KTH Electrical Engineering

Minimizing Regret in Combinatorial Bandits and Reinforcement Learning

MOHAMMAD SADEGH TALEBI MAZRAEH SHAHI

Doctoral Thesis
Stockholm, Sweden 2017

KTH Royal Institute of Technology
School of Electrical Engineering
Department of Automatic Control
SE-100 44 Stockholm
SWEDEN

TRITA-EE 2017:168
ISSN 1653-5146
ISBN 978-91-7729-618-8

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av doktorsexamen i elektro och systemteknik tisdagen den 19 december 2017 klockan 10.00 i Q2, Kungliga Tekniska högskolan, Osquldasväg 10, Stockholm, Sweden.

Tryck: Universitetsservice US AB

Abstract

This thesis investigates sequential decision making tasks that fall in the framework of reinforcement learning (RL). These tasks involve a decision maker repeatedly interacting with an environment modeled by an unknown finite Markov decision process (MDP), who wishes to maximize a notion of reward accumulated during her experience. Her performance can be measured through the notion of regret, which compares her accumulated expected reward against that achieved by an oracle algorithm always following an optimal behavior. In order to maximize her accumulated reward, or equivalently to minimize the regret, she needs to face a trade-off between exploration and exploitation.

The first part of this thesis investigates combinatorial multi-armed bandit (MAB) problems, which are RL problems whose state-space is a singleton. It also addresses some applications that can be cast as combinatorial MAB problems. The number of arms in such problems generically grows exponentially with the number of basic actions, but the rewards of various arms are correlated. Hence, the challenge in such problems is to exploit the underlying combinatorial structure. For these problems, we derive asymptotic (i.e., when the time horizon grows large) lower bounds on the regret of any admissible algorithm and investigate how these bounds scale with the dimension of the underlying combinatorial structure. We then propose several algorithms and provide finite-time analyses of their regret. The proposed algorithms efficiently exploit the structure of the problem, provide better performance guarantees than existing algorithms, and significantly outperform these algorithms in practice.

The second part of the thesis concerns RL in an unknown and discrete MDP under the average-reward criterion. We develop some variations of the transportation lemma that could serve as novel tools for the regret analysis of RL algorithms. Revisiting existing regret lower bounds allows us to derive alternative bounds, which motivate that the local variance of the bias function of the MDP, i.e., the variance with respect to next-state transition laws, could serve as a notion of problem complexity for regret minimization in RL. Leveraging these tools also allows us to report a novel regret analysis of the KL-Ucrl algorithm for ergodic MDPs. The leading term in our regret bound depends on the local variance of the bias function, thus coinciding with observations obtained from our presented lower bounds. Numerical evaluations in some benchmark MDPs indicate that the leading term of the derived bound can provide an order of magnitude improvement over previously known results for this algorithm.

Sammanfattning

I denna avhandling behandlas sekventiella beslutsproblem som faller inom ramverket för förstärkande inlärning (eng: reinforcement learning, RL). Dessa problem involverar en beslutstagare som, upprepade gånger, interagerar med en miljö som kan modelleras enligt en Markoviansk beslutsprocess (eng: Markov decision process, MDP). Beslutstagaren försöker maximera ett mått på den förväntade belöningen som kan ackumuleras vid dessa interaktioner. Hur bra beslutstagaren gör ifrån sig kan kvantifieras genom dess ånger (eng: regret), som jämför beslutstagarens förväntade ackumulerade belöning gentemot den belöning som en orakelalgoritm (som alltid beter sig optimalt) kan åstadkomma. För att beslutstagaren ska maximera sin ackumulerade belöning, eller ekvivalent, minimera sin ånger, så måste vid varje interaktion för- och nackdelarna med att utforska miljön gentemot att exploatera den vägas mot varandra.

Den första delen av denna avhandling behandlar problem modellerade som kombinatoriska flerarmade banditer (eng: multi-armed bandit, MAB). Dessa är RL-problem vars tillståndsrum består av endast ett element. Även tillämpningar som kan modelleras med hjälp av MAB:er behandlas. I allmänhet växer antalet armar exponentiellt med antalet tillgängliga handlingar, men belöningarna för de olika armarna är korrelerade. Utmaningen i dessa problem ligger i att utnyttja den underliggande kombinatoriska strukturen. För dessa problem härleder vi asymptotiska (d.v.s. när tiden låts gå mot oändligheten) undre gränser på ångern för godtyckliga algoritmer, samt studerar hur dessa undre gränser skalas med dimensionen hos den underliggande kombinatoriska strukturen. Vi föreslår sedan flera algoritmer, samt härleder gränser för deras ånger som är giltiga även icke-asymptotiskt. De föreslagna algoritmerna utnyttjar effektivt strukturen hos problemen, har bättre teoretiska garantier än redan existerande algoritmer, samt överträffar prestandamässigt dessa i praktiken.

Den andra delen av avhandlingen behandlar RL i en okänd och diskret MDP under ett medelbelöningskriterium. Vi utvecklar nya verktyg som kan användas för att utföra ångeranalys för RL-algoritmer. Mer specifikt härleder vi variationer på transportlemmat och kombinerar dessa med Kullbeck-Leibler koncentrationsolikheter. Med dessa nya verktyg kan vi härleda alternativa, nya, undre gränser för ångern som påvisar att den lokala variansen hos biasfunktionen av MDP:n (d.v.s. variansen med avseende på dess övergångsfunktion) kan användas som ett mått på problemets komplexitet för ångeranalys i RL. Med hjälp av dessa verktyg utför vi en ny ångeranalys för KL-Ucrl-algoritmen för ergodiska MDP:er. Den ledande termen i analysen beror på den lokala variansen hos biasfunktionen, vilket rättfärdigar varianstermen i vår härleda undre gräns. Numeriska simuleringar för några standardtestfall indikerar att den ledande termen i den härledda undre gränsen kan vara upp till en storleksordning bättre än tidigare kända gränser för denna algoritm.

To my parents, Batool and Esmaeel.

To my wife, Zahra.

Acknowledgement

First and foremost, I would like to express my gratitude to my advisor, Prof. Alexandre Proutiere, for his excellent guidance and encouraging me throughout the journey of my PhD. Without his continuous support and inspiration, I would not have come this far. I wish to thank him for teaching me to think clearly and independently, and to present ideas precisely. I am also very grateful to my co-advisor, Prof. Mikael Johansson, for all exciting discussions and invaluable advices during these years.

I am deeply grateful to Prof. Richard Combes, whose mentoring and collaboration proved invaluable during my research. I have the good fortune to be able to work with him and I owe him a lot. I would also like to greatly thank Prof. Odalric-Ambrym Maillard at INRIA Lille – Nord Europe, France, for giving me the opportunity to visit SequeL team at INRIA and for his kind hospitality during my stay at Lille. We had a really fruitful collaboration and I enjoyed a lot working with him. I also acknowledge Ericsson Research Foundation for financial support of my visit to INRIA Lille – Nord Europe during summer 2017.

My special thanks go to the present and past colleagues in the Department of Automatic Control. To my officemates Pedro, Olle, Riccardo, Demia, Valerio, Mattias, Antonio, and Patricio who contributed to make room C:628 a wonderful office, and to my colleagues Euhanna, Afrooz, Hossein, Matin, Hamid, Ehsan, Mohammad Reza, Burak, Themis, Gabor, Arda, Zhenhua, Mohamed, Rui, Stefan, Vahan, Othmane, Gerd, Niclas, Jaron, Jungseul, and Emma for all nice conversations on various topics. My deep thanks go to Afrooz, Hossein, Matin, and Ehsan for all fun moments during lunch and fika. I also want to thank the administrators Karin, Anneli, Hanna, Silvia, and Felicia for all the support to make the department run smoothly. I also would like to acknowledge Robert and Olle for their help in writing the Swedish abstract.

In the past few years, I enjoyed the friendship of several great people in Stockholm: Euhanna and Elaheh, Hossein and Forough, Afrooz and Iman, Behdad, Sajed and Zahra, Nasser and Ghazaleh, Alireza (Ahmadi), Kaveh, Hamed and Maryam, Zahra (Besharat), Mansoureh, Farhad, Mohammad Reza, and Mohammad (Khodaei). I am glad that I have met you and spent time with you.

Lastly, I am indebted to my family; to my parents, Esmaeel and Batool, for their unconditional love, encouragement, and endless support, to my wonderful wife, Zahra, for all the happiness and joy she brought to my life, to my sisters, Saeedeh, Zohreh, Azadeh, and Fatemeh and to my family-in-law, for their support and encouragements.

Sadegh Talebi
Stockholm, November 2017.

Abbreviations

APF	Approximate PF
iff	if and only if
i.i.d.	independent and identically distributed
KKT	Karush-Kuhn-Tucker
KL	Kullback-Leibler
LP	Linear Program
MAB	Multi-armed Bandit
MDP	Markov Decision Process
PF	Proportionally Fair
RL	Reinforcement Learning
UCB	Upper Confidence Bound

Notations

$\ll$	Absolute continuity relation
$\sim$	Distributed according to
$\mathbb{I}A$	Indicator function of event A: $\mathbb{I}A = 1$ if A is true, and $\mathbb{I}A = 0$ otherwise.
I	Identity matrix
$\mathbf{1}$	The vector of all ones
$A^\top$	The transpose of matrix A
$[n]$	The set $\{1, \dots, n\}$ for $n \in \mathbb{N}$
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of non-negative real numbers
$\mathbb{R}_{++}$	The set of positive real numbers
$\mathbb{E}[\cdot]$	Expectation
$\mathbb{P}(\cdot)$	Probability
$\mathbb{V}_p(\cdot)$	Variance under distribution p
$\mathcal{P}(\mathcal{X})$	The set of probability distributions over alphabet $\mathcal{X}$
$\mathbb{S}(f)$	Span of function f
$\exp$	Exponential function
$\log$	The logarithm to base e
$\mathrm{llnp}(n)$	The function $\log(\log(\max(n, e)))$
$\mathtt{KL}(P, Q)$	The Kullback-Leibler divergence between discrete distributions P and Q
$\mathtt{kl}(u, v)$	The Kullback-Leibler divergence between Bernoulli distributions with parameters u and v

Contents

Chapter 1

Introduction

This thesis investigates sequential decision making tasks that fall in the framework of reinforcement learning (RL). These tasks arise in a broad range of real-world applications including game playing [1, 2], sequential clinical trials [3, 4], communication systems [5, 6], economics [7, 8], recommendation systems [9, 10], robotics [11], power systems [12], to name a few. A sequential decision making task involves a decision maker repeatedly interacting with an environment with unknown dynamics. The decision maker wishes to maximize a notion of reward accumulated during her experience. To this end, she has to learn the optimal behavior, or a near-optimal one, as quickly as possible. At each time (or round), the decision maker selects an action. As a result, she receives an instantaneous reward and the state of the system (or environment) evolves. The decision maker does not know the system dynamics, i.e., the way the state of the system evolves, and the reward functions. In order to maximize her cumulative reward, she has to learn reward functions and system dynamics using her observed rewards and state transitions. She therefore needs to face a trade-off between *exploration* and *exploitation*: On the one hand, she must explore different actions in various states to maintain a precise enough model of the system, and on the other hand, her chosen action in a given state should be consistent with her past experience to maximize the reward.

The performance of the decision maker (or equivalently that of the learning algorithm she uses) can be measured through several metrics. The most important metrics considered in the literature include *(i) asymptotic convergence to optimality, (ii) convergence rate to (near-)optimality, and (iii) regret*; see, e.g., [13]. Metric (i) is concerned with the eventual learning of the optimal behavior, whereas (ii) takes into account the speed at which such a behavior is learnt. It turns out that a more practical notion is captured by convergence rate towards finding a near-optimal behavior. The notion of regret is usually defined as the difference between the accumulated reward of the algorithm when compared to an oracle algorithm always performing the optimal behavior. One of the weaknesses associated to performance metrics (i) and (ii) is that they do not incorporate the potentially large penalties that the decision maker has undergone to learn a (near-)optimal behavior. Therefore, among the aforementioned metrics, regret is considered as the most appealing since it penalizes the performance of the algorithm by taking into account the mistakes during the experience.

RL can also be viewed through the lens of adaptive optimal control, as argued in, e.g., [14, 15], as it amounts to the problem of controlling an unknown dynamical system in order to acquire the maximal cumulative reward. Broadly speaking, there are two classes of control problems: Tracking problems, where the goal is to follow a reference trajectory, and optimal control problems whose objective is to maximize a notion of reward, which itself is a function of the trajectory followed by the system. RL problems fall in the second class and have been studied in the control community under various names including 'adaptive control', 'optimal control', 'robust control of unknown Markov decision processes (MDPs)', etc. From a theoretical perspective, a big class of RL algorithms rely on the tools developed for direct and indirect adaptive optimal control methods.

In this thesis we investigate two classes of RL problems: the first one is combinatorial MAB problem in the stochastic setting, whereas the second one is that of RL in an MDP under average-reward criterion, which we will refer to as *RL in undiscounted MDPs*, or for short, *undiscounted RL*.

1.1 Stochastic Combinatorial MAB Problems

The first part of this thesis investigates online combinatorial optimization problems in the stochastic setting and under bandit feedback. These problems, which are also referred to as *stochastic combinatorial bandits*[1] in the literature, are RL problems where there is no notion of state (or equivalently, the state-space is a singleton), and where the set of actions (or arms, according to MAB terminology) is endowed with a *given* combinatorial structure. The MAB framework was introduced in the seminal paper by Robbins [16] in 1952 to study the sequential design of experiments[2]. Lai and Robbins [18] studied the classical MAB problem and derived a lower bound on the regret of any admissible policy, asymptotically scaling as $\Omega(\log(T))$ after T rounds. They also constructed policies for certain reward distributions and showed that they asymptotically achieve the aforementioned lower bound, that is their regret upper bounds asymptotically grow at most as $\mathcal{O}(\log(T))$, where the constant in $\mathcal{O}(.)$ is the same as that in the lower bound.

The considered setup may be concretely described as follows: Let E be a ground set with cardinality d. The set of arms $\mathcal{A}$ is an arbitrary subset of $\{0,1\}^d$, such that each $a \in \mathcal{A}$ is a subset of at most m *basic actions* taken from E. Arm a is identified with a binary column vector $(a_1, \dots, a_d)^\top$. In each round $n \geq 1$, a decision maker selects an arm $a \in \mathcal{A}$ and receives a reward $X^a(n) := a^\top X(n) = \sum_{i \in E} a_i X_i(n)$. The reward vector $X(n) \in \mathbb{R}_+^d$ is unknown, and for all $i \in E$, $(X_i(n))_{n \geq 1}$ is drawn i.i.d. from an unknown distribution. The reward sequences may be arbitrarily

[1] In this thesis, we will use the terms 'online combinatorial optimization', 'combinatorial MAB', and 'combinatorial bandit', interchangeably.

[2] The first algorithm for MAB problems, however is due to Thompson [17], which dates back to 1933.

correlated across basic actions. After selecting an arm a in round n, the decision maker receives some feedback. We are interested in two types of feedback:

(i) *Semi-bandit feedback* [3] under which after round n, for all $i \in E$, the component $X_i(n)$ of the reward vector is revealed if and only if $a_i = 1$.

(ii) *Bandit feedback* under which only the reward $a^\top X(n)$ is revealed.

Based on the feedback received up to round $n-1$, the decision maker selects an arm for the next round n. Her goal is to identify a policy, amongst all feasible policies, maximizing the cumulative expected reward over T rounds. To this aim, she is required to balance *exploitation* and *exploration*: Arms with higher observed rewards should be selected often whilst all arms should be explored to learn their average rewards. Equivalently, the decision maker aims at designing a policy that minimizes the regret, where the regret of policy π is defined as the gap between the expected reward achieved by π and that achieved by an oracle algorithm always selecting the optimal arm:

$$\mathfrak{R}_{\pi,T} = \max_{a\in\mathcal{A}} \mathbb{E}[\sum_{n=1}^{T} X^a(n)] - \mathbb{E}[\sum_{n=1}^{T} X^{a^\pi(n)}(n)]\,, \tag{1.1}$$

where $a^\pi(n)$ denotes the arm chosen by policy π at time n. The expectation is here taken with respect to the randomness in the rewards and the possible randomization in the policy. The notion of regret quantifies the performance loss due to the need for learning the average rewards of the various arms.

1.1.1 Objectives

In combinatorial MAB problems, one could apply classical sequential arm selection policies, developed in, e.g., [20, 21], as if arms would yield independent rewards. Such policies would have a regret asymptotically scaling as $|\mathcal{A}|\log(T)$. However, since the number of arms $|\mathcal{A}|$ could grow exponentially with d, treating arms as independent would lead to a prohibitive regret. In contrast to classical MAB studied by Lai and Robbins [18], where the random rewards from various arms are *independent*, in combinatorial MAB problems the rewards of the various arms are inherently correlated since arms may share the basic actions. It may then be crucial to exploit these correlations, i.e., the structure of the problem to speed up the exploration of sub-optimal arms. This in turn results in the design of efficient arm selection policies that have a regret scaling as $C\log(T)$, where C is much smaller than $|\mathcal{A}|$.

The objectives in this thesis for combinatorial MABs may be formalized as follows: Firstly, we would like to study the asymptotic (namely when T grows large) regret lower bounds for policies with bandit and semi-bandit feedback. Such

[3] The term ‘semi-bandit feedback’ was introduced by Audibert et al. [19]. Note that this type of feedback is relevant for combinatorial problems only.

lower bounds provide fundamental performance limits that no policy can beat. Correlations significantly complicate the derivation and the expression of the regret lower bounds. However, study of such lower bounds is of great importance as they provide insights into the design of arm selection policies being capable of exploiting the combinatorial structure of the problem. Secondly, we would like to propose arm selection policies whose performance approaches the proposed lower bounds.

1.2 Reinforcement Learning in Undiscounted Markov Decision Processes

The second part of this thesis concerns RL in Markov Decision Processes (MDPs) under the average-reward criterion, when the decision maker interacts with the system in a single stream of observations, starting from an initial state without any reset.

The setup we consider can be formally described as follows. Let $M = (\mathcal{S}, \mathcal{A}, \nu, p)$ denote an MDP, where $\mathcal{S}$ and $\mathcal{A}$ respectively denote the finite set of states and set of actions available at any state. Let S and A denote the respective cardinalities of $\mathcal{S}$ and $\mathcal{A}$. The functions ν and p respectively denote the reward function and transition kernel. For any (s, a), $\nu(s, a)$ has support $[0, 1]$ and mean $\mu(s, a)$. The decision maker does not know ν and p. At each time step $t \in \mathbb{N}$, the decision maker chooses one action $a_t \in \mathcal{A}$ in her current state $s_t \in \mathcal{S}$ based on her past decisions and observations. When executing action a_t in state s_t, the decision maker receives a random reward r_t drawn independently from distribution $\nu(s_t, a_t)$. The state then transits to a next state $s' \in \mathcal{S}$ sampled with probability $p(s'|s_t, a_t)$, and a new decision step begins. As the transition probabilities and reward functions are unknown, the decision maker has to learn them by trying different actions and recording the realized rewards and state transitions.

As already stated, the performance of the decision maker is assessed through the notion of regret, which compares the reward collected by the algorithm to that obtained by an oracle always following an optimal policy, where a policy is a mapping from $\mathcal{S}$ to $\mathcal{A}$. Letting $g^\star$ denote the maximal achievable long-term reward[4], we define the regret of a learning algorithm $\mathbb{A}$ after T steps as

$$\mathrm{Regret}_{\mathbb{A},T} := Tg^\star - \sum_{t=1}^{T} r(s_t, a_t)\,, \tag{1.2}$$

where $a_t = \mathbb{A}(s_t, (\{s_{t'}, a_{t'}, r_{t'}\})_{t'<t})$, and $s_{t+1} \sim p(\cdot|s_t, a_t)$ is a sequence of states generated by $\mathbb{A}$, and $r(s_t, a_t) \sim \nu(s_t, a_t)$.

[4] For a precise definition of $g^\star$, we refer to Chapter 2. Note further that we consider communicating MDPs, for which the optimal gain does not depend on the initial state.

1.2.1 Objective

To date, several algorithms with finite-time regret guarantees have been proposed for the aforementioned RL setup. Although these algorithms enjoy rate-optimal[5] regret bounds that hold uniformly over time (scaling as either $\mathcal{O}(\log(T))$ or $\widetilde{\mathcal{O}}(\sqrt{T})$ depending on the context of analysis), none of them, to the best of our knowledge, is shown to be order-optimal. Hence, finding an algorithm with an order-optimal and finite-time regret guarantee for RL under average-reward criterion is still an open problem.

A big class of these algorithms implement the paradigm of *optimism in the face of uncertainty* mainly through maintaining confidence bounds on the transition kernel and reward function. While the only known, up to our knowledge, tight and problem-dependent lower bound (cf. [22]) suggests that such confidence bounds should be defined using the Kullback-Leibler (KL) divergence of involved distributions, most proposed algorithms rely on confidence bounds defined by the L_1 or total variation norm. The use of L_1 norm, instead of the KL-divergence, allows one to describe the uncertainty of the transition kernel by a *polytope*, which in turn brings computational advantages and ease in the regret analysis. On the other hand, such polytopic models are typically known to provide poor representations of underlying uncertainties [23].

The shortcomings due to use of such polytopic models are avoided by resorting to KL-divergence to define the confidence bounds, as already incorporated into the design of the KL-Ucrl algorithm [24]. Despite such potential benefits as well as superior numerical performance of KL-Ucrl over existing algorithms, the best known regret bound for KL-Ucrl matches that of Ucrl2 [25]. Hence, from a theoretical perspective, the potential gain of use of KL-divergence to define confidence bounds for transition probabilities has remained largely unexplored. Our goal is to investigate these benefits. Our approach towards this end relies on the development of concentration inequalities that enable to decouple the concentration properties of the transition kernel from the specific structure of the involved value functions. As we shall see later, these results allow us to derive a refined regret bound for KL-Ucrl for the class of ergodic MDPs.

1.3 Motivating Examples

Combinatorial MAB and undiscounted RL can be used to model a variety of applications. Here, we provide two examples to motivate the proposed algorithms and their analyses provided in subsequent chapters. The first example considers dynamic spectrum access in wireless systems whereas the second one concerns shortest-path routing in multihop networks.

[5] We a policy is said to be rate-optimal if its regret asymptotically grows at the same rate of the optimal algorithm.

1.3.1 Dynamic Spectrum Access

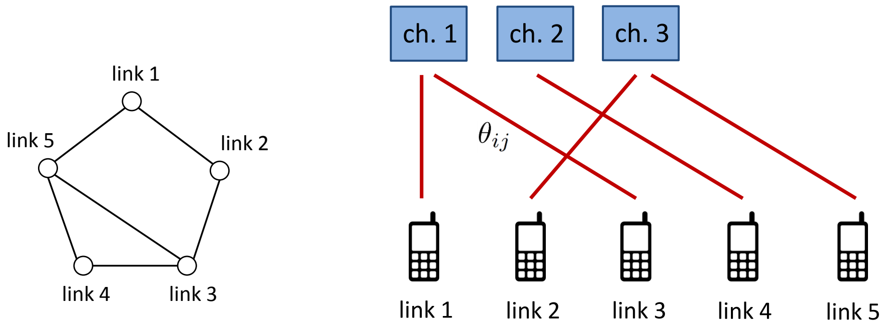
As the first motivating example, we consider a dynamic spectrum access scenario as studied in [26]. Spectrum allocation has attracted considerable attention recently, mainly due to the increasing popularity of cognitive radio systems. In such systems, transmitters have to explore spectrum to find frequency bands free from primary users. The fundamental objective here is to devise an allocation that maximizes the network-wide throughput. In such networks, transmitters should be able to select a channel that (i) is not selected by neighbouring transmitters to avoid interference, and (ii) offers good radio conditions.

Consider a network consisting of L users or links indexed by $i \in [L] = \{1, \dots, L\}$. Each link can use one of the K available radio channels indexed by $j \in [K]$. Interference is represented as an interference graph $G = (V, E)$ [6] where vertices are links and edges indicate interference among links. More precisely, we have $(i, i') \in E$ if links i and i' interfere, i.e., these links cannot be simultaneously active. A spectrum allocation is represented as an allocation matrix $a \in \{0, 1\}^{L \times K}$, where $a_{ij} = 1$ if and only if transmitter of user i uses channel j. Allocation a is feasible if (i) for all i, the corresponding transmitter uses at most one channel, i.e., $\sum_{j \in [K]} a_{ij} \le 1$, and (ii) two interfering links cannot be active on the same channel, i.e., for all $i, i' \in [L]$, $(i, i') \in E$ implies for all $j \in [K]$, $a_{ij} a_{i'j} = 0$. [7] Let $\mathcal{A}$ be the set of all feasible allocation matrices. In the following we denote by $\mathcal{F} = \{\mathcal{F}_\ell, \ell \in [f]\}$ the set of maximal cliques of the interference graph G. We also introduce $F_{\ell i} \in \{0, 1\}$ such that $F_{\ell i} = 1$ if and only if link i belongs to the maximal clique $\mathcal{F}_\ell$. Hence, for any clique ℓ and channel j, $\sum_{i \in [L]} F_{\ell i} a_{ij} \le 1$. An example of an interference graph along with a feasible allocation is shown in Figure 1.1.

We consider a time slotted system, where the duration of a slot corresponds to the transmission of a single packet. We denote by $X_{ij}(n)$ the number of packets successfully transmitted during slot n when user i selects channel j for transmission in this slot and in absence of interference. Depending on the ability of transmitters to switch channels, we consider two settings. In the *stochastic setting*, the number of successful packet transmissions $X_{ij}(n)$ on link i and channel j are independent over i and j, and are i.i.d. across slots n. The average number of successful packet transmissions per slot is denoted by $\mathbb{E}[X_{ij}(n)] = \theta_{ij}$, and is supposed to be unknown initially. $X_{ij}(n)$ is a Bernoulli random variable of mean θ_{ij}. The stochastic setting models scenarios where the radio channel conditions are stationary. In the *adversarial setting*, $X_{ij}(n) \in [0, 1]$ can be arbitrary (as if it was generated by an *adversary*), and unknown in advance. This setting is useful to model scenarios where the duration of a slot is comparable to or smaller than the channel coherence time. In such scenarios, we assume that the channel allocation cannot change at the same pace as the radio conditions on the various links, which is of interest in practice, when the radios cannot rapidly change channels.

[6] In some works, interference graph is referred to as *conflict graph*.

[7] This model assumes that the interference graph is the same over the various channels. This assumption, however, can be relaxed.



(a) Interference graph

(b) An example of a feasible allocation

Figure 1.1: Spectrum allocation in a wireless system with 5 links and 3 channels

If the radio conditions on each (user, channel) pair were known, the problem would reduce to the following combinatorial optimization problem:

$$\begin{aligned} \max_{a\in\mathcal{A}} \quad & \sum_{i\in[L],j\in[K]} X_{ij}a_{ij} && (1.3)\\ \text{subject to:} \quad & \sum_{j\in[K]} a_{ij} \le 1, \quad \forall i\in[L], \\ & \sum_{i\in[L]} F_{\ell i}a_{ij} \le 1, \quad \forall \ell\in[f], j\in[K], \\ & a_{ij}\in\{0,1\}, \quad \forall i\in[L], j\in[K]. && (1.4) \end{aligned}$$

Problem (1.3) is indeed a coloring problem of the interference graph G, which is shown to be NP-complete for general interference graphs. In contrast, if all links interfere each other (i.e., no two links can be active on the same channel), a case referred to as *full interference*, the above problem becomes an instance of a Maximum Weighted Matching in a bipartite graph (vertices on one side correspond to users and vertices on the other side to channels; the weight of an edge, i.e., a (user, channel) pair, represents the radio conditions for the corresponding user and channel). As a consequence, it can be solved in strongly polynomial time [27].

In practice, the radio conditions on the various channels are not known a priori, and they evolve over time in an unpredictable manner. The presented spectrum allocation problem can be cast as a combinatorial MAB problem, where the objective is to identify a policy maximizing the expected number of successfully transmitted packets over T time slots. The corresponding notion of regret, defined similarly to (1.1), quantifies the performance loss due to the need for learning radio channel conditions. Spectrum sharing problems similar to this have been recently investigated in [6, 5, 28, 29, 30].

1.3.2 Shortest-Path Routing

Shortest-path routing is amongst the first instances of combinatorial MAB problems considered in the literature, e.g., in [31, 32]. As our second example, we consider shortest-path routing in the stochastic setting as studied in [33, 34].

Consider a network whose topology is modeled as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of links. Each link $i \in E$ may, for example, represent an unreliable wireless link. Without loss of generality, we assume that time is slotted and that one slot corresponds to the time to send a packet over a single link. At time t, $X_i(t)$ is a binary random variable indicating whether a transmission on link i is successful. $(X_i(t))_{t\geq 1}$ is a sequence of i.i.d. Bernoulli variables with initially unknown mean θ_i. Hence if a packet is sent on link i repeatedly until the transmission is successful, the time D_i to complete the transmission (referred to as the delay on link i) is geometrically distributed with mean $1/\theta_i$ (see Figure 1.2). Let $\theta_{\min} = \min_{i\in E}\theta_i > 0$, and let $\theta = (\theta_i)_{i\in E}$ be the vector representing the packet successful transmission probabilities on the various links. We consider a single source-destination pair $(u, v) \in V^2$, and denote by $\mathcal{A} \subseteq \{0,1\}^d$ the set of loop-free paths from u to v in G, where each path $a \in \mathcal{A}$ is a d-dimensional binary vector; for any $i \in E$, $a_i = 1$ if and only if i belongs to a. Hence, for any $a \in \mathcal{A}$, the length of path a is $\|a\|_1 = \sum_{i\in E} a_i$.

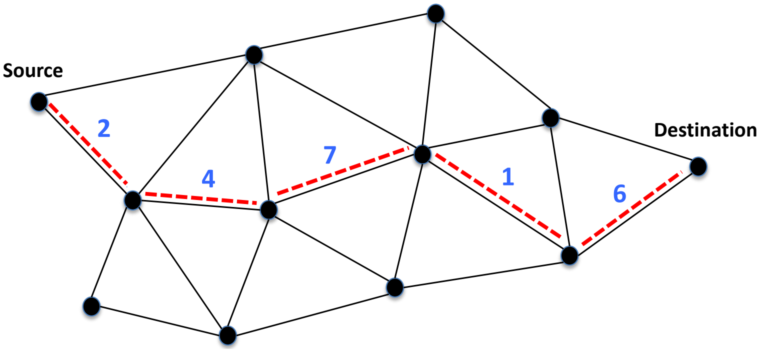


Figure 1.2: Shortest-path routing in a network: A realization of delay for the links along the chosen path (in red)

We assume that the source is fully backlogged (i.e., it always has packets to send), and that the parameter θ is initially unknown. Packets are sent successively from u to v over various paths to estimate θ, and in turn to learn the path $a^\star$, namely the path whose average delay is minimal. After a packet is sent, we assume that the source gathers some feedback from the network (essentially per-link or

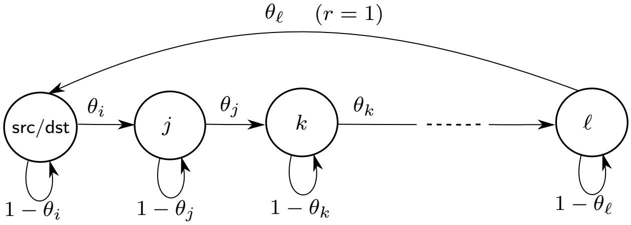


Figure 1.3: The Markov chain induced by a path in the shortest-path routing

end-to-end delays) *before* sending the next packet. If θ were known, one would choose the path $a^\star$ given by

$$a^\star \in \operatorname*{argmin}_{a\in\mathcal{A}} \sum_{i\in E} \frac{a_i}{\theta_i}. \tag{1.5}$$

Our objective is to design and analyze online routing strategies, i.e., strategies that take routing decisions based on the feedback received for the packets previously sent. Depending on the received feedback (per-link or end-to-end delay), we consider two different types of online routing policies: (i) *Source routing with end-to-end (bandit) feedback* in which the path used by a packet is determined at the source based on the observed end-to-end delays for previous packets, and (ii) *source routing with per-link (semi-bandit) feedback*, where the path used by a packet is determined at the source based on the observed per-link delays for previous packets. The variants of online routing problem described above can be cast as:

- a *combinatorial MAB problem* where the rewards are geometrically distributed, and where each path corresponds to an arm;

- and as a *RL problem* where the state-space is the set V of nodes, and where outgoing links at each node correspond to available actions in that state. Moreover, the decision maker receives a non-zero reward (equal to 1) only when the packet is successfully delivered to the destination (see Figure 1.3).

For each case, we can define the corresponding regret definition similarly to (1.1) or (1.2). Furthermore, one can establish the relation between the two. In both setups the regret quantifies the performance loss due to the need to explore sub-optimal paths to learn the path with the minimum delay.

1.4 Thesis Outline and Contributions

Here we present the outline and contributions of this thesis in detail as well as the relation to the corresponding publications.

Chapter 2: Background

This chapter provides background material on classical stochastic MAB and undiscounted RL. In particular, it presents regret lower bounds and some well-known algorithms along with their performance guarantees for both stochastic MAB and undiscounted RL.

Chapter 3: Stochastic Combinatorial MABs

In chapter 3, we study generic stochastic combinatorial MAB with a generic combinatorial structure and bounded rewards. We derive tight and problem-specific lower bounds on the regret of any admissible algorithm under bandit and semi-bandit feedback. These constitute the first lower bounds proposed for generic combinatorial MABs in the literature. Our derivation leverages the theory of optimal control of Markov chains with unknown transition probabilities. We further investigate scaling of the lower bound with the dimension of the underlying combinatorial structure. Furthermore, we propose `ESCB`, an algorithm that efficiently exploits the structure of the problem, and provide a finite-time analysis of its regret. `ESCB` has a better performance guarantee than existing algorithms and significantly outperforms these algorithms in practice as confirmed by our numerical experiments.

The chapter is based on the following publications:

- Marc Lelarge, Alexandre Proutiere, and M. Sadegh Talebi, “Spectrum Bandit Optimization,” in *Information Theory Workshop (ITW)*, 2013.
- Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge, “Combinatorial Bandits Revisited,” in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.
- M. Sadegh Talebi and Alexandre Proutiere, “An Optimal Algorithm for Stochastic Matroid Bandit Optimization,” in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2016.

Chapter 4: Stochastic Online Shortest-Path Routing

In Chapter 4, we study online shortest-path problem discussed in Section 1.3.2. We consider several scenarios that differ in where routing decisions are made and in the feedback available when making the decision. Leveraging similar techniques as in Chapter 3, for each scenario we derive a tight asymptotic lower bound on the regret. For the case of source routing, namely when routing decisions are determined at the source node, we then propose two algorithms: `GeoCombUCB-1` and `GeoCombUCB-2`. Moreover, we improve the regret upper bound of `KL-SR` [33]. These algorithms exhibit a trade-off between computational complexity and performance. Moreover, the regret upper bounds of these algorithm improve over those of state-of-the-art algorithms. Numerical experiments also validated that these policies outperform state-of-the-art algorithms in practice.

The chapter is based on the following work:

- M. Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson, "Stochastic Online Shortest Path Routing: The Value of Feedback," *IEEE Transaction on Automatic Control*, to appear.

Chapter 5: Learning Proportionally Fair Allocations

Chapter 5 addresses a generic sequential resource allocation problem, which is motivated by a fairly large class of applications that arise in, e.g., crowdsouring systems and wireless communication. The considered problem involves a decision maker, who selects in each round an allocation of resources (servers) to a set of tasks consisting of a large number of jobs. A job (or sub-task) is successfully treated by a server with a fixed and task-server dependent probability in a round, and the decision maker is informed on whether this job is completed at the end of the round. The success probabilities are initially unknown and have to be learnt. The objective of the decision maker is to sequentially assign jobs of various tasks to servers so that it rapidly learns and converges to the Proportionally Fair (PF) allocation (or other similar allocations achieving an appropriate trade-off between efficiency and fairness). In that chapter, we provide a formulation of the problem as a MAB problem, for which we devise a sequential assignment algorithm with low regret. The latter is defined as the difference in utility achieved by an oracle algorithm aware of success probabilities and by the proposed algorithm.

The chapter is based on the following work:

- M. Sadegh Talebi and Alexandre Proutiere, "Learning Proportionally Fair Allocations with Semi-bandit Feedback," *in preparation.*

Chapter 6: Variance-Aware Regret Bounds for Undiscounted RL

Chapter 6 concerns RL in MDPs under average-reward criterion. We revisit some concentration inequalities that prove useful for the analysis of this setting. We revisit the analysis of the KL-UCRL algorithm of Filippi et al. [24] and show that under mild assumptions its regret scales with the local variance of the bias function of the MDP. Our regret bounds illuminates the benefit that could be obtained by using Kullback-Leibler divergence. We also provide variance-aware regret lower bounds for the considered RL problem. Our derivation of lower bounds largely relies on the existing ones in the literature.

The chapter relies on the following works:

- M. Sadegh Talebi and Odalric-Ambrym Maillard, "Variance-Aware Regret Bounds for Undiscounted RL in MDPs," *submitted to Algorithmic Learning Theory (ALT).*

- M. Sadegh Talebi, Alexandre Proutiere, and Odalric-Ambrym Maillard, "Revisiting Regret Lower Bounds for Undiscounted Reinforcement Learning," *in preparation.*

Chapter 7: Conclusions and Future Work

Chapter 7 draws some conclusions and provides some directions for the future work.

Appendices

The thesis is concluded with two appendices. The first appendix overviews some important properties of the Kullback-Leibler divergence, whereas the second one presents several important concentration inequalities. The results in both appendices prove useful for the analyses in the various chapters of this thesis.

1.4.1 Additional publications:

- Mohammad Hassan Hajiesmaili, M. Sadegh Talebi, and Ahmad Khonsari, "Multi-Period Network Rate Allocation with End-to-End Delay Constraints," *IEEE Transactions on Control of Network Systems*, to appear.
- Mohammad Hassan Hajiesmaili, M. Sadegh Talebi, and Ahmad Khonsari, "Utility-optimal Dynamic Rate Allocation under Average End-to-end Delay Requirements," in *Decision and Control Conference (CDC)*, 2015.

Chapter 2

Background

This chapter provides background materials on stochastic MABs and undiscounted reinforcement learning in MDPs.

2.1 Stochastic MAB

The multi-armed bandit (MAB) problem was introduced in the seminal paper by Robbins [16] to study the sequential design of experiments. The first bandit algorithm, however, dates back to a paper by Thompson [17] in 1933. In this section, we give an overview of regret lower bounds as well as various algorithms for stochastic MAB.

The classical stochastic MAB is formalized as follows. Let us assume that we have $K \geq 2$ arms. Successive plays of arm i generates the reward sequence $(X_i(n))_{n\geq 1}$. For any i, the sequence of rewards $(X_i(n))_{n\geq 1}$ is drawn i.i.d. from a parametric distribution $\nu(\theta_i)$, where $\theta_i \in \Theta$ is a parameter initially unknown to the decision maker. We let $\mu(\theta)$ denote the expected value of $\nu(\theta)$ for any $\theta \in \Theta$. We assume that the rewards are independent across various arms.

