$

The standard analysis of frequentist index policies bounds the number of draws of a suboptimal rate by considering two events: i) the optimal rate is under-estimated, and ii) the optimal rate is not under-estimated and the suboptimal rate is selected [5, 29, 30]. As discussed in [74], we cannot compare throughput sample $\theta_{p,r}(t)$ to the true mean $\mu(p,r)$, due to the fact that $\theta_{p,r}(t)$ is not an optimistic estimate of $\mu(p,r)$. Hence, we compare $\theta_{p,r}(t)$ with a lower confidence bound given as $\mu(p,r) - \beta_{p,r}(t)$, where $\beta_{p,r}(t) := \sqrt{\frac{6\log(b_{p,r_p^*}(t)+1)}{N^*_{p,r}(t)}}$ and $N^*_{p,r}(t)$ is the number of times rate r was selected when the rate leader was r_p^* and the transmit power was p up to round t. Note that $N_{p,r}(t) \geq N^*_{p,r}(t)$.

From KL-UCB-U [5] and Bayes-UCB [75], we have KL-UCB and Bayes-UCB indices at time t as

$$u_{p,r}(t) := \max_{f \in [\hat{\mu}_{p,r}(t), \frac{r}{r_K}]} \{N_{p,r}(t) d(\frac{\hat{\mu}_{p,r}(t)}{(r/r_K)}, \frac{f}{(r/r_K)}) \leq \log(t) + \log(\log(T))\}$$

$$q_{p,r}(t) := \frac{r}{r_K}\, Q\Big(1 - \frac{1}{t\log(T)}, \pi_{p,r}(t)\Big)$$

where $d(x,y)$ is the Kullback-Leibler divergence between two Bernoulli distributions x and y, and $Q(a,\pi)$ is the quantile of order a of distribution π, i.e., for $X \sim \pi$, $\mathbb{P}_\pi(X \leq Q(a,\pi)) = a$. It is known that $q_{p,r}(t) \leq u_{p,r}(t)$.

Let $y_{p,r}(t) := \text{argmin}_{\{p' \in \overline{\mathcal{M}}_{p,r}(t)\}} \theta_{p',r}(t)$ denote the target transmit power of power-rate pair (p,r) in round t. Next, we introduce quantities dependent on target transmit power: $M_{p,r}(t)$ is the number of times rate r has been selected when the transmit power was $y_{p,r}(t)$ up to round t, and $\eta_{p,r}(t)$ is the sample mean reward of power-rate pair $(y_{p,r}(t), r)$ up to round t. Similarly, $\vartheta_{p,r}(t)$ is the throughput sample, $o_{p,r}(t)$ is the Bayes-UCB index, and $w_{p,r}(t)$ is the KL-UCB index of power-rate pair $(y_{p,r}(t), r)$ at the beginning of round t.

Let $\tau_p(t)$ denote the round in which transmit power p arrives for the tth time. Let $\tilde{N}_{p,r}(t) := N_{p,r}(\tau_p(t))$, $\tilde{N}^*_{p,r}(t) := N^*_{p,r}(\tau_p(t))$, $\tilde{\mu}_{p,r}(t) := \hat{\mu}_{p,r}(\tau_p(t))$, $\tilde{r}_p(t) := r(\tau_p(t))$, $\tilde{u}_{p,r}(t) := u_{p,r}(\tau_p(t))$, $\tilde{q}_{p,r}(t) := q_{p,r}(\tau_p(t))$, $\tilde{b}_{p,r}(t) := b_{p,r}(\tau_p(t))$, $\tilde{\beta}_{p,r}(t) := \beta_{p,r}(\tau_p(t))$, $\tilde{L}_p(t) := L_p(\tau_p(t))$, $\tilde{\theta}_{p,r}(t) := \theta_{p,r}(\tau_p(t))$, $\tilde{S}_{p,r}(t) := S_{p,r}(\tau_p(t))$, $\underline{\tilde{\theta}}_{p,r}(t) := \underline{\theta}_{p,r}(\tau_p(t))$, $\tilde{y}_{p,r}(t) := y_{p,r}(\tau_p(t))$, $\tilde{M}_{p,r}(t) := M_{p,r}(\tau_p(t))$, $\tilde{\eta}_{p,r}(t) := \eta_{p,r}(\tau_p(t))$, $\tilde{\vartheta}_{p,r}(t) := \vartheta_{p,r}(\tau_p(t))$, $\tilde{o}_{p,r}(t) := o_{p,r}(\tau_p(t))$ and $\tilde{w}_{p,r}(t) := w_{p,r}(\tau_p(t))$. Let $N_p(T)$ represent the number of times the context was p by the end of round T.

Furthermore, we introduce $\tau_{p,r}(s)$ to denote the round in which transmit power is p, rate leader is r^*_p and rate r is selected for the sth time. Let $\tilde{y}^s_{p,r}$ denote the target transmit power in round $\tau_{p,r}(s)$, $\tilde{N}^s_{p,r}$ denote the number of times rate r has been selected in rounds when the transmit power was p and rate leader was r^*_p before round $\tau_{p,r}(s)$, $\tilde{V}_{p,r}(s)$ denote the number of times rate r has been selected when the transmit power was p by the end of round $\tau_{p,r}(s)$, $\tilde{\mu}^s_{p,r}$ denote the sample mean reward calculated from rounds when transmit power was p and rate leader was r^*_p and rate r was selected at the beginning of round $\tau_{p,r}(s)$, and $\tilde{v}^s_{p,r}$ denote the sample mean reward of power-rate pair (p, r) at the beginning of round $\tau_{p,r}(s)$. We have $\tilde{N}^s_{p,r} = (s-1)$, $\tilde{\mu}^s_{p,r} = \frac{1}{(s-1)} \sum_{k=1}^{(s-1)} \tilde{X}^k_{p,r}$ and $\tilde{v}^s_{p,r} = \frac{1}{(\tilde{V}_{p,r}(s)-1)} \sum_{k=1}^{(\tilde{V}_{p,r}(s)-1)} X^k_{p,r}$, where $\tilde{X}^k_{p,r}$ is the reward of rate r when it is selected for kth time when the transmit power is p and rate leader is r^*_p, and $X^k_{p,r}$ is the reward of rate r when it is selected for kth time when the transmit power is p. For $s = 1$, we use the convention $\tilde{\mu}^s_{p,r} = 0$ and for $\tilde{V}_{p,r}(s) = 1$, $\tilde{v}^s_{p,r} = 0$. Next, we introduce quantities dependent on target transmit power: $\tilde{M}^s_{p,r}$ is the number of times rate r has been selected when the transmit power was $\tilde{y}^s_{p,r}$ before round $\tau_{p,r}(s)$, and $\tilde{\eta}^s_{p,r}$ is the sample mean reward of power-rate pair $(\tilde{y}^s_{p,r}, r)$ at the beginning of round $\tau_{p,r}(s)$. If there is no target transmit power in a certain round, the quantities $\tilde{M}^s_{p,r}$ and $\tilde{\eta}^s_{p,r}$ are zero for $\gamma_{p,r}$ in (3.8) for that round. We further introduce $\tilde{Z}^{s,*}_{p,r}$ as the number of times rate r^*_p has been selected when transmit power was p and rate leader was r^*_p before round $\tau_{p,r}(s)$, and define $\tilde{g}^{s,*}_{p,r} = \sqrt{6\log(T)/\tilde{Z}^{s,*}_{p,r}}$.

Since $\tilde{N}^*_{p,r^*_p}(t) \geq \lfloor \tilde{b}_{p,r^*_p}(t)/3 \rfloor$, the probability that the algorithm has selected the optimal rate for small number of times, when optimal rate is the rate leader

is itself small and is given as

$$\mathbb{E}\left[\sum_{t=1}^{\infty}\mathbf{1}\{\tilde{L}_p(t)=r_p^*,\ \tilde{N}^*_{p,r_p^*}(t)\leq(\tilde{b}_{p,r_p^*}(t))^b\}\right]\leq C_b'$$

where $b\in(0,1)$ and $C_b'<\infty$ are constants. This reduces the analysis to rounds when algorithm has seen reasonable number of draws of the optimal rate, and thus the posterior distribution is well concentrated.

We define another term dependent on b as $N_0(b):=\inf\{t\in\mathbb{N}:\log(t)t^b\geq(\sqrt{6}-\sqrt{5})^{-2}\}$. For $\frac{x}{(r/r_K)},\frac{y}{(r/r_K)}\in[0,1]$, let $d_r(x,y)=d(\frac{x}{(r/r_K)},\frac{y}{(r/r_K)})$, $d_r^+(x,y)=0$ if $x\geq y$ and $d_r^+(x,y)=d_r(x,y)$ if $x<y$, and $f(t):=\log(t)+\log(\log(T))$. Let $K^T_{p,r}:=\left\lfloor\frac{(1+\epsilon)f(T)}{d_r(\mu(p,r),\mu(p,r_p^*))}\right\rfloor$ for some $\epsilon>0$. If there exists a rate $r'\in\mathcal{R}$ for which $(r'/r_K)<\mu(p,r_p^*)$, we introduce $\mathcal{N}'(p,r_p^*):=\mathcal{N}(r_p^*)\backslash r'$, otherwise $\mathcal{N}'(p,r_p^*):=\mathcal{N}(r_p^*)$ [2, 29, 32].

3.3.2 Main Result

Theorem 2. *For all $\epsilon>0$, there exists a constant $C>0$ such that the expected regret of DRS-TS satisfies:*

$$\mathbb{E}[R(T)]\leq\sum_{p=p_1}^{p_P}\sum_{r\in\mathcal{N}'(p,r_p^*)}\left(\gamma_{p,r}\ \frac{(1+\epsilon)\Delta(p,r)}{d\left(\frac{\mu(p,r)}{r/r_K},\frac{\mu(p,r_p^*)}{r/r_K}\right)}\right)\log(T)+O(\log(\log(T)))+C$$

where $\gamma_{p,r}=\frac{\sum_{s=1}^{K^T_{p,r}+1}\mathbb{P}\left(\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r_p^*)-\tilde{g}^{s,*}_{p,r})<f(T)\right)}{K^T_{p,r}+1}\in[0,1]$.

The term $\gamma_{p,r}$ depends on how well power-rate pair (p,r) has exploited its monotone neighborhood. For negligible selections of rate r by the target transmit power, $\gamma_{p,r}$ is high and close to 1. As the target transmit power increases selections of rate r, $\gamma_{p,r}$ decreases, and the regret of DRS-TS decreases.

3.3.3 Proof of Theorem 2

For p such that $N_p(T) > 0$ and $r \neq r_p^*$, the expectation in (3.4) is decomposed into two terms as in [23, Theorem 2]: $\mathbb{E}[N_{p,r}(T+1)] = \mathbb{E}[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{r}_p(t) = r\}]$ $= \mathbb{E}[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) \neq r_p^*, \tilde{r}_p(t) = r\} + \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\}]$. We say that a rate r is suboptimal rate for a given transmit power p if $\Delta(p,r) > 0$. Since only the rate leader and its neighbors are explored, when the rate leader is r_p^* we only select from rates that lie in $\mathcal{N}(r_p^*) \cup \{r_p^*\}$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}[R(T)] \\
&= \underbrace{\sum_{p=p_1}^{p_P} \Bigg(\sum_{r \neq r_p^*} \Delta(p,r) \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) \neq r_p^*, \tilde{r}_p(t) = r\} \right]}_{(\mathbf{A})} \\
&\quad + \sum_{r \in \mathcal{N}(r_p^*)} \Delta(p,r) \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\} \right] \Bigg).
\end{aligned} \tag{3.5}$$

Note that $(\mathbf{A}) \leq \sum_{p=p_1}^{p_P} \sum_{r \neq r_p^*} \mathbb{E}[b_{p,r}(T+1)]$. Similar to [48, Theorem 5], a suboptimal rate r can be the rate leader for a given transmit power p only for a small number of times, and thus, we have $\mathbb{E}[b_{p,r}(T+1)] \leq C_1$, where C_1 is a positive constant.

For $\mathcal{N}(r_p^*) \neq \mathcal{N}'(p, r_p^*)$, let $r' = \mathcal{N}(r_p^*) \backslash \mathcal{N}'(p, r_p^*)$. We decompose second term in (3.5) as

$$\begin{aligned}
&= \underbrace{\Delta(p,r') \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r'\} \right]}_{(\mathbf{B})} \\
&\quad + \sum_{r \in \mathcal{N}'(p, r_p^*)} \Delta(p,r) \underbrace{\mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\} \right]}_{(\mathbf{C})}.
\end{aligned}$$

Note that $(\mathbf{B})$ is neglected when $\mathcal{N}(r_p^*) \backslash \mathcal{N}'(p, r_p^*) = \emptyset$. We have

$$(\mathbf{B}) \leq \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \frac{r'}{r_K} \geq \underline{\tilde{\theta}}_{p,r_p^*}(t)\} \Big]$$

$$+\mathbb{E}\Big[\sum_{t=1}^{N_p(T)}\mathbf{1}\{\tilde{L}_p(t)=r_p^*,\tilde{r}_p(t)=r',\frac{r'}{r_K}<\underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big]$$
$$\leq\mathbb{E}\Big[\sum_{t=1}^{N_p(T)}\mathbf{1}\{\tilde{L}_p(t)=r_p^*,\frac{r'}{r_K}\geq\underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big]\leq PC_a.$$

The above holds since $\underline{\tilde{\theta}}_{p,r'}(t)\in[0,\frac{r'}{r_K}]$, and for event $\{\tilde{L}_p(t)=r_p^*,\tilde{r}_p(t)=r',\frac{r'}{r_K}<\underline{\tilde{\theta}}_{p,r_p^*}(t)\}$ to happen we need $\underline{\tilde{\theta}}_{p,r'}(t)\geq\underline{\tilde{\theta}}_{p,r_p^*}(t)$ which cannot happen when $\underline{\tilde{\theta}}_{p,r_p^*}(t)>\frac{r'}{r_K}$. C_a is independent of T and comes from [76, Lemma 9], which bounds underestimation of Thompson sample for an optimal arm (rate) from a fixed distance, i.e., $\mu(p,r_p^*)-\delta=r'/r_K$.