A policy or algorithm π is a sequence of random variables $I^\pi(1), I^\pi(2), \dots$ all taking values from $[K]$, where $I^\pi(n)$ denotes the arm chosen at round n under π, such that $\{I^\pi(n) = i\} \in \mathcal{F}_n$ for all $i \in [K]$ and $n \geq 1$. Let Π be the set of all feasible policies. The objective is to identify a policy in Π maximizing the cumulative expected reward over a finite time horizon T. The expectation is here taken with respect to the randomness in the rewards and the possible randomization in the policy. Equivalently, we aim at designing a policy that minimizes regret, where the regret of policy $\pi \in \Pi$ is defined by:

$$\mathfrak{R}_{\pi,T} = \max_{i\in[K]} \mathbb{E}[\sum_{n=1}^{T} X_i(n)] - \mathbb{E}[\sum_{n=1}^{T} X_{I^\pi(n)}(n)].$$

For any $i \in [K]$ introduce $\Delta_i = \max_{j\in[K]} \mu(\theta_j) - \mu(\theta_i)$. Moreover, let $t_i^\pi(n)$ denote the number of times arm i is selected up to round n under policy π, i.e.,

$t_i^\pi(n) = \sum_{s=1}^n \mathbb{I}\{I^\pi(s) = i\}$. Then, the regret $\mathfrak{R}_{\pi,T}$ can be decomposed as follows:

$$\mathfrak{R}_{\pi,T} = \sum_{i \in [K]} \Delta_i \mathbb{E}[t_i^\pi(T)].$$

2.1.1 Lower Bounds on the Regret

In this subsection we present lower bounds on the regret for stochastic MAB problems.

Lai-Robbins Lower Bound

The first lower bound was proposed by Lai and Robbins in their seminal paper [18]. They consider a simple parametric case in which $\Theta \subset \mathbb{R}$. Namely, the distribution of the rewards of a given arm is parameterized by a scalar parameter. To state their result, we first introduce the notion of *uniformly good algorithms.*

Definition 2.1 ([18]). *A policy or algorithm π is* uniformly good *if for all $\theta \in \Theta$, the regret under π satisfies $\mathfrak{R}_{\pi,T} = o(T^\alpha)$ for any $\alpha > 0$.*

Let $i^\star$ be an optimal arm, namely $\mu(\theta_{i^\star}) = \max_{i \in [K]} \mu(\theta_i)$. For the case of distributions parameterized by a single parameter, Lai and Robbins show that the number of times that a sub-optimal arm i is pulled by any uniformly good policy π satisfies:

$$\liminf_{T \to \infty} \frac{\mathbb{E}[t_i^\pi(T)]}{\log(T)} \geq \frac{1}{\mathtt{KL}(\nu(\theta_i), \nu(\theta_{i^\star}))},$$

where $\mathtt{KL}(p, q)$ denotes the Kullback-Leibler divergence between two distributions p and q.[1] From the regret decomposition rule described above, it then follows that the regret satisfies:[2]

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq \sum_{i:\Delta_i > 0} \frac{\Delta_i}{\mathtt{KL}(\theta_i, \theta_{i^\star})}.$$

This result indeed defines *the asymptotic optimality* criterion: An algorithm π is said to be asymptotically optimal if its regret for any $\theta \in \Theta$ satisfies:

$$\limsup_{T \to \infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \leq \sum_{i:\Delta_i > 0} \frac{\Delta_i}{\mathtt{KL}(\theta_i, \theta_{i^\star})}.$$

Lai-Robbins lower bound is generalized in subsequent works, e.g., [36, 37, 38]. Extension to *multiple play*, i.e., the case where multiple arms are pulled at the same time, is studied by Anantharam et al. [36, 37]. Let us assume that arms are enumerated such that $\mu(\theta_1) \geq \mu(\theta_2) \geq \cdots > \mu(\theta_{m+1}) \geq \cdots \geq \mu(\theta_K)$ and that at

[1] With some abuse of notation, hereafter we write $\mathtt{KL}(\theta, \theta')$ to indicate $\mathtt{KL}(\nu(\theta), \nu(\theta'))$.

[2] A simplified proof of this result can be found in [35, Chapter 1].

each round, m arms are played. Anantharam et al. [36] show that the regret of any uniformly good rule π satisfies:

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \leq \sum_{i=m+1}^{K} \frac{\mu(\theta_m)-\mu(\theta_i)}{\mathtt{KL}(\theta_i,\theta_m)}.$$

Furthermore, [37] investigates the case when multiple arms are played and rewards are generated from an aperiodic and irreducible Markov chain with a finite state-space. These results were also extended and generalized by Burnetas and Katehakis [38] to distributions that rely on multiple parameters, and by Graves and Lai [39] to a more general framework of adaptive control of Markov chains.

Regret Lower Bound for Adaptive Control of Markov Chains

Graves and Lai [39] study adaptive control algorithms for controlled Markov chains with unknown transition probabilities. The Markov chain is assumed to have a general state-space and its transition probabilities are parameterized by an unknown parameter belonging to some compact metric space. The framework of Graves and Lai generalizes those of Lai and Robbins [18], Anantharam et al. [37], and Burnetas and Katehakis [38], and plays a pivotal role in the derivation of lower bound on the regret for various problems in this thesis. Here, we give an overview of this general framework.

Consider a controlled Markov chain $(X_n)_{n\geq 0}$ on a finite state-space $\mathcal{S}$ with a control set U. The transition probabilities given control $u \in U$ are parameterized by θ taking values in a compact metric space Θ: The probability to move from state x to state y given the control u and the parameter θ is $p(x, y; u, \theta)$. The parameter θ is not known. The decision maker is provided with a finite set of stationary control laws $G = \{g_1, \dots, g_K\}$, where each control law g_j is a mapping from $\mathcal{S}$ to U: When control law g_j is applied in state x, the applied control is $u = g_j(x)$. It is assumed that if the decision maker always selects the same control law g, the Markov chain is then irreducible with stationary distribution π_θ^g. Now the reward obtained when applying control u in state x is denoted by $r(x, u)$, so that the expected reward achieved under control law g is:

$$\mu_\theta(g) = \sum_{x\in\mathcal{S}} r(x, g(x))\pi_\theta^g(x).$$

Given θ, an optimal control law is optimal if its expected reward equals

$$\mu_\theta^\star := \max_{g\in G} \mu_\theta(g).$$

Letting $J(\theta) = \{j \in [K] : \mu_\theta(g_j) = \mu_\theta^\star\}$, the set of optimal stationary control laws is $\{g_j, j \in J(\theta)\}$. Now the objective of the decision maker is to sequentially select control laws so as to maximize the expected reward up to a given time horizon T.

An *adaptive control algorithm* φ is a sequence of random variables $I(1), I(2), \dots$ that belong to G such that $\{I(n) = g\} \in \mathcal{F}_n$ for all $g \in G$ and $n \geq 1$. An adaptive control algorithm φ is said to be uniformly good if for all $\theta \in \Theta$, we have that $\mathfrak{R}_{T,\varphi} = \mathcal{O}(\log(T))$ and $S(T) = o(\log(T))$, where $S(T)$ denotes the number of switchings between successive control laws such that both are not optimal, up to round T. The performance of an adaptive control algorithm φ can be quantified through the notion of regret which compares the expected reward to that obtained by always applying the optimal control law:

$$\mathfrak{R}_{T,\varphi} = T\mu_\theta^\star - \mathbb{E}[\sum_{n=1}^{T} r(X_n, u_n)] = \sum_{g \in G : \mu_\theta(g) < \mu_\theta^\star} (\mu_\theta^\star - \mu_\theta(g))\mathbb{E}[t_g(T)],$$

where $(X_n)_{n \geq 1}$ denotes the sequence of states generated by φ.

In order to state the lower bound on the regret of a uniformly good (adaptive control) rule, we first introduce some concepts. For control law $g \in G$, the Kullback-Leibler information number is defined by

$$I^g(\theta, \lambda) = \sum_x \sum_y \log \frac{p(x, y; g(x), \theta)}{p(x, y; g(x), \lambda)} p(x, y; g(x), \theta) \pi_\theta^g(x).$$

Next we introduce the notion of bad parameter set. Let us decompose Θ into L subsets $\{\Theta_j, j \in [L]\}$, such that for any $\theta \in \Theta_j$, g_j is the stationary control law, i.e.,

$$\Theta_j = \{\theta \in \Theta : \mu_\theta(g_j) = \max_{g \in G} \mu_\theta(g)\}.$$

Then the set of bad parameters, denoted by $B(\theta)$, is

$$B(\theta) = \Big\{\lambda \in \Theta : \lambda \notin \bigcup_{j \in J(\theta)} \Theta_j \text{ and } I^{g_j}(\theta, \lambda) = 0, \forall j \in J(\theta)\Big\}.$$

Indeed, $B(\theta)$ is the set of bad parameters that are *statistically indistinguishable* from θ under optimal control laws $\{g_j, j \in J(\theta)\}$.

The following theorem asserts that under certain regularity conditions, the regret of any uniformly good rule admits the asymptotic lower bound of $(c(\theta) + o(1)) \log(T)$.

Theorem 2.1 ([39, Theorem 1]). *For every $\theta \in \Theta$ and for any uniformly good algorithm φ,*

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_{T,\varphi}}{\log(T)} \geq c(\theta),$$

where

$$c(\theta) = \inf\Big\{\sum_{j \notin J(\theta)} x_j(\mu^\star - \mu(g_j)) : \ x_j \geq 0, \ \inf_{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} x_j I^{g_j}(\theta, \lambda) \geq 1\Big\}.$$

We remark that $c(\theta)$ is the optimal value of a linear semi-infinite program (LSIP) [40]. Hence, in general it is difficult to compute though in some cases deriving explicit solution is possible.

Theorem 2.1 indicates that within the T first rounds, the total amount of draw of a sub-optimal control law g_j should be of the order of $x_j^\star \log(T)$, where $x_j^\star$ is the optimal solution of the presented optimization problem. Graves and Lai present policies that achieve this objective, but they are unfortunately extremely difficult to implement in practice. Indeed, these policies may require to solve, in each round, a LSIP that might be computationally expensive.

Minimax Lower Bound

We finally present the following theorem from [21], which establishes a problem-independent lower bound on the regret:

Theorem 2.2 ([21, Theorem 5.1]). *We have that:*

$$\inf_{\pi} \sup_{(\nu_k)_{k\in[K]}} \mathfrak{R}_{T,\pi} \geq \frac{1}{20}\sqrt{KT},$$

where sup *is taken over all set of K distributions on* $[0,1]$ *and* inf *is taken over all policies.*

This lower bound implies that for any algorithm there exists a choice of reward sequence such that the expected regret grows at least as $\Omega(\sqrt{KT})$.

2.1.2 Algorithms for Stochastic MAB

In this section we present some of the most important algorithms for the stochastic MAB problem.

Upper Confidence Bound Index Policies

Most of the algorithms we present here are *upper confidence bound index policies*, or index policies for short, whose underlying idea is to select the arm with the largest (high-probability) upper confidence bound for the expected reward. To this end, an index policy maintains an *index function* for each arm, which is a function of the past observations of this arm only (e.g., the empirical average reward, the number of draws, etc.). The index policy then simply consists in selecting the arm with the maximal index at each round. Algorithm 2.1 shows the pseudo-code of a generic index policy that relies on index function ξ.

An index policy relies on constructing an upper confidence bound for the expected reward of each arm[3] in a way that $\mu_i \in [\hat{\mu}_i(n) - \delta_i(n),\ \hat{\mu}_i(n) + \delta_i(n)]$ with high probability, where $\hat{\mu}_i(n)$ denotes the empirical average reward of arm i up to

[3] Of course, for loss minimization we are interested in lower confidence bounds.

Algorithm 2.1 Index policy using index ξ

```
for $n \geq 1$ do
    Select arm $I(n) \in \arg\max_{i\in[K]} \xi_i(n)$.
    Observe the rewards, and update $t_i(n)$ and $\hat{\theta}_i(n), \forall i \in [K]$.
end for
```

time n. A sub-optimal arm will be selected if $\delta_i(n)$ is large or if $\hat{\mu}_i(n)$ is large. Observe that $\delta_i(n)$ quickly decreases if arm i is sampled sufficiently. Moreover, the number of times that i is selected and $\hat{\mu}_i(n)$ is badly estimated is finite. Hence it is expected that after sampling sub-optimal arms sufficiently, the index policy will select the optimal arm most of the time.

Index policies were first introduced in the seminal work of Gittin [41] for the MABs in the Bayesian setting. For non-Bayesian stochastic MAB problems, the first index policy was introduced by Lai and Robbins [18]. This policy constitutes the first asymptotically optimal algorithm for the classic MAB problem. Lai and Robbins' algorithm was very complicated. Hence it motivated developments of simpler index policies in subsequent works, e.g., in [42, 20, 43, 44, 45]. Agrawal [42] proposed simple index policies in explicit form for some distributions such as Bernoulli, Poisson, Gaussian, etc. He further showed that these policies are asymptotically optimal and achieve $\mathcal{O}(\log(T))$ regret.

The `UCB1` Algorithm [20]. It wasn't until the paper by Auer et al. [20] that a finite-time analysis of index policies was presented. Auer et al. consider rewards drawn from distributions with (known) bounded supports. Without loss of generality assume that the support of rewards is $[0,1]$. Under this assumption, Auer et al. propose the following index

$$b_i(n) = \hat{\mu}_i(n) + \sqrt{\frac{\alpha \log(n)}{t_i(n)}}.$$

To simplify the presentation, in what follows we assume that the first arm $i = 1$ is the unique optimal arm. In the following theorem, we present a regret upper bound for `UCB1` for $\alpha = 3/2$. [4]

Theorem 2.3 ([20])**.** *The regret under $\pi =$`UCB1` satisfies*

$$\mathfrak{R}_{\pi,T} \leq 6 \sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i} + \frac{K\pi^2}{6} + \sum_{i>1} \frac{4}{\Delta_i}.$$

[4] In their paper, Auer et al. originally chose $\alpha = 2$ and provided the following regret upper bound for `UCB1`:

$$\mathfrak{R}_{\texttt{UCB1},T} \leq 8 \sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i>1} \Delta_i.$$

Observe that UCB1 achieves a sub-optimal regret in view of Lai and Robbins' lower bound since $\texttt{kl}(\theta_i, \theta_1) > 2\Delta_i^2$.

The UCB-V Algorithm [44]. Incorporating variance estimates into index function allows to have superior algorithms. One such index policy is UCB-Tuned [20], for which no theoretical guarantee is proposed. Audibert et al. [44] proposed UCB-V (UCB1 with Variance estimates) index which incorporates variance estimates (empirical variance) in the index. UCB-V index is defined as

$$\hat{\theta}_i(n) + \sqrt{\frac{2\alpha V_i(n)\log(n)}{t_i(n)}} + 3\alpha\frac{\log(n)}{t_i(n)},$$

where $V_i(n)$ is the empirical variance of arm i up to round n:

$$V_i(n) = \frac{1}{t_i(n)}\sum_{n=1}^{t_i(n)}(X_i(n) - \hat{\theta}_i(n))^2.$$

Let σ_i^2 denote the variance of arm i. It is shown that UCB-V achieves the following regret upper bound [44]:

$$\mathfrak{R}_{\texttt{UCB-V},T} \le 10\Big(\sum_{i:\Delta_i>0}\frac{\sigma_i^2}{\Delta_i} + 2\Big)\log(T).$$

The KL-UCB Algorithm [43]. The KL-UCB algorithm is an optimal algorithm for stochastic MABs with bounded rewards proposed by Garivier and Cappé [43] (see also [45, 46]). KL-UCB relies on the following index:

$$b_i(n) = \sup\big\{q \in \Theta : t_i(n)\texttt{kl}(\hat{\theta}_i(n), q) \le \log(n) + 3\log(\log(n))\big\}.$$

The following theorem provides the regret bound of KL-UCB.

Theorem 2.4 ([43]). *The regret under $\pi =$KL-UCB satisfies*

$$\mathfrak{R}_{\pi,T} \le (1+\varepsilon)\sum_{i:\Delta_i>0}\frac{\Delta_i}{\texttt{kl}(\theta_i,\theta_1)}\log(T) + C_1\log(\log(T)) + \frac{C_2(\varepsilon)}{T^{\beta(\varepsilon)}},$$

where C_1 is a positive constant and where $C_2(\varepsilon)$ and $\beta(\varepsilon)$ denote positive functions of ε. Hence,

$$\limsup_{T\to\infty}\frac{\mathfrak{R}_{\pi,T}}{\log(T)} \le \sum_{i:\Delta_i>0}\frac{\Delta_i}{\texttt{kl}(\theta_i,\theta_1)}.$$

It is noted that the regret upper bound of KL-UCB matches the lower bound of Burnetas and Katehakis [38].

Algorithm 2.2 `TS`

```
Initialization: For each arm $i \in [K]$ set $S_i = 0, F_i = 0$.
for $n \geq 1$ do
  For each arm $i$, sample $z_i(n)$ from Beta$(S_i + 1, F_i + 1)$.
  Play arm $I(n) = \arg\max_{i \in [K]} z_i(n)$ and receive the reward $X_{I(n)}$.
  if $X_{I(n)} = 1$ then
    Set $S_{I(n)} = S_{I(n)} + 1$.
  else
    Set $F_{I(n)} = F_{I(n)} + 1$.
  end if
end for
```

The `Thompson Sampling` Algorithm

`Thompson Sampling` (`TS`) was proposed by Thompson [17] in 1933. However, it was not until very recently that its regret analysis was presented by Agrawal and Goyal [47, 48] and Kaufmann et al. [49].

In contrast to previously described index policies, `TS` belongs to the family of *randomized probability matching algorithms* and selects an arm based on posterior samples. The underlying idea in `TS` is to assume a prior distribution on the parameters of the reward distribution of every arm. Then at any time step, `TS` plays an arm according to its posterior probability of being the best arm. Algorithm 2.2 presents the pseudo-code of `TS` for the case of Bernoulli rewards, for which the appropriate prior distribution is the Beta distribution (see, e.g., [48] for details).

The first regret analysis for `TS` was proposed by Agrawal and Goyal [47]. Later, Kaufmann et al. [49] improved this regret analysis and proved the asymptotic optimality of `TS` for classical stochastic MABs. Optimality of `TS` was also addressed by Agrawal and Goyal [48] with a different regret analysis. In the following theorem, we provide the regret upper bound for `TS` with Beta priors.

Theorem 2.5 ([48, Theorem 1]). *The regret under* `TS` *using Beta priors satisfies:*

$$\mathfrak{R}_{\texttt{TS},T} \leq (1+\varepsilon) \sum_{i:\Delta_i>0} \frac{\Delta_i}{\texttt{kl}(\theta_i, \theta_1)} \log(T) + C(\varepsilon, \theta_1, \dots, \theta_K),$$

where $C(\varepsilon, \theta_1, \dots, \theta_K)$ is a problem-dependent constant independent of T. In particular, $C(\varepsilon, \theta_1, \dots, \theta_K) = \mathcal{O}(K\varepsilon^{-2})$.

2.2 Theory of Markov Decision Processes

In this section, we briefly overview the background material on the theory of MDPs. These results can be found in standard textbooks, e.g., [50].

Consider a finite MDP M as a tuple $M = (\mathcal{S}, \mathcal{A}, \nu, p)$ where $\mathcal{S}$ is a finite set of states and $\mathcal{A}$ is a finite set of actions available at any state, with respective cardinalities S and A. The functions ν and p respectively denote the reward function and the transition kernel. Taking action a in state s results in a random instantaneous reward drawn from $\nu(s,a)$ with mean $\mu(s,a)$, as well as a transition to state s' with probability $p(s'|s,a)$.

A policy in MDP M is a mapping from the set of states to the set of actions. More formally, let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ denote a possibly stochastic policy. We further introduce the notation $p(s'|s,\pi(s)) = \mathbb{E}_{Z\sim\pi(s)}[p(s'|s,Z)]$, and $P_\pi f$ to denote the function $s \mapsto \sum_{s'\in\mathcal{S}} p(s'|s,\pi(s))f(s')$. Likewise, let $\mu_\pi(s) = \mathbb{E}_{Z\sim\pi(s)}[\mu(s,Z)]$ denote the mean reward after choosing action $\pi(s)$ in step s.

Definition 2.2 (Expected Cumulative Reward). *The expected cumulative reward of policy π when run for T steps from initial state s_1 is defined as*

$$R_{\pi,T}(s_1) = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t,a_t)\right] = \mu_\pi(s_1) + (P_\pi\mu_\pi)(s_1) + \cdots = \sum_{t=1}^{T}(P_\pi^{t-1}\mu_\pi)(s_1)\,.$$

where $a_t \sim \pi(s_t)$, $s_{t+1} \sim p(\cdot|s_t,a_t)$, and finally $r(s,a) \sim \nu(s,a)$.

Definition 2.3 (Average Gain and Bias). *Let us introduce the average transition operator $\overline{P}_\pi = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} P_\pi^{t-1}$. The average gain g_π and bias function b_π are defined by*

$$\begin{array}{ll} g_\pi(s_1) & = \lim_{T\to\infty}\frac{1}{T}R_{\pi,T}(s_1) = (\overline{P}_\pi\mu_\pi)(s_1)\,,\\ b_\pi(s) & = \sum_{t=1}^{\infty}\Big((P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi\Big)(s)\,. \end{array}$$

The previous definition requires some mild assumption on the MDP for the limits to make sense. It is shown (see, e.g., [50]) that the average gain achieved by executing a stationary policy π in a communicating MDP M is well-defined and further does not depend on the initial state, i.e., $g_\pi(s_1) = g_\pi$. For this reason, in the sequel, we restrict our attention to such MDPs. Let $\star$ denote an optimal policy, that is [5] $g^\star = \max_\pi g_\pi$.

Lemma 2.1 (Bias and Gain). *The gain and bias function satisfy the following relations*

$$\begin{array}{lrcl} (\textit{Bellman equation}) & b_\pi + g_\pi & = & \mu_\pi + P_\pi b_\pi \\ (\textit{Fundamental matrix}) & b_\pi & = & (I - P_\pi + \overline{P_\pi})^{-1}(I - \overline{P}_\pi)\mu_\pi\,. \end{array}$$

This result is an easy consequence of the fact that $\overline{P}_\pi$ (see Definition 2.3) satisfy $\overline{P}_\pi P_\pi = P_\pi\overline{P}_\pi = \overline{P}_\pi\overline{P}_\pi = \overline{P}_\pi$ (see [50] for details).

According to the standard terminology, we say a policy is $b^\star$-improving policy if it satisfies $\pi(s) = \mathrm{argmax}_{a\in\mathcal{A}}\,\mu(s,a) + (P_a b^\star)(s)$. Applying the theory of MDPs

[5] The maximum is reached since there are only finitely many deterministic policies.

(see, e.g., [50]), it can be shown that any $b^\star$-improving policy is optimal and thus that we can choose $\star$ to satisfy[6] the following fundamental identity

$$\forall s \in \mathcal{S}, \quad b^\star(s) + g^\star = \max_{a\in\mathcal{A}}\Big(\mu(s,a) + \sum_{y\in\mathcal{S}} p(y|s,a) b^\star(y)\Big) \,.$$

This equation is referred to as *Bellman optimality equation.*

In the sequel, we recall the definitions of diameter and mixing time:

Definition 2.4 (Diameter [25])**.** *Let $T_\pi(s'|s)$ denote the first hitting time of state s' when following stationary policy π from initial state s. The diameter D of an MDP M is defined as*

$$D := \max_{s\neq s'} \min_{\pi} \mathbb{E}[T_\pi(s'|s)].$$

Definition 2.5 (Mixing Time [51])**.** *Let $\mathcal{C}_\pi$ denote the Markov chain induced by policy π in an ergodic MDP M and let $T_{\mathcal{C}_\pi}$ represent the hitting time of $\mathcal{C}_\pi$. The mixing time T_M of MDP M is defined as*

$$T_M := \max_\pi T_{\mathcal{C}_\pi} \ .$$

2.2.1 Value Iteration

Now we introduce *Value Iteration (VI)*, also known as successive approximation, which is an iterative procedure to find an optimal policy (as well as bias function) via solving Bellman optimality equation.

VI defines a sequence of functions $(u_n)_{n\in\mathbb{N}}$ and policies $(\pi_n)_{n\in\mathbb{N}}$, where $u_0 = 0$, and for all $n \in \mathbb{N}$,

$$\begin{cases} u_{n+1}(s) = \max_{a\in\mathcal{A}} \mu(s,a) + (P_a u_n)(s)\,, \\ \pi_{n+1}(s) = \mathcal{U}\Big(\operatorname{Argmax}_{a\in\mathcal{A}} \mu(s,a) + (P_a u_n)(s)\Big)\,, \end{cases}$$

where $\mathcal{U}(\mathcal{B})$ denotes the uniform distribution over a set $\mathcal{B}$.

One can stop VI when $\mathbb{S}(u_{n+1} - u_n) \le \varepsilon$, where $\varepsilon > 0$ is an input parameter, and where for any function f defined on $\mathcal{S}$, $\mathbb{S}(f) := \max_{s\in\mathcal{S}} f(s) - \min_{s\in\mathcal{S}} f(s)$ denotes the span of f.[7] If n is such that the stopping criterion above is met, then it holds that

$$g^\star - g_{\pi_{n+1}} \le \varepsilon, \quad |u_{n+1} - u_n - g^\star| \le \varepsilon \quad \text{and} \quad |u_{n+1} - u_n - g_{\pi_n+1}| \le \varepsilon\,.$$

[6] The solution to this fixed-point equation is defined only up to an additive constant.

[7] Span operator actually acts as a semi-norm (see [50]).

2.3 Undiscounted RL in MDPs

In this section, we provide background material on RL under average-reward criterion in MDPs, which we refer to as undiscounted RL. The problem involves a decision maker interacting with an unknown environment that can be modeled by an unknown and discrete MDP. She interacts with the system in a single stream of observations, starting from an initial state without any reset. The game goes as follows: The decision maker starts in some state $s_1 \in \mathcal{S}$ at time $t = 1$. At each time step $t \in \mathbb{N}$, the decision maker chooses one action $a \in \mathcal{A}$ in her current state $s \in \mathcal{S}$ based on her past decisions and observations. When executing action a in state s, she receives a random reward r drawn independently from distribution $\nu(s, a)$ with support $[0, 1]$ and mean $\mu(s, a)$. The state then transits to a next state $s' \in \mathcal{S}$ sampled with probability $p(s'|s, a)$, and a new decision step begins. As the transition probabilities and reward functions are unknown, the decision maker has to learn them by trying different actions and recording the realized rewards and state transitions.

The performance of the decision maker can be quantified through the notion of regret, which compares the reward collected by the algorithm to that obtained by an oracle always following an optimal policy. Given a learning algorithm $\mathbb{A}$, consider the following quantity that compares the cumulative reward after T steps obtained by an optimal algorithm to that obtained by $\mathbb{A}$:

$$\mathrm{Reg}_{\mathbb{A},T} := \sum_{t=1}^{T} r(s_t^\star, \star(s_t^\star)) - \sum_{t=1}^{T} r(s_t, a_t)\,,$$

where $a_t = \mathbb{A}(s_t, (\{s_{t'}, a_{t'}, r_{t'}\})_{t'<t})$ and $s_{t+1}^\star \sim p(\cdot|s_t^\star, \star(s_t^\star))$ with $s_1^\star = s_1$ is a sequence of states generated by the optimal strategy.

By an application of Azuma-Hoeffding's inequality for bounded martingales, it is immediate to show that with probability higher than $1 - \delta$,

$$\begin{aligned}\mathrm{Reg}_{\mathbb{A},T} &\leq \sum_{t=1}^{T} \Big(P_\star^{t-1}\mu_\star - r(s_t, a_t)\Big) + \sqrt{2T\log(1/\delta)} \\ &= \sum_{t=1}^{T} (P_\star^{t-1} - \overline{P}_\star)\mu_\star + \left[Tg^\star - \sum_{t=1}^{T} r(s_t, a_t)\right] + \sqrt{2T\log(1/\delta)}\,.\end{aligned}$$

Thus, following [25], it makes sense to focus on the control of the middle term in brackets only, which we now call the *effective regret*:

$$\mathrm{Regret}_{\mathbb{A},T} := Tg^\star - \sum_{t=1}^{T} r(s_t, a_t)\ .$$

We finally introduce the following quantity that appears in the known problem-dependent lower-bounds on the regret, and plays the analogue of the mean gap in the bandit literature.

Definition 2.6 (Sub-optimality gap)**.** *The sub-optimality of action a at state s is*

$$\varphi(s,a) = \mu(s,\star(s)) - \mu(s,a) + (p(\cdot|s,\star(s)) - p(\cdot|s,a))^\top b^\star\,. \tag{2.1}$$

Note importantly that φ is defined in terms of the bias $b^\star$ of the optimal policy $\star$. Indeed, it can be shown that minimizing the effective regret is essentially equivalent to minimizing the quantity $\sum_{s,a}\varphi(s,a)\mathbb{E}[N_T(s,a)]$, where $N_T(s,a)$ is the total number of steps when action a has been played in state s. More precisely, it is not difficult to show that for any stationary policy π and all t

Lemma 2.2 (Effective Regret to Pseudo-regret Reduction [52])**.** *Let π be any stationary policy. Then, it holds for all T and any initial state,*

$$\begin{aligned}\mathfrak{R}_{\pi,T}(s_1) &= \big([P_\pi^{T-1} - I]b^\star\big)(s_1) + \sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi(s,a)\\ &\le D + \sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi(s,a)\,.\end{aligned}$$

2.3.1 Regret Lower Bounds

In this section we present lower bounds on the regret for undiscounted RL. The first bound is an asymptotic problem-dependent lower bound, whereas the second one is a non-asymptotic one that holds in a minimax sense.

Burnetas-Katehakis Lower Bound

Burnetas and Katehakis [22] derive a tight lower bound on the regret for the class of ergodic MDPs. They consider RL in ergodic MDPs, where the decision maker does not know the transition probabilities except for their support. However, she knows the average rewards. In other words, she knows $\mu(s,a)$ and $\mathcal{S}_{s,a}^+ := \{s' : p(s'|s,a) > 0\}$ for all pairs (s,a).

To present their lower bound, we introduce some notations. For any state-action pair (s,a), we denote by $\Delta_{s,a}$ the parameter space for the probability vector $p(\cdot|s,a)$:

$$\Delta_{s,a} = \Big\{q \in \mathbb{R}_+^S : \sum_{y\in\mathcal{S}} q(y) = 1 \text{ and } q(y) > 0, \forall y \in \mathcal{S}_{s,a}^+\Big\}.$$

We further define $\Delta := \prod_{(s,a)}\Delta_{s,a}$.

We now introduce the notion of *critical state-action pairs.* To this aim, for a given state-action pair (s,a) and probability vector $q \in \Delta_{s,a}$, we define a modified MDP $M' = (\mathcal{S},\mathcal{A},\nu,Q)$, where the modified transition law $Q \in \mathcal{P}$ satisfies: $q(\cdot|s',a') = q$ if $(s',a') = (s,a)$, and $Q(\cdot|s',a') = p(\cdot|s',a')$ otherwise. Note that MDP M' depends on the true MDP M and on (s,a,q): $M' = M'(M,s,a,q)$. For simplicity, we omit its dependence on these quantities.

For any state-action pair (s, a), we let $\Lambda(s, a)$ be the set of all distributions that make action a the unique optimal action at state s in the modified MDP M':

$$\Lambda(s, a) := \{q \in \Delta_{s,a} : O(s, M') = \{a\}\},$$

where $O(s, M')$ denotes the set of optimal actions in state s in MDP M'. Introduce for any state-action pair (s, a):

$$\mathcal{K}(s, a) = \inf\Big\{\mathtt{KL}(p(\cdot|s, a), q) : q \in \Lambda(s, a)\Big\}.$$

A state-action pair (s, a) is said to be *critical* if a is not optimal in state s, yet there exists some modified MDP M' in which a is optimal in s. Furthermore, we let $\mathcal{C}_M$ denote the set of all critical state-action pairs of M:

$$\mathcal{C}_M = \big\{(s, a) : a \notin O(s, M),\ \Lambda(s, a) \neq \emptyset\big\}.$$

We note that definition of $\mathcal{K}$ implies that $0 < \mathcal{K}(s, a) < \infty$ iff (s, a) is critical. Moreover, if a is sub-optimal at state s, $\mathcal{K}(s, a) = 0$, and when $\Lambda(s, a) = \emptyset$, $\mathcal{K}(s, a) = \infty$.

The following theorem provides the asymptotic regret lower bound on the regret of any *uniformly fast convergent (UF)* algorithm for ergodic MDPs. By definition, an algorithm $\mathbb{A}$ is *uniformly fast convergent (UF)* if from all starting states $s_0 \in \mathcal{S}$, and all ergodic MDPs, the regret under $\mathbb{A}$ satisfies $\mathfrak{R}_{\mathbb{A},T}(s_0) = o(T^\alpha)$ as T grows large, for all $\alpha > 0$ (see [22]). Let Π_F denote the set of all UF algorithms.

Theorem 2.6 (Burnetas-Katehakis Lower Bound [22, Theorem 2]). *For all algorithms $\mathbb{A} \in \Pi_F$, any ergodic MDP M, and any initial states $s_0 \in \mathcal{S}$:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}(s_0)]}{\log T} \geq c_{\mathrm{bk}}(M) := \sum_{(s,a)\in\mathcal{C}_M} \frac{\varphi(s, a)}{\mathcal{K}(s, a)}. \tag{2.2}$$

Minimax Lower Bound

Now we turn to the minimax lower bound presented by Jaksch et al. [25]. To present a lower bound, they consider a family of *hard-to-learn* MDPs in the class of communicating MDPs. Leveraging techniques for deriving minimax lower bounds for MAB problems as in [21], they show that:

Theorem 2.7 ([25, Theorem 5]). *For any algorithm $\mathbb{A}$, there exists an MDP M with $S \geq 10$ states, $A \geq 10$, and diameter $D \geq 10\log_A(S)$, such that for any initial state, the expected regret under $\mathbb{A}$ satisfies:*

$$\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}] \geq 0.015\sqrt{DSAT}\ .$$

The lower bound in Theorem 2.7 implies that the expected regret scales at least as $\Omega(\sqrt{DSAT})$.

2.3.2 Algorithms for Undiscounted RL

In this section, we present two algorithms that could be used for the considered RL setup.

The Burnetas-Katehakis Algorithm

Burnetas and Katehakis [22] propose one of the first algorithms for RL under average-reward criterion for the class of ergodic MDPs. Under the assumption that the decision maker knows average reward function μ and the support of transition kernel, they propose an algorithm, which we refer to as Burnetas-Katehakis.

To present Burnetas-Katehakis, we introduce the set of *relatively frequently sampled* actions for any state s at time t as

$$\mathcal{D}_{t,s} := \{a \in \mathcal{A} : N_t(s,a) \geq \log^2 N_t(s)\} .$$

Any action $a \notin \mathcal{D}_{t,s}$ is referred to as relatively under-sampled in state s at time t. Given MDP M, the associated *restricted empirical MDP* to M at time t is an MDP M_t that excludes relatively under-sampled actions in various states, and whose transition kernel is the empirical kernel derived by the observations so far; namely $M_t = (\mathcal{S}, \mathcal{D}_t, \mu, \hat{p}_t)$, where $\mathcal{D}_t = \cup_s \mathcal{D}_{t,s}$. Burnetas-Katehakis also relies on the following index function: for all (s,a) and $t \geq 1$,

$$U_t(s,a) = \sup_{q \in \Lambda(s,a)} \left\{ \mu(s,a) + q^\top \hat{b}_t : N_t(s,a) \mathrm{KL}(\hat{p}_t(\cdot|s,a), q) \leq \log t \right\} , \tag{2.3}$$

where $\hat{b}_t$ is a bias function satisfying the Bellman optimality equation for restricted empirical MDP M_t.

The Burnetas-Katehakis algorithm can be described as follows. At each time step t, the algorithm forms a $M_t = (\mathcal{S}, \mathcal{D}_t, \nu, \hat{p}_t)$ and then finds $\hat{b}_t$ by solving Bellman optimality equation for M_t. As such a solution might be misleading due to estimation errors, the algorithm computes the index U_t for all actions. If all the optimal actions in M_t may become under-sampled in the next time step, the algorithm takes one of them arbitrarily. Otherwise, it chooses the action with the highest index $a_t \in \mathrm{argmax}_{a \in \mathcal{A}} U_t(s_t, a)$. We refer to Algorithm 2.3 for the pseudo-code of Burnetas-Katehakis.

The following theorem provides the regret upper bound for Burnetas-Katehakis:

Theorem 2.8 ([22, Theorem 1]). *For any starting state, the regret under algorithm* $\mathbb{A} =$ Burnetas-Katehakis *satisfies:*

$$\limsup_{T \to \infty} \frac{\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}]}{\log(T)} \leq c_{\mathrm{bk}}(M) = \sum_{(s,a) \in \mathcal{C}_M} \frac{\varphi(s,a)}{\mathcal{K}(s,a)} .$$

In view of the lower bound of Theorem 2.6, Burnetas-Katehakis is asymptotically optimal in the class of ergodic MDPs. We remark that the above regret bound only holds asymptotically (i.e., as $T \to \infty$) and to the best of our knowledge, no finite-time analysis for Burnetas-Katehakis is provided in the literature.

Algorithm 2.3 Burnetas-Katehakis [22]

```
for $t \geq 1$ do
    Let $\mathcal{D}_{t,s} := \{a \in \mathcal{A} : N_t(s,a) \geq \log^2 N_t(s)\}$ and $\mathcal{D}_t = \cup_s \mathcal{D}_{t,s}$
    Find $\hat{b}_t$ by solving the Bellman equations for the restricted empirical MDP $M_t = (\mathcal{S}, \mathcal{D}_t, \mu, \hat{p}_t)$
    Let $O(s, M_t)$ be the set of optimal actions in state $s$ in $M_t$
    Let $\Gamma_t = \{a \in O(s_t, M_t) : N_t(s_t, a) < \log^2(N_t(s_t) + 1)\}$
    if $\Gamma_t = O(s_t, M_t)$ then
        Choose any $a_t \in \Gamma_t$ arbitrarily
    else
        Choose $a_t \in \mathrm{argmax}_{a \in \mathcal{A}} U_t(s_t, a)$
    end if
end for
```

The Ucrl2 Algorithm

Ucrl2 [25] is an algorithm designed for the class of communicating MDPs, which implements the principle of "optimism in face of uncertainty".

The algorithm works in episodes of increasing lengths. To compute the optimistic policy for the k-th episode, Ucrl2 first defines a set of plausible MDPs $\mathcal{M}_k$, which is the set of all MDPs M' with state-space $\mathcal{S}$ and action space $\mathcal{A}$, whose reward function and transition kernel satisfies:

$$|\tilde{\mu}(s,a) - \hat{\mu}_k(s,a)| \leq \sqrt{\frac{3.5 \log(2SAt_k/\delta)}{N_k(s,a)^+}} \, , \tag{2.4}$$

$$\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S \log(2At_k/\delta)}{N_k(s,a)^+}} \, , \tag{2.5}$$

where t_k denotes the time where episode k starts.

Extended Value Iteration. To find an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ as well as a near-optimal policy in $\tilde{M}_k$, one can use the following iterative procedure, referred to as *Extended Value Iteration*: for all $s \in \mathcal{S}$,

$$\begin{aligned} u_0(s) &= 0 \, , \\ u_{i+1}(s) &= \max_{a \in \mathcal{A}} \Big\{ \tilde{\mu}(s,a) + \max_{p \in \mathcal{P}(s,a)} u_i^\top p \Big\} \, . \end{aligned} \tag{2.6}$$

Computing the inner maximization can be done in $\mathcal{O}(S)$ steps by an algorithm due to [53]; for details see Algorithm 2 in [25]. The pseudo-code of Ucrl2 is provided in Algorithm 2.4.