For $r\in\mathcal{N}'(p,r_p^*)$, we have $(\mathbf{C})\leq(\mathbf{D})+(\mathbf{E})$, where

$$(\mathbf{D}):=\mathbb{E}\Big[\sum_{t=1}^{N_p(T)}\mathbf{1}\{\tilde{L}_p(t)=r_p^*,\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)>\underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big]$$
$$(\mathbf{E}):=\mathbb{E}\Big[\sum_{t=1}^{N_p(T)}\mathbf{1}\{\tilde{L}_p(t)=r_p^*,\ \tilde{r}_p(t)=r,\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)\leq\underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big].$$

Next, we bound $(\mathbf{E})$. Let $Q_t:=\{\tilde{\theta}_{p,r}(t)\leq\tilde{q}_{p,r}(t),\tilde{\vartheta}_{p,r}(t)\leq\tilde{o}_{p,r}(t)\}$. Note that $\tilde{L}_p(t)=r_p^*$ and $\tilde{r}_p(t)=r$ together imply that $\underline{\tilde{\theta}}_{p,r_p^*}(t)\leq\underline{\tilde{\theta}}_{p,r}(t)$ and recall that $\underline{\tilde{\theta}}_{p,r}(t)=\min\{\tilde{\theta}_{p,r}(t),\ \tilde{\vartheta}_{p,r}(t)\}$. Thus, we have

$$\begin{aligned}(\mathbf{E})&\leq\sum_{t=1}^{N_p(T)}\Big(\mathbb{P}(\tilde{L}_p(t)=r_p^*,\tilde{r}_p(t)=r,\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)\leq\tilde{\theta}_{p,r}(t),\\&\qquad\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)\leq\tilde{\vartheta}_{p,r}(t),Q_t)+\mathbb{P}(Q_t^c)\Big)\\&\leq\sum_{t=1}^{N_p(T)}\mathbb{P}(\tilde{L}_p(t)=r_p^*,\tilde{r}_p(t)=r,\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)\leq\tilde{u}_{p,r}(t),\\&\qquad\mu(p,r_p^*)-\tilde{\beta}_{p,r_p^*}(t)\leq\tilde{w}_{p,r}(t))+\sum_{t=1}^{N_p(T)}\mathbb{P}(Q_t^c).\end{aligned}\tag{3.6}$$

The last inequality comes from the fact that the event Q_t and fact $q_{p,r}(t)\leq u_{p,r}(t)$ together ensure that $\tilde{\theta}_{p,r}(t)\leq\tilde{q}_{p,r}(t)\leq\tilde{u}_{p,r}(t)$ and $\tilde{\vartheta}_{p,r}(t)\leq\tilde{o}_{p,r}(t)\leq\tilde{w}_{p,r}(t)$. Since, the samples $\tilde{\theta}_{p,r}(t)$ and $\tilde{\vartheta}_{p,r}(t)$ are not very likely to exceed the corresponding quantiles of the posterior distribution [74], we have

$$\sum_{t=1}^{N_p(T)}\mathbb{P}(Q_t^c)=\sum_{t=1}^{N_p(T)}\mathbb{P}(\tilde{\theta}_{p,r}(t)>\tilde{q}_{p,r}(t)\cup\tilde{\vartheta}_{p,r}(t)>\tilde{o}_{p,r}(t))$$

$$\leq P \sum_{t=1}^{T} \frac{1}{t \log(T)} \leq 2P.$$

Let $\tau^*_{p,t'}$ represent the round at which rate r^*_p is the rate leader for t'th time, when the transmit power is p. We bound the first term in (3.6) for $r \in \mathcal{N}'(p, r^*_p)$ as

$$\begin{aligned}
&\leq \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{L}_p(t) = r^*_p,\ \tilde{r}_p(t) = r,\ \tilde{N}^*_{p,r^*_p}\big(t\big) > (\tilde{b}_{p,r^*_p}(t))^b, \\
&\quad \mu(p, r^*_p) - \tilde{\beta}_{p,r^*_p}(t) \leq \tilde{u}_{p,r}(t), \mu(p, r^*_p) - \tilde{\beta}_{p,r^*_p}(t) \leq \tilde{w}_{p,r}(t)) \\
&\quad + \sum_{t=1}^{N_p(T)} \mathbb{P}\big(\tilde{L}_p(t) = r^*_p, \tilde{N}^*_{p,r^*_p}(t) \leq (\tilde{b}_{p,r^*_p}(t))^b\big) \\
&\quad \leq \sum_{t'=1}^{N_p(T)} \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{L}_p(t) = r^*_p,\ \tilde{b}_{p,r^*_p}(t) = t' - 1, \tilde{r}_p(t) = r, \tilde{N}^*_{p,r^*_p}\big(t\big) > (t'-1)^b, \\
&\qquad \mu(p, r^*_p) - \sqrt{\frac{6\log(t')}{\tilde{N}^*_{p,r^*_p}\big(t\big)}} \leq \tilde{u}_{p,r}(t), \mu(p, r^*_p) - \sqrt{\frac{6\log(t')}{\tilde{N}^*_{p,r^*_p}\big(t\big)}} \leq \tilde{w}_{p,r}(t)) + C'_b \\
&\quad \leq \sum_{t'=1}^{N_p(T)} \mathbb{P}(r(\tau^*_{p,t'}) = r, N^*_{p,r^*_p}\big(\tau^*_{p,t'}\big) > (t'-1)^b, \\
&\qquad \mu(p, r^*_p) - \sqrt{\frac{6\log(t')}{N^*_{p,r^*_p}\big(\tau^*_{p,t'}\big)}} \leq u_{p,r}(\tau^*_{p,t'}), \mu(p, r^*_p) - \sqrt{\frac{6\log(t')}{N^*_{p,r^*_p}\big(\tau^*_{p,t'}\big)}} \leq w_{p,r}(\tau^*_{p,t'})) \\
&\qquad + C'_b. \qquad (3.7)
\end{aligned}$$

Let $\alpha^{t'}_{p,r^*_p} = \sqrt{6\log(t')/N^*_{p,r^*_p}\big(\tau^*_{p,t'}\big)}$. Since $u_{p,r}(\tau^*_{p,t'}) \geq \mu(p, r^*_p) - \alpha^{t'}_{p,r^*_p}$, we have

$$\begin{aligned}
N^*_{p,r}(\tau^*_{p,t'}) d_r^+(\hat{\mu}_{p,r}(\tau^*_{p,t'}), \mu(p, r^*_p) - \alpha^{t'}_{p,r^*_p}) &\leq N_{p,r}(\tau^*_{p,t'}) d_r(\hat{\mu}_{p,r}(\tau^*_{p,t'}), u_{p,r}(\tau^*_{p,t'})) \\
&\leq \log(T) + \log(\log(T)) = f(T).
\end{aligned}$$

Similarly, condition $w_{p,r}(\tau^*_{p,t'}) \geq \mu(p, r^*_p) - \alpha^{t'}_{p,r^*_p}$ implies that

$$\begin{aligned}
&M_{p,r}(\tau^*_{p,t'}) d_r^+(\eta_{p,r}(\tau^*_{p,t'}), \mu(p, r^*_p) - \alpha^{t'}_{p,r^*_p}) \\
&\qquad \leq M_{p,r}(\tau^*_{p,t'}) d_r(\eta_{p,r}(\tau^*_{p,t'}), w_{p,r}(\tau^*_{p,t'})) \leq f(T).
\end{aligned}$$

Using the inequalities above, we bound the summation term in (3.7) as

$$\sum_{t'=1}^{N_p(T)} \mathbb{P}\big(r(\tau^*_{p,t'}) = r, N^*_{p,r^*_p}\big(\tau^*_{p,t'}\big) > (t'-1)^b,$$

$$\mu(p,r_p^*)-\alpha_{p,r_p^*}^{t'}\leq u_{p,r}(\tau_{p,t'}^*),\mu(p,r_p^*)-\alpha_{p,r_p^*}^{t'}\leq w_{p,r}(\tau_{p,t'}^*)\Big)$$
$$\leq E\Bigg[\sum_{t'=1}^{N_p(T)}\mathbf{1}\Big(r(\tau_{p,t'}^*)=r,N_{p,r_p^*}^*\big(\tau_{p,t'}^*\big)>(t'-1)^b,$$
$$N_{p,r}^*(\tau_{p,t'}^*)d_r^+(\hat{\mu}_{p,r}(\tau_{p,t'}^*),\mu(p,r_p^*)-\alpha_{p,r_p^*}^{t'})\leq f(T),$$
$$M_{p,r}(\tau_{p,t'}^*)d_r^+(\eta_{p,r}(\tau_{p,t'}^*),\mu(p,r_p^*)-\alpha_{p,r_p^*}^{t'})\leq f(T)\Big)\Bigg].$$

We introduce s and bound the sum above as follows with the fact that $N_p(T)\leq T$:

$$\leq E\Bigg[\sum_{t'=1}^{T}\sum_{s=1}^{t'}\mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*)=(s-1),r(\tau_{p,t'}^*)=r,\tilde{Z}_{p,r}^{s,*}>(t'-1)^b,$$
$$(s-1)d_r^+\big(\tilde{v}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T),$$
$$\tilde{M}_{p,r}^s d_r^+\big(\tilde{\eta}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T)\big)\Bigg].$$

By change of variables, and separating outer sum into two intervals, we have

$$\leq E\Bigg[\sum_{s=1}^{K_{p,r}^T+1}\sum_{t'=s}^{T}\mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*)=(s-1),r(\tau_{p,t'}^*)=r,\tilde{Z}_{p,r}^{s,*}>(t'-1)^b,$$
$$(s-1)d_r^+\big(\tilde{v}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T),$$
$$\tilde{M}_{p,r}^s d_r^+\big(\tilde{\eta}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T)\big)$$
$$+\sum_{s=K_{p,r}^T+2}^{T}\sum_{t'=s}^{T}\mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*)=(s-1),r(\tau_{p,t'}^*)=r,\tilde{Z}_{p,r}^{s,*}>(t'-1)^b,$$
$$(s-1)d_r^+\big(\tilde{v}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T),$$
$$\tilde{M}_{p,r}^s d_r^+\big(\tilde{\eta}_{p,r}^s,\mu(p,r_p^*)-\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big)\leq f(T)\big)\Bigg].$$

For first summation we use $\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\leq\tilde{g}_{p,r}^{s,*}$ because $t'\leq T$ and $\tilde{g}_{p,r}^{s,*}=\sqrt{6\log(T)/\tilde{Z}_{p,r}^{s,*}}$. For second summation we utilize the condition $\tilde{Z}_{p,r}^{s,*}>(t'-1)^b$ and have $\sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}<\sqrt{6\log(t')/(t'-1)^b}$. We introduce $h_t=\sqrt{\frac{6\log(t+1)}{(t)^b}}$, and thus

$$\leq E\Bigg[\sum_{s=1}^{K_{p,r}^T+1}\sum_{t'=s}^{T}\mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*)=(s-1),r(\tau_{p,t'}^*)=r,$$

$$
\begin{aligned}
&(s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) \le f(T), \\
&\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) \le f(T)\big) \\
+ \sum_{s=K_{p,r}^T+2}^{T} \sum_{t'=s}^{T} &\mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), r(\tau_{p,t'}^*) = r, \\
&(s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{t'-1}) \le f(T), \\
&\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - h_{t'-1}) \le f(T)\big)\Big].
\end{aligned}
$$

Note that $q \mapsto d_r^+(p,q)$ is nondecreasing, and for large enough t' (i.e., $t' > e^{1/b}$), $t' \mapsto h_{t'}$ is decreasing. Thus, for the second term above, for T such that $K_{p,r}^T > e^{1/b}$, we have $h_{t'-1} \le h_{K_{p,r}^T+1} < h_{K_{p,r}^T}$ due to the fact that $t'-1 \ge s-1 \ge K_{p,r}^T+1$. We separate t' and s terms and obtain

$$
\begin{aligned}
\le E\Bigg[\sum_{s=1}^{K_{p,r}^T+1} &\mathbf{1}\big((s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) \le f(T), \\
&\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) \le f(T)\big) \\
&\sum_{t'=s}^{T} \mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), r(\tau_{p,t'}^*) = r\big) \\
+ \sum_{s=K_{p,r}^T+2}^{T} &\mathbf{1}\big((s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \le f(T), \\
&\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \le f(T)\big) \\
&\sum_{t'=s}^{T} \mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), r(\tau_{p,t'}^*) = r\big)\Bigg].
\end{aligned}
$$

Since, $\sum_{t'=s}^{T} \mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), r(\tau_{p,t'}^*) = r\big) \le 1$ for every $s \in [1,T]$, and using the fact $\mathbb{P}(A,B) \le \mathbb{P}(A)$ for events A and B, we have

$$
\begin{aligned}
&\le \sum_{s=1}^{K_{p,r}^T+1} \mathbb{P}\big(\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) \le f(T)\big) \\
&+ \sum_{s=K_{p,r}^T+2}^{T} \mathbb{P}\big((s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \le f(T)\big). \qquad (3.8)
\end{aligned}
$$

We let $\gamma_{p,r} := \frac{\sum_{s=1}^{K_{p,r}^T+1} \mathbb{P}\left(\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) < f(T)\right)}{K_{p,r}^T+1}$, and write the first term in (3.8) as $\gamma_{p,r}(K_{p,r}^T+1)$.

We bound the second term in (3.8) as

$$\sum_{s=K_{p,r}^T+2}^{T} \mathbb{P}((s-1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \leq f(T))$$
$$\leq \sum_{s=K_{p,r}^T+2}^{T} \mathbb{P}((K_{p,r}^T+1)d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \leq f(T))$$
$$\leq \sum_{s=K_{p,r}^T+2}^{T} \mathbb{P}\Big(d_r^+(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - h_{K_{p,r}^T}) \leq \frac{d_r(\mu(p,r), \mu(p,r_p^*))}{1+\epsilon}\Big) \leq C_2$$

where C_2 is a constant from [74, Lemma 2] and utilizing the fact that $\mathbb{P}(\tilde{V}_{p,r}(s) \geq s) = 1$.