The following theorem provides a finite-time upper bound on the regret of Ucrl2:

Algorithm 2.4 UCRL2 [25]

```
Set $\delta \in (0,1]$,
Initialize: For all $(s,a)$, set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. Set $t = 1$, $k = 1$, and observe initial state $s_1$
for episodes $k \geq 1$ do
    Set $t_k = t$
    Set $N_k(s,a) = N_{k-1}(s,a) + v_{k-1}(s,a)$ for all $(s,a)$
    Compute empirical estimates $\hat{\mu}_k(s,a)$ and $\hat{p}_k(\cdot|s,a)$ for all $(s,a)$, and form the set of plausible MDPs $\mathcal{M}_k$
        $$\mathcal{M}_k = \Big\{ M' = (\mathcal{S}, \mathcal{A}, \tilde{\mu}_k, \tilde{p}_k) : \text{ (2.4) and (2.5) hold } \Big\} .$$
    Find an $\frac{1}{\sqrt{t_k}}$-optimal policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ using Extended Value Iteration (see (2.6))
    while $v_k(s_t, a_t) \geq N_k(s_t, a_t)$ do
        Play action $a_t = \tilde{\pi}_k(s_t)$, and observe the next state $s_{t+1}$ and reward $r_t$
        Update $N_k(s,a,x)$ and $v_{t+1}(s,a)$ for all actions $a$ and states $s, x$
    end while
end for
```

Theorem 2.9 ([25, Theorem 2])**.** *Let M be a communicating MDP. Then, starting from any initial state in M, the regret under algorithm $\mathbb{A} =$ UCRL2 after $T \geq 2$ steps is bounded by*

$$\mathrm{Regret}_{\mathbb{A},T} \leq 34DS\sqrt{AT\log(T/\delta)}$$

with probability at least $1 - \delta$.

2.A Proof of Lemma 2.2

Since $g^\star$ is a constant function, it first comes

$$\mathfrak{R}_{\pi,T} = \sum_{t=1}^{T} \Big(g^\star - P_\pi^{t-1}\mu_\pi \Big) = \sum_{t=1}^{T} P_\pi^{t-1} \Big(g^\star - \mu_\pi \Big) .$$

Then we note that by construction it holds that $g^\star - \mu_\star = (P_\star - I)b^\star$. thus, introducing the sub-optimality gap $\varphi_\pi(s) = \mu_\star(s) + (P_\star b^\star)(s) - \mu_\pi(s) - (P_\pi b^\star)(s)$. Then it comes

$$g^\star - \mu_\pi = \varphi_\pi + g^\star - \mu_\star - P_\star b^\star + P_{\tilde{\pi}} b^\star = (P_\pi - I)b^\star + \varphi_\pi .$$

Thus far, we have we obtained that

$$\mathfrak{R}_{\pi,T} = \sum_{t=1}^{T} P_\pi^{t-1}\varphi_\pi + \sum_{t=1}^{T} P_\pi^{t-1}(P_\pi - I)b^\star = \sum_{t=1}^{T} P_\pi^{t-1}\varphi_\pi + (P_\pi^{T-1} - I)b^\star .$$

In order to conclude, we note that

$$\begin{aligned}(\sum_{t=1}^{T} P_{\pi_k}^{t-1}\varphi_{\pi_k})(s_1) &= \sum_{t=1}^{T}\mathbb{E}_{s_{t-1}}[\varphi_{\pi_k}(s_{t-1})]\\ &= \sum_{s,a}\varphi_a(s)\sum_{t=1}^{T}\mathbb{E}_{s_{t-1}}[\mathbb{I}\{s_{t-1}=s,\pi_k(s)=a\}]\\ &= \sum_{s,a}\varphi_a(s)\mathbb{E}[N_T(s,a)]\,.\end{aligned}$$

For the inequality, we use the simple bound $[P_\pi^{T-1}-I]b^\star \leq \|P_\pi^{T-1}-I\|_1\frac{1}{2}\mathbb{S}(b^\star) \leq D\,.$
□

Chapter 3

Stochastic Combinatorial MABs

This chapter investigates generic combinatorial MABs with Bernoulli rewards and relies on the publications [54] and [55]. It begins with an outline of contributions and an overview of related works in Section 3.1. Section 3.2 describes the model and objectives. In Section 3.3, we derive lower bounds on the regret under semi-bandit and bandit feedback. In Section 3.4, we present the `ESCB` algorithm and provide a finite-time analysis of its regret. We provide simulation results in Section 3.5. Finally, Section 3.6 summarizes the chapter. All proofs are presented in the appendix.

3.1 Contributions and Related Work

In this chapter we make the following contributions:

(a) We derive asymptotic (as the time horizon T grows large) regret lower bounds satisfied by any algorithm under semi-bandit and bandit feedback (Theorems 3.1 and 3.4). These lower bounds are *problem-specific* and *tight*: There exists an algorithm that attains the bound on all problem instances, although the algorithm might be computationally expensive. To our knowledge, such lower bounds have not been proposed in the case of stochastic combinatorial bandits. The dependency of the lower bound in terms of problem dimensions (m, d) is unfortunately not explicit (recall that d denotes the number of basic actions and m is the maximal number of basic actions per arm). For semi-bandit feedback, we further provide a simplified lower bound (Theorem 3.3) and derive its scaling in (m, d) in specific examples.

(b) In the case of semi-bandit feedback, we propose `ESCB` (Efficient Sampling for Combinatorial Bandits), an algorithm whose regret scales at most as $\mathcal{O}(\frac{\sqrt{m}d}{\Delta_{\min}}\log(T))$ (Theorem 3.8), where $\Delta_{\min}$ denotes the expected reward difference between the best and the second-best arm. `ESCB` assigns an index to each arm. Our proposed indexes are the natural extensions of `KL-UCB` and `UCB` indexes defined for unstructured bandits [43, 20]. We present numerical experiments for some specific combinatorial problems, which show that `ESCB` significantly outperforms existing algorithms.

3.1.1 Related Work

Previous contributions on stochastic combinatorial MABs mainly considered semi-bandit feedback. Most of these contributions focused on specific combinatorial structures, e.g., fixed-size subsets [36, 56], matroids [57, 58], or permutations [30, 59]. Generic combinatorial problems were investigated in [60], [61], and [62]. Gai et al. [60] propose `LLR`, a variant of the `UCB` algorithm that assigns index to basic actions. Gai et al. [60] establish a loose regret bound of $\mathcal{O}(\frac{m^3 d\Delta_{\max}}{\Delta_{\min}}\log(T))$ for `LLR`, where $\Delta_{\max}$ denotes the expected reward difference between the best and the worst arm. Chen et al. [61] present a general framework for combinatorial optimization problems in the semi-bandit setting that covers a large class of problems. Under mild regularity conditions, their proposed framework also allows for nonlinear reward functions. The proposed algorithm in [61], `CUCB`, is a variant of `UCB` that assigns index to basic actions. For linear combinatorial problems, `CUCB` achieves a regret of order $\mathcal{O}(\frac{m^2 d}{\Delta_{\min}}\log(T))$, which improves over the regret bound of `LLR` by a factor of $m\Delta_{\max}/\Delta_{\min}$. For linear combinatorial problems, Kveton et al. [62] improve the regret upper bound of `CUCB`[1] to $\mathcal{O}(\frac{md}{\Delta_{\min}}\log(T))$. However, the constant in the leading term of this regret bound is fairly large. They also derive another regret bound scaling as $\mathcal{O}(\frac{m^{4/3}d}{\Delta_{\min}}\log(T))$ with better constants[2]. Our algorithms improve over `LLR` and `CUCB` by a multiplicative factor of (at least) $\sqrt{m}$. We also remark that for combinatorial MABs under semi-bandit feedback, Wen et al. [63] provide algorithms with problem-independent regret bounds of order $\mathcal{O}(\sqrt{T})$. The performance guarantees of these algorithms are presented in Table 3.1.

In spite of specific lower bound examples, problem-dependent regret lower bounds that hold for all problem instances have not been reported in existing works so far. Such specific results are mainly proposed to examine the tightness of regret bounds. For instance, to prove that a regret of $\mathcal{O}(\frac{md}{\Delta_{\min}}\log(T))$ is order-optimal[3] in terms of d and m, and cannot be beaten in general, Kveton et al. [62] artificially create an instance of shortest-path routing problem, where the rewards of the basic actions of the same arm are identical, or in other words, they consider a classical bandit problem where the rewards of the various arms are either 0 or equal to m. This does not contradict our regret bounds scaling as $\mathcal{O}(\frac{\sqrt{m}d}{\Delta_{\min}}\log(T))$ [4] since we assume independence among the rewards of various basic actions.

Linear combinatorial MABs may be viewed as linear optimization over a polyhedral set. Dani et al. [65] consider stochastic linear optimization over compact and convex sets under bandit feedback. They propose algorithms with high-probability

[1] In [62], the proposed algorithm is `CombUCB1`, which is essentially identical to `CUCB`.

[2] A similar regret scaling for the case of matching problem is provided independently in [59].

[3] A policy π is order-optimal in terms of d and m, if it satisfies the following: For all problem instances, $\mathfrak{R}_{\pi,T} = \mathcal{O}(C_1 g(d,m)\log(T))$ with C_1 independent of d, m, and T, and there exists a problem instance and a constant $C_2 > 0$, independent of d, m, and T, such that $\liminf_{T\to\infty}\mathfrak{R}_{\pi',T}/\log(T) \ge C_2 g(d,m)$ for any uniformly good algorithm π'.

[4] We mention that using a refined analysis one can show that `ESCB` enjoys a regret upper bound of $\mathcal{O}(\frac{\log^2(m)d}{\Delta_{\min}}\log(T))$; see [64].

Algorithm	Regret
`LLR` [60]	$\mathcal{O}\left(\frac{m^3 d\Delta_{\max}}{\Delta_{\min}^2}\log(T)\right)$
`CUCB` [61]	$\mathcal{O}\left(\frac{m^2 d}{\Delta_{\min}}\log(T)\right)$
`CombUCB1` (`CUCB`) [62]	$\mathcal{O}\left(\frac{m^{4/3} d}{\Delta_{\min}}\log(T)\right)$
`CombUCB1` (`CUCB`) [62]	$\mathcal{O}\left(\frac{md}{\Delta_{\min}}\log(T)\right)$
`ESCB` (Theorem 3.8)	$\mathcal{O}\left(\frac{\sqrt{m}d}{\Delta_{\min}}\log(T)\right)$
`ESCB` [64]	$\mathcal{O}\left(\frac{\log^2(m)d}{\Delta_{\min}}\log(T)\right)$

Table 3.1: Regret upper bounds for stochastic combinatorial bandits under semi-bandit feedback.

regret bounds scaling as $\mathcal{O}(\log^3(T))$. We stress, however, that Dani et al. [65] assume that the set of arms $\mathcal{A}$ is full rank and therefore, their algorithms are not applicable to all classes of $\mathcal{A}$.

Finally, we mention that some studies addressed combinatorial MABs under Markovian rewards in the semi-bandit feedback setting. While generic problems are investigated by Tekin et al. [66], most of existing works focused on specific problems, e.g., fixed-size subsets [37] and permutations [67, 29].

3.2 Model and Objectives

We consider MAB problems where each arm a is a subset of at most m basic actions taken from a set E with cardinality d. For $i \in E$, $X_i(n)$ denotes the reward of basic action i in round n. For each i, the sequence of rewards $(X_i(n))_{n\geq 1}$ is i.i.d. with Bernoulli distribution with mean θ_i. Rewards are assumed to be independent across actions. We denote by $\theta = (\theta_1, \dots, \theta_d)^\top \in \Theta = [0,1]^d$ the vector of unknown expected rewards of the various basic actions.

The set of arms $\mathcal{A}$ is an arbitrary subset of $\{0,1\}^d$, such that each of its elements a has at most m basic actions. Arm a is identified with a binary column vector $(a_1, \dots, a_d)^\top$, and we have $\|a\|_1 \leq m$, $\forall a \in \mathcal{A}$. At the beginning of each round n, an algorithm or policy π, selects an arm $a^\pi(n) \in \mathcal{A}$ based on the arms chosen in previous rounds and their observed rewards. The reward of arm $a^\pi(n)$ selected in round n is $X^{a^\pi(n)}(n) = \sum_{i\in E} a_i^\pi(n) X_i(n) = a^\pi(n)^\top X(n)$.

We consider both semi-bandit and bandit feedback. Under semi-bandit feedback and policy π, at the end of round n, the outcome of basic actions $X_i(n)$ for all $i \in a^\pi(n)$ [5] are revealed to the decision maker, whereas under bandit feedback, $a^\pi(n)^\top X(n)$ can only be observed. Let Π_{s} and Π_{b} be respectively the set of all feasible policies with semi-bandit and bandit feedback. The objective is to identify

[5] Throughout for simplicity of the notation, for any binary vector z, we write $i \in z$ to denote $z_i = 1$, and similarly $i \notin z$ to imply $z_i = 0$.

a policy in Π_{s} and Π_{b} maximizing the cumulative expected reward over a finite time horizon T. The expectation is here taken with respect to the randomness in the rewards and the possible randomization in the policy. Equivalently, we aim at designing a policy that minimizes regret, where the regret of policy π is defined by:

$$\mathfrak{R}_{\pi,T} = \max_{a\in\mathcal{A}} \mathbb{E}[\sum_{n=1}^{T} X^{a}(n)] - \mathbb{E}[\sum_{n=1}^{T} X^{a^{\pi}(n)}(n)].$$

Finally, we denote by $\mu_a(\theta) = a^\top\theta$ the expected reward of arm a, and let $a^\star(\theta) \in \mathcal{A}$, or $a^\star$ for short, be any arm with the maximum expected reward: $a^\star(\theta) \in \arg\max_{a\in\mathcal{A}} \mu_a(\theta)$. In what follows, to simplify the presentation, we assume that $a^\star$ is unique. We further define: $\mu^\star(\theta) = a^{\star\top}\theta$, $\Delta_{\min} = \min_{a\neq a^\star} \Delta_a$ where $\Delta_a = \mu^\star(\theta) - \mu_a(\theta)$, and $\Delta_{\max} = \max_a \Delta_a$.

3.3 Regret Lower Bounds

3.3.1 Semi-bandit Feedback

Given θ, define the set of parameters that cannot be distinguished from θ when selecting action $a^\star(\theta)$, and for which arm $a^\star(\theta)$ is sub-optimal:

$$B_{\mathrm{s}}(\theta) = \big\{\lambda \in \Theta : \lambda_i = \theta_i,\ \forall i \in a^\star(\theta),\ \mu^\star(\lambda) > \mu^\star(\theta)\big\}.$$

Let $\mathtt{kl}(u,v)$ be the Kullback-Leibler divergence between Bernoulli distributions of respective means u and v, i.e., $\mathtt{kl}(u,v) = u\log(u/v) + (1-u)\log((1-u)/(1-v))$. We derive a regret lower bound valid for any *uniformly good* algorithm in Π_{s} (see Definition 2.1). The proof of this result relies on a general result on controlled Markov chains due to Graves and Lai [39]; we refer to Chapter 2 for an overview of their result.

Theorem 3.1. *For all $\theta \in \Theta$ and for any uniformly good policy $\pi \in \Pi_{\mathrm{s}}$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq c_{\mathrm{s}}(\theta), \tag{3.1}$$

where $c_{\mathrm{s}}(\theta)$ is the optimal value of the following optimization problem:

$$\inf_{x\geq 0} \sum_{a\in\mathcal{A}} x_a \Delta_a \tag{3.2}$$

$$\text{subject to: } \sum_{i\in E} \mathtt{kl}(\theta_i,\lambda_i) \sum_{a\in\mathcal{A}} x_a a_i \geq 1,\ \ \forall \lambda \in B_{\mathrm{s}}(\theta). \tag{3.3}$$

Observe first that optimization problem (3.2) is a linear semi-infinite program [40], which can be solved for any fixed θ, but its optimal value is difficult to compute explicitly. Determining how $c_{\mathrm{s}}(\theta)$ scales as a function of the problem dimensions

d and m is not obvious. Also note that (3.2) has the following interpretation: Assume that (3.2) has a unique solution $x^\star$. Then any uniformly good algorithm must select action a at least $x_a^\star \log(T) + o(\log(T))$ times over the T first rounds. From [39], we know that there exists an algorithm which is asymptotically optimal, namely its regret matches the lower bound of Theorem 3.1. However this algorithm suffers from two problems: It is computationally infeasible for large problems since it involves solving problem (3.2) T times. Furthermore, the algorithm has no finite-time performance guarantees, and numerical experiments suggest that its finite-time performance on typical problems is rather poor.

Remark 3.1. *Theorem 3.1 can be generalized in a straightforward manner for when rewards belong to a one-parameter exponential family of distributions (e.g., Gaussian, Exponential, Gamma, etc.) by replacing* `kl` *by the appropriate divergence measure.*

Specific Cases

The lower bound presented in Theorem 3.1 is unfortunately implicit. Here we consider two specific cases, where we can provide explicit expressions for $c(\theta)$.

Matroids. Consider a weighted matroid $M = (E, \mathcal{I}, \theta)$, where E and θ respectively denote the ground set and weight function, namely each $i \in E$ has weight θ_i. Moreover, $\mathcal{I} \subset 2^E$ is the set of independent sets (for background materials on matroids, we refer to Appendix 3.K). Here each arm corresponds to a *basis* of matroid M, i.e., an inclusion-wise maximal element of $\mathcal{I}$. Equivalently, the set of arms $\mathcal{A}$ corresponds to the set of bases of matroid M.

To present the lower bound for the combinatorial MAB problem defined by matroid M, we introduce mapping $\sigma_M : E \setminus a^\star \to a^\star$ with

$$\sigma_M(i) = \operatorname*{arg\,min}_{j \in \mathcal{K}_i} \theta_j, \quad \forall i \in E \setminus a^\star,$$

where $\mathcal{K}_i := \{\ell \in a^\star : (a^\star \setminus \ell) \cup \{i\} \in \mathcal{A}\}$. Figure 3.1 shows an example of $\mathcal{K}_i$ for the case of graphic matroids.

By Proposition 3.1 (see Appendix 3.K), we have that $\mathcal{K}_i \neq \emptyset$ for any $i \notin a^\star$. Moreover, for any $i \notin a^\star$, if $\ell \in \mathcal{K}_i$, then $\theta_\ell > \theta_i$. We show this claim by contradiction: Assume this does not hold, namely $\theta_\ell < \theta_i$ since θ comprises distinct elements. Consider $a' = (a^\star \setminus \ell) \cup \{i\}$. Then, by Proposition 3.1, $a' \in \mathcal{A}$. Moreover,

$$\mu_{a'}(\theta) - \mu^\star(\theta) = \sum_{k \in a'} \theta_k - \sum_{k \in a^\star} \theta_k = \theta_i - \theta_\ell > 0,$$

which contradicts the optimality of $a^\star$. Hence, $\theta_\ell > \theta_i$ for any $\ell \in \mathcal{K}_i$.

The next theorem provides a regret lower bound for the policies in Π_{s}, which may be viewed as the specialization of Theorem 3.1 for the case of matroids.

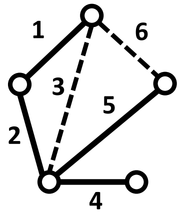


Figure 3.1: An example for the set $\mathcal{K}_i$ in the case of graphic matroids: Edges shown with solid line correspond to optimal actions. Two sub-optimal actions are shown in dashed line, where $\mathcal{K}_3 = \{1, 2\}$ and $\mathcal{K}_6 = \{1, 2, 5\}$.

Theorem 3.2. *For any matroid M and for any uniformly good algorithm $\pi \in \Pi_{\rm s}$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq \sum_{i\in E\setminus a^\star} \frac{\theta_{\sigma_M(i)} - \theta_i}{\texttt{kl}(\theta_i, \theta_{\sigma_M(i)})}.$$

Remark 3.2. *When the underlying matroid is the uniform matroid $U_{m,d}$, the problem reduces to MAB with multiple plays as studied in [36, 56]. Assume that basic actions are enumerated such that $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_m > \cdots \geq \theta_d$. Then $a^\star = \{1, 2, \dots, m\}$ and $\sigma_M(i) = m$ for all $i \notin a^\star$. Hence, the regret lower bound of Theorem 3.2 reduces to the lower bound of Anantharam et al. [36].*

For the case of semi-bandit feedback, a specific lower bound example for the case of a partition matroid is presented in Kveton et al. [57] to support the claim that regret scaling of $\Omega(\frac{d-m}{\Delta_{\min}} \log(T))$ is tight. In contrast to their lower bound, the one in Theorem 3.2 is problem-dependent and tight, i.e., it holds for any matroid M, and cannot be improved.

Matchings. As a second example, we consider a specific case of the matching problem, where $\mathcal{A}$ is the set of perfect matchings in the complete bipartite graph $\mathcal{K}_{m,m}$. We choose parameter θ as follows: Let $0 < \beta < \alpha < 1$. Let $\theta \in \Theta$ be defined such that $\theta_i = \alpha$ if $i \in a^\star$ and $\theta_i = \beta$ otherwise. The following corollary to Theorem 3.1 provides the regret lower bound for the aforementioned matching problem:

Corollary 3.1. *For all integer $m \geq 2$ and all $0 < \beta < \alpha < 1$, we have:*

$$c_{\rm s}(\theta) \geq \frac{m(m-1)(\alpha - \beta)}{2\texttt{kl}(\beta, \alpha)} .$$

Let us remark that for the considered instance of matching problem, $\Delta_{\min} = 2(\alpha - \beta)$. Hence, the lower bound presented in Corollary 3.1 implies, in view of inequality $\texttt{kl}(x, y) \leq \frac{(x-y)^2}{y(1-y)}$, the existence of a class of problems whose regret is at least $\Omega(\frac{d}{\Delta_{\min}} \log(T))$.

A Simplified Lower Bound

We now return back to the generic case and study how the coefficient $c_{\mathrm{s}}(\theta)$ in our proposed regret lower bound scales as a function of the problem dimensions d and m. To this aim, we present a simplified regret lower bound.

Definition 3.1. *Given θ, we say that a set $\mathcal{H} \subset \mathcal{A} \setminus a^\star$ has property $P(\theta)$ iff, for all $(a, a') \in \mathcal{H}^2$ with $a \neq a'$, we have $(a \setminus a^\star) \cap (a' \setminus a^\star) = \emptyset$.*

Theorem 3.3. *Let $\mathcal{H}$ be a maximal (inclusion-wise) subset of $\mathcal{A}$ satisfying the property $P(\theta)$. Define $\beta(\theta) = \min_{a \neq a^\star} \frac{\Delta_a}{|a \setminus a^\star|}$. Then:*

$$c_{\mathrm{s}}(\theta) \geq \sum_{a \in \mathcal{H}} \frac{\beta(\theta)}{\max_{i \in a \setminus a^\star} \mathtt{kl}\left(\theta_i, \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \theta_j\right)}.$$

Corollary 3.2. *Let $\theta \in [\frac{\alpha}{2}, \alpha]^d$ for some constant $0 < \alpha < \frac{1}{2}$ and $\mathcal{A}$ be such that each arm $a \in \mathcal{A}, a \neq a^\star$ has at most k sub-optimal basic actions. Then: $c_{\mathrm{s}}(\theta) = \Omega(|\mathcal{H}|/k)$.*

Theorem 3.3 provides explicit regret lower bound and Corollary 3.2 states that $c_{\mathrm{s}}(\theta)$ has to scale at least with the size of $\mathcal{H}$. As will be discussed next, for most combinatorial structures of interest, $|\mathcal{H}|$ is proportional to $d - m$, which implies that in these cases one cannot obtain a regret smaller than $\mathcal{O}((d-m)\Delta_{\min}^{-1} \log(T))$. This result is intuitive since $d - m$ is the number of parameters not observed when selecting the optimal arm. The algorithm proposed below has a regret of $\mathcal{O}(d\sqrt{m}\Delta_{\min}^{-1} \log(T))$, which is acceptable since typically, $\sqrt{m}$ is much smaller than d.

Next we examine Theorem 3.3 for some concrete classes of $\mathcal{A}$.

Matchings. In the first example, we assume that $\mathcal{A}$ is the set of perfect matchings in the complete bipartite graph $\mathcal{K}_{m,m}$, with $|\mathcal{A}| = m!$ and $d = m^2$. A maximal subset $\mathcal{H}$ of $\mathcal{A}$ satisfying property $P(\theta)$ can be constructed by adding all matchings that differ from the optimal matching by only two edges; see Figure 3.2 for illustration in the case of $m = 4$. Here $|\mathcal{H}| = \binom{m}{2}$ and thus, $|\mathcal{H}|$ scales as $d - m$.

Spanning trees (matroids revisited). Consider the problem of finding the minimum spanning tree in a complete graph $\mathcal{K}_N$.[6] This corresponds to letting $\mathcal{A}$ be the set of all spanning trees in $\mathcal{K}_N$, where $|\mathcal{A}| = N^{N-2}$ (Cayley's formula). In this case, we have $d = \binom{N}{2} = \frac{N(N-1)}{2}$, which is the number of edges of $\mathcal{K}_N$, and $m = N - 1$. A maximal subset $\mathcal{H}$ of $\mathcal{A}$ satisfying property $P(\theta)$ can be constructed by composing all spanning trees that differ from the optimal tree by one edge only; see Figure 3.3. In this case, $\mathcal{H}$ has $d - m = \frac{(N-1)(N-2)}{2}$ elements.

[6] Let us remark that spanning trees in a given graph are bases of the corresponding graphic matroid, for which Theorem 3.2 already provides a tight lower bound on the regret. Nonetheless, we present this case here for the sake of illustration.

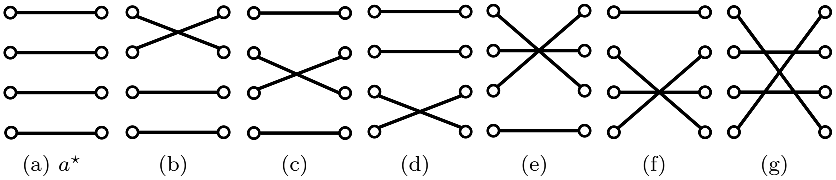


Figure 3.2: Matchings in $\mathcal{K}_{4,4}$: (a) The optimal matching $a^\star$, (b)-(g) elements of $\mathcal{H}$.

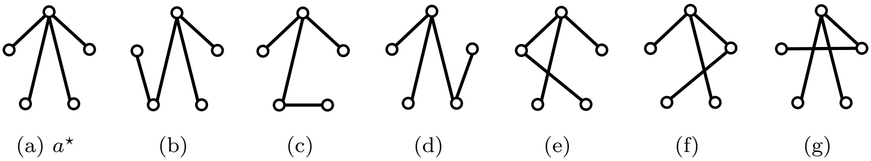
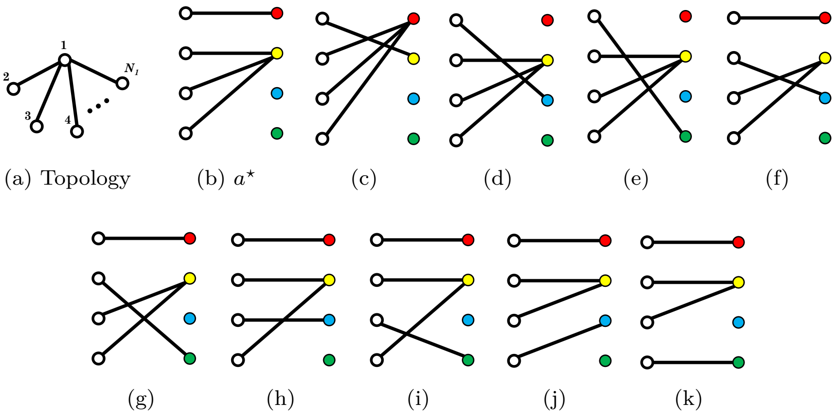


Figure 3.3: Spanning trees in $\mathcal{K}_5$: (a) The optimal spanning tree $a^\star$, (b)-(g) elements of $\mathcal{H}$.

A coloring problem. Consider the problem of coloring a star graph with N_1 nodes with $N_2 \geq N_1$ available colors. The task is to color nodes such that every two adjacent nodes have different colors. Let w_{ij} denote the weight of assignment of color $j \in [N_2]$ to node $i \in [N_1]$. The goal is to find a coloring scheme maximizing the sum of the weights. There are $d = N_1 N_2$ basic actions and each arm (coloring) has at most $m = N_1$ basic actions. For simplicity we assume $N_1 = N_2 = N$ (so $d = N^2$ and $m = N$). An arm or coloring in this case can be represented by a bipartite graph (see, e.g., Figure 3.4), whose left-hand (resp. right-hand) side vertices correspond to nodes of the star (resp. colors). Figure 3.4 displays the elements of $\mathcal{H}$ for the case of $N = 4$. One can easily verify that $\mathcal{H}$ has $(N-1)^2$ elements, and so $|\mathcal{H}| = \Omega(N^2) = \Omega(d)$.

Below we consider a particular instance of routing problem in which the underlying topology is a grid. This instance demonstrates that Theorem 3.3 may prove inapplicable for routing problems.

Routing in a grid. Consider routing in an K-by-K directed grid, whose topology is shown in Figure 3.5(a), where the source (resp. destination) node is shown in red (resp. blue). Here $\mathcal{A}$ is the set of all $\binom{2K-2}{K-1}$ paths with $m = 2(K-1)$ edges. We further have $d = 2K(K-1)$. In this example, elements of any maximal set $\mathcal{H}$ satisfying $P(\theta)$ do not cover all sub-optimal links. For instance, for the grid shown in Figure 3.5(a), there are 6 links that do not appear in any arm in $\mathcal{H}$ shown



(a) Topology (b) $a^\star$ (c) (d) (e) (f)

(g) (h) (i) (j) (k)

Figure 3.4: Coloring star graph: (a) Star topology, (b) optimal coloring $a^\star$ with minimal number of colors, (c)-(k) elements of $\mathcal{H}$.

here. Moreover, one may easily prove that in this case, $|\mathcal{H}|$ scales as K rather than $K^2 = d$.

3.3.2 Bandit Feedback

Now we consider the case of bandit feedback. Consider $a \in \mathcal{A}$ and introduce for all $k = 0, 1, \dots, m$:

$$\psi_\theta^a(k) = \sum_{X \subseteq a, |X| = k} \prod_{i \in X} \theta_i \prod_{i \in a \setminus X} (1 - \theta_i). \tag{3.4}$$

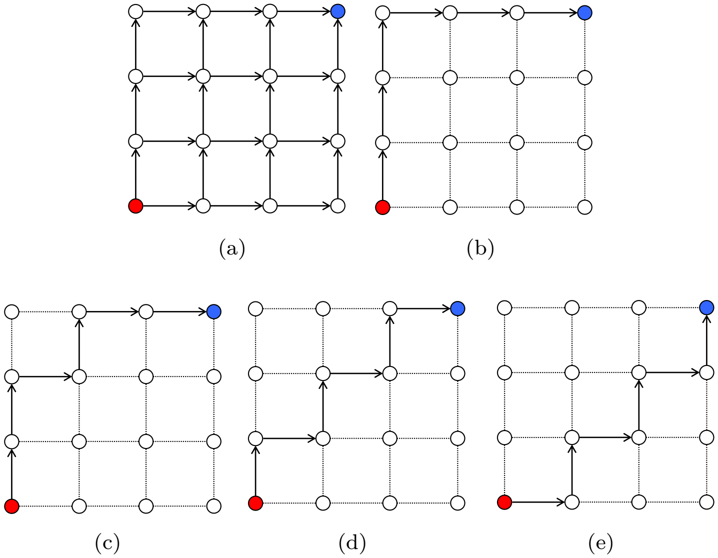
For two sets of parameters $\theta, \lambda \in \Theta$, we define the KL information number under arm a as:

$$I^a(\theta, \lambda) = \sum_{k=0}^{m} \psi_\theta^a(k) \log \frac{\psi_\theta^a(k)}{\psi_\lambda^a(k)}. \tag{3.5}$$

Now we define the set of bad parameters for a given θ, i.e., parameters for which arm $a^\star(\theta)$ is sub-optimal, yet the distribution of the reward of the optimal arm $a^\star(\theta)$ is the same under θ or λ:

$$B_{\mathrm{b}}(\theta) = \Big\{ \lambda \in \Theta : \sum_{i \in a^\star} \lambda_i = \sum_{i \in a^\star} \theta_i, \ \mu^\star(\lambda) > \mu^\star(\theta) \Big\}.$$

The slight difference between the definitions of $B_{\mathrm{b}}(\theta)$ and $B_{\mathrm{s}}(\theta)$ comes from the difference of feedback (bandit vs. semi-bandit). It is also noted that the set of bad



(a) (b)

(c) (d) (e)

Figure 3.5: Routing in a grid: (a) Grid topology with source (red) and destination (blue) nodes, (b) optimal path $a^\star$, (c)-(e) elements of $\mathcal{H}$.

parameters in the case of bandit feedback *contains* that of semi-bandit feedback, i.e., $B_{\rm s}(\theta) \subset B_{\rm b}(\theta)$.

In the following theorem, we derive an asymptotic regret lower bound. This bound is different than that derived in Theorem 3.1, due to the different nature of the feedback considered. Comparing the two bounds may indicate the price to pay by restricting the set of policies to those based on bandit feedback only.

Theorem 3.4. *For all $\theta \in \Theta$, for any uniformly good policy $\pi \in \Pi_{\rm b}$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \ge c_{\rm b}(\theta), \tag{3.6}$$

where $c_{\rm b}(\theta)$ is the optimal value of the following optimization problem:

$$\inf_{x\ge 0} \quad \sum_{a\in\mathcal{A}} x_a \Delta_a \tag{3.7}$$

$$\text{subject to:} \quad \sum_{a\in\mathcal{A}} x_a I^a(\theta,\lambda) \ge 1, \quad \forall \lambda \in B_{\rm b}(\theta). \tag{3.8}$$

The variables $x_a^\star, a \in \mathcal{A}$ solving (3.7) have the same interpretation as that given previously in the case of semi-bandit feedback. Similarly to the lower bound of

Theorem 3.1, the above lower bound is implicit. In this case, it is however much more complicated to see how $c_{\mathrm{b}}(\theta)$ scales with m and d, and we let if for future work.

Remark 3.3. *Of course, we know that $c_{\mathrm{b}}(\theta) \geq c_{\mathrm{s}}(\theta)$, since the lower bounds we derive are tight and getting semi-bandit feedback can be exploited to design smarter arm selection policies than those we can devise using bandit feedback (i.e., $\Pi_{\mathrm{b}} \subset \Pi_{\mathrm{s}}$).*

We conclude this section by providing an specialization of Theorem 3.4 to the case of matroids:

Theorem 3.5. *For any matroid M and for any uniformly good algorithm $\pi \in \Pi_{\mathrm{b}}$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq \sum_{i\in E\setminus a^\star} \frac{\theta_{\sigma_M(i)} - \theta_i}{\max_{a:i\in a} I^a(\theta, \zeta_M^i)},$$

where ζ_M^i is a vector of parameters defined as $\zeta_{M,j}^i = \theta_j$ if $j \neq i$, and $\zeta_{M,i}^i = \theta_{\sigma_M(i)}$.

The proof of the above theorem involves decomposing the set of bad parameters as $B_{\mathrm{b}}(\theta) = \cup_{i\notin a^\star} B_{\mathrm{b}}^i(\theta)$, where $B_{\mathrm{b}}^i(\theta)$ is the set of parameters $\lambda \in B_{\mathrm{b}}(\theta)$ such that $\lambda_i > \theta_{\sigma_M(i)}$; see Appendix 3.G for details.

3.4 Algorithms

In this section, we present ESCB, an algorithm for the case of semi-bandit feedback that relies on arm indexes. First we introduce two new index functions that will be used by ESCB.

In general, an index function for a given arm a is defined as a function of properties of arm a but also depends on the round n. These properties could well include the empirical estimate of mean reward, empirical variance of reward, and number of pulls of a. Moreover, it should be defined so that it exceeds $\mu_a(\theta) = a^\top\theta$ with high probability.

To present the index functions, we introduce the following notations. Under a given algorithm, at time n, we define $t_i(n) = \sum_{s=1}^n a_i(s)$ the number of times basic action i has been sampled. The empirical mean reward of action i is then defined as $\hat\theta_i(n) = (1/t_i(n))\sum_{s=1}^n X_i(s)a_i(s)$ if $t_i(n) > 0$ and $\hat\theta_i(n) = 0$, otherwise. Finally, we define the corresponding vectors $t(n) = (t_i(n))_{i\in E}$ and $\hat\theta(n) = (\hat\theta_i(n))_{i\in E}$.

3.4.1 Indexes

Our first index function is an extension of KL-UCB index to the case of combinatorial arms. Let $\lambda \in \Theta$, $t \in \mathbb{N}^d$, and $n \in \mathbb{N}$. Our first index for arm a, denoted by $b_a(n, \lambda, t)$, is defined as the optimal value of the following optimization problem:

$$\max_{q\in\Theta} \; a^\top q \tag{3.9}$$

$$\text{subject to: } \sum_{i\in E} a_i t_i \mathtt{kl}(\lambda_i, q_i) \le f(n) \ ,$$

with $f(n) = \log(n) + 4m\log(\log(n))$. As we shall see later, b_a may be computed efficiently using a line search procedure similarly to that used to compute KL-UCB index.

Our second index $c_a(n, \lambda, t)$ is a generalization of the UCB1 and UCB-Tuned indexes:

$$c_a(n,\lambda,t) = a^\top \lambda + \sqrt{\frac{f(n)}{2} \sum_{i\in E} \frac{a_i}{t_i}}$$

Note that, in the classical bandit problems with independent arms, i.e., when $m = 1$, b_a reduces to the KL-UCB index (which yields an asymptotically optimal algorithm) and c_a reduces to the UCB-Tuned index [20]. The next theorem provides generic properties of our indexes. An important consequence of these properties is that the expected number of times where $b_{a^\star}(n, \hat\theta(n), t(n))$ or $c_{a^\star}(n, \hat\theta(n), t(n))$ underestimates $\mu^\star$ is *finite*, as stated in the corollary below.

Theorem 3.6. *(i) For all $n \ge 1$, $a \in \mathcal{A}$ and $\lambda \in [0,1]^d$, we have $b_a(n,\lambda,t) \le c_a(n,\lambda,t)$. (ii) There exists $C_m > 0$ depending on m only such that, for all $a \in \mathcal{A}$ and $n \ge 2$:*

$$\mathbb{P}(b_a(n,\hat\theta(n),t(n)) \le a^\top\theta) \le C_m n^{-1}(\log(n))^{-2}.$$

Corollary 3.3. *We have:*

$$\sum_{n\ge1} \mathbb{P}(b_{a^\star}(n,\hat\theta(n),t(n)) \le \mu^\star) \le 1 + C_m \sum_{n\ge 2} n^{-1}(\log(n))^{-2} < \infty.$$

Statement (i) in the above theorem is obtained combining Pinsker's and Cauchy-Schwarz inequalities. The proof of statement (ii) is based on a concentration inequality on sums of empirical KL-divergences proven in [68] (see Appendix B). It enables to control the fluctuations of multivariate empirical distributions for exponential families. It should also be observed that indexes b_a and c_a can be extended in a straightforward manner to the case of continuous linear bandit problems, where the set of arms is the unit sphere and one wants to maximize the dot product between the arm and an unknown vector. Index function b_a can also be extended to the case where reward distributions are not Bernoulli but lie within an exponential family (e.g., Gaussian, Exponential, Gamma, etc.), replacing kl by a suitably chosen divergence measure.