Finally, we bound (**D**):

$$(\mathbf{D}) \leq \sum_{t'=1}^{N_p(T)} \mathbb{P}(\mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{N_{p,r_p^*}^*(\tau_{p,t'}^*)}} > \theta_{p,r_p^*}(\tau_{p,t'}^*))$$
$$+ \sum_{t'=1}^{N_p(T)} \mathbb{P}(\mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{N_{p,r_p^*}^*(\tau_{p,t'}^*)}} > \vartheta_{p,r_p^*}(\tau_{p,t'}^*)) \leq P(N_0(b) + 3 + C_b') = C_3$$

where the last inequality comes from [74, Lemma 1]. Using above bounds, we obtain

$$\mathbb{E}[R(T)] = \sum_{p=p_1}^{p_P} \sum_{r \neq r_p^*} \Delta(p,r)\mathbb{E}[N_{p,r}(T+1)]$$
$$\leq \sum_{p=p_1}^{p_P} \sum_{r \in \mathcal{N}'(p,r_p^*)} \Big(\gamma_{p,r} \frac{\Delta(p,r)(1+\epsilon)}{d\big(\frac{\mu(p,r)}{(r/r_K)}, \frac{\mu(p,r_p^*)}{(r/r_K)}\big)}\Big) \log(T) + O(\log(\log(T))) + C$$

where $C < \infty$ is a constant given as $C := (P)(R-1)(C_1) + P^2 C_a + \sum_{p=p_1}^{p_P} \sum_{r \in \mathcal{N}'(p,r_p^*)} \Delta(p,r)\Big(C_2 + C_3 + \gamma_{p,r} + 2P + C_b'\Big)$.

3.4 Illustrative Results

In this section, we numerically evaluate the performance of DRS-TS for dynamic rate selection under time-varying transmit power and compare its performance with the state-of-the-art algorithms.

3.4.1 Competitor Learning Algorithms

1) *Dynamic Rate Selection via Thompson Sampling Without Contexts* (DRS-TS-NC): This is the non-contextual version of DRS-TS. It decouples the rate from throughput and exploits unimodality of the expected reward similar to MTS [31] and UTS [23].

2) *Contextual Upper Confidence Bound* (CUCB): This runs a separate instance of UCB1 [44] for each transmit power. It does not use rate unimodality.

3) *Dynamic Rate Selection via Kullback–Leibler Upper Confidence Bound* (DRS-KLUCB): This is the frequentist variant of DRS-TS, which is a simplified version of CUL [2], where the modified neighborhood is set beforehand similar to DRS-TS. This benchmark is used to test if the Bayesian approach is superior to the frequentist approach in practice.

4) *Contextual Unimodal Thompson Sampling* (CUTS): This is the contextual version of unimodal Thompson sampling (UTS) proposed in [23]. It runs a separate instance of UTS for each transmit power.

5) *Dynamic Rate Selection via Thompson Sampling without contextual monotonicity* (DRS-TS-NM): This runs a separate instance of DRS-TS-NC for each transmit power. It is a variant of DRS-TS that decouples the rate from the throughput, exploits the unimodality over rates and contextual information but ignores the monotonicity over the contexts. This benchmark evaluates the effect of exploiting the monotonicity over transmit powers on the regret.

3.4.2 Regret and Complexity Comparison

The expected regret of DRS-TS is compared with the other learning algorithms in Table 3.1. It is evident that exploiting unimodality reduces the search space to the neighborhood of the optimal rate. Additionally, exploiting the context monotonicity introduces the context dependent factor $\gamma_{p,r} \in [0, 1]$, which ensures

Table 3.1: Comparison of DRS-TS with state-of-the-art algorithms

Algorithm	Arm unimodality	Contexts	Context monotonicity	The Expected Regret
DRS-TS-NC	✓	×	×	$O(\lvert\mathcal{N}'(r^*)\rvert \log(T))$
CUCB [44]	×	✓	×	$O(\sum_{p\in\mathcal{P}} \sum_{r\neq r_p^*} \log(T))$
DRS-KLUCB [2]	✓	✓	✓	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T))$, $\gamma_{p,r} \in [0,1]$
CUTS [23]	✓	✓	×	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}(r_p^*)} \log(T))$
DRS-TS-NM	✓	✓	×	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \log(T))$
DRS-TS (this chapter)	✓	✓	✓	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T))$, $\gamma_{p,r} \in [0,1]$

the fact that $\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T) \leq \sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \log(T)$. While the frequentist analogue of DRS-TS achieves a regret bound similar to that of DRS-TS, our experimental results illustrate that the performance of DRS-TS is superior in practice.

DRS-TS-NC has the lowest complexity, since it neglects the context. However, it performs poorly. Contextual algorithms have higher complexity but they achieve a better performance than the non-contextual algorithms. DRS-TS has the highest computational complexity, since it needs to refine the throughput sample via exploiting the neighborhood contexts, whereas the performance achieved is greater than all of the competitor algorithms. Nevertheless, computational complexity is linear in terms of number of rates and transmit powers, and hence, in practice the algorithm can be deployed in a device with limited computational resources.

3.4.3 Experimental Setup

We consider a single-link system with 4 available rates and 18 transmit powers. The rate set is $\{2, 4, 6, 8\}$ bits per symbol (bps), which corresponds to modulation

Table 3.2: Throughput of power-rate pairs.

Throughput	r_1	r_2	r_3	r_4
p_1	**0.1233**	0.0602	0.0042	0.0000
p_2	**0.1236**	0.0623	0.0052	0.0000
p_3	**0.1247**	0.0630	0.0052	0.0000
p_4	**0.1264**	0.0658	0.0052	0.0000
p_5	**0.1288**	0.0672	0.0052	0.0000
p_6	**0.1292**	0.0699	0.0073	0.0000
p_7	0.1970	**0.2625**	0.1558	0.0305
p_8	0.1977	**0.2652**	0.1589	0.0332
p_9	0.1988	**0.2673**	0.1610	0.0332
p_{10}	0.1994	**0.2722**	0.1662	0.0332
p_{11}	0.2005	**0.2742**	0.1662	0.0332
p_{12}	0.2008	**0.2742**	0.1735	0.0346
p_{13}	0.2396	0.4301	**0.5287**	0.4571
p_{14}	0.2396	0.4307	**0.5287**	0.4584
p_{15}	0.2400	0.4321	**0.5287**	0.4584
p_{16}	0.2403	0.4328	**0.5308**	0.4626
p_{17}	0.2403	0.4335	**0.5319**	0.4640
p_{18}	0.2403	0.4356	**0.5319**	0.4709

schemes QPSK, 16QAM, 64QAM, and 256QAM. Contexts (transmit powers) correspond to average signal-to-noise-ratio (SNR) values ranging in $[2, 25]$dBs, and we set $T = 7.2 \times 10^4$. A snapshot of throughput calculated for various power-rate pairs is provided in Table 3.2. Packet success probabilities are obtained via sending packets over a *tapped delay line* (TDL) channel with Rayleigh Fading using 5G toolbox in MATLAB. For each experiment run, Bernoulli random numbers are generated by using the obtained probabilities and random rewards are generated by multiplying the obtained random number with the corresponding rate. Results are averaged over 50 runs to reduce the effect of randomness due to rate selections, transmit power arrivals and reward generation. We evaluate performance of the algorithms under three different types of context arrivals that are explained in the following subsections.

3.4.4 Type I Arrivals

We consider a sequence of ordered contexts $\{p_{18}, \ldots, p_1\}$, where each context arrives for a block of $T/18$ samples. Therefore, the expected reward for rate $r \in \mathcal{R}$ decreases monotonically, i.e., $\mu(p(t+1), r) \leq \mu(p(t), r)$. The monotonicity of throughput over transmit power is maximally utilized for Type I arrivals due to the fact that contexts with higher values of expected reward arrive early, and monotone neighborhood for most of the contexts provides a refined *throughput sample*. Fig. 3.2 compares the performance of DRS-TS-NC, CUCB, DRS-KLUCB, CUTS, DRS-TS-NM and DRS-TS for Type I arrivals. DRS-TS-NC tries to find a global optimal rate for all transmit powers, and hence, it incurs the largest regret. Although CUCB obtains the optimal rate for each transmit power, it still incurs a large regret due to not exploiting unimodality in the rates. Another point worth noting is that CUTS and DRS-TS-NM outperforms DRS-KLUCB, which provides evidence for the fact that the Bayesian approach results in faster learning than the frequentist approach. Also DRS-TS-NM outperforms CUTS, since DRS-TS-NM decouples the rate and packet success probability in addition to rate unimodality exploitation. Finally, DRS-TS achieves considerably smaller regret than DRS-TS-NC since it additionally utilizes the transmit power monotonicity. Also DRS-TS outperforms its frequentist counterpart DRS-KLUCB with a significant margin.

3.4.5 Type II Arrivals

We again consider a sequence of ordered contexts $\{p_1, \ldots, p_{18}\}$, where each context arrives for a block of $T/18$ samples. However, in this scenario the expected reward for rate $r \in \mathcal{R}$ increases monotonically i.e., $\mu(p(t+1), r) \geq \mu(p(t), r)$. The monotonicity of throughput over transmit power is not utilized for Type II arrivals, due to the fact that contexts with lower expected rewards arrive first, and hence, the monotone neighborhood for all of the contexts is an empty set. In Fig. 3.2 we show that similar to Type I arrivals, DRS-TS-NM outperforms

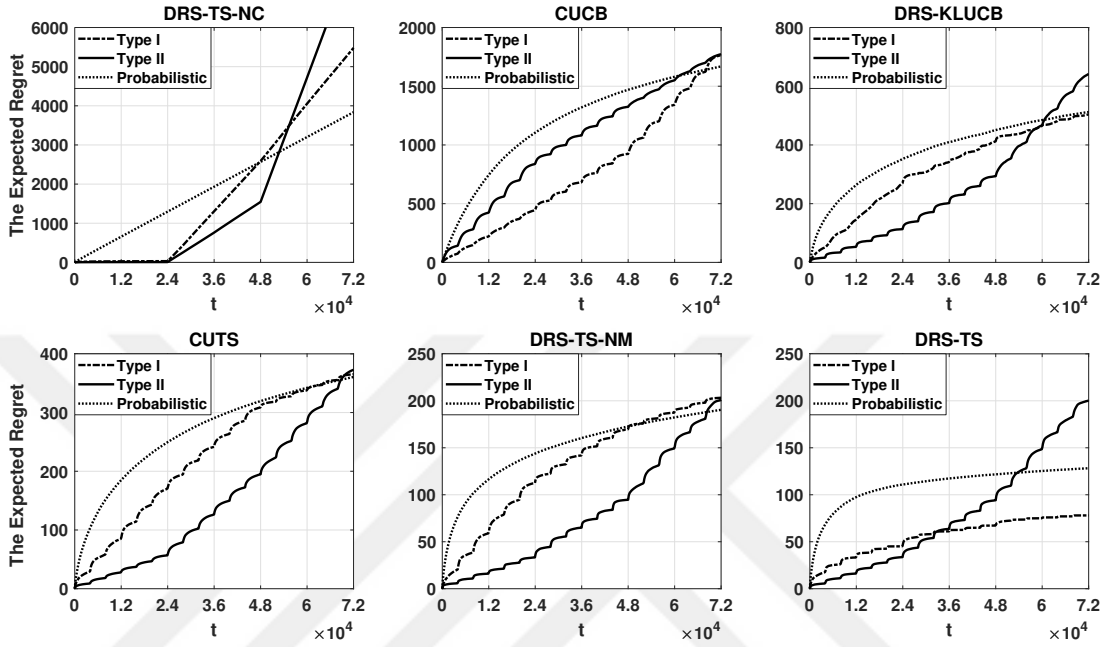


Figure 3.2: The expected regrets of DRS-TS-NC, CUCB, DRS-KLUCB, CUTS, DRS-TS-NM and DRS-TS.

DRS-TS-NC, CUCB, DRS-KLUCB and CUTS thanks to exploiting the rate decomposition, contextual information and rate unimodality. However, DRS-TS behaves similarly to DRS-TS-NM, since transmit power monotonicity does not have impact for Type II arrivals. Similar to Type I arrivals, a staircase pattern is observed for the contextual algorithms in Fig. 3.2. This pattern is due to the fact that the context arrivals follow a monotone trend. In contrast, the regret due to exploring suboptimal arms is growing logarithmic in time as seen by the form of each staircase step.

3.4.6 Probabilistic Arrivals

For this case, there are three sets of 6 transmit powers each for which the optimal rate is same as given in Table 3.2. Each time a set is selected uniformly at random and then sorted (descending) transmit powers inside the selected set are selected with the following probabilities $\{\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}, \frac{1}{21}\}$. The monotonicity

of throughput over transmit powers is considerably utilized for probabilistic arrivals, due to the fact that contexts with higher values of expected reward arrive more frequently within each set, and monotone neighborhood for most of the contexts can provide refined a *throughput sample*. The utilization of transmit power monotonicity in probabilistic case is in between the two extreme cases (Type I and II arrivals). Fig. 3.2 shows that similar to Type I and Type II arrivals, DRS-TS-NM outperforms DRS-TS-NC, CUCB, DRS-KLUCB and CUTS. DRS-TS outperforms DRS-TS-NM since transmit power monotonicity is considerably utilized.

Chapter 4

Exploiting Structure in Multi-Armed Bandits under Time-Varying Constraints

This section was published in [4].[1]

In this chapter, we study the dynamic rate and channel adaption under time-varying constraints for cognitive radio networks. In cognitive radio networks, a SU usually has access to only those resources (e.g., wireless channels) which are free from their predefined owners i.e., PUs at a given time. Thus, the SU opportunistically accesses these resources and makes cognitive decisions aiming to maximize its performance (e.g., throughput) [5]. In heterogeneous networks, there is a multitude of channels each offering time-varying transmission opportunities. In addition to spectrum sharing, rate adaptation is also an important problem in wireless networks, where a transmitter has access to a finite set of transmission rates to choose from at each decision epoch. As the optimal transmission rate and the channel conditions dynamically change over time, traditional RA algorithms (e.g., SampleRate, MiRA, etc.) are outperformed by AI-based learning strategies

[1] © [2020] IEEE. Reprinted, with permission, from M. A. Qureshi and C. Tekin, "Rate and Channel Adaptation in Cognitive Radio Networks under Time-Varying Constraints", in Communications Letters, 2020.