Remark 3.4. *A close look at c_a reveals that the indexes proposed in [61], [62], and [60] are too conservative to be optimal in our setting: There the "confidence bonus" $\sum_{i\in E} \frac{a_i}{t_i}$ was replaced by (at least) $m\sum_{i\in E}\frac{a_i}{t_i}$. We remark that [61], [62] assumed that the various basic actions are arbitrarily correlated, while we assume independence among basic actions.*

3.4.2 Index Computation

While the index c_a is explicit, b_a is defined as the optimal value of an optimization problem. We show that it may be computed by a simple line search.

Consider arm a. Fix $n \in \mathbb{N}$, $\lambda \in \Theta$, and $t \in \mathbb{N}^d$. Define $J_a(\lambda) = \{i \in a : \lambda \neq 1\}$, and for $\gamma > 0$, define:

$$F(\gamma, \lambda, n, t) = \sum_{i \in J_a(\lambda)} t_i \mathtt{kl}(\lambda_i, g(\gamma, \lambda_i, t_i)) \ , \quad \text{with}$$

$$g(\gamma, \lambda_i, t_i) = \frac{1}{2}\left(1 - \gamma t_i + \sqrt{(1-\gamma t_i)^2 + 4\gamma\lambda_i t_i}\right).$$

Theorem 3.7. *If $J_a(\lambda) = \emptyset$, $b_a(n, \lambda, t) = \|a\|_1$. Otherwise:*
(i) $\gamma \mapsto F(\gamma, \lambda, n, t)$ is strictly increasing, and $F(\mathbb{R}_+, \lambda, n, t) = \mathbb{R}_+$.
(ii) Define $\gamma^\star$ as the unique solution to $F(\gamma, \lambda, n, t) = f(n)$. Then

$$b_a(n, \lambda, t) = \|a\|_1 - |J_a(\lambda)| + \sum_{i \in J_a(\lambda)} g(\gamma^\star, \lambda_i, t_i).$$

Theorem 3.7 shows that b_a can be computed using a line search procedure such as bisection, as this computation amounts to solving the non-linear equation $F(\gamma, \lambda, n, t) = f(n)$, where F is a strictly increasing function. The proof of Theorem 3.7 follows from KKT conditions and the convexity of the KL divergence (see Appendix A for a summary of the properties of the KL divergence).

3.4.3 The ESCB Algorithm

Having introduced the index function, we are now in a position to present the ESCB algorithm. Following the principle of "optimism in face of uncertainty" as in UCB1 and KL-UCB, ESCB consists in selecting in each round the arm with the largest index. More precisely, in round n it selects the arm $a(n) \in \mathrm{argmax}_{a \in \mathcal{A}}\, b_a(n)$ or $a(n) \in \mathrm{argmax}_{a \in \mathcal{A}}\, c_a(n)$, where we define $b_a(n) := b_a(n, \hat{\theta}(n), t(n))$ and $c_a(n) := c_a(n, \hat{\theta}(n), t(n))$.

The pseudo-code of ESCB is presented in Algorithm 3.1. We consider two variants of the algorithm based on the choice of the index ξ_a: ESCB-1 when $\xi_a = b_a$ and ESCB-2 if $\xi_a = c_a$.

In practice, ESCB-1 outperforms ESCB-2, as verified by numerical results in Section 3.5. Introducing ESCB-2 is however instrumental in the regret analysis of ESCB-1 (in view of Theorem 3.6 (i)). The following theorem provides a finite-time analysis of our ESCB algorithms.

Theorem 3.8. *The regret under algorithm $\pi \in \{$ESCB-1, ESCB-2$\}$ satisfies for any time horizon $T > 1$:*

$$\mathfrak{R}_{\pi,T} \leq \frac{16d\sqrt{m}}{\Delta_{\min}}(\log(T) + 4m\log(\log(T))) + \frac{4dm^3}{\Delta_{\min}^2} + C'_m,$$

Algorithm 3.1 ESCB

```
for $n \geq 1$ do
    Select arm $a(n) \in \arg\max_{a\in\mathcal{A}} \xi_a(n)$.
    Observe the rewards, and update $t_i(n)$ and $\hat{\theta}_i(n), \forall i \in a(n)$.
end for
```

where $C'_m \geq 0$ does not depend on θ, d, and T. As a consequence $\mathfrak{R}_{\pi,T} = \mathcal{O}(d\sqrt{m}\Delta_{\min}^{-1}\log(T))$ when $T \to \infty$.

ESCB with time horizon T has a complexity of $\mathcal{O}(|\mathcal{A}|T)$ as neither b_a nor c_a can be written as $a^\top y$ for some vector $y \in \mathbb{R}^d$. Assuming that the offline (static) combinatorial problem is solvable in $\mathcal{O}(V(\mathcal{A}))$ time, the complexity of CUCB in [61] and [62] after T rounds is $\mathcal{O}(V(\mathcal{A})T)$. Thus, if the offline problem is efficiently implementable, i.e., $V(\mathcal{A}) = \mathcal{O}(\text{poly}(d))$, CUCB is efficient, whereas ESCB is not since $\mathcal{A}$ may generically have exponentially (in d) many elements. Next, we provide an extension to ESCB, which we may call Epoch-ESCB, that may attain almost the same regret as ESCB while enjoying much lower computational complexity.

3.4.4 Epoch-ESCB: An Algorithm with Lower Computational Complexity

Epoch-ESCB algorithm works in epochs of varying lengths. Epoch k comprises rounds $\{N_k, \dots, N_{k+1} - 1\}$, where N_{k+1} (and thus the length of the k-th epoch) is determined at time $n = N_k$, i.e., at the start of the k-th epoch. The Epoch-ESCB algorithm simply consists in playing the arm with the maximal index at the beginning of every epoch, and playing the current leader (i.e., the arm with the highest empirical average reward) in the rest of rounds. If the leader is the arm with the maximal index, the length of epoch k will be set twice as long as the previous epoch $k-1$, i.e., $N_{k+1} = N_k + 2(N_k - N_{k-1})$. Otherwise, it will be set to 1. In contrast to ESCB, Epoch-ESCB computes the maximal index infrequently, and more precisely (almost) at an exponentially decreasing rate. Thus, one might expect that after T rounds, the maximal index will be computed $\mathcal{O}(\log(T))$ times. The pseudo-code of Epoch-ESCB is presented in Algorithm 3.2.

We assess the performance of Epoch-ESCB through numerical experiments in Section 3.5, and leave the analysis of its regret as a future work. These experiments corroborate our conjecture that the complexity of Epoch-ESCB after T rounds will be $\mathcal{O}(V(\mathcal{A})T + \log(T)|\mathcal{A}|)$. Compared to CUCB, the complexity is penalized by $|\mathcal{A}|\log(T)$, which may become dominated by the term $V(\mathcal{A})T$ as T grows large.

3.5 Numerical Experiments

In this section, we compare the performance of ESCB against existing algorithms through numerical experiments for some classes of $\mathcal{A}$. When implementing ESCB,

Algorithm 3.2 EPOCH-ESCB

```
Initialization: Set $k = 1$ and $N_0 = N_1 = 1$.
for $n \geq 1$ do
  Compute $L(n) \in \arg\max_{a \in \mathcal{A}} a^\top \hat{\theta}(n)$.
  if $n = N_k$ then
    Select arm $a(n) \in \arg\max_{a \in \mathcal{A}} \xi_a(n)$.
    if $a(n) = L(n)$ then
      Set $N_{k+1} = N_k + 2(N_k - N_{k-1})$.
    else
      Set $N_{k+1} = N_k + 1$.
    end if
    Increment $k$.
  else
    Select arm $a(n) = L(n)$.
  end if
  Observe the rewards, and update $t_i(n)$ and $\hat{\theta}_i(n), \forall i \in a(n)$.
end for
```

we replace $f(n)$ by $\log(n)$, ignoring the term proportional to $\log(\log(n))$, as is done when implementing KL-UCB in practice.

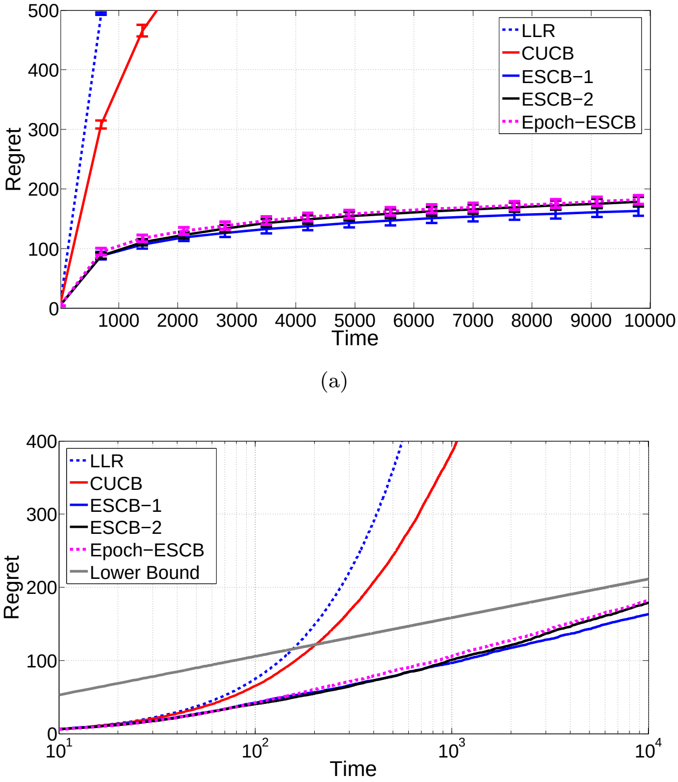
Experiment 1: Matching

In our first experiment, we consider the matching problem in complete bipartite graph $\mathcal{K}_{5,5}$, for which $d = 5^2 = 25$ and $m = 5$. Furthermore, we consider parameter θ defined in Corollary 3.1.

Figure 3.6(a)-(b) depicts the regret of various algorithms for the case of $\alpha = 0.7$ and $\beta = 0.5$. The curves in Figure 3.6(a) are shown with a 95% confidence intervals. We observe that ESCB-1 has the smallest regret. Moreover, ESCB-2 significantly outperforms CUCB and LLR, and its regret is close to that of ESCB-1. Moreover, we observe that the regret of EPOCH-ESCB is quite close to that of ESCB-2.

Figures 3.7(a)-(b) presents the regret of various algorithms for the case of $\alpha = 0.95$ and $\beta = 0.3$. The difference compared to the former case is that ESCB-1 significantly outperforms ESCB-2. The reason is that in the former case, mean rewards of most basic actions were close to $\frac{1}{2}$, for which the performance of UCB-type algorithms are closer to their KL-based counterparts. On the other hand, when mean rewards are not close to $\frac{1}{2}$, there exists a significant performance gap between ESCB-1 and ESCB-2. Comparing the results with the 'lower bound' curve, we highlight that ESCB-1 gives close-to-optimal performance in both cases. Furthermore, similarly to the previous case, EPOCH-ESCB attains a regret whose curve is almost indistinguishable from that of ESCB-2.

The number of epochs in EPOCH-ESCB vs. time for the two examples is displayed in Figure 3.8(a)-(b), where the curves are shown with 95% confidence intervals. We observe that in both cases, the number of epochs grows at a rate proportional to



(a)

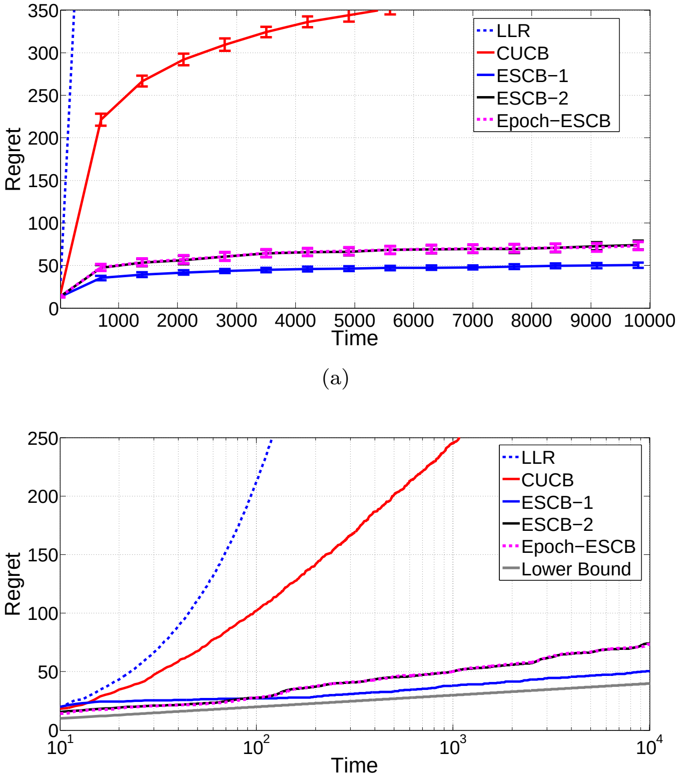
(b)

Figure 3.6: Regret of various algorithms for matchings with $\alpha = 0.7$ and $\beta = 0.5$.

$\log(n)/n$ at round n. Since the number of times Epoch-ESCB computes the index c_a is equal to the number of epochs, these curves suggest that the computational complexity of index computations in Epoch-ESCB after n rounds scales as $|\mathcal{A}|\log(n)$.

Experiment 2: Spanning Trees

In the second experiment, we consider spanning trees problem for the case of $N = 5$. In this case, we have $d = \binom{5}{2} = 10$, $m = 4$, and $|\mathcal{A}| = 5^3 = 125$. We generate parameter θ uniformly at random from $[0,1]^{10}$. Figure 3.9 portrays the regret of various algorithms with 95% confidence intervals, for a case with $\Delta_{\min} = 0.54$. The results show that our algorithms outperform CUCB and LLR.



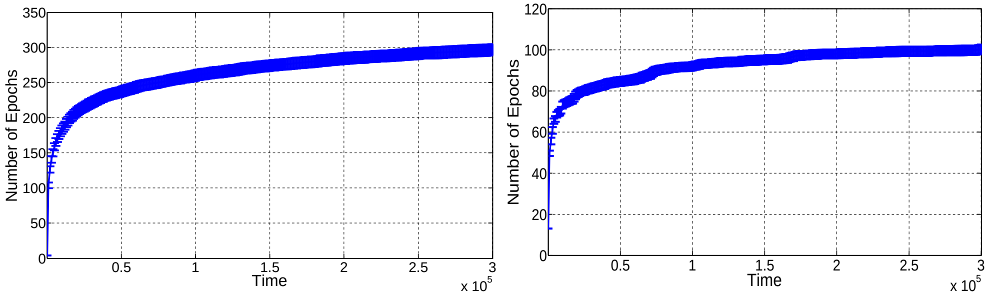
(a)

(b)

Figure 3.7: Regret of various algorithms for matchings with $\alpha = 0.95$ and $\beta = 0.3$.

3.6 Summary

In this chapter we investigated stochastic combinatorial MABs with Bernoulli rewards. We derived asymptotic regret lower bounds for both bandit and semi-bandit feedback. The proposed lower bounds are not explicit, and hence we further examined its scaling in terms of the dimension of the decision space for the case of semi-bandit feedback. We then proposed the ESCB algorithm and provided a finite-time analysis of its regret. ESCB achieves lower regret compared to state-of-the-art algorithms and outperforms these algorithms in practice. We also proposed EPOCH-ESCB that has lower computational complexity than ESCB. The regret analysis of EPOCH-ESCB is much more complicated than that of ESCB, and hence is let for future work.



(a) $\alpha = 0.7$ and $\beta = 0.5$ (b) $\alpha = 0.95$, $\beta = 0.3$

Figure 3.8: Number of epochs in Epoch-ESCB vs. time for Experiment 1 and 2 (%95 confidence interval).

3.A Proof of Theorem 3.1

To derive regret lower bounds, we apply the techniques used by Graves and Lai [39] to investigate efficient adaptive decision rules in controlled Markov chains; we refer to Chapter 2 for a brief summary of their general framework.

To this end, we construct a controlled Markov chain as follows. The state-space is $\mathcal{S} = \{0,1\}^d$. The set of controls corresponds to the set of arms $\mathcal{A}$, and the set of control laws is also $\mathcal{A}$. These laws are constant in the sense that the control applied by control law $a \in \mathcal{A}$ does not depend on the state of the Markov chain, and corresponds to selecting arm a. The parameter θ takes values in $[0,1]^d$ and the transition probabilities are given as follows: For all $x, y \in \mathcal{S}$,

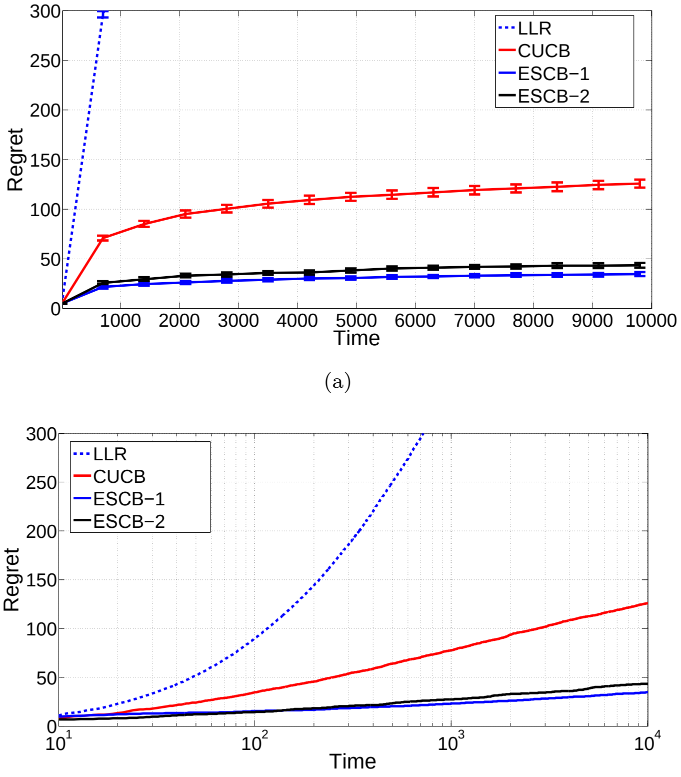
$$p(x,y;a,\theta) = p(y;a,\theta) = \prod_{i \in E} p_i(y_i;a,\theta),$$

where for all $i \in E$, if $a_i = 0$, $p_i(0;a,\theta) = 1$, and if $a_i = 1$, $p_i(y_i;a,\theta) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$. Finally, the reward $r(y,a)$ is defined by $r(y,a) = a^\top y$. Note that the state-space of the Markov chain is here finite, and so, we do not need to impose any cost associated with switching control laws (see the discussion on page 718 in [39]). Note that the KL divergence under arm a is

$$I^a(\theta,\lambda) = \sum_{i \in E} a_i \mathtt{kl}(\theta_i, \lambda_i).$$

From [39, Theorem 1], we conclude that for any uniformly good algorithm $\pi \in \Pi_{\mathrm{s}}$,

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq c_{\mathrm{s}}(\theta),$$



(a)

(b)

Figure 3.9: Regret of various algorithms for spanning trees with $N = 5$ and $\Delta_{\min} = 0.54$.

where $c_{\rm s}(\theta)$ is the optimal value of the following optimization problem:

$$\inf_{x \geq 0} \sum_{a \neq a^\star} x_a \Delta_a, \tag{3.10}$$

$$\text{subject to:} \quad \inf_{\lambda \in B_{\rm s}(\theta)} \sum_{a \neq a^\star} x_a I^a(\theta, \lambda) \geq 1. \tag{3.11}$$

Substituting the expression of $I^a(\theta, \lambda)$ into (3.11) completes the proof. □

3.B Proof of Theorem 3.2

Let $M = (E, \mathcal{I}, \theta)$ be a weighted matroid. To ease notation, we use the abbreviations $B(\theta)$ and σ to respectively denote $B_{\rm s}(\theta)$ and σ_M. Applying Theorem 3.1 and

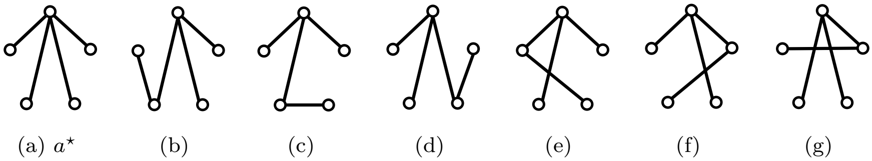


Figure 3.10: Spanning trees in $\mathcal{K}_5$: (a) The optimal spanning tree $a^\star$, (b)-(g) $a^{(i)}$.

using similar lines as in the proof of Theorem 3.3, we have

$$\begin{aligned}
&\inf_{x\ge 0} \sum_{a\neq a^\star} x_a\Delta_a,\\
&\text{subject to: } \inf_{\lambda\in B_a(\theta)} \sum_{i\in a\setminus a^\star} \mathtt{kl}(\theta_i,\lambda_i)\sum_{k\in\mathcal{A}} k_i x_k \ge 1\ , \quad \forall a\neq a^\star,
\end{aligned} \tag{3.12}$$

where $B_a(\theta)=\big\{\lambda\in\Theta:\lambda_i=\theta_i,\forall i\in a^\star,\ \mu^\star(\theta)<\mu^\star(\lambda)\big\}$.

Let $i\in E\setminus a^\star$ and consider $a^{(i)}:=(a^\star\setminus\sigma(i))\cup\{i\}$. Proposition 3.1 implies that $a^{(i)}\in\mathcal{A}$ (see Figure 3.10 that portrays an instance of $\{a^{(i)},i\in E\setminus a^\star\}$ for the case of graphic matroids). We may simplify the left-hand side of the constraint (3.12) corresponding to arm $a^{(i)}$ as follows:

$$\begin{aligned}
\inf_{\lambda\in B_{a^{(i)}}(\theta)} \sum_{j\in a^{(i)}\setminus a^\star} \mathtt{kl}(\theta_j,\lambda_j)\sum_k k_j x_k &= \inf_{\lambda\in B_{a^{(i)}}(\theta)} \mathtt{kl}(\theta_i,\lambda_i)\sum_k k_i x_k\\
&= \inf_{\lambda\in\Theta:\lambda_i>\theta_{\sigma(i)}} \mathtt{kl}(\theta_i,\lambda_i)\sum_k k_i x_k\\
&= \mathtt{kl}(\theta_i,\theta_{\sigma(i)})\sum_k k_i x_k.
\end{aligned}$$

Hence, the constraint (3.12) for $a=a^{(i)}$ may be equivalently written as

$$\sum_k k_i x_k \ge \frac{1}{\mathtt{kl}(\theta_i,\theta_{\sigma(i)})}.$$

Letting $\mathcal{A}^-=\mathcal{A}\setminus(\{a^\star\}\cup\{a^{(i)},i\in E\setminus a^\star\})$, it then follows that

$$\begin{aligned}
c_{\mathrm{s}}(\theta) \ = \ &\inf_{x\ge 0} \sum_{a\in\mathcal{A}} \Delta_a x_a\\
&\text{subject to: } \sum_{k\neq a^\star} k_i x_k \ge \frac{1}{\mathtt{kl}(\theta_i,\theta_{\sigma(i)})},\ \ \forall i\in E\setminus a^\star,\\
&\qquad \inf_{\lambda\in B_a(\theta)} \sum_{k\in\mathcal{A}} x_k \sum_{i\in E} k_i\mathtt{kl}(\theta_i,\lambda_i)\ge 1,\ \ \forall a\in\mathcal{A}^-.
\end{aligned} \tag{3.13}$$

Now we bound the objective function of problem (3.13) from below. Let $a \neq a^\star$ and further define a bijection $\tau_a : E \to E$ defined as follows: If $i \in a \setminus a^\star$, then $\tau_a(i) = j$ for some $j \in \mathcal{K}_i$. Otherwise, $\tau_a(i) = i$. We have:

$$\Delta_a = \sum_{i \in a} (\theta_{\tau_a(i)} - \theta_i) = \sum_{i \in E \setminus a^\star} a_i(\theta_{\tau_a(i)} - \theta_i) \geq \sum_{i \in E \setminus a^\star} a_i(\theta_{\sigma(i)} - \theta_i).$$

Hence, introducing $z_i = \sum_a a_i x_a$ for any $i \in E \setminus a^\star$, we obtain:

$$\sum_a x_a \Delta_a \geq \sum_a x_a \sum_{i \in E \setminus a^\star} a_i(\theta_{\sigma(i)} - \theta_i) = \sum_{i \in E \setminus a^\star} (\theta_{\sigma(i)} - \theta_i) z_i.$$

As a result,

$$\begin{aligned} c_{\mathrm{s}}(\theta) \geq \inf_{z \geq 0} & \sum_{i \in E \setminus a^\star} (\theta_{\sigma(i)} - \theta_i) z_i \\ \text{subject to: } & z_i \geq \frac{1}{\mathtt{kl}(\theta_i, \theta_{\sigma(i)})}, \quad \forall i \in E \setminus a^\star, \end{aligned}$$

which yields $c_{\mathrm{s}}(\theta) = \sum_{i \in E \setminus a^\star} \frac{\theta_{\sigma(i)} - \theta_i}{\mathtt{kl}(\theta_i, \theta_{\sigma(i)})}$ and thus concludes the proof. □

3.C Proof of Corollary 3.1

By Theorem 3.1, the regret of any uniformly good policy in Π_{s} is at least $\Omega(c_{\mathrm{s}}(\theta)\log(T))$ as T grows large, where

$$\begin{aligned} c_{\mathrm{s}}(\theta) = \inf_{x \geq 0} & \sum_{a \neq a^\star} x_a \Delta_a, \\ \text{subject to: } & \inf_{\lambda \in B_a(\theta)} \sum_{i \in a \setminus a^\star} \mathtt{kl}(\beta, \lambda_i) \sum_{a' \in \mathcal{A}} a'_i x_{a'} \geq 1 \ , \quad \forall a \neq a^\star, \end{aligned}$$

with $B_a(\theta) = \big\{\lambda \in \Theta : \lambda_i = \alpha, \forall i \in a^\star,\ \mu^\star(\theta) < \mu^\star(\lambda)\big\}$.

To derive an explicit lower bound on the regret for the considered parameter θ, in the sequel we simplify the objective and constraints of the above problem.

Let $a \neq a^\star$ and consider the constraint corresponding to arm $a \neq a^\star$. Let $\rho > 0$. Since $x \mapsto \mathtt{kl}(\beta, z)$ is continuous for $z > \beta$, we can choose $\xi > \alpha$ such that

$$|\mathtt{kl}(\beta, \xi) - \mathtt{kl}(\beta, \alpha)| \leq \rho \mathtt{kl}(\beta, \alpha) \ .$$

Now consider $\tilde{\lambda}^a \in \Theta$ such that $\tilde{\lambda}^a_i = \alpha$ if $i \in a^\star$, and $\tilde{\lambda}^a_i = \xi$ if $i \in a \setminus a^\star$, and $\tilde{\lambda}^a_i = \beta$ otherwise. As $\tilde{\lambda}^a \in B_a(\theta)$, we have

$$\inf_{\lambda \in B_a(\theta)} \sum_{i \in a \setminus a^\star} \mathtt{kl}(\beta, \lambda_i) \sum_{a'} a'_i x_{a'} \leq \sum_{i \in a \setminus a^\star} \mathtt{kl}(\beta, \tilde{\lambda}^a_i) \sum_{a'} a'_i x_{a'}$$

$$= \mathtt{kl}(\beta, \xi) \sum_{i \in a \setminus a^\star} \sum_{a'} a'_i x_{a'} \; .$$

Defining $\varepsilon = \frac{\rho}{1+\rho}$ and noting that $\Delta_a = |a \setminus a^\star|(\alpha - \beta)$ for any $a \neq a^\star$, we thus get that

$$\begin{aligned} c_{\mathrm{s}}(\theta) \geq \inf_{x \geq 0} \quad & (\alpha - \beta) \sum_{a \neq a^\star} |a \setminus a^\star| x_a \\ \text{subject to:} \quad & \sum_{i \in a \setminus a^\star} \sum_{a'} a'_i x_{a'} \geq \frac{1-\varepsilon}{\mathtt{kl}(\beta, \alpha)} \; , \quad \forall a \neq a^\star \end{aligned}$$

for every $\varepsilon \in (0,1)$. Now, applying Lemma 3.1, proven next, and letting $\varepsilon \to 0$ give the desired result

$$c_{\mathrm{s}}(\theta) \geq \frac{m(m-1)(\alpha - \beta)}{\mathtt{kl}(\beta, \alpha)} \; ,$$

and conclude the proof. □

Lemma 3.1. *Let $\mathcal{A}$ be the set of perfect matchings in $\mathcal{K}_{m,m}$ for a given $m \geq 2$, and consider parameter θ defined in Corollary 3.1. Define*

$$\begin{aligned} g^\star(\theta) = \min_{x \geq 0} \quad & \sum_{a \neq a^\star(\theta)} |a \setminus a^\star(\theta)| x_a \\ \text{subject to:} \quad & \sum_{i \in k \setminus a^\star(\theta)} \sum_{a \in \mathcal{A}} a_i x_a \geq 1, \;\; \forall k \neq a^\star(\theta). \end{aligned}$$

Then $g^\star(\theta) = m(m-1)$.

Proof. For any $2 \leq j \leq m$, introduce $\mathcal{D}_j \in \mathcal{A}$ as the set of arms that have j sub-optimal basic actions:

$$\mathcal{D}_j = \{a \in \mathcal{A} : |a \setminus a^\star| = j\},$$

and let D_j be its cardinality. Hence, $\mathcal{A} = \cup_{j=2}^m \mathcal{D}_j \cup \{a^\star\}$ and $\mathcal{D}_j \cup \mathcal{D}_{j'} = \emptyset$ for all $j \neq j'$. The symmetry in the problem implies the existence of at least one solution $x^\star$ satisfying the following property: For any j, for all $a, a' \in \mathcal{D}_j$, $x^\star_a = x^\star_{a'}$. Observe that the dependence of x_a on a is captured by $|a \setminus a^\star|$, and hence for all $a \in \mathcal{D}_j$, we introduce $y_j = x^\star_a$. We then get:

$$\sum_{a \neq a^\star} |a \setminus a^\star| x_a = \sum_{j=2}^m j \sum_{a \in \mathcal{D}_j} x^\star_a = \sum_{j=2}^m j D_j y_j. \tag{3.14}$$

Let $k \neq a^\star$ and consider $i \in k \setminus a^\star$. We have that:

$$\sum_a a_i x^\star_a = \sum_{j=2}^m \sum_{a \in \mathcal{D}_j} a_i x^\star_a = \sum_{j=2}^m y_j \sum_{a \in \mathcal{D}_j} a_i.$$

Note that $\sum_{i\notin a^\star}\sum_{a\in\mathcal{D}_j} a_i = jD_j$ since there are a total number of jD_j sub-optimal basic actions in $\mathcal{D}_j$. Then, symmetry in $\mathcal{A}$ implies:

$$\sum_{a\in\mathcal{D}_j} a_i = \frac{1}{m(m-1)}\sum_{i\notin a^\star}\sum_{a\in\mathcal{D}_j} a_i = \frac{jD_j}{m(m-1)},$$

so that the constraint corresponding to arm k becomes: $\sum_{j=2}^m jD_j y_j \ge m(m-1)$. Putting this together with (3.14) concludes the proof of the lemma. □

3.D Proof of Theorem 3.3

The proof proceeds in three steps. In the subsequent analysis, given the optimization problem P, we use $\mathrm{val}(\mathsf{P})$ to denote its optimal value. Moreover, for brevity we use the short-hand $B(\theta)$ to denote $B_{\mathrm{s}}(\theta)$.

Step 1. We begin with introducing an equivalent formulation for problem (3.10) by simplifying its constraints. In particular, we show that constraint (3.11) is equivalent to:

$$\inf_{\lambda\in B_a(\theta)}\sum_{i\in a\setminus a^\star}\mathtt{kl}(\theta_i,\lambda_i)\sum_{k\in\mathcal{A}}k_i x_k \ge 1,\ \ \forall a\neq a^\star,$$

where $B_a(\theta)=\big\{\lambda\in\Theta:\lambda_i=\theta_i,\ \forall i\in a^\star,\ \mu^\star(\theta)<\mu_a(\lambda)\big\}$. Fix $a\neq a^\star$. In view of the definition of $B_a(\theta)$, we can find $\lambda\in B_a(\theta)$ such that $\lambda_i=\theta_i,\forall i\in(E\setminus a)\cup a^\star$. Thus, for the right-hand side of the a-th constraint in (3.11), we get:

$$\begin{aligned}\inf_{\lambda\in B_a(\theta)}\sum_{k\neq a^\star}x_k I^k(\theta,\lambda) &= \inf_{\lambda\in B_a(\theta)}\sum_{i\in E}\mathtt{kl}(\theta_i,\lambda_i)\sum_{k\neq a^\star}k_i x_k\\ &= \inf_{\lambda\in B_a(\theta)}\sum_{i\in a\setminus a^\star}\mathtt{kl}(\theta_i,\lambda_i)\sum_k k_i x_k,\end{aligned}$$

and therefore problem (3.10) can be equivalently written as:

$$c_{\mathrm{s}}(\theta) = \inf_{x\ge 0}\sum_{a\neq a^\star}\Delta_a x_a, \tag{3.15}$$

$$\text{subject to:}\quad \inf_{\lambda\in B_a(\theta)}\sum_{i\in a\setminus a^\star}\mathtt{kl}(\theta_i,\lambda_i)\sum_k k_i x_k \ge 1,\ \ \forall a\neq a^\star. \tag{3.16}$$

Next, we formulate an LP whose value gives a lower bound for $c_{\mathrm{s}}(\theta)$. Consider $a\neq a^\star$. Let $\varepsilon>0$ and define $\tilde\lambda^a(\varepsilon)=(\tilde\lambda^a_i(\varepsilon))_{i\in E}$ with

$$\tilde\lambda^a_i(\varepsilon)=\begin{cases}\frac{1}{|a\setminus a^\star|}\sum_{j\in a^\star\setminus a}\theta_j+\varepsilon & \text{if } i\in a\setminus a^\star,\\ \theta_i & \text{otherwise.}\end{cases}$$

Clearly $\tilde{\lambda}^a(\varepsilon) \in B_a(\theta)$ and therefore:

$$\inf_{\lambda \in B_a(\theta)} \sum_{i \in a \setminus a^\star} \mathtt{kl}(\theta_i, \lambda_i) \sum_k k_i x_k \le \sum_{i \in a \setminus a^\star} \mathtt{kl}(\theta_i, \tilde{\lambda}_i^a(\varepsilon)) \sum_k k_i x_k,$$

Then:

$$c_{\mathrm{s}}(\theta) \ge \quad \inf_{x \ge 0} \sum_{a \ne a^\star} \Delta_a x_a \tag{3.17}$$

$$\text{subject to:} \sum_{i \in a \setminus a^\star} \mathtt{kl}(\theta_i, \tilde{\lambda}_i^a(\varepsilon)) \sum_k k_i x_k \ge 1, \quad \forall a \ne a^\star. \tag{3.18}$$

Introducing $g_a(\varepsilon) = \max_{i \in a \setminus a^\star} \mathtt{kl}(\theta_i, \tilde{\lambda}_i^a(\varepsilon))$ for any $a \ne a^\star$, we form P1 as follows:

$$\mathsf{P1}: \quad \inf_{x \ge 0} \sum_{a \ne a^\star} \Delta_a x_a \tag{3.19}$$

$$\text{subject to:} \sum_{i \in a \setminus a^\star} \sum_k k_i x_k \ge \frac{1}{g_a(\varepsilon)}, \quad \forall a \ne a^\star. \tag{3.20}$$

Observe that $c_{\mathrm{s}}(\theta) \ge \mathrm{val}(\mathsf{P1})$ since the feasible set of problem (3.17) is contained in that of P1.

Step 2. In this step, we formulate an LP to give a lower bound for val(P1). To this end, for any sub-optimal basic action $i \in E$, we define $z_i = \sum_a a_i x_a$. Further, we let $z = (z_i)_{i \in E}$. Next, we represent the objective of P1 in terms of z, and give a lower bound for it as follows:

$$\begin{aligned}\sum_{a \ne a^\star} \Delta_a x_a &= \sum_{a \ne a^\star} x_a \sum_{i \in a \setminus a^\star} \frac{\Delta_a}{|a \setminus a^\star|} \\ &= \sum_{a \ne a^\star} x_a \sum_{i \in E \setminus a^\star} \frac{\Delta_a}{|a \setminus a^\star|} a_i \\ &\ge \min_{a \ne a^\star} \frac{\Delta_a}{|a \setminus a^\star|} \cdot \sum_{i \in E \setminus a^\star} \sum_{a' \ne a^\star} a'_i x_{a'} \\ &= \min_{a \ne a^\star} \frac{\Delta_a}{|a \setminus a^\star|} \cdot \sum_{i \in E \setminus a^\star} z_i \ .\end{aligned}$$

Recalling the definition $\beta(\theta) = \min_{a \ne a^\star} \frac{\Delta_a}{|a \setminus a^\star|}$ and defining

$$\mathsf{P2}: \quad \inf_{z \ge 0} \beta(\theta) \sum_{i \in E \setminus a^\star} z_i$$

$$\text{subject to:} \quad \sum_{i \in a \setminus a^\star} z_i \geq \frac{1}{g_a(\varepsilon)}, \quad \forall a \neq a^\star,$$

yields: val(P1) $\geq$ val(P2).

Step 3. Introduce a set $\mathcal{H}$ satisfying property $P(\theta)$ as stated in Definition 3.1. Now define

$$\mathcal{Z} = \Big\{ z \in \mathbb{R}_+^d : \sum_{i \in a \setminus a^\star} z_i \geq 1/g_a(\varepsilon), \;\; \forall a \in \mathcal{H} \Big\},$$

and

$$\mathsf{P3:} \quad \inf_{z \in \mathcal{Z}} \beta(\theta) \sum_{i \in E \setminus a^\star} z_i.$$

Observe that val(P2) $\geq$ val(P3) since the feasible set of P2 is contained in $\mathcal{Z}$. The definition of $\mathcal{H}$ implies that $\sum_{i \in E \setminus a^\star} z_i = \sum_{a \in \mathcal{H}} \sum_{i \in a \setminus a^\star} z_i$. Letting $\varepsilon \to 0$, we thus get

$$\mathrm{val}(\mathsf{P3}) = \sum_{a \in \mathcal{H}} \frac{\beta(\theta)}{\max_{i \in a \setminus a^\star} \mathtt{kl}\left(\theta_i, \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \theta_j\right)}\, .$$

The proof is completed by observing that: $c_{\mathrm{s}}(\theta) \geq$ val(P1) $\geq$ val(P2) $\geq$ val(P3). □

3.E Proof of Corollary 3.2

Fix $a \neq a^\star$. For any $i \in a \setminus a^\star$, we have:

$$\begin{aligned}
\mathtt{kl}\Big(\theta_i, \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \theta_j\Big) &\leq \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \mathtt{kl}\left(\theta_i, \theta_j\right) \quad \text{(By convexity of } \mathtt{kl}\text{)} \\
&\leq \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \frac{(\theta_i - \theta_j)^2}{\theta_j(1-\theta_j)} \\
&\leq \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \frac{\alpha^2/4}{\alpha/2(1-\alpha/2)} \\
&\leq \frac{\alpha}{2-\alpha} \leq \frac{1}{3},
\end{aligned}$$

where the second inequality follows from the inequality $\mathtt{kl}(p,q) \leq \frac{(p-q)^2}{q(1-q)}$ for all $(p,q) \in [0,1]^2$, and where the third inequality uses the fact that $x \mapsto x(1-x)$ is increasing over $x \in (0, \frac{1}{2})$. Moreover, we have that

$$\beta(\theta) = \min_{a \neq a^\star} \frac{\Delta^a}{|a \setminus a^\star|} \geq \frac{\Delta_{\min}}{\max_a |a \setminus a^\star|} = \frac{\Delta_{\min}}{k}.$$

Applying Theorem 3.3, we get:

$$c_{\mathrm{s}}(\theta) \geq \sum_{a \in \mathcal{H}} \frac{\beta(\theta)}{\max_{i \in a \setminus a^\star} \mathtt{kl}\left(\theta_i, \frac{1}{|a \setminus a^\star|} \sum_{j \in a^\star \setminus a} \theta_j\right)} \geq \frac{\Delta_{\min}}{3k}|\mathcal{H}|,$$

which gives the required lower bound and completes the proof. □

3.F Proof of Theorem 3.4

To prove the theorem, we apply the techniques used by Graves and Lai [39] as used in the case of semi-bandit feedback. To this end, we construct a controlled Markov chain as follows. The state-space is $\mathcal{S} = \{0, \dots, m\}$. The set of controls corresponds to the set of arms $\mathcal{A}$, and the set of control laws is also $\mathcal{A}$. The parameter θ takes values in $[0,1]^d$. The probability that the reward under arm a is equal to k is then $\psi_\theta^a(k)$ defined in (3.4), and so:

$$p(k', k; a, \theta) = \psi_\theta^a(k), \;\; \forall k, k' \in \mathcal{S}.$$

From [39, Theorem 1], we conclude that for any uniformly good rule π,

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq c_{\mathrm{b}}(\theta),$$

where $c_{\mathrm{b}}(\theta)$ is the optimal value of the following optimization problem:

$$c_{\mathrm{b}}(\theta) = \inf_{x \geq 0} \sum_{a \neq a^\star} x_a \Delta_a, \tag{3.21}$$

$$\text{subject to: } \inf_{\lambda \in B_{\mathrm{b}}(\theta)} \sum_{k \neq a^\star} x_k I^k(\theta, \lambda) \geq 1, \tag{3.22}$$

where $I^k(\theta, \lambda)$ is defined in (3.5). This concludes the proof. □

3.G Proof of Theorem 3.5

Let $M = (E, \mathcal{I}, \theta)$ be a weighted matroid. To ease notation, we use the abbreviations $B(\theta)$ and σ to respectively denote $B_{\mathrm{b}}(\theta)$ and σ_M. Recall from Theorem 3.4 that the regret of any uniformly good policy $\pi \in \Pi_{\mathrm{b}}$ for any $\theta \in \Theta$ satisfies

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq c_{\mathrm{b}}(\theta),$$

where $c_{\mathrm{b}}(\theta)$ is the optimal value of the optimization problem:

$$\inf_{x \geq 0} \sum_{a \in \mathcal{A}} \Delta_a x_a \tag{3.23}$$

$$\text{subject to:} \quad \sum_{a \neq a^\star} x_a I^a(\theta, \lambda) \geq 1, \ \ \forall \lambda \in B(\theta) \ .$$

We argue that $\mu^\star(\lambda) > \mu^\star(\theta)$ implies that there exists at least one sub-optimal action i with $\lambda_i > \theta_{\sigma(i)}$. Hence, we decompose $B(\theta)$ into sets where in each set, action i is better than action $\sigma(i)$ under λ. For any $i \notin a^\star$, define

$$A_i(\theta) = \Big\{ \lambda : \sum_{\ell \in a^\star} \lambda_\ell = \sum_{\ell \in a^\star} \theta_\ell, \ \lambda_i > \theta_{\sigma(i)} \Big\}.$$

Then, $B(\theta) = \bigcup_{i \notin a^\star} A_i(\theta)$ and problem (3.23) reads

$$\begin{aligned} c_{\mathrm{b}}(\theta) = & \inf_{x \geq 0} \ \sum_a x_a \Delta_a \\ & \text{subject to:} \ \inf_{\lambda \in A_i(\theta)} \sum_{a \neq a^\star} x_a I^a(\theta, \lambda) \geq 1, \ \ \forall i \notin a^\star. \end{aligned} \tag{3.24}$$

Consider ζ^i with $\zeta^i_i = \theta_{\sigma(i)}$ and $\zeta^i_j = \theta_j$ for $j \neq i$. Since $\zeta^i \in A_i(\theta)$, we have

$$\begin{aligned} \inf_{\lambda \in A_i(\theta)} \sum_{a \neq a^\star} x_a I^a(\theta, \lambda) &\leq \sum_a x_a I^a(\theta, \zeta^i) \\ &= \sum_a a_i x_a I^a(\theta, \zeta^i) \\ &\leq \max_{a : i \in a} I^a(\theta, \zeta^i) \sum_a a_i x_a . \end{aligned}$$

Hence, problem (3.24) is lower bounded as follows:

$$\begin{aligned} c_{\mathrm{b}}(\theta) \geq & \inf_{x \geq 0} \ \sum_a x_a \Delta_a \\ & \text{subject to:} \ \max_{a : i \in a} I^a(\theta, \zeta^i) \sum_a a_i x_a \geq 1, \ \ \forall i \notin a^\star. \end{aligned} \tag{3.25}$$

Recall from the proof of Theorem 3.2 that $\sum_a x_a \Delta_a \geq \sum_{i \in E \setminus a^\star} (\theta_{\sigma(i)} - \theta_i) z_i$. Hence, problem (3.25) is further lower bounded as

$$\begin{aligned} c_{\mathrm{b}}(\theta) \geq & \inf_{z \geq 0} \ \sum_{i \in E \setminus a^\star} (\theta_{\sigma(i)} - \theta_i) z_i \\ & \text{subject to:} \ z_i \geq \frac{1}{\max_{a : i \in a} I^a(\theta, \zeta^i)}, \ \ \forall i \notin a^\star, \end{aligned}$$

which further gives

$$c_{\mathrm{b}}(\theta) \geq \sum_{i \in E \setminus a^\star} \frac{\theta_{\sigma(i)} - \theta_i}{\max_{a : i \in a} I^a(\theta, \zeta^i)}$$

and concludes the proof. □

3.H Proof of Theorem 3.6

Proof of the first statement. Consider arm $a \in \mathcal{A}$ and let $q, \lambda \in \Theta$, $t \in \mathbb{N}^d$, and $n \in \mathbb{N}$. Applying Cauchy-Schwarz inequality gives

$$a^\top(q-\lambda) = \sum_{i \in a} \sqrt{t_i}(q_i - \lambda_i)\frac{1}{\sqrt{t_i}} \le \sqrt{\sum_{i \in a} t_i (q_i - \lambda_i)^2}\sqrt{\sum_{i \in a}\frac{1}{t_i}}.$$

By Pinsker's inequality (see Lemma A.2),

$$a^\top(q-\lambda) \le \sqrt{\frac{1}{2}\sum_{i \in a} t_i \mathtt{kl}(\lambda_i, q_i)}\sqrt{\sum_{i \in a}\frac{1}{t_i}}.$$

Hence, $\sum_{i \in a} t_i \mathtt{kl}(\lambda_i, q_i) \le f(n)$ implies:

$$a^\top q = a^\top \lambda + a^\top(q-\lambda) \le a^\top\lambda + \sqrt{\frac{f(n)}{2}\sum_{i \in a}\frac{1}{t_i}} = c_a(n, \lambda, t),$$

so that by definition of $b_a(n, \lambda, t)$, we deduce $b_a(n, \lambda, t) \le c_a(n, \lambda, t)$.

Proof of the second statement. If $\sum_{i \in a} t_i(n)\mathtt{kl}(\hat\theta_i(n), \theta_i) \le f(n)$, then by definition of b_a we have $b_a(n, \hat\theta(n), t(n)) \ge a^\top\theta$. Therefore, using [68, Theorem 2] (see Corollary B.1), there exists a constant C_m such that for all $n \ge 2$ we have:

$$\begin{aligned}\mathbb{P}(b_a(n, \hat\theta(n), t(n)) < a^\top\theta) &\le \mathbb{P}\Big(\sum_{i \in a} t_i(n)\mathtt{kl}(\hat\theta_i(n), \theta_i) > f(n)\Big)\\ &\le C_m n^{-1}(\log(n))^{-2},\end{aligned}$$

which concludes the proof. □

3.I Proof of Theorem 3.7

Consider a and n fixed throughout the proof. Recall that $J_a(\lambda) = \{i \in a : \lambda_i \ne 1\}$. Consider $q^\star \in \Theta$ the optimal solution of the following optimization problem:

$$\begin{aligned}\max_{q \in \Theta} \;& a^\top q\\ \text{subject to: } & \sum_{i \in a} t_i \mathtt{kl}(\lambda_i, q_i) \le f(n),\end{aligned}$$

so that $b_a(n, \lambda, t) = a^\top q^\star$. Consider $i \notin a$. Then $a^\top q$ does not depend on q_i and from Lemma A.1 (statement (i)), we get $q_i = \lambda_i$. Now consider $i \in a$. From Lemma A.1 (statement (i)), we get that $1 \ge q_i^\star \ge \lambda_i$. Hence $q_i^\star = 1$ if $\lambda_i = 1$. If $J_a(\lambda)$ is empty, then $q_i^\star = 1$ for all $i \in a$, so that $b_a(n, \lambda, t) = \|a\|_1$.

Now consider the case where $J_a(\lambda) \neq \emptyset$. From Lemma A.1 (statement (iii)) and the fact that $\sum_{i\in a} t_i \mathtt{kl}(\lambda_i, q_i^\star) < \infty$, we get $\lambda_i \le q_i^\star < 1$. From the KKT conditions (see, e.g., [69]), there exists Lagrange multiplier $\gamma^\star > 0$ such that for all $i \in J_a(\lambda)$:

$$1 = \gamma^\star t_i \mathtt{kl}'(\lambda_i, q_i^\star).$$

For $\gamma > 0$ define $\lambda_i \le \overline{q}_i(\gamma) < 1$ as a solution to the following equation:

$$1 = \gamma t_i \mathtt{kl}'(\lambda_i, \overline{q}_i(\gamma)).$$

It follows from Lemma A.1 (statement (i)) that $\gamma \mapsto \overline{q}_i(\gamma)$ is uniquely defined, strictly decreasing, and $\lambda_i < \overline{q}_i(\gamma) < 1$. From Lemma A.1 (statement (iii)), we deduce that $\overline{q}_i(\mathbb{R}_+) = [\lambda_i, 1]$. Define the function:

$$F(\gamma, \lambda, n, t) = \sum_{i\in J_a(\lambda)} t_i \mathtt{kl}(\lambda_i, \overline{q}_i(\gamma)).$$

From the reasoning above, F is well-defined, strictly increasing, and $F(\mathbb{R}_+, \lambda, n, t) = \mathbb{R}_+$. Therefore, $\gamma^\star$ is the unique solution to $F(\gamma^\star, \lambda, n, t) = f(n)$, and $q_i^\star = \overline{q}_i(\gamma^\star)$. Furthermore, replacing $\mathtt{kl}'$ (see, e.g., Lemma A.1) by its expression we obtain the following quadratic equation:

$$\overline{q}_i(\gamma)^2 + \overline{q}_i(\gamma)(\gamma t_i - 1) - \gamma t_i \lambda_i = 0.$$

Solving for $\overline{q}_i(\gamma)$, we obtain that $\overline{q}_i(\gamma) = g(\gamma, \lambda_i, t_i)$, which concludes the proof. □

3.J Proof of Theorem 3.8

To prove Theorem 3.8, we borrow some ideas from the proof of [62, Theorem 3].

For any $n \in \mathbb{N}$, $s \in \mathbb{R}^d$, and $a \in \mathcal{A}$ define $h_{n,s,a} = \sqrt{\frac{f(n)}{2}\sum_{i\in E}\frac{a_i}{s_i}}$, and introduce the following events:

$$\begin{aligned}
G_n &= \Big\{\sum_{i\in a^\star} t_i(n)\mathtt{kl}(\hat\theta_i(n), \theta_i) > f(n)\Big\},\\
H_{i,n} &= \{a_i(n) = 1,\ |\hat\theta_i(n) - \theta_i| \ge m^{-1}\Delta_{\min}/2\},\ \ H_n = \bigcup_{i\in E} H_{i,n},\\
F_n &= \{\Delta_{a(n)} \le 2h_{T,t(n),a(n)}\}.
\end{aligned}$$

Then the regret can be bounded as:

$$\begin{aligned}
\mathfrak{R}_{\pi,T} = \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}] &\le \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}(\mathbb{I}\{G_n\} + \mathbb{I}\{H_n\})] + \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}\mathbb{I}\{\overline{G_n},\ \overline{H_n}\}]\\
&\le m\mathbb{E}[\sum_{n=1}^T(\mathbb{I}\{G_n\} + \mathbb{I}\{H_n\})] + \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}\mathbb{I}\{\overline{G_n},\ \overline{H_n}\}],
\end{aligned}$$

since $\Delta_{a(n)} \le m$.

Next we show that for any n such that $a(n) \neq a^\star$, it holds that $\overline{G_n \cup H_n} \subset F_n$. Recall that $c_a(n) \ge b_a(n)$ for any a and n (Theorem 3.6). Moreover, if $\overline{G_n}$ holds, we have $\sum_{i\in a^\star} t_i(n)\mathtt{kl}(\hat\theta_i(n), \theta_i) \le f(n)$, which by definition of b_a implies: $b_{a^\star}(n) \ge {a^\star}^\top\theta$. Hence we have:

$$\begin{aligned}
\mathbb{I}\{\overline{G_n},\ \overline{H_n},\ a(n)\neq a^\star\} &= \mathbb{I}\{\overline{G_n},\ \overline{H_n},\ \xi_{a(n)}(n) \ge \xi_{a^\star}(n)\}\\
&\le \mathbb{I}\{\overline{H_n},\ c_{a(n)}(n) \ge {a^\star}^\top\theta\}\\
&= \mathbb{I}\{\overline{H_n},\ a(n)^\top\hat\theta(n) + h_{n,t(n),a(n)} \ge {a^\star}^\top\theta\}\\
&\le \mathbb{I}\{a(n)^\top\theta + \Delta_{a(n)}/2 + h_{n,t(n),a(n)} \ge {a^\star}^\top\theta\}\\
&= \mathbb{I}\{2h_{n,t(n),a(n)} \ge \Delta_{a(n)}\}\\
&\le \mathbb{I}\{2h_{T,t(n),a(n)} \ge \Delta_{a(n)}\}\\
&= \mathbb{I}\{F_n\},
\end{aligned}$$

where the second inequality follows from the fact that event $\overline{G_n}$ implies: $a(n)^\top\hat\theta(n) \le a(n)^\top\theta + \Delta_{\min}/2 \le a(n)^\top\theta + \Delta_{a(n)}/2$.

Hence, the regret is upper bounded by:

$$\mathfrak{R}_{\pi,T} \le m\mathbb{E}[\sum_{n=1}^T \mathbb{I}\{G_n\}] + m\mathbb{E}[\sum_{n=1}^T \mathbb{I}\{H_n\}] + \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}\mathbb{I}\{F_n\}].$$

We will prove the following inequalities:

$$\begin{aligned}
&(i) \quad \mathbb{E}[\sum_{n=1}^T \mathbb{I}\{G_n\}] \le m^{-1}C'_m,\\
&(ii) \quad \mathbb{E}[\sum_{n=1}^T \mathbb{I}\{H_n\}] \le 4dm^2\Delta_{\min}^{-2},\\
&(iii) \quad \mathbb{E}[\sum_{n=1}^T \Delta_{a(n)}\mathbb{I}\{F_n\}] \le 16d\sqrt{m}\Delta_{\min}^{-1}f(T)\ ,
\end{aligned}$$

with $C'_m \ge 0$ independent of θ, d, and T.

Hence as announced:

$$\mathfrak{R}_{\pi,T} \le 16d\sqrt{m}\Delta_{\min}^{-1}f(T) + 4dm^3\Delta_{\min}^{-2} + C'_m.$$

Analysis of inequality (i). An application of Theorem B.6 gives

$$\begin{aligned}
\mathbb{E}[\sum_{n=1}^T \mathbb{I}\{G_n\}] &= \sum_{n=1}^T \mathbb{P}\Big(\sum_{i\in a^\star} t_i(n)\mathtt{kl}(\hat\theta_i(n), \theta_i) > f(n)\Big)\\
&\le 1 + \sum_{n\ge 2} C_m n^{-1}(\log(n))^{-2} \equiv m^{-1}C'_m < \infty.
\end{aligned}$$

Analysis of inequality (ii). Fix i and n. Define $s = \sum_{n'=1}^{n} \mathbb{I}\{H_{n',i}\}$. Observe that $H_{n',i}$ implies $a_i(n') = 1$, hence $t_i(n) \geq s$. Therefore, applying Applying [70, Lemma B.1] (see Corollary B.2 in Appendix B), we have that $\sum_{n=1}^{T} \mathbb{P}(H_{n,i}) \leq 4m^2\Delta_{\min}^{-2}$. Using the union bound: $\sum_{n=1}^{T} \mathbb{P}(H_n) \leq 4dm^2\Delta_{\min}^{-2}$.

Analysis of inequality (iii). Let $\ell > 0$. For any n introduce the following events:

$$\begin{aligned}
S_n &= \{i \in a(n) : t_i(n) \leq 4mf(T)\Delta_{a(n)}^{-2}\},\\
A_n &= \{|S_n| \geq \ell\},\\
B_n &= \{|S_n| < \ell,\ [\exists i \in a(n) : t_i(n) \leq 4\ell f(T)\Delta_{a(n)}^{-2}]\}.
\end{aligned}$$

We claim that for any n such that $a(n) \neq a^\star$, we have $F_n \subset (A_n \cup B_n)$. To prove this, we show that when F_n holds and $a(n) \neq a^\star$, the event $\overline{A_n \cup B_n}$ cannot happen. Let n be a time instant such that $a(n) \neq a^\star$ and F_n holds, and assume that $\overline{A_n \cup B_n} = \{|S_n| < \ell,\ [\forall i \in a(n) : t_i(n) > 4\ell f(T)\Delta_{a(n)}^{-2}]\}$ happens. Then F_n implies:

$$\begin{aligned}
\Delta_{a(n)} \leq 2h_{T,t(n),a(n)} &= 2\sqrt{\frac{f(T)}{2}}\sqrt{\sum_{i\in E\setminus S_n} \frac{a_i(n)}{t_i(n)} + \sum_{i\in S_n} \frac{a_i(n)}{t_i(n)}}\\
&< 2\sqrt{\frac{f(T)}{2}}\sqrt{m\frac{\Delta_{a(n)}^2}{4mf(T)} + |S_n|\frac{\Delta_{a(n)}^2}{4\ell f(T)}} < \Delta_{a(n)}, \qquad (3.26)
\end{aligned}$$

where the last inequality uses the observation that $\overline{A_n \cup B_n}$ implies $|S_n| < \ell$. Clearly, (3.26) is a contradiction. Thus $F_n \subset (A_n \cup B_n)$ and consequently:

$$\sum_{n=1}^{T} \Delta_{a(n)}\mathbb{I}\{F_n\} \leq \sum_{n=1}^{T} \Delta_{a(n)}\mathbb{I}\{A_n\} + \sum_{n=1}^{T} \Delta_{a(n)}\mathbb{I}\{B_n\}. \qquad (3.27)$$

To further bound the right-hand side of the above, we introduce the following events for any i:

$$\begin{aligned}
A_{i,n} &= A_n \cap \{i \in a(n),\ t_i(n) \leq 4mf(T)\Delta_{a(n)}^{-2}\},\\
B_{i,n} &= B_n \cap \{i \in a(n),\ t_i(n) \leq 4\ell f(T)\Delta_{a(n)}^{-2}\}.
\end{aligned}$$

It is noted that:

$$\sum_{i\in E} \mathbb{I}\{A_{i,n}\} = \mathbb{I}\{A_n\}\sum_{i\in E} \mathbb{I}\{i \in S_n\} = |S_n|\mathbb{I}\{A_n\} \geq \ell\mathbb{I}\{A_n\},$$

and hence: $\mathbb{I}\{A_n\} \leq \frac{1}{\ell}\sum_{i\in E} \mathbb{I}\{A_{i,n}\}$. Moreover $\mathbb{I}\{B_n\} \leq \sum_{i\in E} \mathbb{I}\{B_{i,n}\}$. Let each basic action i belong to K_i sub-optimal arms, ordered based on their gaps as:

$\Delta^{i,1} \geq \cdots \geq \Delta^{i,K_i} > 0$. Also define $\Delta^{i,0} = \infty$. Plugging the above inequalities into (3.27), we have

$$\begin{aligned}
\sum_{n=1}^{T} \Delta_{a(n)} \mathbb{I}\{F_n\} &\leq \sum_{n=1}^{T} \sum_{i\in E} \frac{\Delta_{a(n)}}{\ell} \mathbb{I}\{A_{i,n}\} + \sum_{n=1}^{T} \sum_{i\in E} \Delta_{a(n)} \mathbb{I}\{B_{i,n}\} \\
&= \sum_{n=1}^{T} \sum_{i\in E} \frac{\Delta_{a(n)}}{\ell} \mathbb{I}\{A_{i,n},\ a(n) \neq a^\star\} + \sum_{n=1}^{T} \sum_{i\in E} \Delta_{a(n)} \mathbb{I}\{B_{i,n},\ a(n) \neq a^\star\} \\
&\leq \sum_{n=1}^{T} \sum_{i\in E} \sum_{k\in[K_i]} \frac{\Delta^{i,k}}{\ell} \mathbb{I}\{A_{i,n},\ a(n) = k\} + \sum_{n=1}^{T} \sum_{i\in E} \sum_{k\in[K_i]} \Delta^{i,k} \mathbb{I}\{B_{i,n},\ a(n) = k\} \\
&\leq \sum_{i\in E} \sum_{n=1}^{T} \sum_{k\in[K_i]} \frac{\Delta^{i,k}}{\ell} \mathbb{I}\{i \in a(n),\ t_i(n) \leq 4mf(T)(\Delta^{i,k})^{-2},\ a(n) = k\} \\
&+ \sum_{i\in E} \sum_{n=1}^{T} \sum_{k\in[K_i]} \Delta^{i,k} \mathbb{I}\{i \in a(n),\ t_i(n) \leq 4\ell f(T)(\Delta^{i,k})^{-2},\ a(n) = k\} \\
&\leq \frac{8df(T)}{\Delta_{\min}} \left(\frac{m}{\ell} + \ell\right),
\end{aligned}$$

where the last inequality follows from Lemma 3.2, which is proven next. The proof is completed by setting $\ell = \sqrt{m}$. □

Lemma 3.2. *Let $C > 0$ be a constant independent of n. Then for any i such that $K_i \geq 1$:*

$$\sum_{n=1}^{T} \sum_{k=1}^{K_i} \mathbb{I}\{i \in a(n),\ t_i(n) \leq C(\Delta^{i,k})^{-2},\ a(n) = k\} \Delta^{i,k} \leq \frac{2C}{\Delta_{\min}}.$$

Proof. Borrowing some techniques from [61], we have:

$$\begin{aligned}
&\sum_{n=1}^{T} \sum_{k=1}^{K_i} \mathbb{I}\{i \in a(n),\ t_i(n) \leq C(\Delta^{i,k})^{-2},\ a(n) = k\} \Delta^{i,k} \\
&\quad = \sum_{n=1}^{T} \sum_{k=1}^{K_i} \sum_{j=1}^{k} \mathbb{I}\{i \in a(n),\ t_i(n) \in (C(\Delta^{i,j-1})^{-2}, C(\Delta^{i,j})^{-2}],\ a(n) = k\} \Delta^{i,k} \\
&\quad \leq \sum_{n=1}^{T} \sum_{k=1}^{K_i} \sum_{j=1}^{k} \mathbb{I}\{i \in a(n),\ t_i(n) \in (C(\Delta^{i,j-1})^{-2}, C(\Delta^{i,j})^{-2}],\ a(n) = k\} \Delta^{i,j} \\
&\quad \leq \sum_{n=1}^{T} \sum_{k=1}^{K_i} \sum_{j=1}^{K_i} \mathbb{I}\{i \in a(n),\ t_i(n) \in (C(\Delta^{i,j-1})^{-2}, C(\Delta^{i,j})^{-2}],\ a(n) = k\} \Delta^{i,j}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{n=1}^{T}\sum_{j=1}^{K_i} \mathbb{I}\{i \in a(n),\ t_i(n) \in (C(\Delta^{i,j-1})^{-2}, C(\Delta^{i,j})^{-2}],\ a(n) \neq a^\star\}\Delta^{i,j} \\
&\leq \frac{C}{\Delta^{i,1}} + \sum_{j=2}^{K_i} C((\Delta^{i,j})^{-2} - (\Delta^{i,j-1})^{-2})\Delta^{i,j} \\
&\leq \frac{C}{\Delta^{i,1}} + \int_{\Delta^{i,K_i}}^{\Delta^{i,2}} Cx^{-2}\mathrm{d}x \leq \frac{2C}{\Delta^{i,K_i}} \leq \frac{2C}{\Delta_{\min}},
\end{aligned}
$$

which completes the proof. □

3.K Background on Matroids

In this section we give a formal definition of matroids and state some useful related results. More details can be found in, e.g., [71, 27].

Definition 3.2. *Let E be a finite set and $\mathcal{I} \subset 2^E$. The pair $M = (E, \mathcal{I})$ is called a* matroid *if the following conditions hold: (i) $\emptyset \in \mathcal{I}$, (ii) if $X \in \mathcal{I}$ and $Y \subseteq X$, then $Y \in \mathcal{I}$, and (iii) if $X, Y \in \mathcal{I}$ with $|X| > |Y|$, then there is some element $\ell \in X \setminus Y$ such that $Y \cup \{\ell\} \in \mathcal{I}$.*

The set E is usually referred to as the *ground set* and the elements of $\mathcal{I}$ are called the *independent sets*. Any system satisfying conditions (i) and (ii) in Definition 3.2 is called an *independence system*. Condition (iii) is referred to as the *augmentation property*. Any (inclusion-wise) maximal independent set is called a *basis* for matroid M. In other words, if $X \in \mathcal{I}$ is a basis for M, then $X \cup \{\ell\} \notin \mathcal{I}$ for all $\ell \in E \setminus X$.

Proposition 3.1 ([71])**.** *Let $M = (E, \mathcal{I})$ be a matroid. Then*

(i) all bases of M have the same cardinality (referred to as rank *of M),*

(ii) for all bases X, Y of M, if $\ell \in X \setminus Y$, then there exists $k \in Y \setminus X$ such that $(X \setminus \ell) \cup \{k\}$ is a basis for M. [7]

(iii) for all bases X, Y of M, if $\ell \in X \setminus Y$ then there exists $k \in Y \setminus X$ such that $(Y \setminus k) \cup \{\ell\}$ is a basis for M.

Next we provide some examples of matroids.

Uniform matroid. Let E be a set with cardinality d. Given a positive integer $m \leq d$, the uniform matroid of rank m is $U_{m,d} = (E, \mathcal{I})$, where $\mathcal{I}$ is the collection of subsets of E with at most m elements, i.e., $\mathcal{I} = \{X \subseteq E : |X| \leq m\}$. Hence, every subset of E with cardinality m is a basis for the uniform matroid $U_{m,d}$.

[7] For any set X and element ℓ, by a slight abuse of notation, we write $X \setminus \ell$ to imply $X \setminus \{\ell\}$.

Partition matroid. Let E be a finite set. Assume that $\{E_i\}_{i\in[l]}$ is a partition of E, i.e., $E_i, i \in [l]$ are disjoint sets and $\cup_{i\in[l]} E_i = E$. Given integers $k_1, \dots, k_l$, define $\mathcal{I} = \{X \subseteq E : |X \cap E_i| \le k_i,\ \forall i \in [l]\}$. Then $(E, \mathcal{I})$ is a partition matroid of rank $\sum_{i\in[l]} k_i$. [8]

Graphic matroid. Given an undirected graph $G = (V, H)$ (that may contain loops), define $\mathcal{I} = \{F \subseteq H : (V, F) \text{ is a forest}\}$. Then, it can be shown that $M(G) = (H, \mathcal{I})$ is a matroid, referred to as graphic matroid. Every spanning forest of G is a basis for matroid $M(G)$.

[8] In some papers, the notion of partition matriod is defined with $k_i = 1$ for every $i \in [l]$.

Chapter 4

Stochastic Online Shortest-Path Routing

In most real-world networks, link delays vary stochastically due to unreliable links and random access protocols (e.g., in wireless networks), mobility (e.g., in mobile ad-hoc networks), randomness of demand (e.g., in overlay networks for peer-to-peer applications), etc. In many cases, the associated parameters to links, e.g., the packet transmission success probabilities in wireless sensor networks, are initially unknown and must be estimated by transmitting packets and observing the outcomes. When designing routing policies, we therefore need to address a challenging trade-off between exploration and exploitation: On the one hand, it is important to route packets on new or poorly known links to explore the network and ensure that the optimal path is eventually found; on the other hand, it is critical that the accumulated knowledge on link parameters is exploited so that paths with low expected delays are preferred. When designing practical routing schemes, one is mostly concerned about the finite-time behaviour of the system and it is crucial to design algorithms that quickly learn link parameters so as to efficiently track the optimal path.

The design of such routing policies is often referred to as an online shortest-path routing problem in the literature [31, 72, 32, 34, 73], and is a particular instance of a combinatorial MAB problem. In this chapter, we study the *stochastic* version of this problem. More precisely, we consider a network, in which the transmission of a packet on a given link is successful with an unknown but fixed probability. A packet is sent on a given link repeatedly until the transmission is successful; the number of time slots to complete the transmission is referred to as the *delay* on this link. We wish to route N packets from a given source to a given destination in a minimum amount of time. A routing policy selects a path to the destination on a packet-by-packet basis. The path selection can be done at the source (source routing), or in the network as the packet progresses towards the destination (hop-by-hop routing). In the case of source routing, some feedback is available when the packet reaches the destination. This feedback can be either the end-to-end delay, or the delays on each link on the path from the source to the destination. In the MAB literature, the former type of feedback is referred to as *bandit* feedback, whereas the latter is called *semi-bandit* feedback (as introduced in [74]). The routing policy then selects the path for the next packet based on the feedback gathered from previously transmitted packets. In the case of hop-by-hop routing, routing decisions are taken for each transmission and the packet is sent over a link selected based on all transmission successes and failures observed so far (for the current packet, and all previously sent packets) on the various links.

The performance of a routing policy is assessed through its expected total delay, i.e., the expected time required to send all N packets to the destination. Equivalently, it can be measured through the notion of *regret*, defined as the difference between the expected total delay under the policy considered and the expected total delay of an oracle policy that would be aware of all link parameters, and would hence always send the packets on the optimal path. Regret conveniently quantifies the loss in performance due to the fact that link parameters are initially unknown and need to be learnt.

In this chapter we study the online shortest-path routing problem in the stochastic setting as described above. Using the machinery of Chapter 3, we derive problem-specific lower bounds on the regret. We present three algorithms for this class of problems and provide upper bounds on their regret. These upper bounds are the best ones proposed so far in the literature for the considered problem. We also provide numerical experiments, which show that our algorithms outperform existing ones.

This chapter is based on the work [75] and is organized as follows: Section 4.1 outlines our contributions of the chapter and discusses related works. Section 4.2 describes the network model, feedback models, and objectives. In Section 4.3, we present regret lower bounds for various types of feedback. In Section 4.4, we present routing policies for the case of source routing with semi-bandit feedback along with their regret analysis. Section 4.5 presents numerical experiments. In Section 4.6, we give a brief summary of the materials presented in this chapter.

4.1 Contributions and Related Work

The first part of this chapter is motivated by the following fundamental questions: (i) what is the benefit of allowing routing decisions at every node, rather than only at the source? and (ii) what is the added value of feeding back the observed delay for every link that a packet has traversed compared to only observing the end-to-end delay?[1] To answer these questions, we derive tight regret lower bounds satisfied by any routing policy in the different scenarios, depending on where routing decisions are made and what information is available to the decision maker when making these decisions. By comparing the different lower bounds, we are able to quantify the value of having semi-bandit feedback rather than bandit feedback, and the improvements that can possibly be achieved by taking routing decisions hop by hop. We then propose routing policies in the semi-bandit feedback setting exhibiting better regret upper bounds than existing algorithms.

[1] The effect of different forms of feedback in the adversarial setting was studied in, e.g., [72, 32].

More precisely, our contributions are the following:

1. Regret lower bounds. We derive tight asymptotic (when N grows large) regret lower bounds. The first two bounds concern source routing policies under bandit and semi-bandit feedback, respectively, whereas the third bound is satisfied by any hop-by-hop routing policy. As we shall see later, these regret bounds are tight in the sense that one can design actual routing policies, despite being complex and impractical, that achieve these bounds. As it turns out, the regret lower bound for source routing policies with semi-bandit feedback and that for hop-by-hop routing policies are identical, indicating that taking routing decisions hop by hop does not bring any advantage. On the contrary, the regret lower bounds for source routing policies with bandit and semi-bandit feedback can be significantly different, illustrating the importance of having information about per-link delays.

2. Routing policies. In the case of semi-bandit feedback, we propose three online source routing policies, namely `GeoCombUCB-1`, `GeoCombUCB-2`, and `KL-SR` (KL-based Source-Routing). `Geo` refers to the fact that the delay on a given link is geometrically distributed, `Comb` stands for combinatorial, and `UCB` (Upper Confidence Bound) indicates that these policies are based on the same "optimism in face of uncertainty" principle as in the celebrated `UCB` algorithm designed for classical MAB problems [20]. `KL-SR` already appears in [33]. Here we improve its regret analysis, and show that its regret scales at most as $\mathcal{O}(dm\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$, [2] where d is the number of links in the network, m denotes the length (number of links) of the longest path in the network from the source to the destination, $\theta_{\min}$ is the success transmission probability of the link with the worst quality, and $\Delta_{\min}$ is the minimal gap between the average end-to-end delays of a sub-optimal and of the optimal path (formal definitions of $\theta_{\min}$ and $\Delta_{\min}$ are provided in Section 4.2). We further show that the regret under `GeoCombUCB-1` and `GeoCombUCB-2` scales at most as $\mathcal{O}(d\sqrt{m}\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$.

The trade-off between computational complexity and performance (regret) of online routing policies is certainly hard to characterize, but our policies provide a first insight into such a trade-off. Furthermore, they exhibit better regret upper bounds than that of the `CUCB` (Combinatorial `UCB`) algorithm [61], which is, to our knowledge, the state-of-the-art algorithm for stochastic online shortest-path routing. Furthermore, we conduct numerical experiments, showing that our routing policies perform significantly better than `CUCB`. We also mention that the `Thompson Sampling (TS)` algorithm of [76] is applicable to the shortest-path problem, but its analysis for general topologies is an open problem. While `TS` performs slightly better than our algorithms on average, its regret sometimes has a large variance according to our experiments. The regret guarantees of various algorithms and their computational complexities are summarized in Table 4.1.[3]

[2] This improves over the regret upper bound scaling as $\mathcal{O}(\Delta_{\max}dm^3\Delta_{\min}^{-1}\theta_{\min}^{-3}\log(N))$ derived in [33], where $\Delta_{\max}$ denotes the maximal gap between the average end-to-end delays of a sub-optimal and of the optimal path.

[3] In Table 4.1, V and $\mathcal{A}$ respectively denote the set of all nodes and the set of all possible paths between the source and the destination.

Algorithm	Regret	Complexity
CUCB [61]	$\mathcal{O}\left(\frac{dm}{\Delta_{\min}\theta_{\min}^3}\log(N)\right)$	$\mathcal{O}(d\|V\|)$
GeoCombUCB-1 (Theorem 4.5)	$\mathcal{O}\left(\frac{d\sqrt{m}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(\|\mathcal{A}\|)$
GeoCombUCB-2 (Theorem 4.5)	$\mathcal{O}\left(\frac{d\sqrt{m}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(\|\mathcal{A}\|)$
KL-SR (Theorem 4.6)	$\mathcal{O}\left(\frac{dm}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$	$\mathcal{O}(d\|V\|)$

Table 4.1: Comparison of various algorithms for shortest-path routing under semi-bandit feedback.

It is worth noting that this chapter is concerned about a single decision maker or agent learning to route her traffic on the optimal path. The agent learns to interact with a stochastic environment that is not influenced by the agent's decisions. This setting is relevant in wireless systems as explained above, but also in scenarios where the agent competes with *many* other similar agents strategically routing their traffic (see the literature on mean-field games). Our results do not apply to cases where a *few* selfish agents compete for the network resources. This scenario, often referred to as *adversarial* in the literature, has attracted a lot attention over the few past decades; see, e.g., [77]. In the adversarial setting, there are algorithms approaching Nash Equilibria (NE) under some fairly mild assumptions. To the best of our knowledge, when link qualities are stochastically varying, the convergence to NEs has not been investigated.

The analysis presented in this chapter can be easily extended to more general link models, provided that the (single-link) delay distributions are taken within one-parameter exponential families of distributions.

4.1.1 Related Work

We summarize existing results for generic stochastic combinatorial bandits that could be applied to online shortest-path routing. In [61], the authors present CUCB, an algorithm for generic stochastic combinatorial MAB problems under semi-bandit feedback. When applied to the online routing problem, the best regret upper bound for CUCB scales as $\mathcal{O}(\frac{dm}{\Delta_{\min}\theta_{\min}^3}\log(N))$ (see Appendix 4.H for details). This upper bound constitutes the best existing result for our problem, where the delay on each link is geometrically distributed. It is important to note that most proposed algorithms for combinatorial bandits [60, 62, 54] deal with bounded rewards, i.e., here bounded delays, and are not applicable to geometrically distributed delays.

Stochastic online shortest-path routing problems have been addressed in [78, 34, 79]. Liu and Zhao [78] consider routing with bandit (end-to-end) feedback and propose a forced-exploration algorithm with $\mathcal{O}(d^3m\log(N))$ regret in which a random barycentric spanner[4] path is chosen for exploration. He et al. [34] consider

[4] A barycentric spanner is a set of paths from which the delay of other paths can be computed

routing under semi-bandit feedback, where the source chooses a path for routing and a possibly different path for probing. Our model coincides with the coupled probing/routing case in their paper, for which they derive an asymptotic lower bound on the regret growing logarithmically with time. As we shall see later, their lower bound is not tight.

Finally, it is worth noting that the papers cited above considered source-routing only. To the best of our knowledge, the work presented here is the first to consider online routing problems with hop-by-hop decisions. Such a problem can be formulated as a classical MDP, in which the states are the packet locations and the actions are the outgoing links of each node. However, most studies consider MDP problems under stricter assumptions than ours and/or targeted different performance measures. Burnetas and Katehakis [22] derive the asymptotic lower bound on the regret and propose an asymptotically optimal index policy. Their result can be applied only to the so-called ergodic MDPs [50], where the induced Markov chain by any policy is irreducible and consists of a single recurrent class. In hop-by-hop routing, however, the policy that routes packets on a fixed path results in a Markov chain with reducible states that are not in the chosen path. [25] and [24] study the bigger class of communicating MDPs and present algorithms with finite-time regret upper bounds scaling logarithmically with time. Nevertheless, these algorithms perform badly when applied to hop-by-hop routing due to loose confidence intervals and due to the fact that the routing policy is not updated at each time slot.