Table 4.1: Comparison of V-CoTS with related works

Algorithm	KL-UCB -U [5]	RATS [77]	G-ORS [30]	MTS [31]	CoTS [32]	CUL [2] (ch. 2)	DRS-TS [3] (ch. 3)	V-CoTS (This chapter)
Rate adaptation	✓	✓	✓	✓	✓	✓	✓	✓
Channel selection (MIMO mode)	✓	✓	✓	×	×	×	×	✓
Structure exploitation	✓	×	✓	×	✓	✓	✓	✓
Bayesian learning	×	✓	×	✓	✓	×	✓	✓
Dynamic rate and channel sets	×	×	×	×	×	×	×	✓

[5, 30–32, 77].

Furthermore, serving heterogeneous applications with diverse quality of service (QoS) requirements may result in different feasible transmission rates set (e.g., minimum transmission rate constraint or maximum transmission rate limit) for each application. Non-identical traffic models are studied in [78], where a high-end application (e.g., uncompressed video) requires higher constellation (e.g., higher than QPSK), low-end applications are limited to lower constellation, and some applications are subject to medium constellation size (e.g., 16-QAM) to achieve good transmission rate as well as good success probability. Given a modulation scheme, different coding schemes (e.g., 1/2, 3/4, etc.) are available for exploration to satisfy the desired QoS requirement. Therefore, every transmission rate is not a candidate to be explored for the search of the optimal transmission rate [77]. Last but not least, the expected rewards over these resources (e.g., transmission rates) are related and follow a predefined structure. One notable example is the monotonicity of the success probability over transmission rates, which indicates that higher transmission rates result in lower transmission success probabilities [32].

In [5], authors propose a non-contextual frequentist algorithm for rate and

channel selection, based on the Kullback-Leibler upper confidence bound (KL-UCB), which exploits the unimodal structure of the expected rewards over rate-channel pairs in a steep throughput scenario. Since, the Bayesian approach is proved to be more efficient [3, 74], authors in [31] utilize Thompson sampling for rate selection in wireless networks. The Bayesian approach is extended to exploit the monotone structure of the success probability in the transmission rates in [32]. Authors in [77] propose RATS for rate adaptation in IEEE 802.11ac MIMO systems to efficiently explore combinations of MCS, channel width, and MIMO mode triple from a predefined reduced set. In [2], authors presented a contextual version of the transmission rate selection problem under time-varying transmit power using the frequentist approach, by exploiting the structure over both the transmission rates and the transmit powers. Later, authors presented a Bayesian counterpart of the proposed frequentist algorithm in [3].

Departing form earlier works, in this chapter we consider rate-channel pair selection in unknown, time-varying channel conditions, and dynamically varying rate-channel sets. The model we consider captures majority of the challenges posed in cognitive radio networks over the rapidly varying channels, in which heterogeneous applications are served by the SU under time-varying PU activity. We propose V-CoTS, which exploits the problem structure to make sequential cognitive decisions for applications with diverse QoS metrics (see Section 4.2). Performance of V-CoTS significantly exceed that of well-known learning methods in such highly dynamic and volatile communication scenarios (see Section 4.3). Table 4.1 lists the key differences of our work from the earlier works.

4.1 Problem Formulation

4.1.1 System Model

The SU has K different rates to choose from, represented by the set $\mathcal{R} = \{r_1, \ldots, r_K\}$. There are C channels, represented by the set $\mathcal{C} = \{c_1, \ldots, c_C\}$.

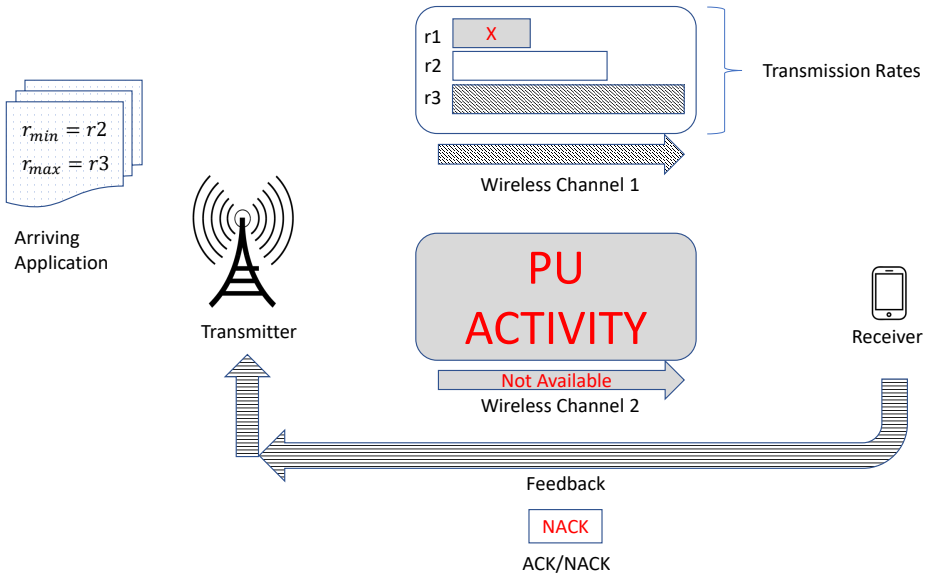


Figure 4.1: Graphical demonstration of dynamic rate and channel adaptation under time-varying constraints. An arriving application has the feasible transmission rate set as $\{r2, r3\}$, the wireless channel 2 is not available due to primary activity. At a given time slot $r3$ is selected and feedback of NACK is received.

The set of rates $\mathcal{R}$ is ordered such that $r_1 < \ldots < r_K$. For $c \in \mathcal{C}$ and $r \in \mathcal{R}$, j_c and i_r represent the indices of channel c and rate r, i.e., $c_{j_c} = c$ and $r_{i_r} = r$. The SU serves heterogeneous applications, and each application has its feasible transmission rate set ranging from the minimum rate to be guaranteed to the maximum allowable transmission rate. We also assume that the number of available channels vary with time, due to PU activity. The SU performs channel sensing at the beginning of each round to determine the set of available channels. Graphical demonstration is provided in Fig. 4.1.

4.1.2 MAB Formulation and Reward Structure

In the MAB formulation, the SU makes decisions sequentially over rounds indexed by $t \in [T]$, where T represents the time horizon. At the beginning of round t, the SU observes the application to be served in that round and its feasible rate set[2] $\mathcal{R}_t$ and performs channel sensing to determine the set of free channels[3] $\mathcal{C}_t$. Then, the SU chooses a rate from $\mathcal{R}_t$ and channel from $\mathcal{C}_t$, and transmits at rate $r(t) \in \mathcal{R}_t$ bits over $c(t) \in \mathcal{C}_t$. At the end of the round, it receives ACK/NACK feedback $x_{r(t),c(t)}(t)$, which indicates that whether the packet transmission was successful or not, and then, collects the reward as $r(t) \cdot x_{r(t),c(t)}(t)$. Here, $x_{r,c}(t)$ is a Bernoulli random variable with expected value $\psi_{r,c}$ that takes value 1 if the transmission given rate-channel pair (r, c) in round t is successful and value 0 otherwise. We call $\psi_{r,c}$ the transmission success probability and $\mu_{r,c} = r \cdot \psi_{r,c}$ the throughput associated with rate-channel pair (r, c) [3]. We call $\mu_{r,c}/r_K$ the normalized throughput of rate-channel pair (r, c).

Transmission success probabilities exhibit a monotonically decreasing structure over the set of rates [32], given as $\psi_{r_1,c} > \psi_{r_2,c} \ldots > \psi_{r_K,c}$. The optimal rate-channel pair at round t is denoted by $(r^*(t), c^*(t)) = \text{argmax}_{r \in \mathcal{R}_t, c \in \mathcal{C}_t} \mu_{r,c}$. Without loss of generality we assume that $(r^*(t), c^*(t))$ is unique for every t.

[2]The feasible set at round t is represented as $\mathcal{R}_t = \{r_{m_t}, r_{m_t+1}, \ldots, r_{n_t}\}$, where $m_t \leq n_t$ and $m_t, n_t \in [K]$, and m_t represents the minimum rate to be guaranteed and n_t represents the maximum allowable transmission rate in round t. Our algorithm will also work even when $\mathcal{R}_t$ is an arbitrary subset of $\mathcal{R}$.

[3]When channels are non-volatile, then $\mathcal{C}_t = \mathcal{C}, \forall t \leq T$.

4.1.3 Regret Definition

Let $\mathcal{A}_t = \mathcal{R}_t \times \mathcal{C}_t$ represent the available action set in round t. For any T round available action sequence $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_T\}$, the expected regret is defined as

$$R_{\mathcal{A}}(T) = \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{r^*(t),c^*(t)} - \mu_{r(t),c(t)}\right)\Big|\mathcal{A}\right]. \tag{4.1}$$

It is obvious that a good learning algorithm should incur small regret. Our goal is to design a learning algorithm for the SU that minimizes the growth rate of the regret, which accounts for maximizing the cumulative throughput.

4.2 The Learning Algorithm

V-CoTS takes into account the monotone structure of the success probability in transmission rates and time-varying availability of rate-channel pairs while minimizing the regret (pseudocode is given in Algorithm 3, graphical demonstration in Fig. 4.2). Its main novelty lies in optimizing channel selections together with rate selections under dynamically varying action sets. It is important to note that the SU does not have any control over the available action set, which makes efficient learning much more challenging compared to prior works. In addition, V-CoTS also utilizes monotonicity of $\psi_{r,c}$ in rates to learn fast [32].

V-CoTS keeps a posterior distribution $\pi_{r,c} = Beta(1 + S_{r,c}, 1 + N_{r,c} - S_{r,c})$ of $\psi_{r,c}$ for each (r, c), where $N_{r,c}(t)$ represents the number of times rate-channel pair (r, c) was selected before round t, and $S_{r,c}(t)$ represents the number of times transmission was successful in rounds where (r, c) was selected before round t. These distributions are used to form sample estimates $\phi_{r,c}$ of $\psi_{r,c}$ for each $(r, c) \in \mathcal{A}_t$. When forming these samples, V-CoTS ensures that the samples are consistent with the monotone structure of $\psi_{r,c}$ in rates. Let $\Phi_{t,c} = \{(\phi_{r_{m_t},c}, \ldots, \phi_{r_{n_t},c}) \mid \phi_{r,c} \geq \phi_{r',c}, \forall r < r',\ r, r' \in \mathcal{R}_t\}$ represent the set of samples for channel c that satisfy monotonicity in round t. Basically, for each $c \in \mathcal{C}_t$, V-CoTS takes samples $\boldsymbol{\phi}_{t,c} =$

Algorithm 3 V-CoTS

Input: K, C
Initialize: $t = 1$
Counters: $S_{r,c} = 0, N_{r,c} = 0, \forall r \in \mathcal{R}, \forall c \in \mathcal{C}$
while $t \geq 1$ **do**
    Observe available channels $\mathcal{C}_t$
    Observe feasible rate set $\mathcal{R}_t$
    for $c \in \mathcal{C}_t$
        Draw $\boldsymbol{\phi}_{t,c} \sim \mathbf{1}(\boldsymbol{\phi}_{t,c} \in \Phi_{t,c}) \times \prod_{r \in \mathcal{R}_t} \pi_{r,c}$
        $\theta_{r,c} = r \cdot \phi_{r,c}$, $\forall r \in \mathcal{R}_t$
    end for
    $[r(t), c(t)] = \text{argmax}_{r \in \mathcal{R}_t, c \in \mathcal{C}_t} \ \theta_{r,c}$
    Transmit using $[r(t), c(t)]$, observe feedback $x_{r(t),c(t)}(t)$
    $S_{r(t),c(t)} = S_{r(t),c(t)} + x_{r(t),c(t)}(t)$
    $N_{r(t),c(t)} = N_{r(t),c(t)} + 1$
    $t = t + 1$

$(\phi_{r_{m_t},c}, \ldots, \phi_{r_{n_t},c})$ such that

$$\boldsymbol{\phi}_{t,c} \sim \mathbf{1}(\boldsymbol{\phi}_{t,c} \in \Phi_{t,c}) \times \prod_{r \in \mathcal{R}_t} \pi_{r,c} \tag{4.2}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

The above selection ensures that non-zero probability is assigned only to the samples which belong to the set $\Phi_{t,c}$. Sampling from (4.2) can be achieved by first sampling $\boldsymbol{\phi}_{t,c} \sim \prod_{r \in \mathcal{R}_t} \pi_{r,c}$, and then, rejecting the samples until the obtained samples belong to $\Phi_{t,c}$.[4] A more efficient sampling method that performs rejection by exploiting the inverse transform property, called *sequential inverse transform sampling*, is discussed in [32].

Obtained sample success probabilities are then multiplied by rates $r \in \mathcal{R}_t$ to obtain throughput samples denoted by $\theta_{r,c}$ for all available rates in channel c. The process is repeated for each available channel $c \in \mathcal{C}_t$. Finally, at the end of round t, the rate-channel pair in $\mathcal{A}_t$ which has the maximum throughput sample is selected for transmission. At the end of the round, the ACK/NACK

[4] Any other structure can be incorporated to V-CoTS simply by defining appropriate constraint sets $\Phi_{t,c}$ and performing rejection sampling. When there is no structure, we can simply set $\Phi_{t,c} = [0,1]^{|\mathcal{R}_t|}$.

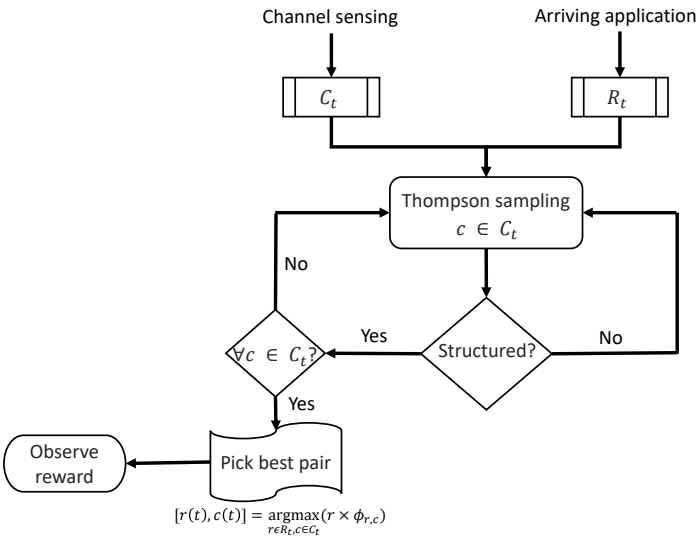


Figure 4.2: Flow chart of V-CoTS algorithm.

feedback $x_{r(t),c(t)}(t)$ and the throughput $r(t) \cdot x_{r(t),c(t)}(t)$ are observed. Finally, the posterior distribution of rate-channel pair $(r(t), c(t))$ is updated using the observations gathered in round t.

Additional Remark

Contextual information can be incorporated to V-CoTS simply by running a separate instance of V-CoTS for each context.

4.3 Simulation Results

4.3.1 Setup

We have $K = 10$ rates and $C = 9$ channels. Transmission success probabilities for each rate and channel are given in Table 4.2, where the optimal rate for a

Table 4.2: Success Probabilities over rate-channel pairs

r_k (Mbit/s)	MCS	ψ_{r_k,c_1}	ψ_{r_k,c_2}	ψ_{r_k,c_3}	ψ_{r_k,c_4}	ψ_{r_k,c_5}	ψ_{r_k,c_6}	ψ_{r_k,c_7}	ψ_{r_k,c_8}	ψ_{r_k,c_9}
1386.00	(QPSK,$\frac{1}{2}$)	0.95	0.85	0.99	0.90	0.85	0.98	0.95	0.80	0.95
1732.50	(QPSK,$\frac{5}{8}$)	0.90	0.75	0.96	0.80	0.80	0.93	0.85	0.70	0.90
2079.00	(QPSK,$\frac{3}{4}$)	0.85	0.70	0.93	0.75	0.75	**0.87**	0.80	0.65	0.85
2772.00	(16-QAM,$\frac{1}{2}$)	0.75	0.55	**0.90**	0.65	0.65	0.22	0.65	0.52	0.80
3465.00	(16-QAM,$\frac{5}{8}$)	0.65	**0.46**	0.18	0.60	**0.60**	0.16	0.60	**0.45**	0.75
4158.00	(16-QAM,$\frac{3}{4}$)	**0.60**	0.35	0.13	**0.55**	0.45	0.12	**0.55**	0.35	0.70
4504.50	(16-QAM,$\frac{13}{16}$)	0.45	0.30	0.10	0.45	0.35	0.10	0.45	0.27	0.68
5197.50	(64-QAM,$\frac{5}{8}$)	0.25	0.20	0.07	0.35	0.15	0.07	0.30	0.20	0.66
6237.00	(64-QAM,$\frac{3}{4}$)	0.15	0.10	0.04	0.20	0.10	0.04	0.20	0.15	0.64
6756.75	(64-QAM,$\frac{13}{16}$)	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.10	**0.62**

given channel is shown in bold. The expected reward (throughput) is calculated as the product of success probability and corresponding transmission rate. We consider three types of channels. i) Gradual channel: the optimal rate has success probability greater than 0.5, ii) Lossy channel: the optimal rate has success probability less than 0.5, and iii) Steep channel: rates have success probabilities either close to 1 or 0 [30,32]. All channels used in the simulation are generated by adding perturbation to the three main categories. We normalize the throughput by r_K to have the expected rewards in $[0, 1]$. We set $T = 2.5 \times 10^4$ and present results by averaging over 20 runs.

We assume each application served by the SU remains active for a random amount of time, and the lifetime (in number of rounds) of an application is sampled from the uniform distribution in $[1, T/25]$. For each application, feasible rate set is also sampled uniformly from 3 categories, i) QPSK and 16-QAM (i.e., $\{r_1, \ldots, r_7\}$) for low-end applications, ii) 16-QAM and 64-QAM for high-end applications, and iii) 16-QAM only for others. For simplicity, we assume channel 1 is available all the time. For the remaining channels, we set the availability probabilities as $\{0.8, 0.7, 0.6, 0.7, 0.7, 0.6, 0.7, 0.5\}$, respectively. Let p_i be the availability probability of channel i. We model the PU activity on channel i as follows. First, we sample $\chi_i \sim$ Bernoulli(p_i) and $\tau_i \in$ Uniform$([1, T/50])$. If $\chi_i = 1$, then channel

i will be available in the next τ_i rounds. Else, it will be unavailable in the next τ_i rounds. This model generates bursty PU activity similar to the Gilbert-Elliot channel model [79].

4.3.2 Competitor Algorithms

i) Oracle: Clairvoyant benchmark that always selects the optimal rate-channel pair in each round (not feasible in practice).

ii) V-TS: A variant of vanilla Thompson Sampling (TS) [80], which runs on the volatile action set $\mathcal{A}_t$.

iii) V-UCB: A variant of UCB1 [44], which runs on the volatile action set $\mathcal{A}_t$.

iv) CoTS: A constrained Thompson sampling benchmark [32] that exploits the monotonicity of the success probability in transmission rates but ignores the volatility of both the channels and the transmission rates.

v) CV-CoTS: A variant of the proposed algorithm V-CoTS, which exploits the monotonicity of the success probability in transmission rates as well as the volatility of the channels but ignores the volatility of transmission rates.

A non-volatile algorithm updates the success probabilities for an available channel, however, zero reward (throughput) is obtained for an unavailable rate selection (i.e., when QoS is not satisfied), or when the selected channel is busy due to PU activity.

4.3.3 Results

Fig. 4.3 shows that the optimal (Oracle's) average throughput is around 2944 *Mbit/s*. V-CoTS follows very closely to this and has average throughput around 2885 *Mbit/s*. On the other hand, the average throughputs achieved by V-TS,

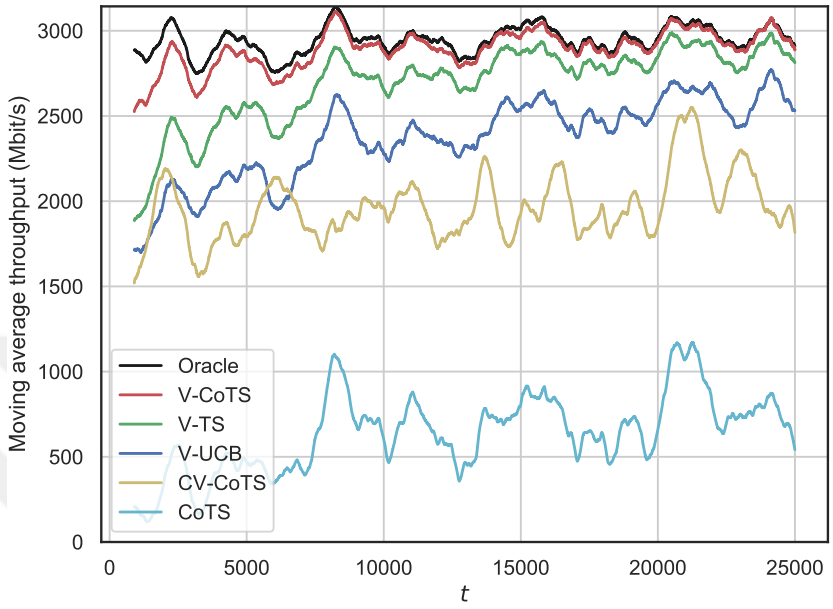


Figure 4.3: Moving average throughput: each value at t is averaged over 20 repetitions of the experiment and previous 900 packet transmissions.

V-UCB, CV-CoTS and CoTS are around 2683, 2362, 1952 and 632 *Mbit/s*, respectively. The average throughput achieved by V-CoTS is around 98% of the optimal strategy, and is around 7.5%, 22%, 48% and 356% higher than that of V-TS, V-UCB, CV-CoTS and CoTS, respectively.

Fig. 4.4 compares the expected regrets of V-CoTS, V-TS, V-UCB, CV-CoTS and CoTS. It is observed that V-CoTS has the minimum expected regret, and its expected regret at the final round is less than 1/4th of the expected regret of the closest competitor, which is V-TS. This improvement is mainly due to constrained sampling. It is also observed that V-TS outperforms its frequentist counterpart V-UCB by significant margin. It is evident that exploiting only the monotone structure is not sufficient and CoTS performs poorly. A substantial reduction in the regret is achieved by exploiting volatility in the channels in addition to monotonicity in the transmission rates by CV-CoTS, however, this reduction still proves to be non-optimal. Fig. 4.5 compares the total accumulated number of successfully transmitted Mbits for Oracle, V-CoTS, V-TS, V-UCB, CV-CoTS and CoTS. V-CoTS is the one closest to Oracle. V-TS, V-UCB and CV-CoTS

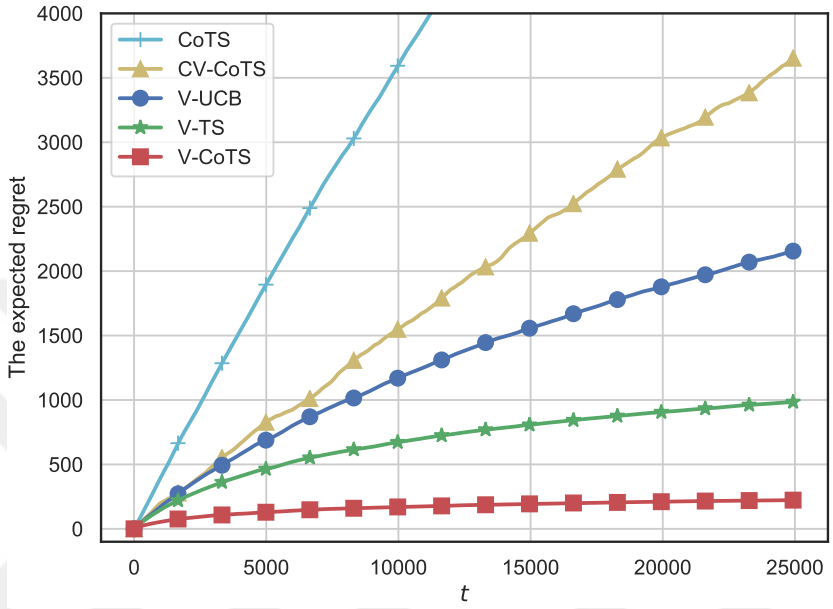


Figure 4.4: The expected regret by round t.

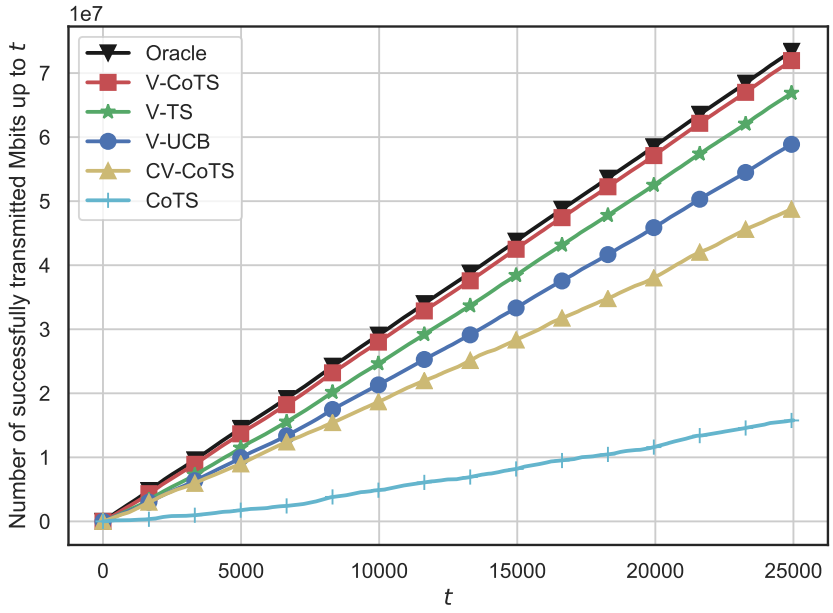


Figure 4.5: Number of successfully transmitted Mbits up to round t.

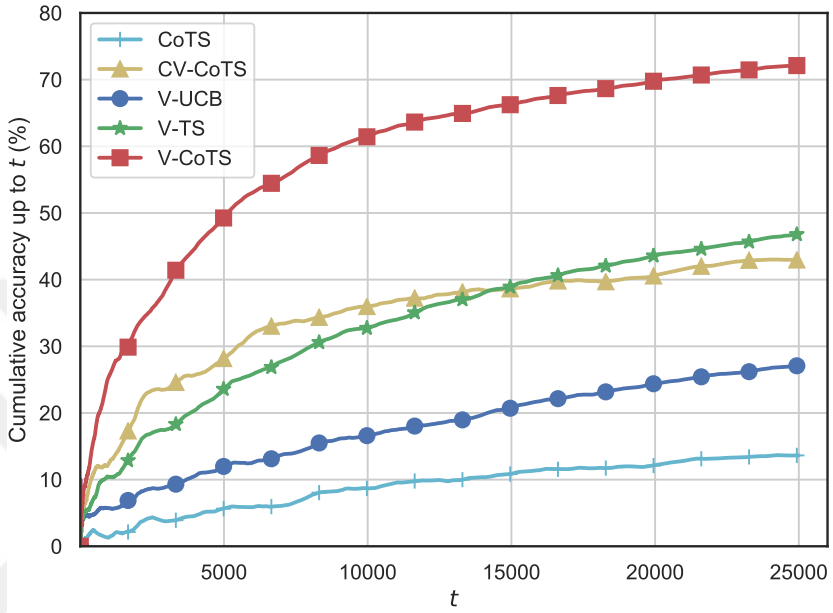


Figure 4.6: Accuracy as a function of t.