4.2 Model and Objectives

4.2.1 Network Model

The network is modeled as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of links with cardinality d. Each link $i \in E$ may, for example, represent an unreliable wireless link. Without loss of generality, we assume that time is slotted and that one slot corresponds to the time to send a packet over a single link. Let $X_i(t)$ be a binary random variable indicating whether a transmission on link i at time t is successful. $(X_i(t))_{t\geq 1}$ is a sequence of i.i.d. Bernoulli variables with initially unknown mean θ_i. Hence, if a packet is sent on link i repeatedly until the transmission is successful, the time to complete the transmission (referred to as the delay on link i) is geometrically distributed with mean $1/\theta_i$. Let $\theta = (\theta_i)_{i\in E}$ be the vector representing the packet successful transmission probabilities on the various links. We consider a single source-destination pair $(\mathsf{src}, \mathsf{dst}) \in V^2$, and denote by $\mathcal{A} \subseteq \{0,1\}^d$ the set of loop-free paths from src to dst in G, where each path $a \in \mathcal{A}$ is a d-dimensional binary vector; for any $i \in E$, $a_i = 1$ if and only if i belongs to a. Let m denote the maximum length of the paths in $\mathcal{A}$, i.e., $m = \max_{a\in\mathcal{A}} \sum_{i\in E} a_i$. For brevity, in what follows, for any binary vector z, we write $i \in z$ to denote $z_i = 1$. Moreover, for any vector z, we use the convention that $z^{-1} = (z_i^{-1})_i$.

as its linear combination with coefficients in $[-1, 1]$ [31].

For any path a, $D_\theta(a) = \sum_{i\in a} \frac{1}{\theta_i}$ is the average packet delay through path a given link success rates θ. The path with minimal delay is: $a^\star \in \arg\min_{a\in\mathcal{A}} D_\theta(a)$. Moreover, for any path $a \in \mathcal{A}$, we define $\Delta_a = D_\theta(a) - D_\theta(a^\star) = (a - a^\star)^\top \theta^{-1}$. Let $\Delta_{\min} = \min_{\Delta_a \neq 0} \Delta_a$. We let $\theta_{\min} = \min_{i\in E} \theta_i$ and assume that $\theta_{\min} > 0$. Finally define $D^\star = D_\theta(a^\star)$ and $D^+ = \max_{a\in\mathcal{A}} D_\theta(a)$ the delays of the shortest and longest paths, respectively.

4.2.2 Objectives and Feedback

We assume that the source is fully backlogged (i.e., it always has packets to send) and that the parameter θ is initially unknown. Packets are sent successively from source (src) to destination (dst) over various paths, and the outcome of each packet transmission is used to estimate θ, and in turn to learn the path $a^\star$ with the minimum average delay. After a packet is sent, we assume that the source gathers feedback from the network (essentially per-link or end-to-end delays) before sending the next packet.

We consider and compare three different types of online routing policies, depending (i) on where routing decisions are taken (at the source or at each node), and (ii) on the received feedback (per-link or end-to-end path delay). The corresponding policy sets are defined below:

- Policy Set Π_1: The path used by a packet is determined at the source based on the observed end-to-end delays for previous packets. More precisely, for the n-th packet, let $a^\pi(n)$ be the path selected under policy π, and let $D^\pi(n)$ denote the corresponding end-to-end delay. Then $a^\pi(n)$ depends on $a^\pi(1), \dots, a^\pi(n-1), D^\pi(1), \dots, D^\pi(n-1)$.

- Policy Set Π_2: The path used by a packet is determined at the source based on the observed per-link delays for previous packets. In other words, under policy π, $a^\pi(n)$ depends on $a^\pi(1), \dots, a^\pi(n-1), (d_i^\pi(1), i \in a^\pi(1)), \dots, (d_i^\pi(n-1), i \in a^\pi(n-1))$, where $d_i^\pi(k)$ is the delay experienced on link i for the k-th packet (if this packet uses link i at all).

- Policy Set Π_3: Routing decisions are taken at each node in an adaptive manner. At a given time t, the packet is sent over a link selected based on all successes and failures observed on the various links before time t.

In the case of source-routing policies (in $\Pi_1 \cup \Pi_2$), if a transmission on a given link fails, the packet is retransmitted on the same link until it is successfully received (per-link delays are geometric random variables). On the contrary, in the case of hop-by-hop routing policies (in Π_3), the routing decisions at a given node can be adapted to the observed failures on a given link. For example, if transmission attempts on a given link failed, one may well decide to switch link and select a different next-hop node.

The regret $\mathfrak{R}_{\pi,N}$ of policy π up to the N-th packet is the expected difference of delays for the first N packets under π and under the policy that always selects the optimal path $a^\star$ for transmission:

$$\mathfrak{R}_{\pi,N} := \mathbb{E}\left[\sum_{n=1}^{N} D^\pi(n)\right] - ND_\theta(a^\star),$$

where $D^\pi(n)$ denotes the end-to-end delay of the n-th packet under policy π and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the random transmission outcomes and possible randomization in the policy π. The regret quantifies the performance loss due to the need to explore sub-optimal paths to learn the path with the minimum delay.

Objectives. The goal is to design online routing policies in Π_1, Π_2, and Π_3 that minimize regret over the first N packets. As it turns out, there are policies in any Π_j, $j = 1, 2, 3$, whose regrets scale as $\mathcal{O}(\log(N))$ when N grows large, and no policy can have a regret scaling as $o(\log(N))$. More specifically, our objective is to derive, for each $j = 1, 2, 3$, an asymptotic regret lower bound $c_j(\theta)\log(N)$ for policies in Π_j, and then propose simple policies whose regret upper bounds asymptotically approach that of the *optimal* algorithm, i.e., an algorithm whose regret matches the lower bound in Π_j. As we shall discuss later, such an algorithm exists. Therefore, by comparing $c_1(\theta)$, $c_2(\theta)$, and $c_3(\theta)$, we can quantify the potential performance improvements taking routing decisions at each hop rather than at the source only, and observing per-link delays rather than end-to-end delays.

4.3 Regret Lower Bounds

4.3.1 Source-Routing with Bandit Feedback

Consider routing polices in Π_1 that make routing decisions at source. Denote by $\psi_\theta^a(k)$ the probability that the delay of a packet sent on path a is k slots, and by $h(a)$ the length (or number of links) of path a. The end-to-end delay is the sum of several independent random geometric variables. If we assume that $\theta_i \neq \theta_j$ for $i \neq j$, according to [80], we have for all $k \geq h(a)$,

$$\psi_\theta^a(k) = \sum_{i\in a}\Bigg(\prod_{j\in a, j\neq i}\frac{\theta_j}{\theta_j - \theta_i}\Bigg)\theta_i(1-\theta_i)^{k-1},$$

i.e., the path delay distribution is a weighted average of the individual link delay distributions, where the weights can be negative but always sum to one.

The next theorem provides the fundamental performance limit of online routing policies in Π_1.

Theorem 4.1. *For all θ and for any uniformly good policy $\pi \in \Pi_1$,*

$$\liminf_{N\to\infty} \frac{\mathfrak{R}_{\pi,N}}{\log(N)} \geq c_1(\theta),$$

where $c_1(\theta)$ is the infimum of the following optimization problem:

$$\inf_{x\geq 0} \sum_{a\in\mathcal{A}} x_a \Delta_a \tag{4.1}$$

$$\text{subject to:} \quad \inf_{\lambda\in B_1(\theta)} \sum_{a\neq a^\star} x_a \sum_{k=h(a)}^{\infty} \psi_\theta^a(k) \log \frac{\psi_\theta^a(k)}{\psi_\lambda^a(k)} \geq 1,$$

$$\textit{with} \quad B_1(\theta) = \left\{\lambda : \{\lambda_i, i\in a^\star\} = \{\theta_i, i \in a^\star\},\ \min_{a\in\mathcal{A}} D_\lambda(a) < D_\lambda(a^\star)\right\}.$$

It is important to observe that in the definition of $B_1(\theta)$, the equality $\{\lambda_i, i \in a^\star\} = \{\theta_i, i \in a^\star\}$ is a set equality, i.e., order does not matter (e.g., if $a^\star = \{1,2\}$, the equality means that either $\lambda_1 = \theta_1, \lambda_2 = \theta_2$ or $\lambda_1 = \theta_2, \lambda_2 = \theta_1$). The proof of Theorem 4.1 follows similar steps as in the proof of Theorem 3.4, but also requires a property for geometric random variables established in Lemma 4.4 in Appendix 4.I.

Remark 4.1. *The difference between set of bad parameters in Theorem 4.1 and Theorem 3.4 comes from the different nature of Bernoulli and geometric distributions. By comparing bad parameter sets in the two theorems, we may conclude that bandit feedback in the case of geometrically distributed rewards provides relatively more information than in the case of Bernoulli rewards.*

4.3.2 Source-Routing with Semi-Bandit (Per-Link) Feedback

We now consider routing policies in Π_2 that make decisions at the source, but receive feedback on the individual link delays on the chosen path. Let $\texttt{KLG}(u, v)$ denote the KL-divergence between two geometric distributions with parameters u and v:

$$\texttt{KLG}(u,v) := \sum_{k\geq 1} u(1-u)^{k-1} \log \frac{u(1-u)^{k-1}}{v(1-v)^{k-1}}.$$

The next theorem provides the regret lower bound for online routing policies in Π_2:

Theorem 4.2. *For all θ and for any uniformly good policy $\pi \in \Pi_2$,*

$$\liminf_{N\to\infty} \frac{\mathfrak{R}_{\pi,N}}{\log(N)} \geq c_2(\theta),$$

where $c_2(\theta)$ is the infimum of the following optimization problem:

$$\begin{aligned} &\inf_{x\geq 0} \sum_{a\in\mathcal{A}} x_a \Delta_a && (4.2)\\ &\text{subject to:} \inf_{\lambda\in B_2(\theta)} \sum_{a\neq a^\star} x_a \sum_{i\in a} \mathtt{KLG}(\theta_i,\lambda_i) \geq 1, \end{aligned}$$

$$\textit{with} \qquad B_2(\theta) = \Big\{\lambda : \lambda_i = \theta_i, \forall i \in a^\star,\ \min_{a\in\mathcal{A}} D_\lambda(a) < D_\lambda(a^\star)\Big\}.$$

Similarly to Theorem 3.1, Theorem 4.2 can be seen as a direct consequence of [39, Theorem 1] (the problem can be easily mapped to a controlled Markov chain). We therefore omit its proof.

We also note that similarly to the case of Bernoulli rewards in Chapter 3, it holds that $c_1(\theta) \geq c_2(\theta)$, since the lower bounds we derive are tight and getting semi-bandit feedback can be exploited to design smarter routing policies than those we can devise using bandit feedback (i.e., $\Pi_1 \subset \Pi_2$).

Remark 4.2. *The asymptotic lower bound proposed in [34] has a similar expression to ours, but the set $B_2(\theta)$ is replaced by $B_2'(\theta) = \bigcup_{i\in E}\{\lambda : \lambda_j = \theta_j, \forall j \neq i, \min_{a\in\mathcal{A}} D_\lambda(a) < D_\lambda(a^\star)\}$. Note that $B_2'(\theta) \subset B_2(\theta)$, which implies that the lower bound derived in [34] is smaller than ours. In other words, we propose a regret lower bound that improves that in [34]. Furthermore, our bound is tight (it cannot be improved further).*

4.3.3 Hop-by-hop Routing

Finally, we consider routing policies in Π_3. These policies are more involved to analyze as the routing choices may change at any intermediate node in the network, and they are also more complex to implement. Surprisingly, the next theorem states that the regret lower bound for hop-by-hop routing policies is the same as that derived for strategies in Π_2 (source-routing with semi-bandit feedback). In other words, we cannot improve the performance by taking routing decisions at each hop.

Theorem 4.3. *For all θ and for any uniformly good policy $\pi \in \Pi_3$,*

$$\liminf_{N\to\infty} \frac{\mathfrak{R}_{\pi,N}}{\log(N)} \geq c_3(\theta) = c_2(\theta).$$

The proof of Theorem 4.3 is more involved than those of previous theorems, since in the hop-by-hop case, the chosen path could change at intermediate nodes. To overcome this difficulty, we introduce another notion of regret corresponding to the achieved throughput (i.e., the number of packets successfully received by the destination per unit time), which we refer to as the *throughput regret.* The proof uses the results of [39] for throughput regret, but also relies on Lemma 4.2, which provides an asymptotic relationship between $\mathfrak{R}_{\pi,N}$ and the throughput regret.

Remark 4.3. *Theorem 4.3, together with the tightness of our regret lower bounds, implies that in spite of exploiting additional feedback, hop-by-hop routing policies cannot yield better regret than source routing policies at least asymptotically when the number of transmissions grows large. Hop-by-hop routing could provide a significant advantage if the link qualities change on a fast time scale. To achieve a low regret in this situation, an efficient algorithm should not try to retransmit many times over a link whose quality went from good to bad.*

4.3.4 Numerical Example

There are examples of network topologies where the above asymptotic regret lower bounds can be explicitly computed. One such example is the line network[5]; see Figure 4.1(a) for an instance of line network. Notice that in line networks, the optimal routing policy consists in selecting the best link in each hop. The following lemma is immediate:

Lemma 4.1. *For any line network with m hops, we have:*

$$c_1(\theta) \geq \sum_{i \notin a^\star} \Big(\max_{p: i \in a} \sum_{k=m}^{\infty} \psi_\theta^a(k) \log \frac{\psi_\theta^a(k)}{\psi_{\vartheta^i}^a(k)} \Big)^{-1} \left(\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}} \right), \tag{4.3}$$

$$c_2(\theta) = c_3(\theta) = \sum_{i \notin a^\star} \frac{1}{\mathtt{KLG}(\theta_i, \theta_{\zeta(i)})} \left(\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}} \right),$$

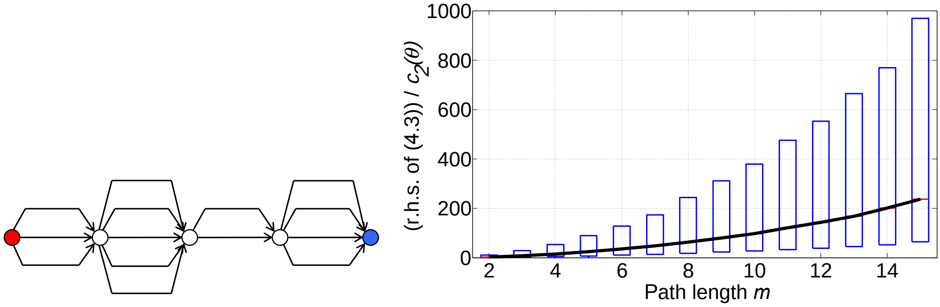
where $\zeta(i)$ is the best link on the same hop as link i, and where ϑ^i is a vector of link parameters defined as $\vartheta^i_j = \theta_j$ if $j \neq i$, and $\vartheta^i_i = \theta_{\zeta(i)}$.

As a consequence of the above lemma, we can establish the following result on the scaling of regret with problem parameters:

Proposition 4.1. *There exist problem instances in line networks with arbitrarily small $\theta_{\min}$, for which the regret of any uniformly good policy in $\Pi_2 \cup \Pi_3$ is $\Omega\Big(\frac{d-m}{\Delta_{\min}\theta_{\min}^2} \log(N)\Big)$.*

For line networks, both $c_1(\theta)$ and $c_2(\theta)$ scale linearly with the number of links in the network. In Figure 4.1(b), we present the median, along with 25% and 75% quantiles, for the ratio of the lower bound of $c_1(\theta)$ (i.e., the right-hand side of (4.3)) to $c_2(\theta)$ for various values of θ (we randomly generated 10^4 link parameters θ) as a function of the network diameter m in a simple line network, which has two links in the first hop and one link in the rest of hops and hence $d = m + 1$. These results suggest that collecting semi-bandit feedback (per-link delays) can significantly improve the performance of routing policies. The gain is significant even for fairly small networks.

[5] A line network is a graph of vertices $\{1, \ldots, n\}$, where the neighbors of i are $i-1$ and $i+1$.



(a) (b)

Figure 4.1: (a) A line network, (b) Semi-bandit vs. bandit feedback: Box-plot for the ratio of the lower bound of $c_1(\theta)$ (the right-hand side of (4.3)) to $c_2(\theta)$ for a line network. The plot shows the median (black curve), 25% quantile, and 75% quantile.

4.4 Routing Policies for Semi-bandit Feedback

Theorems 4.1-4.2-4.3 indicate that within the first N packets, the total amount of packets routed on a sub-optimal path a should be of the order of $x_a^\star \log(N)$, where $x_a^\star$ is the optimal solution of the optimization problems in (4.1) and (4.2). In [39], the authors present policies that achieve the regret bounds of Theorems 4.1-4.2-4.3 (see [39, Theorem 2]). These policies suffer from two problems: firstly, they are computationally infeasible for large problems since their implementation involves solving in each round a semi-infinite linear program [40] similar to those providing the regret lower bounds (defined in (4.1) and (4.2)). Secondly, these policies have no finite-time performance guarantees, and numerical experiments suggest that their finite-time performance on typical problems is rather poor.

In this section, we present online routing policies for semi-bandit feedback, which are simple to implement, yet approach the performance limits identified in the previous section. We further analyze their regret and show that they outperform existing algorithms. To present our policies, we introduce additional notations. Under a given policy, we let $t_i(n)$ be the total number of transmission attempts (including retransmissions) on link i before the n-th packet is sent. We define $\hat\theta_i(n)$ the empirical success rate of link i estimated over the transmissions of the first $(n-1)$ packets. Furthermore, we define the corresponding vectors $t(n) = (t_i(n))_{i\in E}$ and $\hat\theta(n) = (\hat\theta_i(n))_{i\in E}$.

Note that the proposed policies and regret analysis presented in this section directly apply for generic combinatorial optimization problems with linear objective function and geometrically distributed rewards.

Index	Type	Computation	Algorithm
b_a	Path	Line search	GeoCombUCB-1
c_a	Path	Explicit	GeoCombUCB-2
ω_i	Link	Line search	KL-SR

Table 4.2: Summary of indexes.

4.4.1 Path and Link Indexes

The proposed policies rely on indexes attached either to individual links or paths. Next we introduce three indexes used in our policies. They depend on the round, i.e., on the number n of the packet to be sent, the number of times a link has been sampled, and the estimated link parameters $\hat{\theta}(n)$. The three indexes and their properties (i.e., in which policy they are used and how one can compute them) are summarized in Table 4.2. Let $n \geq 1$ and assume that the n-th packet is to be sent. The indexes are defined as follows.

Path Indexes

Let $\lambda \in (0,1]^d$, $t \in \mathbb{N}^d$, and $n \in \mathbb{N}$. The first path index, denoted by $b_a(n, \lambda, t)$ for path $a \in \mathcal{A}$, is motivated by the index b_a presented in Chapter 3. $b_a(n, \lambda, t)$ is defined as the infimum of the following optimization problem:

$$\begin{aligned} \inf_{u \in (0,1]^d} \quad & a^\top u^{-1} \\ \text{subject to:} \quad & \sum_{i \in a} t_i \texttt{kl}(\lambda_i, u_i) \leq f_1(n), \\ & u_i \geq \lambda_i, \ \ \forall i \in E, \end{aligned}$$

where $f_1(n) = \log(n) + 4m \log(\log(n))$, and for all $u, v \in [0,1]$, $\texttt{kl}(u, v)$ is the KL information number between two Bernoulli distributions with respective means u and v, i.e., $\texttt{kl}(u, v) = u \log(u/v) + (1-u) \log((1-u)/(1-v))$.

The second index is denoted by $c_a(n, \lambda, t)$ and defined for path $a \in \mathcal{A}$ as:

$$c_a(n, \lambda, t) = a^\top \lambda^{-1} - \sqrt{\sum_{i \in a} \frac{2 f_1(n)}{t_i \lambda_i^3}}.$$

Similarly to Theorem 3.6, the next theorem provides generic properties of the two indexes b_a and c_a.

Theorem 4.4. *(i) For all $n \geq 1$, $a \in \mathcal{A}$, $\lambda \in (0,1]^d$, and $t \in \mathbb{N}^d$, we have $b_a(n, \lambda, t) \geq c_a(n, \lambda, t)$.*

(ii) There exists a constant $K_m > 0$ depending on m only such that for all $a \in \mathcal{A}$ and $n \geq 2$:

$$\mathbb{P}(b_a(n, \hat{\theta}(n), t(n)) > a^\top \theta) \leq K_m n^{-1} (\log(n))^{-2}.$$

Algorithm 4.1 GeoCombUCB

```
for $n \geq 1$ do
    Select path $a(n) \in \arg\min_{a\in\mathcal{A}} \xi_a(n)$ (ties are broken arbitrarily), where $\xi_a(n) = b_a(n)$ for GeoCombUCB-1, and $\xi_a(n) = c_a(n)$ for GeoCombUCB-2.
    Collect feedback on links $i \in a(n)$, and update $\hat{\theta}_i(n)$ for $i \in a(n)$.
end for
```

Algorithm 4.2 KL-SR [33]

```
for $n \geq 1$ do
    Select path $a(n) \in \arg\min_{a\in\mathcal{A}} a^\top \omega(n)$ (ties are broken arbitrarily).
    Collect feedback on links $i \in a(n)$, and update $\hat{\theta}_i(n)$ for $i \in a(n)$.
end for
```

Corollary 4.1. *We have:*

$$\sum_{n\geq 1} \mathbb{P}(b_{a^\star}(n, \hat{\theta}(n), t(n)) > \ a^{\star\top}\theta^{-1}) \leq 1 + K_m \sum_{n\geq 2} n^{-1}(\log(n))^{-2} < \infty.$$

Link Index

Our third index is a link index. For $n, t \in \mathbb{N}$ and $\lambda \in (0, 1]$, the index $\omega_i(n, \lambda, t)$ of link $i \in E$ is defined as:

$$\omega_i(n, \lambda, t) = \min\Big\{\frac{1}{u} : \ u \in [\lambda, 1], \ \ t\texttt{kl}\big(\lambda, u\big) \leq f_2(n)\Big\},$$

where $f_2(n) = \log(n) + 4\log(\log(n))$.

4.4.2 Routing policies

We present three routing policies, referred to as GeoCombUCB-1, GeoCombUCB-2, and KL-SR, respectively. For the transmission of the n-th packet, GeoCombUCB-1 (resp. GeoCombUCB-2) selects the path a with the smallest index $b_a(n) := b_a(n, \hat{\theta}(n), t(n))$ (resp. $c_a(n) := c_a(n, \hat{\theta}(n), t(n))$). KL-SR was initially proposed in [33] and for the transmission of the n-th packet, it selects the path $a(n) \in \arg\min_{a\in\mathcal{A}} a^\top \omega(n)$, where $\omega(n) = (\omega_i(n))_{i\in E}$ and $\omega_i(n) := \omega_i(n, \hat{\theta}_i(n), t_i(n))$. The pseudo-code of GeoCombUCB and KL-SR are presented in Algorithm 4.1 and Algorithm 4.2, respectively.

In the following theorems, we provide a finite-time analysis of GeoCombUCB and KL-SR, and show the optimality of KL-SR in line networks. Define $\varepsilon = (1-2^{-\frac{1}{4}})\frac{\Delta_{\min}}{D^+}$.

Theorem 4.5. *For all $N \geq 1$, under policies $\pi \in \{\texttt{GeoCombUCB-1}, \texttt{GeoCombUCB-2}\}$ we have:*

$$\mathfrak{R}_{\pi,N} \leq \frac{32d\sqrt{m}f_1(N)}{\Delta_{\min}\theta_{\min}^2} + 2D^+ \left(2K_m + \sum_{i\in E} \frac{1}{\varepsilon^2\theta_i^2}\right).$$

Hence, $\mathfrak{R}_{\pi,N} = \mathcal{O}\left(\frac{d\sqrt{m}}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$ when $N \to \infty$.

Theorem 4.6. *For all $N \geq 1$, under policy $\pi = \texttt{KL-SR}$ we have:*

$$\mathfrak{R}_{\pi,N} \leq \frac{360dmf_2(N)}{\Delta_{\min}\theta_{\min}^2} + 2D^+ \left(4m + \sum_{i\in E} \frac{1}{\varepsilon^2\theta_i^2}\right).$$

Hence, $\mathfrak{R}_{\pi,N} = \mathcal{O}\left(\frac{dm}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$ when $N \to \infty$.

Remark 4.4. *The regret bound of* KL-SR *scales as $\mathcal{O}(m)$ while that of* GeoCombUCB *scales as $\mathcal{O}(\sqrt{m})$. Indeed, the index ω_i used in* KL-SR *ignores the statistical independence of delays of various links in a given path, whereas the indexes b_a and c_a in* GeoCombUCB *use this independence and yield smaller confidence intervals.*

The proof of Theorem 4.6 is completely different from the regret analysis of KL-SR in [33]; it relies on Lemma 4.3, which provides a sharp lower bound for the index ω_i, and borrows some ideas from [62, Theorem 5].

Remark 4.5. *Theorem 4.6 holds even when the delays on the various links are not independent, as in [62].*

The proposed policies have better performance guarantees than existing routing algorithms. Indeed, as shown in Appendix 4.H, the best known regret upper bound for the CUCB algorithm [61] is $\mathcal{O}\left(\frac{dm}{\Delta_{\min}\theta_{\min}^3}\log(N)\right)$, which constitutes a weaker performance guarantee than those of our routing policies. The numerical experiments presented in the next section will confirm the superiority of GeoCombUCB and KL-SR over CUCB.

The next proposition states that KL-SR is asymptotically optimal in line networks:

Proposition 4.2. *In line networks, the regret under $\pi = \texttt{KL-SR}$ satisfies*

$$\limsup_{N\to\infty} \frac{\mathfrak{R}_{\pi,N}}{\log(N)} \leq c_2(\theta).$$

Hence, $\mathfrak{R}_{\pi,N} = \mathcal{O}\left(\frac{d-m}{\Delta_{\min}\theta_{\min}^2}\log(N)\right)$ when $N \to \infty$.

Remark 4.6. *When the link parameters smoothly evolve over time, we can modify the proposed routing policies so that routing decisions are based on past choices and observations over a sliding window consisting of a fixed number of packets, as considered in [81] and [70].*

4.4.3 Implementation

Next we discuss the implementation of our routing policies and give simple methods to compute $b_a(n,\lambda,t)$, $c_a(n,\lambda,t)$, $\omega_i(n,\lambda,t)$ given a, i, n, λ, and t. The path index c_a is explicit and easy to compute. The computation of link index ω_i is also straightforward as it amounts to finding the roots of a strictly convex and increasing function in one variable (note that $v \mapsto \mathtt{kl}(u,v)$ is strictly convex and increasing for $v \geq u$). Hence, the index ω_i can be computed by a simple line search. The path index $b_a(n,\lambda,t)$ can also be computed using a simple line search, as shown below.

Introduce $J_a(\lambda) = \{i \in a : \lambda_i \neq 1\}$, and for $\gamma > 0$ define:

$$F(\gamma,\lambda,n,t) = \sum_{i\in J_a(\lambda)} t_i \mathtt{kl}(\lambda_i, g(\gamma,\lambda_i,t_i)),$$

$$\text{with} \quad g(\gamma,\lambda_i,t_i) = \frac{1}{2\gamma t_i}\Big(\gamma\lambda_i t_i - 1 + \sqrt{(1-\gamma\lambda_i t_i)^2 + 4\gamma t_i}\Big).$$

Now, following similar lines as in the proof of Theorem 3.7, we can prove that:

Proposition 4.3. *(i) $\gamma \mapsto F(\gamma,\lambda,n,t)$ is strictly increasing, and $F(\mathbb{R}_+,\lambda,n,t) = \mathbb{R}_+$. (ii) If $J_a(\lambda) = \emptyset$, $b_a(n,\lambda,t) = \sum_{i\in E} a_i$. Otherwise, let $\gamma^\star$ be the unique solution to $F(\gamma,\lambda,n,t) = f_1(n)$. Then,*

$$b_a(n,\lambda,t) = \sum_{i\in E} a_i - |J_a(\lambda)| + \sum_{i\in I_a(\lambda)} g(\gamma^\star,\lambda_i,t_i).$$

As stated in Proposition 4.3, $\gamma^\star$ can be computed efficiently by a simple line search and b_a is easily deduced. We thus have efficient methods to compute the three indexes. To implement our policies, we then need to find in each round, the path minimizing the index (or the sum of link indexes along the path for KL-SR). KL-SR can be implemented (in a distributed fashion) using the Bellman-Ford algorithm, and its complexity is $\mathcal{O}(|V|d)$ in each round. GeoCombUCB-1 and GeoCombUCB-2 are more computationally involved than KL-SR and have complexity $\mathcal{O}(|\mathcal{A}|)$ in each round.

4.5 Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of the proposed source-routing policies to that of the CUCB algorithm [61] and TS applied to our online routing problem. The CUCB algorithm is an index policy in Π_2 (the set of source-routing policies with semi-bandit feedback) that selects path $a(n)$ for the transmission of the n-th packet:

$$a(n) \in \arg\min_{a\in\mathcal{A}} \sum_{i\in a} \frac{1}{\hat\theta_i(n) + \sqrt{1.5\log(n)/t_i(n)}}.$$

We consider a grid network whose topology is depicted in Figure 4.3(a), where the node in red (resp. blue) is the source (resp. the destination). In this network, there are $\binom{6}{3} = 20$ possible paths from the source to the destination. Let us compare these algorithms in terms of their per-packet complexity. The complexity of GeoCombUCB-1 and GeoCombUCB-2 is $\mathcal{O}(|\mathcal{A}|)$, whereas that of KL-SR, CUCB, and TS is $\mathcal{O}(|V|d)$.

In Figures 4.2(a)-(b), we plot the regret against the number of the packets N under the various routing policies, and for two sets of link parameters θ. For each set, we choose a value of $\theta_{\min}$ and generate the values of θ_i independently, uniformly at random in $[\theta_{\min}, 1]$. The results are averaged over 100 independent runs, and the 95% confidence intervals are shown using the grey area around curves. The three proposed policies outperform CUCB, and GeoCombUCB-1 attains the smallest regret amongst the proposed policies. The comparison between GeoCombUCB-2 and KL-SR is more subtle and depends on the link parameters: While in Figure 4.2(a) KL-SR significantly outperforms GeoCombUCB-2, they attain regrets growing similarly for the link parameter of Figure 4.2(b). Yet there are some parameters for which KL-SR is significantly outperformed by GeoCombUCB-2. KL-SR seems to perform better than GeoCombUCB-2 in scenarios where $\Delta_{\min}$ is large. TS performs slightly better than GeoCombUCB-1 on average. Its regret, however may not be well concentrated around the mean for some link parameters, as in Figure 4.2(b). Furthermore, the regret analysis of TS for shortest-path routing with general topologies is an open problem.

4.5.1 A distributed hop-by-hop routing policy

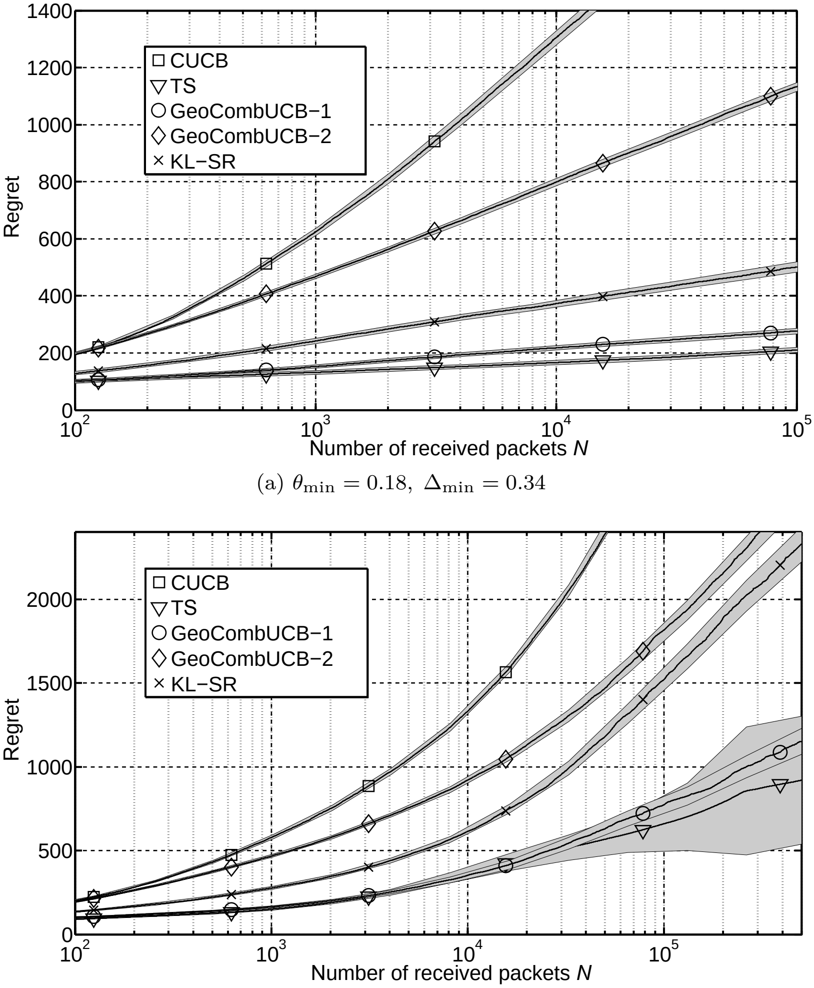
Motivated by the Bellman-Ford implementation of the KL-SR algorithm, we propose KL-HHR, a distributed routing policy that is a hop-by-hop version of the KL-SR algorithm and hence belongs to the set of policies Π_3. We first introduce the necessary notations. For any node $v \in V$, we let $\mathcal{A}_v$ denote the set of loop-free paths from node v to the destination. For any time slot τ, we denote by $n(\tau)$ the packet number that is about to be sent or is already in the network. For any link i, let $\tilde{\theta}_i(\tau)$ be the empirical success rate of link i *up to time slot* τ, that is $\tilde{\theta}_i(\tau) = s_i(n(\tau))/t'_i(\tau)$, where $t'_i(\tau)$ denotes the total number of transmission attempts on link i up to time slot τ. Moreover, with slight abuse of notation, we denote the index of link i at time τ by $\omega_i(\tau, \tilde{\theta}_i(\tau))$. Note that by definition $t'_i(\tau) \geq t_i(n)$ and $\tilde{\theta}_i(\tau)$ is a more accurate estimate of θ_i than $\hat{\theta}_i(n(\tau))$.

We define $\ell_v(\tau)$ as the minimum *cumulative index* from node v to the destination:

$$\ell_v(\tau) = \min_{a \in \mathcal{A}_v} \sum_{i \in a} \omega_i(\tau, \tilde{\theta}_i(\tau)).$$

The index $\ell_v(\tau)$ can be computed using the Bellman-Ford algorithm. KL-HHR (Algorithm 4.3) works based on the following idea: Assuming that the current packet is at node v at time τ, send it to node v' such that $\omega_{(v,v')}(\tau, \tilde{\theta}_v(\tau)) + \ell_{v'}(\tau)$ is minimal over all outgoing links of node v.

We compare the performance of KL-HHR, KL-SR, and TS through numerical experiments for a network shown in Figure 4.3(b), in which there are 40 links and



(a) $\theta_{\min} = 0.18,\ \Delta_{\min} = 0.34$

(b) $\theta_{\min} = 0.1,\ \Delta_{\min} = 0.08$

Figure 4.2: Regret versus number of received packets

413 possible paths between the source (in red) and the destination (in blue). Figures 4.4(a)-(b) display the regret under KL-HHR, KL-SR, and TS, averaged over 100 independent runs, against the number of the packets N for two sets of link parameters θ. These parameters are generated similarly to the previous experiments. As expected, KL-HHR outperforms KL-SR in both scenarios since it can change routing decisions dynamically at intermediate nodes thus avoiding retransmissions on bad links when they are discovered. Note however that, in both scenarios, the regret of

Algorithm 4.3 KL-HHR for node v

```
for $\tau \geq 1$ do
    Select link $(v, v') \in E$, where
        $v' \in \arg \min_{w \in V : (v,w) \in E} \left( \omega_{(v,w)}(\tau, \tilde{\theta}_v(\tau)) + \ell_w(\tau) \right).$
    Update index of the link $(v, v')$.
end for
```

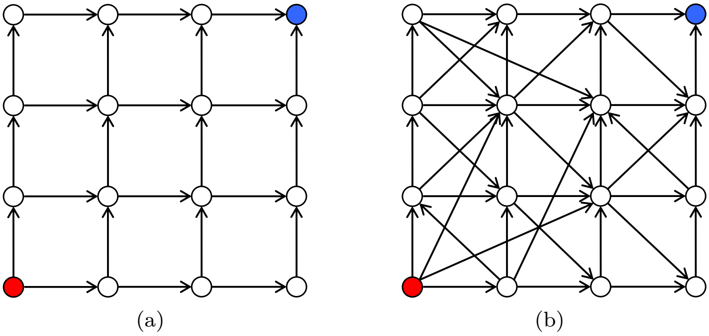


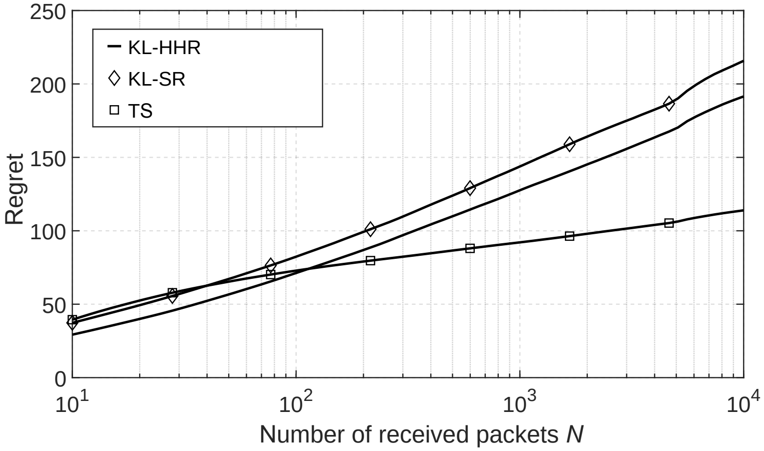
Figure 4.3: Network topologies

both KL-HHR and KL-SR seem to grow at the same rate as the number of received packets grows large. Moreover, TS would outperform KL-HHR asymptotically, i.e., as the number of received packets grows large. On the other hand, in scenarios where $\theta_{\min}$ is very small, if the number of total packets N is not very large, KL-HHR could outperform TS; see, e.g., Figure 4.4(b).

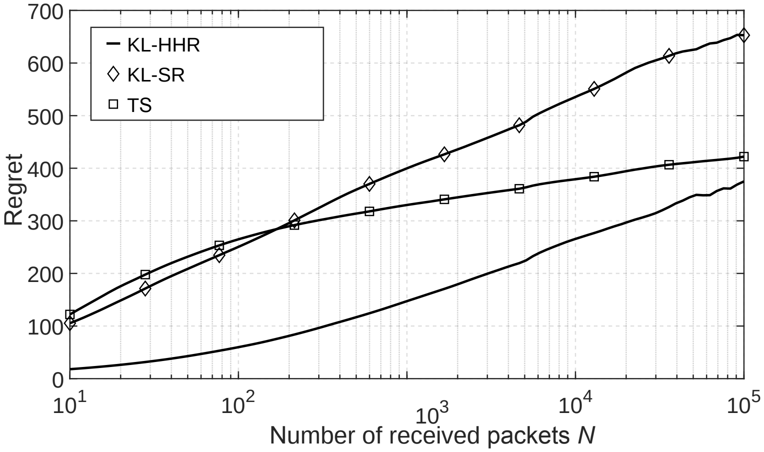
The regret analysis of KL-HHR is left for future work.