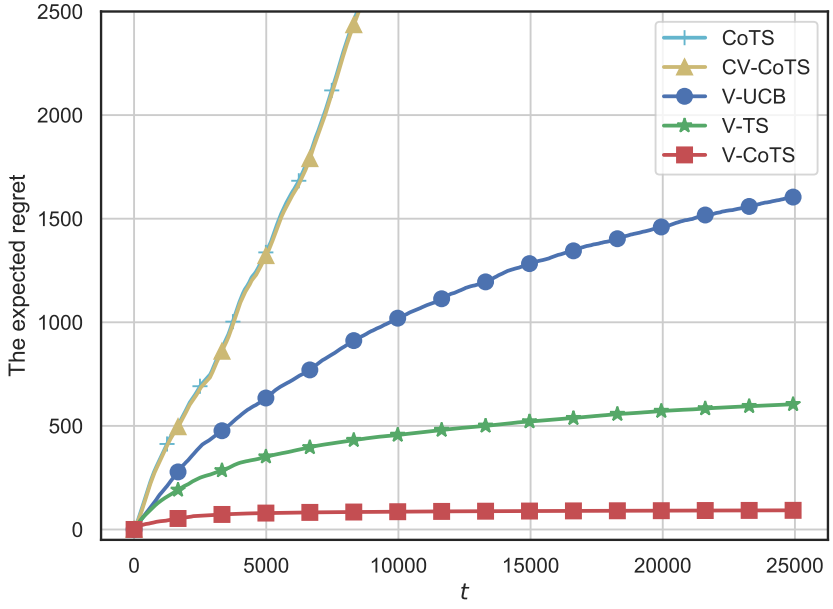


Figure 4.7: The expected regret by round t (non-volatile channels).

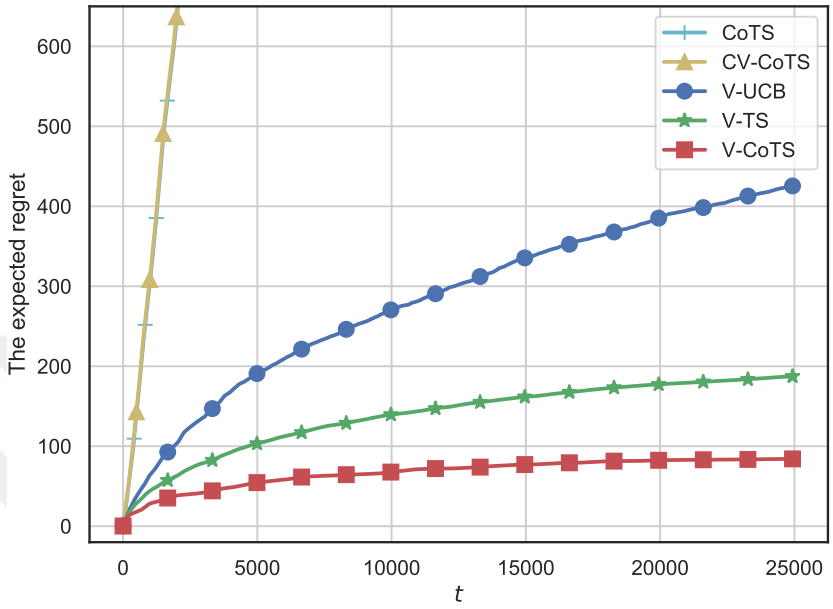


Figure 4.8: The expected regret by round t (one-dimensional actions).

are slow in learning due to the fact that they ignore the monotone (or volatile) structure, while CoTS is the worst.

Fig. 4.6 compares the accuracies of V-CoTS, V-TS, V-UCB, CV-CoTS and CoTS. Accuracy is defined as the fraction of times the selected rate-channel is the optimal one. We observe that by the end, V-CoTS is able to select the optimal rate-channel pair more than 70% of the time, while the closest competitor can only reach around 50%. CV-CoTS and V-TS achieve accuracies close to each other, however, the regret of CV-CoTS is higher than V-TS due to the fact that CV-CoTS receives zero reward when QoS is not satisfied and worst possible regret is incurred in those rounds. Fig. 4.7 and Fig. 4.8 provide regret plots for non-volatile channels ($\mathcal{C}_t = \mathcal{C} = \{c_1, c_2, c_4, c_5, c_9\}, \forall t \in [T]$) and transmission rate only actions space ($\mathcal{C} = \{c_9\}$), respectively. The reason behind the similar regret of CV-CoTS and CoTS is the non-volatility of the channels. It is evident that V-CoTS is able to achieve the minimum expected regret in both these simplified scenarios.

Chapter 5

Exploiting Structure in Contextual Multi-Armed Bandits for Continuous Arms and Contexts

In this chapter, we consider a variant of contextual MAB (with context set given as $[0, 1]$) in which the arm set is continuous (taken to be $[0, 1]$ for simplicity). We assume that for a given context, the expected reward is a unimodal function over the set of arms. Moreover, the optimal arms are monotonically increasing over the set of contexts. At the beginning of each round, a context $x \in [0, 1]$ arrives, and then, the learner selects an arm $a \in [0, 1]$, and observes a random reward from a fixed but unknown probability distribution parameterized by x and a. Continuous intervals for both the arms and the contexts exhibiting structural properties make the proposed setting significantly different than earlier works on non-contextual versions of unimodal bandits [21], and contextual versions over discrete set of arms and contexts [2,3]. Basically, we adapt *line search elimination* (LSE) algorithm in [21] to work in the proposed contextual setting by utilizing the structure of the optimal arm over contexts to learn fast.

5.1 Problem Formulation

5.1.1 System Model

Let $\mathcal{A} := [0, 1]$ denote the set of arms and $\mathcal{X} := [0, 1]$ denote the set of contexts. Our algorithm can also work under a discrete real-valued context set $\mathcal{X}$ by normalizing the minimum and maximum context to be 0 and 1. In the MAB formulation, the learner makes decisions sequentially over time slots indexed by $t \in [T]$, where T represents the time horizon. At the beginning of t, the learner receives the context, denoted by $x(t)$. Then, the learner chooses an arm denoted by $a(t)$. At the end of the round it collects the sample reward from a probability distribution with expected value $\mu(x, a)$. The optimal arm given a context x is denoted by $a_x^* = \text{argmax}_{a \in \mathcal{A}} \mu(x, a)$. Without loss of generality we assume that a_x^* is unique. The optimal arm over the contexts is monotonically increasing, i.e., for given contexts $x_i, x_j \in \mathcal{X}$ such that $x_i < x_j$, we have $a_{x_i}^* \leq a_{x_j}^*$.

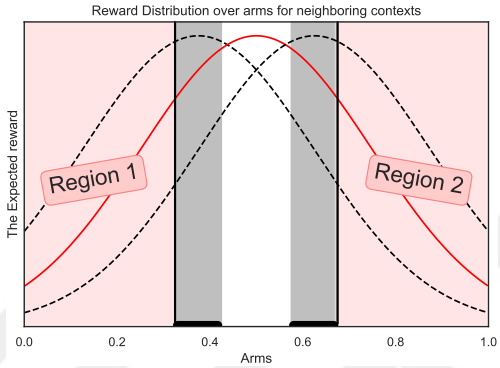
Assumption 1. *[21] The expected reward function $\mu(x, a)$ is called unimodal in the arms if for any given x there exist an optimal arm $a_x^* \in \mathcal{A}$ such that $\mu(x, a)$ is monotonically increasing in the interval $[0, a_x^*]$, and monotonically decreasing in the interval $[a_x^*, 1]$.*

Assumption 2. *The optimal arm a_x^* exhibits a monotone structure over the contexts, i.e., $\forall x_i, x_j \in \mathcal{X}$, such that for $x_i < x_j$ we have $a_{x_i}^* \leq a_{x_j}^*$.*

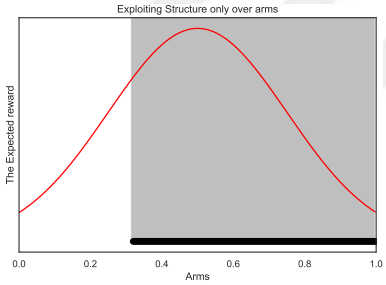
Assumption 3. *For a given arm a, $L > 0$, and $x, x' \in \mathcal{X}$ we have $|\mu(x, a) - \mu(x', a)| \leq L|x - x'|$.*

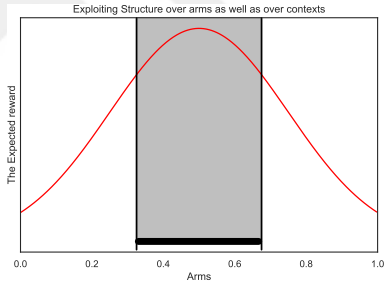
Assumption 4. *[21] For a given context x, $C_H > 0$, and $a, a' \in \mathcal{A}$ we have $|\mu(x, a) - \mu(x, a')| \leq C_H|a - a'|$. Furthermore, there exists $C_L > 0$ for $a, a' \in [a_x^* - C_L, a_x^* + C_L]$ such that $|\mu(x, a) - \mu(x, a')| \leq C_L|a - a'|$ and $|\mu(x, a) - \mu(x, a')| \geq \frac{C_L}{\varphi}|a - a'|$ for $a, a' \in [0, a_x^* - C_L]$ or $a, a' \in [a_x^* + C_L, 1]$, where $\varphi \approx 1.618$.*

Assumptions 3 and 4 indicate that the expected reward is similar for similar contexts given an arm and similar for similar arms given a context.



(a) The expected rewards over arms for 3 neighboring contexts





(b) Utilizing structure over arms (c) Utilizing structure over arms and contexts

Figure 5.1: Demonstration of structure exploitation in 3 neighboring contexts for an example expected reward distributions and continuous arms set in $[0, 1]$: (a) The expected reward distribution for center context is shown in red solid line and is under-explored, whereas its left and right neighboring contexts with black color dashed lines are well-explored, and thus, both have reduced intervals for optimal arms as shown by gray shaded intervals (i.e., $[0.33, 0.42]$ and $[0.58, 0.67]$, respectively). Due to monotonicity of optimal arms in the contexts, optimal arm for center context cannot lie in Region 1 (i.e., $[0, 0.33]$) and Region 2 (i.e., $[0.67, 1]$) with high probability. (b) The structure over the arms is exploited, and optimal arm interval is reduced by an elimination algorithm (e.g., LSE in [21]) from one side only. (c) In addition to utilizing the structure over arms, structure of optimal arms over the contexts is utilized, which results in further reduction of the optimal arm search interval.

For a given sequence of contexts $x(1), \ldots, x(T)$, the (pseudo) regret after the first T rounds is defined as

$$R(T) = \sum_{t=1}^{T} \Big(\mu(x(t), a^*_{x(t)}) - \mu(x(t), a(t)) \Big). \tag{5.1}$$

The advantage of exploiting the structure of optimal arms over the contexts is explained in Fig. 5.1.

5.2 Algorithm

We propose Structured Contextual Unimodal Bandits (S-CUB), which is an online reinforcement learning algorithm that takes into account unimodality of $\mu(x, a)$ in arms and monotonicity of the optimal arm in contexts to maximize the expected reward (equivalently minimize the expected regret). S-CUB exploits unimodality of $\mu(x, a)$ in arms somewhat similar to LSE. The main novelty of S-CUB comes from exploiting the contextual information and monotonicity of the optimal arm in contexts. It is worth noting that the learner does not have any control over the context arrivals, and thus, exogenous context arrivals makes the problem much more complex and challenging than the non-contextual state-of-the-art.

We partition the context space $\mathcal{X}$ into m equal intervals to form a partition set $\mathcal{P}$. For each $p \in \mathcal{P}$, we select a set of four arms i.e., $\mathcal{A}_p = \{a_p^L, a_p^M, a_p^N, a_p^H\}$ which are initialized as follows:

$$\begin{aligned} a_p^L &= 0 \\ a_p^H &= 1 \\ a_p^M &= (\varphi \cdot a_p^L + a_p^H)/(1+\varphi) \\ a_p^N &= (a_p^L + \varphi \cdot a_p^H)/(1+\varphi)\ . \end{aligned}$$

For each $p \in \mathcal{P}$ and arms $a \in \mathcal{A}_p$, S-CUB keeps the counter $N_{p,a}(t)$, which counts the number of selections of arm a in rounds in which arriving context belongs to $p \in \mathcal{P}$ before round t. It also keeps sample mean estimate of the rewards $\hat{\mu}_{p,a}(t)$

that are obtained from rounds in which arriving context belongs to $p \in \mathcal{P}$ and arm a was selected prior to the round t.

We denote the lower monotone neighborhood of context interval $p_j \in \mathcal{P}$ as $\mathcal{L}_j$, and it contains all the contexts intervals lower or equal to the context interval p as

$$\mathcal{L}_j = \begin{cases} \{p_1, \ldots, p_j\} & \text{if } j > 1 \\ \emptyset & \text{otherwise.} \end{cases}$$

We also denote the upper monotone neighborhood of context interval $p_j \in \mathcal{P}$ as $\mathcal{H}_j$, and it contains all the contexts intervals greater or equal to the context interval p as

$$\mathcal{H}_j = \begin{cases} \{p_j, \ldots, p_m\} & \text{if } j < m \\ \emptyset & \text{otherwise.} \end{cases}$$

We define the under-explored arms set for context interval $p \in \mathcal{P}$ as

$$\mathcal{M}_p^{ue}(t) = \Big\{ a' \in \mathcal{A}_p : N_{p,a'}(t) < D \Big\} \tag{5.2}$$

where D is chosen in a way that achieves a certain degree of certainty in the estimation of the rewards and the optimality of the best arm (e.g., Theorem 4.1 in [21]: For a given $x \in \mathcal{X}$ with probability $1 - \delta$, an arm $a'^*_x \in \mathcal{A}_p$ is ϵ-optimal i.e., $\mu(x, a'^*_x) \geq \max_{a' \in \mathcal{A}_p} \mu(x, a') - \epsilon$, when we set $D = (4/\epsilon^2) \log(2|\mathcal{A}_p|/\delta)$).

Unimodality of $\mu(x, a)$ in arms implies that when a_p^L, a_p^M, a_p^N, a_p^H are sufficiently sampled, one portion of the search interval can be eliminated as shown in [21, Fig. 3]. If the maximum of these four points is either a_p^N or a_p^H, then the interval $[a_p^L, a_p^M]$ is eliminated since optimal arm cannot lie in this interval with high probability. Similarly, if the maximum of these four points is either a_p^L or a_p^M, then the interval $[a_p^N, a_p^H]$ is eliminated since optimal arm cannot lie in this interval with high probability.

Using the above facts, S-CUB simply refines its lower point of the search interval as

$$a_p^L = \max_{p' \in \mathcal{L}_p} a_{p'}^L \tag{5.3}$$

Algorithm 4 S-CUB

Input: $\mathcal{X}, \mathcal{K}$
Initialize: Create partition set $\mathcal{P}$ for context space $\mathcal{X}$ into m equal intervals.
Initialize: $a_p^L = 0, a_p^H = 1, a_p^M = (\varphi \cdot a_p^L + a_p^H)/(1+\varphi), a_p^N = (a_p^L + \varphi \cdot a_p^H)/(1+\varphi), \forall p \in \mathcal{P}$.
Counters: $t = 1$, $t_p = 1, \forall p \in \mathcal{P}, N_{p,a} = 0, \hat{\mu}_{p,a} = 0, \forall p \in \mathcal{P}, \forall a \in \mathcal{A}_p$
while $t \geq 1$ **do**
    Observe context $x(t)$
    Find interval $p \in \mathcal{P}$ the arriving context belongs to
    Find the set of under-explored arm $\mathcal{M}_p^{ue}(t)$ by 5.2
    if $t_p = 1$
        $a_p^L = \max_{p' \in \mathcal{L}_j} a_{p'}^L$ //Refine interval
        $a_p^H = \min_{p' \in \mathcal{H}_j} a_{p'}^H$ //Refine interval
        $a_p^M = (\varphi \cdot a_p^L + a_p^H)/(1+\varphi)$
        $a_p^N = (a_p^L + \varphi \cdot a_p^H)/(1+\varphi)$
    end if
    if $|\mathcal{M}_p^{ue}(t)| = 0$
        $a_p^* = \text{argmax}_{a' \in \mathcal{A}_p} \hat{\mu}_{p,a'}$
        if $a_p^* = a_p^N$ or $a_p^* = a_p^H$
            $a_p^L = a_p^M$ //Update interval
            $a_p^L = \max_{p' \in \mathcal{L}_j} a_{p'}^L$ //Refine interval
            $a_p^H = \min_{p' \in \mathcal{H}_j} a_{p'}^H$ //Refine interval
            $a_p^M = (\varphi \cdot a_p^L + a_p^H)/(1+\varphi)$
            $a_p^N = (a_p^L + \varphi \cdot a_p^H)/(1+\varphi)$
        else
            $a_p^H = a_p^N$ //Update interval
            $a_p^L = \max_{p' \in \mathcal{L}_j} a_{p'}^L$ //Refine interval
            $a_p^H = \min_{p' \in \mathcal{H}_j} a_{p'}^H$ //Refine interval
            $a_p^M = (\varphi \cdot a_p^L + a_p^H)/(1+\varphi)$
            $a_p^N = (a_p^L + \varphi \cdot a_p^H)/(1+\varphi)$
        end if
        $a(t) = a_p^*$
    else
        Randomly select $a(t) \in \mathcal{M}_p^{ue}(t)$
    end if
    Observe reward $g_{x(t),a(t)}(t)$
    $N_{p,a(t)} = N_{p,a(t)} + 1$
    $\hat{\mu}_{p,a(t)} = \frac{\hat{\mu}_{p,a(t)}(N_{p,a(t)}-1) + g_{x(t),a(t)}(t)}{N_{p,a(t)}}$
    $t_p = t_p + 1$, $t = t + 1$
end while

and refines its higher point of the search interval as

$$a_p^H = \min_{p' \in \mathcal{H}_p} a_{p'}^H \ . \tag{5.4}$$

Finally, at the end of round t, arm $a(t)$ is played, a (random) reward $g_{x(t),a(t)}(t)$ with mean $\mu(x(t), a(t))$ is observed, and the sample mean estimates and counters that correspond to the pair $(p, a(t))$ are updated.

5.3 Illustrative Results

We numerically analyze the performance of S-CUB and compare the obtained performance with the earlier state-of-the-art in this section.

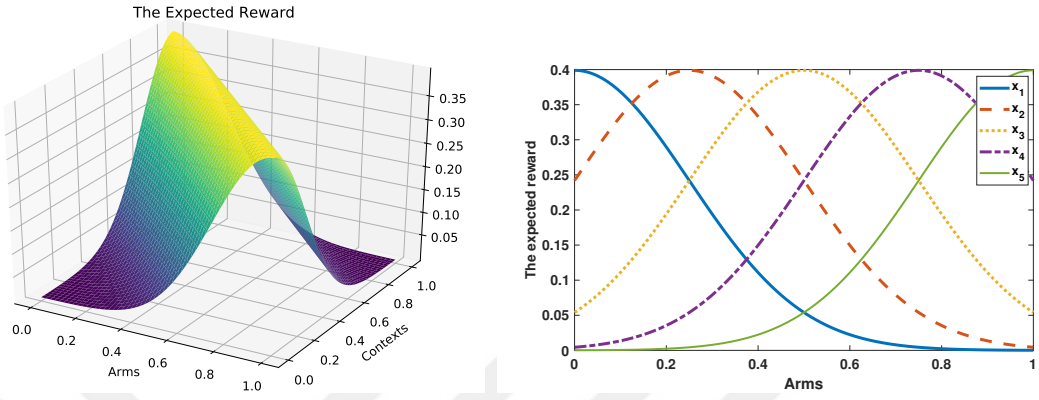
5.3.1 Competitor Learning Algorithms

We compare the performance of the S-CUB with the following algorithms:

1) *Line Search Algorithm* (LSE): This is the non-contextual search interval elimination algorithm proposed in [21]. It sequentially eliminates the portion of continuous arm set by sampling based elimination criterion.

2) *Contextual Upper Confidence Bound* (C-UCB): The context space is partitioned into $m = 25$ equal intervals, and then, for each interval we run an instance of UCB1 on a discretized set of arms (400 discretized arm sets).

3) *Strucutured Contextual Unimodal Bandits Variant* (S-CUB-V): This runs a separate instance of LSE for each context interval (Similar to C-UCB, the context space is partitioned into $m = 25$ equal intervals). It is a variant of S-CUB that exploits the contexts but ignores the strucuture over the contexts.



(a) (b)

Figure 5.2: (a) Expected reward as a function of arms and contexts. (b) Expected reward as a function of arms for 5 distant and ordered contexts fetched from (a).

5.3.2 Simulation Setup

The expected reward as a function of the contexts and the arms is given in Fig. 5.2(a). In addition, the expected reward as a function of the arms for a set of 5 distant and ordered contexts fetched from Fig. 5.2(a) is given in Fig. 5.2(b). It is clear that the expected reward is a unimodal function over arms for a given context, and the optimal arm increases monotonically with the context. We assume that the rewards are Bernoulli distributed, where the expected reward at a given context and arm represents the parameter of the Bernoulli distribution for that context-arm pair. Rewards are independent over arms and contexts.

We evaluate performance of the algorithms under three different types of context arrivals that are explained in the following subsections. Reported results are averaged over 50 runs to reduce the effect of randomness due to arm selections, context arrivals and reward generation.

5.3.3 Unimodal Context Arrivals

In this experiment, we allow contexts to arrive in a unimodal sequence of blocks, where each blocks contains $T/400$ samples, where the context space is partitioned

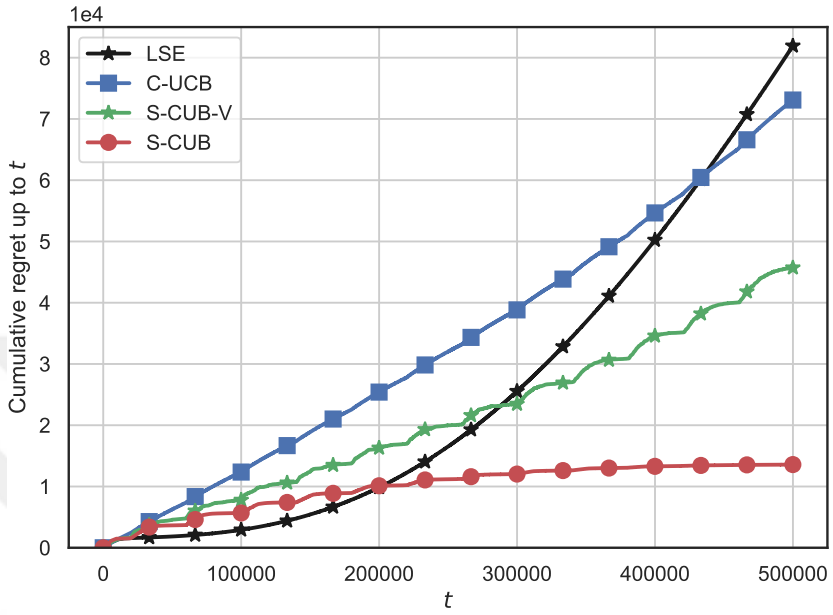


Figure 5.3: The expected regret comparison for unimodal context arrivals

into 400 equal context intervals. The starting block contains the contexts arriving uniformly at random from center context interval, while later arrivals are from left and right side of the centered context, alternatively, and this process is repeated till both sides of the arrivals reach the lowest index and highest index contexts interval (e.g., for a context space partitioned into 5 context intervals and $T = 500$, the arrivals pattern will be $\{X_3, X_4, X_2, X_5, X_1\}$, i.e., the first 100 arrivals are sampled uniformly at random from X_3, the second 100 arrivals are sampled uniformly at random from X_4, and so on). Therefore, the expected reward for a context has the learned neighboring context on at least one side. The monotonicity of optimal arm over context is utilized for this type of context arrivals due to the fact that neighboring contexts have reduced search intervals. Fig. 5.3 compares the performance of LSE, C-UCB, S-CUB-V, and S-CUB for unimodal context arrivals. LSE eliminates search intervals to find a global optimal arm for all the contexts, due to which it incurs the largest regret. Although C-UCB obtains the optimal arm for each context partitioned interval, it still incurs a large regret due to large number of discretized arms. S-CUB-V outperforms LSE and C-UCB, which demonstrate the benefit of interval elimination for each context. Finally, S-CUB achieves considerably smaller regret than S-CUB-V since it

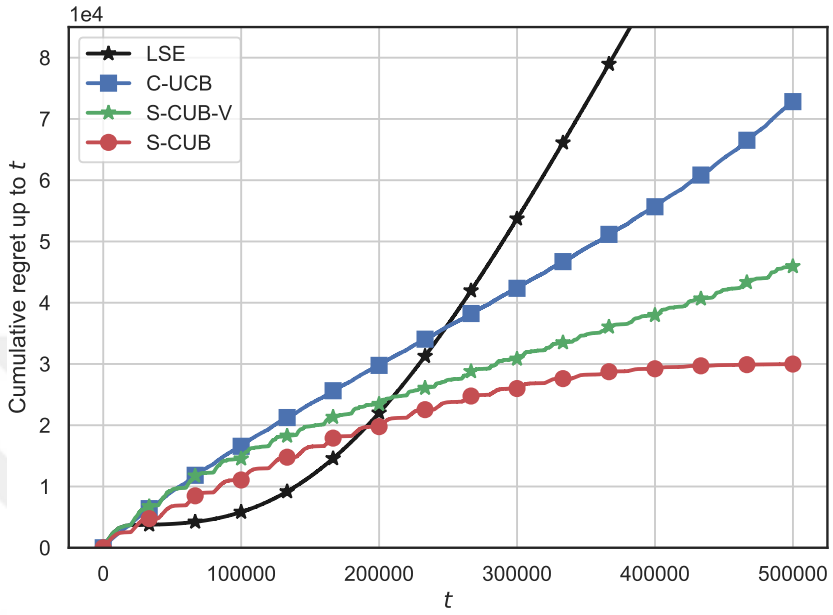


Figure 5.4: The expected regret comparison for monotone context arrivals

additionally utilizes the structure of optimal arms over the contexts.

5.3.4 Monotone Context Arrivals

In this experiment, we consider a sequence of ordered contexts, where each context interval lasts for a block of $T/400$ samples. In this scenario, the optimal arm $a_p^* \in \mathcal{A}$ increases monotonically i.e., $a^*(t) \geq a^*(t-1)$. The monotonicity of optimal arm over contexts is able to provide reduced intervals for incoming contexts. In Fig. 5.4 we show that S-CUB-V outperforms C-UCB and LSE. Furthermore, S-CUB achieves considerably smaller regret than S-CUB-V since it additionally utilizes the structure of optimal arms over the contexts.

5.3.5 Uniformly Random Context Arrivals

For this case, contexts arrive uniformly at random. The monotonicity of optimal arm over contexts may still able to provide reduced intervals for probabilistic

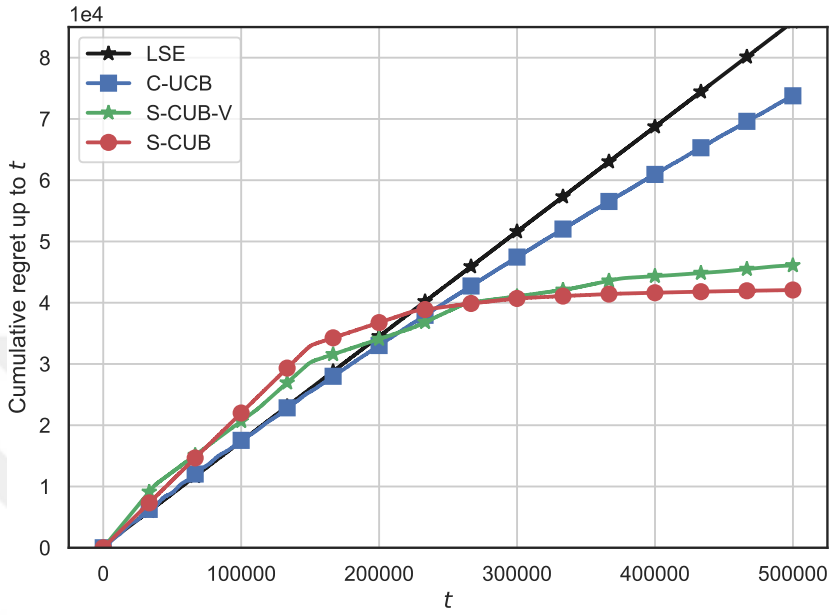


Figure 5.5: The expected regret comparison for uniformly random context arrivals

arrivals, but due to randomness in the nature the gain is not significant. Fig. 5.5 shows that S-CUB-V outperforms C-UCB and LSE. S-CUB slightly outperforms S-CUB-V by exploiting the monotonicity.