4.6 Summary

In this chapter we investigated online shortest-path routing problems in networks with stochastic link delays. We derived asymptotic regret lower bounds for source routing policies under bandit and semi-bandit feedback, and for hop-by-hop routing policies. We further showed that the regret lower bounds for source routing policies with semi-bandit feedback and that for hop-by-hop routing policies are identical. We then proposed two online source routing policies, namely GeoCombUCB-1 and GeoCombUCB-2, and provided a finite-time analysis of their regret. Moreover, we improve the regret upper bound of KL-SR [33]. These routing policies strike an interesting trade-off between computational complexity and performance, and exhibit better regret upper bounds than state-of-the-art algorithms. Furthermore, through



(a) $\theta_{\min} = 0.014$



(b) $\theta_{\min} = 0.0056$

Figure 4.4: Regret versus number of received packets

numerical experiments we demonstrated that these policies outperform state-of-the-art algorithms in practice. As future work, we plan to propose practical algorithms with provable performance bounds for hop-by-hop routing and source-routing with bandit feedback. Furthermore, we would like to study the effect of delayed feedback on the performance as studied in, e.g., [82].

4.A Proof of Theorem 4.3

To prove the theorem, we first define another notion of regret corresponding to the achieved throughput (i.e., the number of packets successfully received by the destination per unit time). The throughput regret is introduced to ease the analysis since computing the throughput regret is easier in the hop-by-hop case. Define $\mu_\theta(a)$ as the average throughput on path a given link success rates θ: $\mu_\theta(a) = 1/D_\theta(a)$. The fontswitch*t*hroughput regret $\mathfrak{S}_{\pi,T}$ of π over time horizon T is: $\mathfrak{S}_{\pi,T} := T\mu_\theta(a^\star) - \mathbb{E}\left[N^\pi(T)\right],$ where $N^\pi(T)$ is the number of packets received up to time T under policy π. Lemma 4.2, stated at the end of the proof, provides the relation between asymptotic bound on $\mathfrak{R}_{\pi,N}$ and $\mathfrak{S}_{\pi,T}$.

Now we are ready to prove Theorem 4.3. The proof relies on the framework of Graves and Lai [39]. To apply their result, we construct the following controlled Markov chain. We let the state of the Markov chain be the packet location. The action is the selected outgoing link. The transitions between two states take one time slot – the time to make a transmission attempt. Hence, the transition probability between state x and y with the action of using link i is denoted by (where $y \neq x$) $P_\theta^i(x, y) = \theta_i$ if link i connects node x and y, and is zero otherwise. On the other hand, the probability of staying at the same state is the transmission failure probability on link i if link i is an outgoing link, that is $P_\theta^i(x, x) = 1 - \theta_i$ if link i is an outgoing link, and is zero otherwise.

We assume that the packet is injected at the source immediately after the previous packet is successfully delivered, and we are interested in counting the number of successfully delivered packets. In order not to count the extra time slot we will spend at the destination, we use a single Markov chain state to represent both the source and the destination.

We give a reward of 1 whenever the packet is successfully delivered to the destination. Let $r(x, y, i)$ be the immediate reward after the transition from node x to node y under the action i, i.e., $r(x, y, i) = 1$ if y is the destination node and is zero otherwise (see Figure 4.5 for an example). Hence, $r(x, i)$ (i.e., the reward at state x with action i) is

$$r(x, i) = \begin{cases} \theta_i & \text{if link } i \text{ connects node } x \text{ and the destination;} \\ 0 & \text{otherwise.} \end{cases}$$

The stationary control law prescribes the action at each state, i.e., the outgoing link at each node. A stationary control law of this Markov chain is then a path a in the network, and we assign arbitrary actions to the nodes that are not on the path a. The maximal irreducibility measure is then to assign measure zero to the nodes that are not on the path a, and a counting measure to the nodes on the path a. The Markov chain is irreducible with respect to this maximal irreducibility measure, and the stationary distribution of the Markov chain under path a is

$$\pi_\theta^a(x) = \frac{\frac{1}{\theta_{a(x)}}}{\sum_{i\in a} \frac{1}{\theta_i}} \mathbb{I}\{\text{if node } x \text{ is on the path } a\},$$

where $a(x)$ denotes the link we choose at node x. The long-run average reward of the Markov chain under control law a is $\sum_x \pi_\theta^a(x) r(x, a(x)) = 1/\sum_{i \in a} \frac{1}{\theta_i} = \mu_\theta(a)$. The optimal control law is then $a^\star$ with long run average reward $\mu_\theta(a^\star)$.

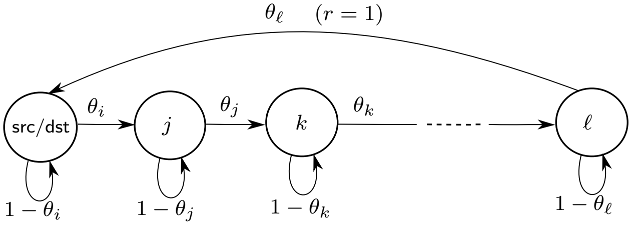


Figure 4.5: A Markov chain example under a control law a

The throughput regret of a policy $\pi \in \Pi_3$ for this controlled Markov chain at time T is

$$\mathfrak{S}_{\pi,T} = T\mu_\theta(a^\star) - \mathbb{E}_\theta[\sum_{t=1}^{T} r(x_t, \pi(t, x_t))], \tag{4.4}$$

where x_t is the state at time t and $\pi(t, x_t)$ is the corresponding action for state x_t at time t. To this end, we construct a controlled Markov chain that corresponds to the hop-by-hop routing in the network. Now define $I^a(\theta, \lambda)$ as the KL information number for a control law a:

$$\begin{aligned} I^a(\theta, \lambda) &= \sum_x \pi_\theta^a(x) \sum_y P_\theta^{a(x)}(x, y) \log \frac{P_\theta^{a(x)}(x, y)}{P_\lambda^{a(x)}(x, y)} \\ &= \sum_x \pi_\theta^a(x) \Big(\theta_{a(x)} \log \frac{\theta_{a(x)}}{\lambda_{a(x)}} + (1 - \theta_{a(x)}) \log \frac{1 - \theta_{a(x)}}{1 - \lambda_{a(x)}} \Big) \\ &= \mu_\theta(a) \sum_{i \in a} \frac{\mathtt{kl}(\theta_i, \lambda_i)}{\theta_i} = \mu_\theta(a) \sum_{i \in a} \mathtt{KLG}(\theta_i, \lambda_i), \end{aligned}$$

where we used Lemma A.4 in the last equality. Since $I^a(\theta, \lambda) = 0$ iff $\theta_i = \lambda_i$ for all $i \in a$, the set $B_2(\theta)$ of bad parameters is:

$$\begin{aligned} B_2(\theta) &= \Big\{ \lambda : \lambda_i = \theta_i, \forall i \in a^\star, \ \max_{a \in \mathcal{A}} \mu_\lambda(a) > \mu_\lambda(a^\star) \Big\} \\ &= \Big\{ \lambda : \lambda_i = \theta_i, \forall i \in a^\star, \ \min_{a \in \mathcal{A}} D_\lambda(a) < D_\lambda(a^\star) \Big\}. \end{aligned}$$

Applying [39, Theorem 1], we get: $\liminf_{T \to \infty} \mathfrak{S}_{\pi,T}/\log(T) \geq c_3'(\theta)$, with

$$c_3'(\theta) = \inf \Big\{ \sum_{a \in \mathcal{A}} x_a(\mu_\theta(a^\star) - \mu_\theta(a)) : x \geq 0,$$

$$\inf_{\lambda\in B_2(\theta)}\sum_{a\neq a^\star} x_a\mu_\theta(a)\sum_{i\in a}\mathtt{KLG}(\theta_i,\lambda_i)\geq 1\Big\}.$$

By Lemma 4.2, $c_3(\theta)\geq c_3'(\theta)/\mu_\theta(a^\star)$. Lastly, observe that $\mu_\theta(a^\star)-\mu_\theta(a)=\mu_\theta(a^\star)\mu_\theta(a)(D_\theta(a)-D_\theta(a^\star))$. It then follows that $c_3'(\theta)/\mu_\theta(a^\star)=c_2(\theta)$ and therefore, $c_3(\theta)\geq c_2(\theta)$. On the other hand, $c_3(\theta)\leq c_2(\theta)$ since $\Pi_2\subset\Pi_3$. As a result, $c_3(\theta)=c_2(\theta)$ and the proof is completed. □

In Lemma 4.2 below we provide the connection between the throughput regret $\mathfrak{S}_{\pi,T}$ and delay regret $\mathfrak{R}_{\pi,N}$.

Lemma 4.2. *For any $\pi\in\Pi_i$, $i=1,2,3$, and any $\beta>0$ we have:*

$$\liminf_{T\to\infty}\frac{\mathfrak{S}_{\pi,T}}{\log(T)}\geq\beta \implies \mu_\theta(a^\star)\liminf_{N\to\infty}\frac{\mathfrak{R}_{\pi,N}}{\log(N)}\geq\beta.$$

Proof. Define $\mu^\star=\mu_\theta(a^\star)$ and $r_t=\sum_{n=1}^t(D^\pi(n)-D^\star)$. Define $\mathcal{F}_t$ the σ-algebra generated by $(a^\pi(n),(d_i^\pi(n),i\in a^\pi(n)))_{1\leq n\leq t}$, where $d_i^\pi(k)$ is the delay experienced on link i for the k-the packet under policy π. Then $a^\pi(t)$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[r_t-r_{t-1}|\mathcal{F}_{t-1}]$ equals

$$\mathbb{E}[D^\pi(t)-D^\star|\mathcal{F}_{t-1}]=D_\theta(a^\pi(t))-D^\star\geq 0,$$

so $(r_t)_{0\leq t\leq T}$ is a $\mathcal{F}_t$-submartingale. Since $T\leq\sum_{n=1}^{N^\pi(T)+1}D^\pi(n)$ and $\mu^\star=1/D^\star$, we have

$$T\mu^\star-N^\pi(T)\leq 1+\sum_{n=1}^{N^\pi(T)+1}(\mu^\star D^\pi(n)-1)=1+\mu^\star r_{N^\pi(T)+1}.$$

Since $(r_t)_{0\leq t\leq T}$ is a submartingale, $N^\pi(T)\leq T$ is a bounded stopping time, Doob's stopping theorem [83, Theorem 5.4.1] gives: $\mathbb{E}(r_{N^\pi(T)+1})\leq\mathbb{E}(r_{T+1})=\mathfrak{R}_{\pi,T+1}$. Taking expectations above yields:

$$\frac{\mathfrak{S}_{\pi,T}}{\log(T)}\leq\frac{1}{\log(T)}+\mu^\star\frac{\mathfrak{R}_{\pi,T+1}}{\log(T)}.$$

Letting $T\to\infty$ proves the result since $\frac{\log(T)}{\log(T+1)}\to 1$. □

4.B Proof of Proposition 4.1

Consider a problem instance with line topology in which $\theta_i=\alpha$ for all $i\notin a^\star$, and $\theta_i=\alpha+\alpha^2$ for all $i\in a^\star$ for some $\alpha\in(0,0.36]$. Hence, $\theta_i<0.5$ for all $i\in a^\star$. For any uniformly good policy $\pi\in\Pi_2\cup\Pi_3$, by Lemma 4.1 we have:

$$\liminf_{N\to\infty}\frac{\mathfrak{R}_{\pi,N}}{\log(N)}\geq\sum_{i\notin a^\star}\frac{1}{\mathtt{KLG}(\theta_i,\theta_{\zeta(i)})}\Big(\frac{1}{\theta_i}-\frac{1}{\theta_{\zeta(i)}}\Big)$$

$$
\begin{aligned}
&\geq \sum_{i \notin a^\star} \frac{1}{2(\theta_{\zeta(i)} - \theta_i)} = \sum_{i \notin a^\star} \frac{1}{2\theta_i \theta_{\zeta(i)}(\theta_i^{-1} - \theta_{\zeta(i)}^{-1})} \\
&= \frac{d-m}{2\alpha(\alpha+\alpha^2)(\alpha^{-1} - (\alpha+\alpha^2)^{-1})} \\
&= \frac{d-m}{2\alpha(\alpha+\alpha^2)\Delta_{\min}} \geq \frac{d-m}{4\alpha^2\Delta_{\min}} = \frac{d-m}{4\theta_{\min}^2\Delta_{\min}},
\end{aligned}
$$

where in the second inequality we used Lemma A.4 and

$$
\mathtt{kl}(u,v) \leq \frac{(u-v)^2}{v(1-v)} \leq \frac{2(u-v)^2}{v}
$$

for $v \leq 0.5$. This implies that the regret of any uniformly good policy $\pi \in \Pi_2 \cup \Pi_3$ for this problem instance is at least $\Omega\Big(\frac{d-m}{\Delta_{\min}\theta_{\min}^2}\log(N)\Big)$. □

4.C Proof of Theorem 4.4

Proof of statement (i). Let $a \in \mathcal{A}$, $n \in \mathbb{N}$, $t \in \mathbb{N}^d$, and $u, \lambda \in (0,1]^d$ with $u_i \geq \lambda_i$ for all i. By Cauchy-Schwarz inequality we have:

$$
\begin{aligned}
a^\top \lambda^{-1} - a^\top u^{-1} &= \sum_{i \in a} \frac{u_i - \lambda_i}{u_i \lambda_i} \\
&= \sum_{i \in a} \frac{\sqrt{t_i}(u_i - \lambda_i)}{\sqrt{u_i}} \frac{1}{\lambda_i \sqrt{t_i u_i}} \\
&\leq \sqrt{\sum_{i \in a} \frac{t_i (u_i - \lambda_i)^2}{u_i}} \sqrt{\sum_{i \in a} \frac{1}{t_i u_i \lambda_i^2}} \\
&\leq \sqrt{\sum_{i \in a} \frac{t_i (u_i - \lambda_i)^2}{u_i}} \sqrt{\sum_{i \in a} \frac{1}{t_i \lambda_i^3}},
\end{aligned}
$$

where we used $u_i \geq \lambda_i$ for all i in the last step. Using Lemma A.3, it then follows that

$$
a^\top \lambda^{-1} - a^\top u^{-1} \leq \sqrt{\sum_{i \in a} 2 t_i \mathtt{kl}(\lambda_i, u_i)} \sqrt{\sum_{i \in a} \frac{1}{t_i \lambda_i^3}}.
$$

Thus, $\sum_{i \in a} t_i \mathtt{kl}(\lambda_i, u_i) \leq f_1(n)$ implies:

$$
a^\top \lambda^{-1} - a^\top u^{-1} \leq \sqrt{\sum_{i \in a} \frac{2 f_1(n)}{t_i \lambda_i^3}},
$$

or equivalently, $a^\top u^{-1} \geq c_a(n, \lambda, t)$. Hence, by definition of $b_a(n, \lambda, t)$, we have $b_a(n, \lambda, t) \geq c_a(n, \lambda, t)$.

Proof of statement (ii). If the constraint $\sum_{i\in a} t_i(n)\texttt{kl}(\hat\theta_i(n),\theta_i) \le f_1(n)$ holds, then we have $b_a(n,\hat\theta(n),t(n)) \le a^\top\theta^{-1}$ by definition of b_a. Therefore, using Corollary B.1 ([68, Theorem 2]), there exists K_m such that for all $n \ge 2$:

$$\begin{aligned}\mathbb{P}(b_a(n,\hat\theta(n),t(n)) > a^\top\theta^{-1}) &\le \mathbb{P}\Big(\sum_{i\in a} t_i(n)\texttt{kl}(\hat\theta_i(n),\theta_i) > f_1(n)\Big)\\ &\le K_m n^{-1}(\log(n))^{-2},\end{aligned}$$

which concludes the proof. □

4.D Proof of Theorem 4.5

To prove the theorem, we borrow some ideas from the analysis of [62, Theorem 3].

Define $\kappa = (1-2^{-\frac{1}{4}})$ and $\varepsilon = \kappa\frac{\Delta_{\min}}{D^+}$. Note that the definition of D^+, together with the fact that $D_\theta(a^\star) > 0$, implies that $\varepsilon < \kappa$. For $s \in \mathbb{N}^d$ and $a \in \mathcal{A}$ define $h(s,a) = \sum_{i\in a}\frac{1}{s_i}$. Define $s_i(n) = t_i(n)\hat\theta_i(n)$ the number of packets routed through link i before the n-th packet is sent and $s(n) = (s_i(n))_{i\in E}$. To ease notation, define $h(n) = h(s(n),a(n))$.

Proof of Theorem 4.5. For any n, introduce the following events:

$$\begin{aligned} A_n &= \Big\{\sum_{i\in a^\star} t_i(n)\texttt{kl}(\hat\theta_i(n),\theta_i) > f_1(n)\Big\},\\ B_{n,i} &= \{a_i(n)=1,\ |\hat\theta_i(n)-\theta_i| \ge \varepsilon\theta_i\},\ \ B_n = \bigcup_{i\in E} B_{n,i},\\ F_n &= \{\Delta_{a(n)} \le (1-\kappa)^{-2}\theta_{\min}^{-1}\sqrt{2f_1(N)h(n)}\}.\end{aligned}$$

We first prove that $a(n) \ne a^\star$ implies: $n \in A_n \cup B_n \cup F_n$. Consider n such that $a(n) \ne a^\star$ and $A_n \cup B_n$ does not occur. By design of the algorithm, $\xi_{a(n)}(n) \le \xi_{a^\star}(n)$, and $\xi_{a^\star}(n) \le D^\star$ since A_n does not occur. By Theorem 4.4 we have $c_{a(n)}(n) \le b_{a(n)}(n)$. Hence $c_{a(n)}(n) \le D^\star$. This implies:

$$a(n)^\top\hat\theta(n)^{-1} - \sqrt{\sum_{i\in p}\frac{2f_1(n)}{s_i(n)\hat\theta_i(n)^2}} \le D^\star,$$

so that:

$$\Delta_{a(n)} \le a(n)^\top\theta^{-1} - a(n)^\top\hat\theta(n)^{-1} + \sqrt{\sum_{i\in a(n)}\frac{2f_1(n)}{s_i(n)\hat\theta_i(n)^2}}.$$

Since B_n does not occur, $\hat\theta(n)^{-1} \ge \theta^{-1}/(1+\varepsilon)$ and:

$$a(n)^\top\theta^{-1} - a(n)^\top\hat\theta(n)^{-1} \le \frac{a(n)^\top\theta^{-1}\varepsilon}{(1+\varepsilon)} \le D^+\varepsilon = \kappa\Delta_{\min} \le \kappa\Delta_{a(n)}.$$

Moreover, $\hat{\theta}_i(n) \geq \theta_{\min}(1-\kappa)$ for all $i \in a(n)$, and $f_1(n) \leq f_1(N)$ so:

$$\sum_{i \in a(n)} \frac{2f_1(n)}{s_i(n)\hat{\theta}_i(n)^2} \leq \frac{2f_1(N)h(n)}{(1-\kappa)^2\theta_{\min}^2}.$$

Hence:

$$\Delta_{a(n)} \leq \kappa\Delta_{a(n)} + \frac{\sqrt{2f_1(N)h(n)}}{(1-\kappa)\theta_{\min}},$$

and $\Delta_{a(n)} \leq (1-\kappa)^{-2}\theta_{\min}^{-1}\sqrt{2f_1(N)h(n)}$. Therefore, $n \in F_n$.

The regret $\mathfrak{R}_{\pi,N}$ is upper bounded by:

$$\mathbb{E}\Big[\sum_{n=1}^{N} \Delta_{a(n)}\Big] \leq \mathbb{E}\Big[\sum_{n=1}^{N} \Delta_{a(n)}(\mathbb{I}\{A_n\} + \mathbb{I}\{B_n\} + \mathbb{I}\{F_n\})\Big].$$

Set A. Applying Corollary 4.1, we have:

$$\sum_{n \geq 1} \mathbb{P}(A_n) \leq 1 + K_m \sum_{n \geq 2} n^{-1}(\log(n))^{-2} \leq 4K_m. \tag{4.5}$$

Set B. Define $\tau_i(n) = \sum_{n'=1}^{n} \mathbb{I}\{B_{n',i}\}$. Since $B_{n',i}$ implies $a_i(n') = 1$, we have $s_i(n) \geq \tau_i(n)$. Applying [70, Lemma B.1] (see Corollary B.2 in Appendix B), we have $\sum_{n=1}^{N} \mathbb{P}(B_{n,i}) \leq 2(\varepsilon\theta_i)^{-2}$. A union bound yields:

$$\sum_{n=1}^{N} \mathbb{P}(B_n) \leq 2\varepsilon^{-2} \sum_{i \in E} \theta_i^{-2}. \tag{4.6}$$

Set F. Define $U = \frac{4f_1(N)}{(1-\kappa)^4\theta_{\min}^2}$. Define the set

$$S_n = \{i \in a(n) : s_i(n) \leq mU\Delta_{a(n)}^{-2}\}$$

and events:

$$\begin{aligned} G_n &= \{|S_n| \geq \sqrt{m}\}, \\ L_n &= \{|S_n| < \sqrt{m}, \min_{i \in a(n)} s_i(n) \leq \sqrt{m}U\Delta_{a(n)}^{-2}]\}. \end{aligned}$$

Assume that neither G_n nor L_n occurs. Then:

$$\begin{aligned} h(n) &= \sum_{i \in a(n), i \in S_n} \frac{1}{s_i(n)} + \sum_{i \in a(n), i \notin S_n} \frac{1}{s_i(n)} \\ &\leq \frac{|S_n|\Delta_{a(n)}^2}{\sqrt{m}U} + \frac{(m - |S_n|)\Delta_{a(n)}^2}{mU} < \frac{2\Delta_{a(n)}^2}{U}, \end{aligned}$$

since $|S_n| < \sqrt{m}$. Hence $\Delta^2_{a(n)} > Uh(n)/2$ and F_n does not occur. So $F_n \subset G_n \cup L_n$. Further decompose G_n and L_n as:

$$G_{i,n} = G_n \cap \{i \in a(n),\ s_i(n) \le mU\Delta^{-2}_{a(n)}\},$$
$$L_{i,n} = L_n \cap \{i \in a(n),\ s_i(n) \le \sqrt{m}U\Delta^{-2}_{a(n)}\}.$$

Applying Lemma 3.2 twice, we get:

$$\sum_{n=1}^{N} \Delta_{a(n)}\mathbb{I}\{G_{i,n}\} \le \frac{2mU}{\Delta_{\min}}\ ,$$
$$\sum_{n=1}^{N} \Delta_{a(n)}\mathbb{I}\{L_{i,n}\} \le \frac{2\sqrt{m}U}{\Delta_{\min}}.$$

We have $\sum_{i\in E}\mathbb{I}\{G_{i,n}\} = |S_n|\mathbb{I}\{G_n\} \ge \sqrt{m}\mathbb{I}\{G_n\}$. So:

$$\sum_{n=1}^{N} \Delta_{a(n)}\mathbb{I}\{G_n\} \le \frac{1}{\sqrt{m}}\sum_{n=1}^{N}\sum_{i\in E} \Delta_{a(n)}\mathbb{I}\{G_{i,n}\} \le \frac{2d\sqrt{m}U}{\Delta_{\min}}.$$

Further:

$$\sum_{n=1}^{N} \Delta_{a(n)}\mathbb{I}\{L_n\} \le \sum_{n=1}^{N}\sum_{i\in E} \Delta_{a(n)}\mathbb{I}\{L_{i,n}\} \le \frac{2d\sqrt{m}U}{\Delta_{\min}}.$$

Since $\mathbb{I}\{F_n\} \le \mathbb{I}\{G_n\} + \mathbb{I}\{L_n\}$ we get:

$$\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{a(n)}\mathbb{I}\{F_n\}\right] \le \frac{4d\sqrt{m}U}{\Delta_{\min}}. \tag{4.7}$$

Combining (4.5), (4.6), and (4.7) with $\Delta_{a(n)} \le D^+$, yields the announced result:

$$\mathfrak{R}_{\pi,N} \le \frac{4d\sqrt{m}U}{\Delta_{\min}} + 2D^+\left(2K_m + \varepsilon^{-2}\sum_{i\in E}\theta_i^{-2}\right).$$

□

4.E Proof of Theorem 4.6

The proof technique is similar to the analysis of [62, Theorem 5].

For $s \in \mathbb{N}^d$ and $a \in \mathcal{A}$ define $h'(s,a) = (\sum_{i\in a}\frac{1}{\sqrt{s_i}})^2$, and as before $s_i(n) = t_i(n)\hat{\theta}_i(n)$ and $s(n) = (s_i(n))_{i\in E}$, and $h'(n) = h'(s(n), a(n))$. We will use the following lemma.

Lemma 4.3. *For all $n, t \in \mathbb{N}$, $\lambda \in (0,1]$, and $i \in E$:*

$$\omega_i(n,\lambda,t) \geq \frac{1}{\lambda} - \sqrt{\frac{2f_2(n)}{t\lambda^3}}.$$

Proof. Let $i \in E$, $n, t \in \mathbb{N}$ and $u, \lambda \in (0,1]$ with $u \geq \lambda$. We have:

$$\frac{1}{\lambda} - \frac{1}{u} = \sqrt{\frac{t(u-\lambda)^2}{u}}\frac{1}{\sqrt{tu\lambda^2}} \leq \sqrt{2t\texttt{kl}(\lambda,u)}\frac{1}{\sqrt{t\lambda^3}},$$

where the second inequality follows from Lemma A.3 and $u \geq \lambda$. Hence, $t\texttt{KL}(\lambda,u) \leq f_2(n)$ implies:

$$\frac{1}{u} \geq \frac{1}{\lambda} - \sqrt{\frac{2f_2(n)}{t\lambda^3}}.$$

The above holds for all $u \in [\lambda, 1]$. Thus, the claim of the lemma follows by definition of $\omega_i(n,\lambda,t)$. □

Proof of Theorem 4.6. For any n, we define the following events:

$$\begin{aligned} A_{n,i} &= \{\omega_i(n) > 1/\theta_i\}, \;\; A_n = \bigcup_{i\in a^\star} A_{n,i}, \\ B_{n,i} &= \{a_i(n) = 1, \; |\hat{\theta}_i(n) - \theta_i| \geq \varepsilon\theta_i\}, \;\; B_n = \bigcup_{i \in E} B_{n,i}, \\ F_n &= \{\Delta_{a(n)} \leq (1-\kappa)^{-2}\theta_{\min}^{-1}\sqrt{2f_2(N)h'(n)}\}. \end{aligned}$$

We show that $a(n) \neq a^\star$ implies: $n \in A_n \cup B_n \cup F_n$. Consider n such that $a(n) \neq a^\star$ and $A_n \cup B_n$ does not occur. By design of the algorithm, $a(n)^\top\omega(n) \leq {a^\star}^\top\omega(n)$, and ${a^\star}^\top\omega(n) \leq D^\star$ since A_n does not occur. Hence $a(n)^\top\omega(n) \leq D^\star$. By Lemma 4.3, for all i:

$$\omega_i(n) \geq \frac{1}{\hat{\theta}_i(n)} - \sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}}.$$

Summing over $i \in a(n)$ we get:

$$\Delta_{a(n)} \leq a(n)^\top\theta^{-1} - a(n)^\top\hat{\theta}(n)^{-1} + \sum_{i\in a(n)}\sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}}.$$

As before, when B_n does not occur we have

$$a(n)^\top\theta^{-1} - a(n)^\top\hat{\theta}(n)^{-1} \leq \kappa\Delta_{a(n)}.$$

Furthermore, $\hat{\theta}_i(n) \geq \theta_{\min}(1-\kappa)$ for all $i \in a(n)$, and $f_2(n) \leq f_2(N)$ so that:

$$\sum_{i\in a(n)}\sqrt{\frac{2f_2(n)}{s_i(n)\hat{\theta}_i(n)^2}} \leq \sum_{i\in a(n)}\sqrt{\frac{f_2(N)}{s_i(n)\theta_{\min}^2(1-\kappa)^2}}.$$

Hence:

$$\Delta_{a(n)} \le \kappa \Delta_{a(n)} + \frac{\sqrt{2f_2(N)h'(n)}}{(1-\kappa)\theta_{\min}}$$

and $\Delta_{a(n)} \le (1-\kappa)^{-2}\theta_{\min}^{-1}\sqrt{2f_2(N)h'(n)}$ so that $n \in F_n$.

The regret $\mathfrak{R}_{\pi,N}$ is upper bounded by:

$$\mathbb{E}\Big[\sum_{n=1}^{N}\Delta_{a(n)}\Big] \le \mathbb{E}\Big[\sum_{n=1}^{N}\Delta_{a(n)}(\mathbb{I}\{A_n\}+\mathbb{I}\{B_n\}+\mathbb{I}\{F_n\})\Big].$$

Set A**.** By [43, Theorem 10] (see Theorem B.5) and a union bound:

$$\mathbb{P}(A_n) \le \sum_{i\in a^\star}\mathbb{P}(A_{n,i}) \le m\lceil f_2(n)\log(n)\rceil e^{1-f_2(n)}.$$

Hence:

$$\sum_{n=1}^{N}\mathbb{P}(A_n) \le m\Big(1+e\sum_{n\ge 2}\lceil f_2(n)\log(n)\rceil e^{-f_2(n)}\Big) \le 8m. \tag{4.8}$$

Set B**.** As in the proof of Theorem 4.5:

$$\sum_{n=1}^{N}\mathbb{P}(B_n) \le 2\varepsilon^{-2}\sum_{i\in E}\theta_i^{-2}. \tag{4.9}$$

Set F**.** Define $U' = 2m^2 f_2(N)(1-\kappa)^{-4}\theta_{\min}^{-2}$. Similarly to the proof of [62, Theorem 5], consider $\alpha,\beta>0$, and for $\ell\in\mathbb{N}$ define $\alpha_\ell = \Big(\frac{1-\beta}{\sqrt{\alpha}-\beta}\Big)^2\alpha^\ell$ and $\beta_\ell = \beta^\ell$. Introduce set $S_{\ell,n}$ and event $G_{\ell,n}$:

$$\begin{aligned}
S_{\ell,n} &= \{i\in a(n), s_i(n) \le U'\alpha_\ell\Delta_{a(n)}^{-2}\},\\
G_{\ell,n} &= \{|S_{\ell,n}| \ge \beta_\ell m\} \cap \{|S_{j,n}| < \beta_j m, j=1,...,\ell-1\}.
\end{aligned}$$

If $\overline{\cup_{\ell\ge1}G_{\ell,n}} = \{|S_{\ell,n}| < m\beta_\ell, \ell\ge 1\}$ occurs, then:

$$\begin{aligned}
\sum_{\ell\ge1}\frac{|S_{\ell-1,n}|-|S_{\ell,n}|}{\sqrt{\alpha_\ell}} &= \frac{|S_{0,n}|}{\sqrt{\alpha_1}} + \sum_{\ell\ge1}|S_{\ell,n}|\Big(\frac{1}{\sqrt{\alpha_{\ell+1}}}-\frac{1}{\sqrt{\alpha_\ell}}\Big)\\
&< \frac{m\beta_0}{\sqrt{\alpha_1}} + \sum_{\ell\ge1}m\beta_\ell\Big(\frac{1}{\sqrt{\alpha_{\ell+1}}}-\frac{1}{\sqrt{\alpha_\ell}}\Big)\\
&= m\sum_{\ell\ge1}\frac{\beta_\ell-\beta_{\ell-1}}{\sqrt{\alpha_\ell}} \le m,
\end{aligned}$$

since $\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_\ell}} \geq 0$. Now:

$$|\{i : s_i(n) \in U'\Delta_{a(n)}^{-2}[\alpha_\ell, \alpha_{\ell-1}]\}| = |S_{\ell-1,n}| - |S_{\ell,n}|$$

so that:

$$\sqrt{h'(n)} \leq \sum_{\ell \geq 1} \frac{(|S_{\ell-1,n}| - |S_{\ell,n}|)}{\sqrt{\alpha_\ell}} \frac{\Delta_{a(n)}}{\sqrt{U'}} < m\frac{\Delta_{a(n)}}{\sqrt{U'}}.$$

Hence, $\Delta_{a(n)}^2 > h'(n)U'm^{-2}$ and thus, F_n does not occur. Therefore, $F_n \subset \cup_{\ell \geq 1} G_{\ell,n}$ and:

$$\sum_{n=1}^{N} \Delta_{a(n)} \mathbb{I}\{F_n\} \leq \sum_{n=1}^{N} \sum_{\ell \geq 1} \Delta_{a(n)} \mathbb{I}\{G_{\ell,n}\}.$$

Further decompose $G_{\ell,n}$ as:

$$G_{i,\ell,n} = G_{\ell,n} \cap \{i \in a(n),\ s_i(n) \leq U'\alpha_\ell \Delta_{a(n)}^{-2}\}.$$

Observe that:

$$\mathbb{I}\{G_{\ell,n}\} \leq \frac{|S_{\ell,n}|}{m\beta_\ell}\mathbb{I}\{G_{\ell,n}\} = \frac{1}{m\beta_\ell}\sum_{i \in E}\mathbb{I}\{G_{i,\ell,n}\}.$$

Applying Lemma 3.2, we get:

$$\sum_{n=1}^{N} \Delta_{a(n)} \mathbb{I}\{G_{i,\ell,n}\} \leq \sum_{n=1}^{N} \Delta_{a(n)} \mathbb{I}\left\{ s_i(n) \leq \frac{U'\alpha_\ell}{\Delta_{a(n)}^2} \right\} \leq \frac{2U'\alpha_\ell}{\Delta_{\min}}.$$

Putting everything together:

$$\sum_{n=1}^{N} \Delta_{a(n)} \mathbb{I}\{F_n\} \leq \frac{2dU'}{m\Delta_{\min}} \sum_{\ell \geq 1} \frac{\alpha_\ell}{\beta_\ell} \leq \frac{90dU'}{m\Delta_{\min}} \tag{4.10}$$

by choosing $\alpha = 0.15$ and $\beta = 0.24$ so that $\sum_{\ell \geq 1} \frac{\alpha_\ell}{\beta_\ell} \leq 45$.

The proof is completed by combining (4.8), (4.9), (4.10), and using the fact that $\Delta_{a(n)} \leq D^+$. □

4.F Proof of Proposition 4.2

In the line network, KL-SR simply chooses the link with the smallest index on each hop. Hence, on each hop, KL-SR is equivalent to the KL-UCB algorithm for a classical MAB with geometrically distributed rewards. By [43, Theorem 1 and Lemma 6], the regret of KL-SR on the j-th hop, for any $j \in \{1, \dots, m\}$, asymptotically grows as:

$$\sum_{i \in E_j \setminus a^\star} \frac{\log(N)}{\mathtt{KLG}(\theta_i, \theta_{\zeta(i)})} \left(\frac{1}{\theta_i} - \frac{1}{\theta_{\zeta(i)}} \right),$$

where E_j denotes the set of links in the j-th hop. Since decisions at various hops are decoupled, the regret due to all hops satisfies

$$\begin{aligned}\limsup_{N\to\infty} \frac{\mathfrak{R}_{\mathtt{KL\text{-}SR},N}}{\log(N)} &\le \sum_{j=1}^{m} \sum_{i\in E_j\setminus a^\star} \frac{1}{\mathtt{KLG}(\theta_i,\theta_{\zeta(i)})}\left(\frac{1}{\theta_i}-\frac{1}{\theta_{\zeta(i)}}\right)\\ &= \sum_{i\notin a^\star} \frac{1}{\mathtt{KLG}(\theta_i,\theta_{\zeta(i)})}\left(\frac{1}{\theta_i}-\frac{1}{\theta_{\zeta(i)}}\right) = c_2(\theta).\end{aligned}$$

Furthermore, using Lemma A.4 and Lemma A.3 in Appendix B, we have for any $i \notin a^\star$:

$$\frac{1}{\mathtt{KLG}(\theta_i,\theta_{\zeta(i)})}\left(\frac{1}{\theta_i}-\frac{1}{\theta_{\zeta(i)}}\right) = \frac{\theta_{\zeta(i)}-\theta_i}{\theta_{\zeta(i)}\mathtt{KL}(\theta_i,\theta_{\zeta(i)})} \le \frac{2}{\theta_{\zeta(i)}-\theta_i}.$$

Moreover, in line networks $\Delta_{\min} = \min_{i\notin a^\star}(\theta_i^{-1}-\theta_{\zeta(i)}^{-1})$. Thus,

$$\begin{aligned}c_2(\theta) &\le \sum_{i\notin a^\star} \frac{2}{\theta_{\zeta(i)}-\theta_i} = \sum_{i\notin a^\star} \frac{2}{\theta_i\theta_{\zeta(i)}(\theta_i^{-1}-\theta_{\zeta(i)}^{-1})}\\ &\le \frac{d-m}{\Delta_{\min}}\cdot\frac{2}{\min_{i\notin a^\star}\theta_i\theta_{\zeta(i)}} \le \frac{2(d-m)}{\Delta_{\min}\theta_{\min}^2},\end{aligned}$$

which completes the proof. □

4.G Proof of Proposition 4.3

The proof uses the same ideas as in the proof of Theorem 3.7. Note that if $i \notin J_a(\lambda)$, then the optimal solution satisfies $u_i = 1$ since $\mathtt{kl}(1,v) = \infty$ unless $v = 1$. Thus, if $J_a(\lambda) = \emptyset$, then $u_i = 1, \forall i \in E$, and $b_a(n,\lambda,t) = \sum_{i\in E} a_i$.

If $J_a(\lambda) \neq \emptyset$, let $i \in J_a(\lambda)$. Computing b_a involves solving a convex optimization problem with one inequality constraint, which must hold with equality since $u_i \mapsto \mathtt{kl}(\lambda_i,u_i)$ is monotone increasing for $u_i \ge \lambda_i$. Since $\frac{d}{du_i}\mathtt{kl}(\lambda_i,u_i) = \frac{u-\lambda}{u(1-u)}$, the corresponding KKT conditions are:

$$\frac{1}{{u_i}^2} - \gamma t_i \frac{u_i-\lambda_i}{u_i(1-u_i)} = 0, \qquad \sum_{i\in J_a(\lambda)} t_i\mathtt{kl}(\lambda_i,u_i) - f_1(n) = 0.$$

with $\gamma > 0$ the Lagrange multiplier. The first equation is a quadratic equation:

$$u_i^2 + u_i\left(\frac{1}{\gamma t_i} - \lambda_i\right) - \frac{1}{\gamma t_i} = 0.$$

Solving for u_i, we obtain $u_i(\gamma) = g(\gamma,\lambda_i,t_i)$ and replacing in the second equation, we obtain $F(\gamma,\lambda,n,t) = f_1(n)$. We finally note that one can verify by inspection $g(\gamma,\lambda_i,t_i) \ge \lambda_i$, so that the box constraints in the definition of $b_a(n,\lambda,t)$ are satisfied. This completes the proof. □

4.H Regret Upper Bound of The CUCB Algorithm

CUCB (see [61]) uses the following link index:

$$\gamma_i(n) = \frac{1}{\hat{\theta}_i(n) + \sqrt{1.5\log(n)/t_i(n)}} \quad , \quad \forall i \in E$$

Define $\kappa = (1 - 2^{-\frac{1}{4}})$ and $\varepsilon = \kappa\frac{\Delta_{\min}}{D^+} < \kappa$. For any $s \in \mathbb{N}^d$ and $a \in \mathcal{A}$ define $h'(s,a) = (\sum_{i\in a}\frac{1}{\sqrt{s_i}})^2$, and as in the proof of Theorem 5.4, $s_i(n) = t_i(n)\hat{\theta}_i(n)$ and $s(n) = (s_i(n))_{i\in E}$, and $h'(n) = h'(s(n), a(n))$. We have that:

$$\begin{aligned} a(n)^\top\gamma(n) &= \sum_{i\in a(n)} \frac{1}{\hat{\theta}_i(n) + \sqrt{1.5\hat{\theta}_i(n)\log(n)/s_i(n)}} \\ &= \sum_{i\in a(n)} \frac{1}{\hat{\theta}_i(n)} - \sum_{i\in a(n)} \frac{\sqrt{1.5\log(n)/(s_i(n)\hat{\theta}_i(n)^3)}}{1 + \hat{\theta}_i(n)^{-\frac{1}{2}}\sqrt{1.5\log(n)/s_i(n)}} \\ &\geq a(n)^\top\hat{\theta}(n)^{-1} - \sum_{i\in a(n)}\sqrt{\frac{1.5\log(n)}{s_i(n)\hat{\theta}_i(n)^3}}. \end{aligned} \tag{4.11}$$

For any n, introduce the following events:

$$\begin{aligned} A_{n,i} &= \Big\{|\hat{\theta}_i(n) - \theta_i| > \sqrt{1.5\log(n)/t_i(n)}\Big\}, \;\; A_n = \bigcup_{i\in a^\star} A_{n,i}, \\ B_{n,i} &= \{a_i(n) = 1,\; |\hat{\theta}_i(n) - \theta_i| \geq \varepsilon\theta_i\}, \;\; B_n = \bigcup_{i\in E} B_{n,i}, \\ F_n &= \{\Delta_{a(n)} \leq (1-\kappa)^{-\frac{5}{2}}\theta_{\min}^{-\frac{3}{2}}\sqrt{2\log(N)h'(n)}\}. \end{aligned}$$

We show that if $a(n) \neq a^\star$ then $A_n \cup B_n \cup F_n$ occurs. Consider n such that $a(n) \neq a^\star$ and $A_n \cup B_n$ does not occur. By design of the algorithm, $a(n)^\top\gamma(n) \leq (a^\star)^\top\gamma(n)$, and $(a^\star)^\top\gamma(n) \leq D^\star$ since A_n does not occur. Hence $a(n)^\top\gamma(n) \leq D^\star$.