Chapter 6

Summary and Conclusions

We proposed new methods for dynamic resource allocation in rapidly-varying wireless channels that enables extremely fast learning by exploiting the structure among arms (transmission parameters) and contexts (side-information). In Chapter 2, we investigated the case when the expected rewards are jointly unimodal in arms and contexts, and proposed a new learning algorithm called CUL, which is able to achieve a regret that grows slowly with the number of arms and contexts and logarithmically in time. We tested the effectiveness of the proposed algorithm on three different resource allocation problems related to AI-enabled communication over rapidly-varying wireless channels and proved that our approach results in significant performance gains compared to other state-of-the-art MAB methods. We observed that CUL, which exploits the contextual information as well as the joint unimodality, significantly outperforms its competitors in all of the applications of AI-enabled radio networks studied in Section 2.1. In particular, we observed that in all of the applications considered in Section 2.1, the resources selected by CUL matched with that of the Oracle are at least 48.1% (between $48.1\% - 68.5\%$) more than that of KL-UCB-U which only exploits the unimodality in arms, at least 9.8% (between $9.8\% - 52.4\%$) more than that of SW-G-ORS and at least 7.2% (between $7.2\% - 26.7\%$) more than that of CUCB which exploits the contextual information but not the joint unimodality. Similarly, we also observed that CUL achieves an average throughput that is at

least 25.7% (between $25.7\% - 221.7\%$) more than that of KL-UCB-U, at least 5.8% (between $5.8\% - 28.8\%$) more than that of SW-G-ORS and at least 4.6% (between $4.6\% - 31.9\%$) more than that of CUCB. The takeaway massage from our study is that understanding and utilizing the structure of the wireless environment is essential in designing learning algorithms that learn fast and achieve high throughput under rapidly-varying conditions.

In Chapter 3, we considered the problem of rate selection under time-varying transmit power. We proposed a Bayesian learning algorithm, called DRS-TS, that exploits the structure of the throughput in rates as well as in transmit powers efficiently. We proved upper bounds on the regret of DRS-TS and presented experiments that compare the performance of DRS-TS with the state-of-the-art. Our results indicate that DRS-TS results in significant performance gains across a variety of context arrival patterns.

In Chapter 4, we proposed a Bayesian learning algorithm (V-CoTS) for dynamic rate and channel selection under unknown channel conditions and with time-varying PU user activity and application rate requirements. V-CoTS exploits the monotonicity of transmission success probability in the rates to sample from dynamically changing action sets, and thereby achieving significant gains in throughput compared to other methods.

In Chapter 5, we considered the contextual unimodal bandits where both the context and arms come from continuous intervals and the expected rewards exhibit a certain structure. We proposed a learning algorithm, called S-CUB, that exploits the structure of the expected reward in arms as well as the structure of optimal arms in contexts efficiently. Our results indicate that S-CUB results in significant performance gains across a variety of context arrival patterns.

Finally, we note that our algorithms and techniques can also be applied to other online learning problems that exhibit payoffs unimodal in the arms and the contexts. Our AI-enabled MAB-based approaches differ from the other AI-based techniques mentioned above in the following aspects: (i) They explicitly take into account the unimodal or monotone structure; (ii) Their optimality is

theoretically proven; (iii) They are completely online and do not require access to training data; (iv) They are efficient in terms of computation and memory.

Bibliography

[1] M. A. Qureshi and C. Tekin, "Online cross-layer learning in heterogeneous cognitive radio networks without CSI," in *Proc. 26th IEEE Signal Process. Commun. Applicat. Conf.*, pp. 1–4, 2018.

[2] M. A. Qureshi and C. Tekin, "Fast learning for dynamic resource allocation in AI-Enabled radio networks," *IEEE Tran. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 95–110, 2020.

[3] M. A. Qureshi and C. Tekin, "Online Bayesian learning for rate selection in millimeter wave cognitive radio networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2020.

[4] M. A. Qureshi and C. Tekin, "Rate and channel adaptation in cognitive radio networks under time-varying constraints," *IEEE Commun. Lett.*, 2020.

[5] R. Combes and A. Proutiere, "Dynamic rate and channel selection in cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 5, pp. 910–921, 2014.

[6] S. Park, H. Kim, and D. Hong, "Cognitive radio networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1386–1397, 2013.

[7] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang, *et al.*, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, 2011.

[8] X. Huang, T. Han, and N. Ansari, "On green-energy-powered cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 827–842, 2015.

[9] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Trans. Embedded Comput. Sys.*, vol. 6, no. 4, p. 32, 2007.

[10] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, 2005.

[11] R. Friedman, A. Kogan, and Y. Krivolapov, "On power and throughput tradeoffs of WiFi and bluetooth in smartphones," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1363–1376, 2012.

[12] X. Ruan and H. Chen, "Performance-to-power ratio aware virtual machine (VM) allocation in energy-efficient clouds," in *Proc. IEEE Int. Conf. Cluster Comput.*, pp. 264–273, 2015.

[13] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits–a musical chairs approach," in *Proc. 33rd Int. Conf. Mach. Learn.*, pp. 155–163, 2016.

[14] M. K. Hanawal and S. J. Darak, "Multi-player bandits: A trekking approach." arXiv:1809.06040. [Online]. Available: `https://arxiv.org/abs/1809.06040`, Sept. 2018.

[15] P. Alatur, K. Y. Levy, and A. Krause, "Multi-player bandits: The adversarial case." arXiv:1902.08036. [Online]. Available: `https://arxiv.org/abs/1902.08036`, Feb. 2019.

[16] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[17] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[18] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. 20th Int. Conf. Neural Inform. Process. Syst.*, pp. 817–824, 2007.

[19] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, pp. 485–492, 2010.

[20] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2533–2568, 2014.

[21] J. Y. Yu and S. Mannor, "Unimodal bandits," in *Proc. 28th Int. Conf. Mach. Learn.*, pp. 41–48, 2011.

[22] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *Proc. Int. Conf. Mach. Learn.*, pp. 521–529, 2014.

[23] S. Paladino, F. Trovò, M. Restelli, and N. Gatti, "Unimodal thompson sampling for graph-structured arms," in *Proc. Conf. on Artif. Intell.*, pp. 2457–2463, 2017.

[24] Z. Bnaya, R. Puzis, R. Stern, and A. Felner, "Volatile multi-armed bandits for guaranteed targeted social crawling," in *Proc. 27th AAAI Conf. Artif. Intell. Workshops*, 2013.

[25] A. Chatterjee, G. Ghalme, S. Jain, R. Vaish, and Y. Narahari, "Analysis of Thompson sampling for stochastic sleeping bandits.," in *Proc. Conf. Uncertainty in Artif. Intell.*, 2017.

[26] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, 2018.

[27] D. Tse and P. Viswanath, "Fundamentals of wireless communication," *Cambridge Univ. Press*, 2005.

[28] J. Li, X. Wu, and R. Laroia, "OFDMA mobile broadband communications: A systems approach," *Cambridge Univ. Press*, 2013.

[29] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi, "Optimal rate sampling in 802.11 systems," in *Proc. IEEE Conf. Comput. Commun.*, pp. 2760–2767, 2014.

[30] R. Combes, J. Ok, A. Proutiere, D. Yun, and Y. Yi, "Optimal rate sampling in 802.11 systems: Theory, design, and implementation," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1145–1158, 2018.

[31] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *Proc. IEEE Conf. Comput. Commun.*, pp. 540–548, 2018.

[32] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained Thompson sampling," in *Proc. IEEE Conf. Comput. Commun.*, pp. 739–747, 2019.

[33] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE Conf. Comput. Commun.*, pp. 2393–2401, 2018.

[34] D. E. Papanikolaou, N. E. Papanikolaou, G. T. Pitsiladis, A. D. Panagopoulos, and Ph. Constantinou, "Spectrum sensing in mm-Wave cognitive radio networks under rain fading," in *Proc. 5th IEEE European Conf. Antennas Propag.*, pp. 1684–1687, 2011.

[35] H. Zhao, J. Zhang, L. Yang, G. Pan, and M.-S. Alouini, "Secure mmWave communications in cognitive radio networks," *IEEE Wireless Commun. Lett.*, 2019.

[36] H. Hosseini, A. Anpalagan, K. Raahemifar, S. Erkucuk, and S. Habib, "Joint wavelet-based spectrum sensing and FBMC modulation for cognitive mmWave small cell networks," *IET Communications*, vol. 10, no. 14, pp. 1803–1809, 2016.

[37] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, 2015.

[38] A. G. Marques, L. M. Lopez-Ramos, G. B. Giannakis, and J. Ramos, "Resource allocation for interweave and underlay CRs under probability-of-interference constraints," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1922–1933, 2012.

[39] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[40] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, 2002.

[41] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, 2012.

[42] N. Baldo and M. Zorzi, "Learning and adaptation in cognitive radios using neural networks," in *Proc. 5th IEEE Consumer Commun. Netw. Conf.*, pp. 998–1003, 2008.

[43] L. Zhang, J. Tan, Y.-C. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, 2019.

[44] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[45] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. JMLR Workshop Conf.*, pp. 359–376, 2011.

[46] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proc. Int. Conf. Mach. Learn.*, pp. 757–765, 2014.

[47] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proc. 39th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, pp. 539–548, 2016.

[48] C. Trinh, E. Kaufmann, C. Vernade, and R. Combes, "Solving Bernoulli rank-one bandits with unimodal Thompson sampling," *arXiv preprint arXiv:1912.03074*, 2019.

[49] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, 2015.

[50] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-end simulation of 5G mmWave networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2237–2263, 2018.

[51] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, 2015.

[52] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, 2017.

[53] K. Haneda, L. Tian, H. Asplund, J. Li, Y. Wang, D. Steer, C. Li, T. Balercia, S. Lee, Y. Kim, *et al.*, "Indoor 5G 3GPP-like channel models for office and shopping mall environments," in *Proc. IEEE Int. Conf. Commun. Workshops*, pp. 694–699, 2016.

[54] K. Haneda, J. Zhang, L. Tan, G. Liu, Y. Zheng, H. Asplund, J. Li, Y. Wang, D. Steer, C. Li, *et al.*, "5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments," in *Proc. 83rd IEEE Veh. Technol. Conf.*, pp. 1–7, 2016.

[55] S. H. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in *Proc. ACM MobiCom*, pp. 146–157, 2006.

[56] J. C. Bicket, *Bit-rate Selection in Wireless Networks*. PhD thesis, MIT, Cambridge, MA, 2005.

[57] O. Ozel and S. Ulukus, "AWGN channel under time-varying amplitude constraints with causal information at the transmitter," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput.*, pp. 373–377, 2011.

[58] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. IEEE Conf. Comput. Commun.*, pp. 1961–1969, 2018.

[59] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, 2018.

[60] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, 2016.

[61] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25463–25473, 2018.

[62] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, 2018.

[63] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. Int. Conf. Algorithmic Learning Theory*, pp. 174–188, Springer, 2011.

[64] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, pp. 199–207, 2014.

[65] C. Wu and D. P. Bertsekas, "Distributed power control algorithms for wireless networks," *IEEE Trans. Veh. Technol.*, vol. 50, no. 2, pp. 504–514, 2001.

[66] R. Xie, F. R. Yu, and H. Ji, "Dynamic resource allocation for heterogeneous services in cognitive radio networks with imperfect channel sensing," *IEEE Trans. Veh. Technol.*, vol. 61, no. 2, pp. 770–780, 2011.

[67] K. D. Huang, K. R. Duffy, and D. Malone, "H-RCA: 802.11 collision-aware rate control," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1021–1034, 2013.

[68] R. Kumar, S. J. Darak, A. Yadav, A. K. Sharma, and R. K. Tripathi, "Channel selection for secondary users in decentralized network of unknown size," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2186–2189, 2017.

[69] X. Shen, C. Bo, J. Zhang, S. Tang, X. Mao, and G. Dai, "EFCon: Energy flow control for sustainable wireless sensor networks," *Ad Hoc Netw.*, vol. 11, no. 4, pp. 1421–1431, 2013.

[70] S. Li, Z. Shao, and J. Huang, "ARM: Anonymous rating mechanism for discrete power control," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 326–340, 2016.

[71] N. Devroye, M. Vu, and V. Tarokh, "Cognitive radio networks," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 12–23, 2008.

[72] K. T. Kim and S. K. Oh, "Cognitive ad-hoc networks under a cellular network with an interference temperature limit," in *Proc. 10th IEEE Int. Conf. Adv. Commun. Technol.*, vol. 2, pp. 879–882, 2008.

[73] I. Pardina Garcia, "Interference management in cognitive radio systems," Master's thesis, Universitat Politècnica de Catalunya, 2011.

[74] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Proc. Int. Conf. Algorithmic Learning Theory*, pp. 199–213, Springer, 2012.

[75] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *Proc. Int. Conf. Artif. Intell. Statist.*, pp. 592–600, 2012.

[76] J. Komiyama, J. Honda, and H. Nakagawa, "Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *Proc. Int. Conf. Mach. Learn.*, pp. 1152–1161, 2015.

[77] H. Qi, Z. Hu, X. Wen, and Z. Lu, "Rate adaptation with Thompson sampling in 802.11 ac WLAN," *IEEE Commun. Lett.*, vol. 23, no. 10, pp. 1888–1892, 2019.

[78] L. Lu, X. Zhang, R. Funada, C. S. Sum, and H. Harada, "Selection of modulation and coding schemes of single carrier PHY for 802.11 ad multi-gigabit mmWave WLAN systems," in *Proc. IEEE Symp. Comput. Commun.*, pp. 348–352, 2011.

[79] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[80] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. Conf. Learn. Theory*, pp. 39.1–39.26, 2012.