When B_n does not occur, $(1-\kappa)\theta_{\min} \leq \hat{\theta}_i(n) \leq (1+\varepsilon)\theta_i$ and $a(n)^\top\theta^{-1} - a(n)^\top\hat{\theta}(n)^{-1} \leq \kappa\Delta_{a(n)}$. Hence, using (4.11) we get

$$\begin{aligned} \Delta_{a(n)} &= a(n)^\top\theta^{-1} - D^\star \leq a(n)^\top\theta^{-1} - a(n)^\top\gamma(n) \\ &\leq \kappa\Delta_{a(n)} + (1-\kappa)^{-\frac{3}{2}}\theta_{\min}^{-\frac{3}{2}}\sqrt{1.5\log(N)h'(n)} \end{aligned}$$

so that $\Delta_{a(n)} \leq (1-\kappa)^{-\frac{5}{2}}\theta_{\min}^{-\frac{3}{2}}\sqrt{1.5\log(N)h'(n)}$ and thus $n \in F_n$.

Hence, $\mathfrak{R}_{\pi,N}$ is upper bounded by:

$$\mathbb{E}\Big[\sum_{n=1}^{N}\Delta_{a(n)}\Big] \leq \mathbb{E}\Big[\sum_{n=1}^{N}\Delta_{a(n)}(\mathbb{I}\{A_n\} + \mathbb{I}\{B_n\} + \mathbb{I}\{F_n\})\Big].$$

Firstly note that using a Chernoff bound and a union bound, we have that $\mathbb{P}(A_n) \leq 2mn^{-2}$ (see, e.g., [61, Lemma 3]). Hence

$$\sum_{n=1}^{N} \mathbb{P}(A_n) \leq \sum_{n=1}^{N} \frac{2m}{n^2} \leq \frac{2\pi^2 m}{3}. \tag{4.12}$$

Furthermore, as in the proof of Theorem 4.5:

$$\sum_{n=1}^{N} \mathbb{P}(B_n) \leq 2\varepsilon^{-2} \sum_{i \in E} \theta_i^{-2}. \tag{4.13}$$

Finally, defining $U' = 2m^2 f_2(N)(1-\kappa)^{-\frac{5}{2}} \theta_{\min}^{-3}$ and applying the same techniques as in the proof of Theorem 4.6, we deduce that

$$\sum_{n=1}^{N} \Delta_{a(n)} \mathbb{I}\{F_n\} \leq \frac{278 dm \log(N)}{\Delta_{\min} \theta_{\min}^3}. \tag{4.14}$$

Putting (4.12), (4.13), and (4.14) together, gives the desired result and concludes the proof. □

4.I Technical Lemmas

Lemma 4.4 ([75, Lemma 2]). *Consider $(X_i)_i$ independent with $X_i \sim \mathrm{Geo}(\theta_i)$ and $\theta_i \in (0,1]$. Consider $(Y_i)_i$ independent with $Y_i \sim \mathrm{Geo}(\lambda_i)$ and $\lambda_i \in (0,1]$. Define $\overline{X} = \sum_i X_i$ and $\overline{Y} = \sum_i Y_i$. Then $\overline{X} \overset{d}{=} \overline{Y}$ iff $(\theta_i)_i = (\lambda_i)_i$ up to a permutation*[6].

[6] The symbol $\overset{d}{=}$ denotes equality in distribution.

Chapter 5

Learning Proportionally Fair Allocations

This chapter addresses a generic sequential resource allocation problem, where one strives to find a *fair* and *efficient* operating point of the system. The problem considered involves a decision maker, who selects in each round an allocation of resources (servers) to a set of tasks consisting of a large number of jobs. A job of task i assigned to server j is successfully treated with probability θ_{ij} in a round, and the decision maker is informed on whether this job is completed at the end of the round. The probabilities θ_{ij}'s are initially unknown and have to be learnt. The objective of the decision maker is to sequentially assign jobs of various tasks to servers so that it rapidly learns and converges to the Proportionally Fair (PF) allocation (or other similar allocations achieving an appropriate trade-off between efficiency and fairness). We formulate the problem as a MAB problem, and devise a sequential assignment algorithm with low regret. The latter is defined as the difference in utility achieved by an oracle algorithm aware of the θ_{ij}'s and by the proposed algorithm.

This chapter is organized as follows. Section 5.1 discusses the motivation of studying the aforementioned resource allocation scenario through some real world applications and states the contributions of this chapter. It is followed by an overview of related works in Section 5.2. Section 5.3 provides a precise description of the problem and introduces the notion of approximate PF allocation, referred to as APF, which allows us to cast the problem as a combinatorial MAB. Section 5.4 presents a procedure to compute APF and investigates its properties. A regret lower bound for learning APF is provided in Section 5.5, and Section 5.6 presents an algorithm for learning APF. Finally, Section 5.7 summarizes the chapter.

5.1 Motivation and Contributions

To motivate our resource allocation problem, let us introduce some definitions. There are m tasks and s servers respectively indexed by i and j. The task service rates are defined as follows: If $z = (z_{ij})_{i,j}$ represents the probabilities that at the beginning of each slot, a job of task i is assigned to server j for any (i, j), then the service rate of task i is $\gamma_i(z, \theta) = \sum_{j\in[s]} \theta_{ij} z_{ij}$, for all i. z will be referred to as an allocation distribution, or for short, an allocation. An allocation z is *efficient* if the global service rate of the system is high, and *fair* if all tasks are served at similar rates avoiding starvation. An allocation known to achieve an appropriate trade-off between efficiency and fairness is the Proportionally Fair (PF) allocation [84]: It maximizes the sum of the logarithm of the task service rates. In many systems involving resource allocation problems, PF is often preferred to other allocations for the ease of its implementation [84, 85], and its behavior in dynamical scenarios (here when tasks arrive and leave after completion) [86]. Hence, we assume that the decision maker wishes to sequentially allocate tasks to servers so as to quickly learn and converge to the PF allocation (θ is initially unknown, and so is the PF allocation).

The work presented in this chapter is motivated by the following two important problems in wireless communication systems:

(1) *Dynamic Spectrum Access.* Transmitters are today able to exploit a large part of the radio spectrum, and can switch frequency bands rapidly. The service rate achieved on a link operating on a given band depends on the channel conditions, which in turn depend on the band and the link (this phenomenon is known as frequency-selective fading [87]). The outcomes of packet transmissions are also random as a result of the so-called fast fading [87]. The average successful packet transmission rates of links on the various frequency bands are usually unknown and have to be learnt.

(2) *Access Point Selection.* When users in a wireless network may attach to various access points, we get a similar situation. The throughput experienced by a user depends on the random channel conditions towards the selected access point. In both examples, the system and the corresponding allocation problem can be directly mapped into the generic parallel server system and the resource allocation problem described above: Servers correspond to either frequency bands or access points, tasks are links or users, and jobs are data packets.

5.1.1 Contributions of the Chapter

Keeping these applications in mind, we address our generic resource allocation problem. Our contributions are as follows:

(i) We first provide a precise description of the system model and of our objectives. We introduce the notion of *Approximate PF (APF)* allocation, as an approximation to the PF allocation, and define a notion of expected regret that captures the rate at which the allocation chosen by the decision maker converges towards the APF allocation. Our regret definition enables us to cast the problem as a combinatorial MAB with a non-linear reward function. Furthermore, we provide an efficient algorithm to compute APF and characterize its tightness. To the best of our knowledge, characterization of APF is new and could be independently interesting.

(ii) We derive an asymptotic (as T grows large) regret lower bound for learning APF. This bound is tight in the sense that there exist policies that achieve it. Our regret lower bound is implicit. So we further provide an explicit lower bound valid for specific problem instances indicating the dependence of the best possible regret on $\theta_{\min}$ and $\Delta_{\min}$, where $\theta_{\min}$ is the job success rate of the worst task-server pair and $\Delta_{\min}$ is the minimal gap between the average utility of the optimal and that of a sub-optimal allocation.

(iii) We develop an algorithm for learning APF, which we may call `ES-APF` (Efficient Sampling for APF). We further show that the regret under `ES-APF` scales at most as $\mathcal{O}(m^3\Delta_{\min}^{-1}\theta_{\min}^{-1}\log(T))$ and examine its performance numerically.

5.2 Related Work

There are a few studies that consider allocation of tasks to a set of servers in the bandit setting, for which optimistic algorithms with regret growing as $\mathcal{O}(\log(T))$ or $\mathcal{O}(\sqrt{T})$ are provided (see, e.g., [88, 89, 90]). To the best of our knowledge, [88, 89] have made the first attempt on such resource allocation problems in the bandit setting. In particular, Lattimore et al. [89] consider allocation of m jobs to a set of s heterogenous resources. In their setup, in each round t an allocation matrix $M(t)$ is chosen, whose element $M_{ij}(t)$ denotes the portion of resource type j allocated to job i. An allocation matrix is feasible iff $\sum_{i=1}^{m} M_{ij}(t) \leq 1$ for all j. Job i is successfully completed according to a Bernoulli distribution of mean $\min(1, \sum_{j\in[s]} M_{ij}(t)\theta_{ij})$, where $\theta_{ij} \geq 0$ is an initially unknown but fixed cut-off parameter capturing the difficulty of job i when using resource j. At the end of each round, the decision maker is informed on whether each job i is completed successfully or not, and her goal is to maximize the number of successfully completed jobs in expectation. Inspired by the algorithms for stochastic linear bandits developed in [65], Lattimore et al. [89] present an optimistic algorithm with logarithmic regret.

Johari et al. [90] study a resource allocation scenario where a decision maker wishes to assign a set of clients of various types to a set of heterogenous servers. The outcome of processing job i on server j is a binary random variable sampled i.i.d. from a Bernoulli distribution with parameter θ_{ij}. The decision maker aims at maximizing the number of successfully completed jobs. Her performance is compared against an asymptotically optimal matching policy, where (i) the number of servers and client arrival rates grow large and (ii) the system state has stationary asymptotic behavior. The authors of [90] provide an optimistic algorithm with logarithmic regret against the offline matching policy. We note that Johari et al. only consider the case where server types and success probabilities are known, and the decision maker needs to learn clients' types only.

Despite some similarities between these models and ours, the aim of these works is to learn an allocation maximizing the throughput of the system without providing any guarantee on the fairness. In all these studies, the payoff is accrued to the overall system performance and no notion of fairness for individual users is considered. To the best of our knowledge, the work presented in this chapter is the first to address fair resource allocation in the bandit setting.

A different line of works that accounts for resource constraints under bandit feedback is the budgeted MAB problem as investigated in [91, 92, 93]. These problems are natural generalizations of the classical MAB problem, in which the pull of any arm results in the corresponding random reward, but also consumes some resource that is shared by all arms and is limited in supply. Different arms may consume the resource at different rates and the task terminates when the resource is exhausted. The resource consumption could be deterministic (as in, e.g., [92, 93]) or stochastic, as studied in [91]. To the best of our knowledge, the considered resource allocation in this chapter cannot be mapped into the framework of budgeted MABs. Furthermore, none of these studies take fairness into account.

We also mention that a notion of fairness for MAB problems has recently been proposed by Joseph et al. [94]. Precisely speaking, this notion requires that for any pair of arms i and j, if the average reward of i is greater than that of j, then at all rounds with high probability, the algorithm plays arm j with a smaller probability than it draws arm i. Although the authors of [94] take into account fair allocation for individual players, their notion of fairness is completely different than ours. Clearly, in our setup the imposed fairness changes the offline optimal arm compared to the unfair setting, whereas their notion does not.

We conclude this section by discussing the relation of our problem to bandit convex optimization. Due to concavity of PF utility function, our resource allocation problem resembles convex optimization in the bandit setting as studied in, e.g., [95]. In the latter problem, the decision maker chooses a point from a compact convex decision space, whereas in ours she has to choose an assignment from a finite set. Moreover, she receives richer information than bandit feedback. These facts allow her to obtain a regret growing as $\mathcal{O}(\log(T))$, as opposed to $\Omega(\sqrt{T})$ in the case of bandit convex optimization (see, e.g., [65]).

5.3 Problem Formulation

We consider a system consisting of m tasks indexed by $i \in [m] = \{1, \dots, m\}$, who wish to share $s < m$ available servers indexed by $j \in [s]$. We assume that each task has an unlimited number of sub-tasks or jobs. We consider a time slotted system, where completion of each job takes one slot. When task i is assigned to server j, the completion outcome is drawn i.i.d. from a Bernoulli distribution with success probability θ_{ij}. We assume that processing of jobs occurs independently at various servers and that the matrix of success probabilities $\theta = (\theta_{ij})_{i\in[m],j\in[s]}$ is fixed but unknown to the decision maker. Let $\theta_{\min} = \min_{i,j} \theta_{ij}$ and assume that $\theta_{\min} > 0$.

Assignment of servers to tasks is represented as *an assignment matrix* $M \in \{0,1\}^{m\times s}$, where $M_{ij} = 1$ iff task i is assigned to server j. An assignment matrix M is feasible iff for all j, the server j is allocated to one task only, i.e., $\sum_{i\in[m]} M_{ij} = 1$ for all $j \in [s]$. Let $\mathcal{M}$ be the set of all feasible assignment matrices. The chosen assignment matrix may change at the beginning of each slot. At a given time slot t, we let $X_{ij}(t)$ be an indicator showing whether job of task i is successfully completed at server j when it is assigned to this server in slot t. Thus, $\mathbb{E}[X_{ij}(t)] = \theta_{ij}$.

5.3.1 The PF Allocation

For any task i and server j, let z_{ij} denote the probability that server j is assigned to task i and define the corresponding matrix $z = (z_{ij})_{i\in[m],j\in[s]}$. The set of all feasible allocation matrices denoted by $\mathcal{Z}$ is expressed as

$$\mathcal{Z} = \Big\{ z \in \mathbb{R}_+^{m\times s} : \sum_{i\in[m]} z_{ij} = 1, \ \forall j \in [s] \Big\}.$$

The service rate of task i achieved under z, denoted by $\gamma_i(z,\theta)$, is defined as the expected number of jobs successfully completed when the underlying assignment matrix is distributed according to z:

$$\gamma_i(z,\theta) = \mathbb{E}[\sum_{j\in[s]} M_{ij}X_{ij}] = \sum_{j\in[s]} \theta_{ij} z_{ij}.$$

The Proportionally Fair (PF) allocation is an allocation in $\mathcal{Z}$ maximizing the sum of the logarithm of service rate of all tasks [84]. Formally, it is the solution to the following problem:

$$\mathsf{PF}_1(\theta) : \qquad \max_{z\in\mathcal{Z}} \ f(z,\theta) := \sum_{i\in[m]} \log \gamma_i(z,\theta). \tag{5.1}$$

For a given θ, the PF allocation, i.e., the optimal solution to $\mathsf{PF}_1(\theta)$ will be denoted by $z^{\mathsf{pf}}(\theta)$. We let $S^{\mathsf{pf}}(\theta)$ be binary matrix associated to the support of $z^{\mathsf{pf}}(\theta)$: For all (i,j), $S^{\mathsf{pf}}_{ij}(\theta) = 1$ iff $z^{\mathsf{pf}}_{ij}(\theta) > 0$. Furthermore, let $\gamma^{\mathsf{pf}}_i(\theta) := \gamma_i(z^{\mathsf{pf}}(\theta),\theta)$ be the service rate of task i under $z^{\mathsf{pf}}(\theta)$. The quantity $f^{\mathsf{pf}}(\theta) := \sum \log_i \gamma^{\mathsf{pf}}_i(\theta)$ will be referred to as the *system utility* under the PF allocation. We finally remark that $\mathsf{PF}_1(\theta)$ is a convex problem for all θ and hence can be solved in polynomial time.

5.3.2 The APF Allocation

The decision maker wishes to learn the PF allocation z^{pf} and to choose an assignment drawn according to z^{pf}. Such an implementation may not lead to a well-defined notion for the expected regret for all problem instances. If there are more than one server to be allocated to user i under z^{pf}, then by drawing just one assignment according to z^{pf}, we do not receive feedback for some task-server pair (i,j) in the support of z^{pf}. Hence, this would potentially require more exploration and it is not clear whether one can guarantee logarithmic regret. To accommodate this situation, we introduce the notion of *Approximate PF (APF)* allocation, as an approximation to the PF allocation. APF is an allocation in $\mathcal{Z}$ under which (i) each task is only assigned to one server, and (ii) the system utility is as close to f^{pf} as possible. To formalize this, we introduce

$$\mathcal{A} = \Big\{ A \in \{0,1\}^{m\times s} : \ \sum_{j\in[s]} A_{ij} = 1, \ \forall i \in [m] \Big\}$$

and for any $A \in \mathcal{A}$, we define the matrix $z(A) \in [0,1]^{m \times s}$ with $z_{ij}(A) = A_{ij} / \sum_{k \in [m]} A_{kj}$ for all i and j. Finding APF amounts to solving the following problem:

$$\mathsf{PF}_2(\theta): \quad \max_{A \in \mathcal{A}} \; f(z(A), \theta). \tag{5.2}$$

Let $A^{\mathsf{apf}}(\theta)$ be any optimal solution to $\mathsf{PF}_2(\theta)$. We will refer to $z(A^{\mathsf{apf}}(\theta))$ as an APF allocation. Furthermore, we let $\gamma_i^{\mathsf{apf}}(\theta) := \gamma_i(z(A^{\mathsf{apf}}(\theta)), \theta)$ denote the service rate of task i under APF, and let $f^{\mathsf{apf}}(\theta) := f(z(A^{\mathsf{apf}}(\theta)), \theta)$ denote the utility achieved under APF.

For the sake of brevity, we use the following conventions throughout this chapter. We may omit the dependence of various quantities on θ when it is clear from the context that the underlying parameter is θ. Furthermore, by a slight abuse of notation, we refer to any $A \in \mathcal{A}$ as an allocation, and to A^{apf} as the APF allocation. Finally, for any binary matrix Z, we write $(i,j) \in Z$ to denote $Z_{ij} = 1$, and similarly $(i,j) \notin Z$ to imply $Z_{ij} = 0$.

5.3.3 Online Server Allocation Problem

At the beginning of each time slot t, an algorithm or policy π selects an allocation $A^\pi(t) \in \mathcal{A}$ based on the past decisions $(A^\pi(t'))_{t'=1}^{t-1}$ and their observed rewards. It then draws an assignment $M^\pi(t) \in \mathcal{M}$ according to $z^\pi(t) := z(A^\pi(t))$. At the end of time slot t, the decision maker receives $X_{ij}(t)$ for all (i,j) such that $M_{ij}^\pi(t) = 1$. Let Π be the set of all feasible policies. We measure the performance of a policy through the notion of regret with respect to the APF allocation A^{apf}. The regret of a policy $\pi \in \Pi$ after T rounds is the expected difference of the system utility for the first T time slots under an oracle policy always selecting A^{apf} and under π:

$$\mathfrak{R}_{\pi,T} = T f^{\mathsf{apf}} - \mathbb{E}[\sum_{t=1}^{T} \sum_{i \in [m]} \log \mathbb{E}[Y_i^\pi(t)]] \,.$$

Here $Y_i^\pi(t)$ denotes the number of successfully completed jobs of task i under π at time t. Furthermore, the first expectation is taken with respect to the possible randomization in policy π, whereas the second captures the randomization in the rewards. In particular, for any task i, server j, and time t, we have

$$\mathbb{E}[Y_i^\pi(t)] = \frac{\theta_{ij}}{\sum_{i \in [m]} A_{ij}^\pi(t)} \quad \text{if } A_{ij}^\pi(t) = 1.$$

The regret quantifies the loss in total system utility due to the need to explore sub-optimal allocations to learn A^{apf}. We further remark that our proposed algorithm does not have any randomization, and therefore we can remove the first expectation.

Frame-based policies. We introduce a restricted class of policies $\Pi_f \subset \Pi$, referred to as *frame-based policies*, for which we provide problem-specific regret lower bound in Section 5.5. A frame-based policy π implements z^π in the following way: It proceeds in frames of varying lengths. Let us index a frame by $n \in \mathbb{N}$ and denote by Λ_n the set of time slots in frame n. At the beginning of a frame n, π chooses an allocation $A^\pi(n)$ based on the past decisions and their observed rewards. Let $z^\pi(n) := z(A^\pi(n))$. Then, it plays a minimal set of assignments so that each task-server pair (i, j) is played with frequency $z^\pi_{ij}(n)$ over Λ_n, namely

$$\frac{1}{|\Lambda_n|} \sum_{t \in \Lambda_n} M^\pi_{ij}(t) = z^\pi_{ij}(n), \quad \forall i, \forall j.$$

For any matrix $A \in \mathcal{A}$ and $j \in [s]$, we use a_j to denote the sum of the elements of the j-th column of A, that is $a_j = \sum_i A_{ij}$. Moreover, we let ℓ_A denote the smallest positive integer that is divisible by $a_j, j = 1, \dots, s$. Using this notation, it is easy to confirm (by the design of APF allocation) that we have $|\Lambda_n| = \ell_{A^\pi(n)}$.

5.4 Tightness of APF Allocation

This section is devoted to investigating the tightness of APF allocation. As a by-product, we present an algorithm for computing such an allocation in polynomial time.

We first provide the following lemma, which characterizes the tightness of the APF allocation in terms of utility difference between the PF and APF allocations.

Lemma 5.1. *For all θ, we have $f^{\mathsf{pf}}(\theta) - f^{\mathsf{apf}}(\theta) = V(\theta)$, where $V(\theta)$ is the optimal value of the following problem:*

$$\begin{aligned} \min_{w \in \mathbb{N}^s} \quad & \sum_{j \in [s]} w_j \log(w_j / w^{\mathsf{pf}}_j(\theta)) \\ \text{subject to:} \quad & \sum_{j \in [s]} w_j = m, \\ & w_j \le \sum_{i \in [m]} S^{\mathsf{pf}}_{ij}(\theta), \ \ \forall j \in [s], \end{aligned} \tag{5.3}$$

where $w^{\mathsf{pf}}_j(\theta) := \max_{i \in [m]} \theta_{ij} / \gamma^{\mathsf{pf}}_i(\theta)$ for any $j \in [s]$.

Lemma 5.1 may be interpreted using the projection using the KL-divergence. To this end, first observe that for any θ, using Lemma 5.4 (see Appendix 5.A), we have

$$\sum_j w^{\mathsf{pf}}_j = \sum_j \max_i \frac{\theta_{ij}}{\gamma^{\mathsf{pf}}_i} = \sum_j \sum_i \frac{\theta_{ij}}{\gamma^{\mathsf{pf}}_i} z^{\mathsf{pf}}_{ij} = m\ .$$

Noting that the objective of problem (5.3) can be written as $m\mathsf{KL}\Big(\frac{w}{m}, \frac{w^{\mathsf{pf}}}{m}\Big)$, one may interpret $V(\theta)$ as the distance (in terms of the KL-divergence) between the distribution of tasks to servers under the PF allocation (i.e., w^{pf}/m) and under the APF allocation. Equivalently, the optimal solution to problem (5.3) may be further viewed as the projection of w^{pf}/m, using the KL-divergence, onto the set of feasible distributions that induce APF allocations.[1]

To gain further insight, we numerically evaluate $V(\theta)$. Figure 5.1 shows the relative utility difference between PF and APF $(f^{\mathsf{pf}} - f^{\mathsf{apf}})/f^{\mathsf{apf}}$ for uniformly sampled parameters averaged over 100 independent experiments. These curves show that for a given number of servers s, the relative utility difference decreases as the number of tasks increases. In a similar flavour, the curve in Figure 5.1(b) indicates that the difference tends to zero as the ratio of the number of tasks to the number of servers m/s increases.

5.4.1 Computing the APF Allocation

Problem (5.3) in Lemma 5.1 can be used to compute the APF allocation A^{apf}. Letting w^{apf} denote any maximizer of problem (5.3), we recall that w_j^{apf} corresponds to the number of tasks sharing server j under the APF allocation (we refer to the proof of Lemma 5.1). Hence, in the first step towards finding APF, we determine w^{apf} by solving problem (5.3), which can be carried out using dynamic programming; see Appendix 5.F for details. Now, given w^{apf}, finding A^{apf} can be cast as the following problem:

$$\begin{aligned} \max_{A \in \mathcal{A}} \quad & \sum_{i \in [m]} \sum_{j \in [s]} A_{ij} \log(\theta_{ij}/w_j^{\mathsf{apf}}) \\ \text{subject to:} \quad & \sum_{i \in [m]} A_{ij} = w_j^{\mathsf{apf}}, \quad \forall j \in [s], \\ & A \leq S^{\mathsf{pf}} . \end{aligned} \tag{5.4}$$

We then show that:

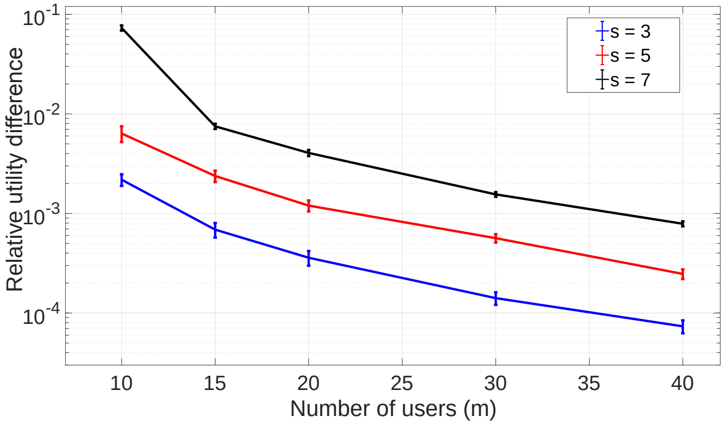
Lemma 5.2. *For any θ, problem* (5.4) *is a matroid intersection problem.*

As a consequence of Lemma 5.2, problem (5.4) can be solved in polynomial time [97, Theorem 10.13]. We therefore have a procedure to compute the corresponding APF allocation of a given PF allocation. The pseudo-code of this procedure is provided in Algorithm 5.1.

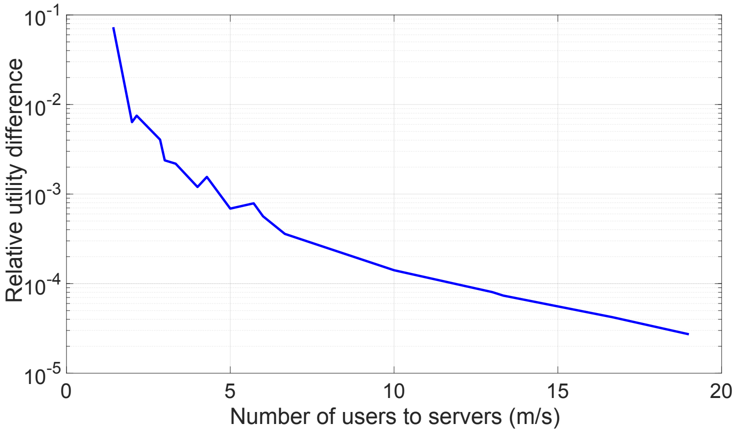
Algorithm 5.1 Computing APF

```
Find $z^{\mathsf{pf}}(\theta)$ any solution to problem $\mathsf{PF}_1(\theta)$.
Let $w_j^{\mathsf{pf}}(\theta) = \max_i \theta_{ij}/\gamma_i^{\mathsf{pf}}(\theta)$ for all $j$.
Find $w^{\mathsf{apf}}(\theta)$ the solution to problem (5.3) using dynamic programming.
Find $A^{\mathsf{apf}}(\theta)$ the solution to problem (5.4).
```

[1] For a thorough description of projection using the KL-divergence, see Chapter 3, I-projections in [96].



(a)



(b)

Figure 5.1: The relative utility difference between PF and APF allocations (see Lemma 5.1)

5.5 Regret Lower Bound

In this section, we provide a fundamental performance limit satisfied by any algorithm in Π_f. Our proposed performance limit is a lower bound on the regret that holds asymptotically, i.e., when the time horizon T grows large. To derive this lower bound, we leverage the result in [39], as used in the previous chapters.

Let us define, by a slight deviation from our notation,

$$f(A,\theta) := f(z(A),\theta) = \sum_{i,j} A_{ij} \log(\theta_{ij}/a_j),$$

and recall that $A^{\mathsf{apf}} \in \mathrm{argmax}_{A\in\mathcal{A}} f(A,\theta)$. Furthermore, for any allocation $A \in \mathcal{A}$, define $\Delta_A = f(A^{\mathsf{apf}},\theta) - f(A,\theta)$. We define $B(\theta)$ as the set of *bad* parameters that cannot be distinguished from θ when the underlying allocation is A^{apf}, and for which A^{apf} is sub-optimal:

$$B(\theta) = \Big\{\lambda : (A^{\mathsf{apf}}_{ij}\lambda_{ij} = A^{\mathsf{apf}}_{ij}\theta_{ij},\ \forall i,j) \text{ and } \max_{A\in\mathcal{A}} f(A,\lambda) > f(A^{\mathsf{apf}},\theta)\Big\}.$$

The following theorem presents a regret lower bound for any uniformly good policy in Π_f. Recall that a policy π is uniformly good if for all θ, the regret satisfies $\mathfrak{R}_{\pi,T} = o(T^\alpha)$ for any $\alpha > 0$.

Theorem 5.1. *For all θ and any uniformly good policy $\pi \in \Pi_f$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \ge c(\theta),$$

where $c(\theta)$ is the infimum of the following optimization problem:

$$\inf_{x\ge 0} \sum_{A\neq A^{\mathsf{apf}}} \Delta_A x_A \tag{5.5}$$

$$\text{subject to:} \quad \inf_{\lambda\in B(\theta)} \sum_{A\neq A^{\mathsf{apf}}} x_A \sum_{i,j} \frac{A_{ij}}{a_j} \mathtt{kl}(\theta_{ij},\lambda_{ij}) \ge 1.$$

In the above theorem, the optimal solution to problem (5.5), denoted by $x^\star$, may be interpreted as the number of explorations of sub-optimal allocations under an *optimal* algorithm, namely an algorithm whose regret asymptotically matches the lower bound of Theorem 5.1.

Theorem 5.1 provides a tight lower bound on the regret of any policy in Π_f in the sense that it can be achieved. Such an implicit bound however fails at providing further insights into the dependence of regret on problem dimension as well as other relevant parameters. In order to gain further insights, we study the following problem instance.

Assume that m is a multiple of s. Let $\alpha \in (0, 0.5)$ and $\beta \in (\frac{\alpha}{2}, \alpha)$. Define parameter θ such that $\theta_{ij} = \alpha$ if $\mathrm{mod}(i-j, s) = 0$ and $\theta_{ij} = \beta$ otherwise. It is straightforward to check that $z^{\mathsf{apf}} = z^{\mathsf{pf}}$, and that $z^{\mathsf{pf}}_{ij} = \frac{s}{m}$ if $\theta_{ij} = \alpha$, and $z^{\mathsf{pf}}_{ij} = 0$ otherwise; see Figure 5.2 for an illustration for the case with $s = 2$.

Using the lower bound in Theorem 5.1, in the following lemma we provide an explicit lower bound on the regret of the aforementioned problem instance:

Lemma 5.3. *For the above problem instance with $m/s \ge 5$, it holds that $c(\theta) \ge \frac{0.15m}{\theta_{\min}\Delta_{\min}}(1-\frac{1}{s})$. Hence, the regret of any uniformly good policy $\pi \in \Pi_f$ for this problem instance satisfies: $\mathfrak{R}_{\pi,T} = \Omega\Big(\frac{m}{\theta_{\min}\Delta_{\min}}\log(T)\Big)$.*

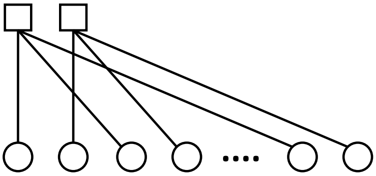


Figure 5.2: Lower bound example: tasks and servers are respectively shown by circles and boxes. Here only connections for task-server pairs under the optimal allocation, namely the ones with $\theta_{ij} = \alpha$, are shown. All other connections have success rate $\theta_{ij} = \beta \in (\frac{\alpha}{2}, \alpha)$.

5.6 The ES-APF Algorithm

In this section we develop an optimistic algorithm for learning APF allocation. Our proposed algorithm is a frame-based policy (a policy in Π_f) that works based on the KL-UCB index [43].

Let $\lambda \in (0, 1]$, and $n, N \in \mathbb{N}$. Define the KL-UCB index function [43]:

$$u(n, \lambda, N) = \sup\big\{q \in [\lambda, 1] : N\mathtt{kl}(\lambda, q) \leq g(n)\big\},$$

where $g(n) = \log(n) + 4\log(\log(n))$. We can now define the index for an allocation $A \in \mathcal{A}$ as follows: Given $\lambda \in (0, 1]^{m\times s}$, $N \in \mathbb{N}^{m\times s}$, and $n \in \mathbb{N}$, define

$$\xi_A(n, \lambda, N) = f\big(A, (u(n, \lambda_{ij}, N_{ij}))_{i\in[m],j\in[s]}\big) \ .$$

Our proposed algorithm, which we refer to as ES-APF (Efficient Sampling for APF), is an index policy relying on ξ_A index function. To present ES-APF, we introduce the following notations. Let $N_{ij}(n)$ be the total number of samples obtained for task-server pair (i, j) before the start of the n-th frame. Define $\hat{\theta}_{ij}(n)$ the empirical success rate of the jobs of task i on server j over the trials of the first $n - 1$ frames. Furthermore, define the corresponding matrices $N(n) = (N_{ij}(n))_{i\in[m],j\in[s]}$, $\hat{\theta}(n) = (\hat{\theta}_{ij}(n))_{i\in[m],j\in[s]}$.

ES-APF works as follows: Let t_n denote the first time slot of frame n. At $t = t_n$, ES-APF computes the index for job success rates of each task-server pair (i, j), denoted by $b_{ij}(n) := u(n, \hat{\theta}_{ij}(n), N_{ij}(n))$. It then selects an allocation with the largest index: $A(n) \in \mathrm{argmax}_{A\in\mathcal{A}}\, \xi_A(n)$, where $\xi_A(n) := \xi_A(n, \hat{\theta}(n), N(n))$. Having determined $A(n)$, ES-APF determines the sequence of assignments to be played in the n-th frame $(M(t))_{t_n\leq t\leq t+\ell_{A(n)}-1}$ to implement $z(A(n))$. The pseudocode of ES-APF is described in Algorithm 5.2.

Algorithm 5.2 ES-APF

```
for $n \geq 1$ do
    Let $t_n = t$
    Find $A(n)$ the solution to problem $\mathsf{PF}_2(b(n))$ using Algorithm 5.1.
    Decompose $A(n)$ into a sequence of assignments $(M(t))_{t_n \leq t \leq t_n + \ell_{A(n)} - 1}$ – see Algorithm 5.4.
    Play assignments $(M(t))_{t_n \leq t \leq t_n + \ell_{A(n)} - 1}$ and collect feedback on task-server pairs.
    Update $\hat{\theta}_{ij}(n)$ and $N_{ij}(n)$ for all $i, j$.
end for
```

Remark 5.1. *We remark that the frame size ℓ_A could grow exponentially in the number of servers s for some $A \neq A^{\mathsf{apf}}$. Although this may seem as a shortcoming of the design of* ES-APF*, we note that this leads to gathering more observations for the task-server pairs belonging to such an allocation A (with large ℓ_A), and thus a more accurate (smaller) index for $f(A, \theta)$. Consequently, A would be explored in fewer frames.*

In the following theorem, we provide a finite-time bound on the regret of ES-APF:

Theorem 5.2. *For all $T > 1$ and under policy $\pi =$*ES-APF*, we have:*

$$\mathfrak{R}_{\pi,T} \leq \frac{360 m^3 g(T)}{\theta_{\min} \Delta_{\min}} + \Big(8m + 2\varepsilon^{-2} \sum_{i,j} \theta_{ij}^{-2}\Big) \max_A \ell_A \Delta_A,$$

where $\varepsilon = (1 - 2^{-\frac{1}{3}}) \frac{\Delta_{\min}}{\max(m, \Delta_{\min})}$. Hence, $\mathfrak{R}_{\pi,T} = \mathcal{O}\Big(\frac{m^3}{\theta_{\min}\Delta_{\min}} \log(T)\Big)$ when $T \to \infty$.

Two comments are in order. (i) In view of Lemma 5.3, the dependence of regret upper bound of ES-APF on $\theta_{\min}$ is tight and cannot be improved further. We remark that such a tight dependence on $\theta_{\min}$ may not be guaranteed if one would use the UCB index for job success rates θ_{ij} instead of the KL-UCB index. Indeed, following similar steps as in the proof of Theorem 5.2, we find out that the best known regret bound with the UCB index, instead of KL-UCB, would grow as $\mathcal{O}(m^3 \Delta_{\min}^{-1} \theta_{\min}^{-2} \log(T))$. In the sequel, we also confirm the superiority of ES-APF over a similar algorithm that relies on UCB through numerical experiments.

(ii) In the case of combinatorial MABs with semi-bandit feedback, the regret of an optimal algorithm scales as $\mathcal{O}(m \times s)$. In contrast, we argue that in our problem and using frame-based policies, the regret should rather scale as m^2. This observation can be intuitively justified as follows. Note that for any chosen allocation A with the corresponding frame size ℓ_A, a given server j is shared among a_j tasks. Since at each time, only one of these tasks are sampled, each task-server pair (i, j) is actually sampled $A_{ij}\ell_A / a_j$ times. This implies that in order to obtain g samples for a given task-server pair (i, j), an allocation A has to be played $a_j g$ times. Hence, to obtain g samples for all task-server pairs, on average one effectively requires $\frac{1}{ms} \sum_{i,j} a_j = m/s$ more samples than the case of combinatorial MABs. Hence, there are effectively $m \times s \times \frac{m}{s} = m^2$ unknown parameters in the system.

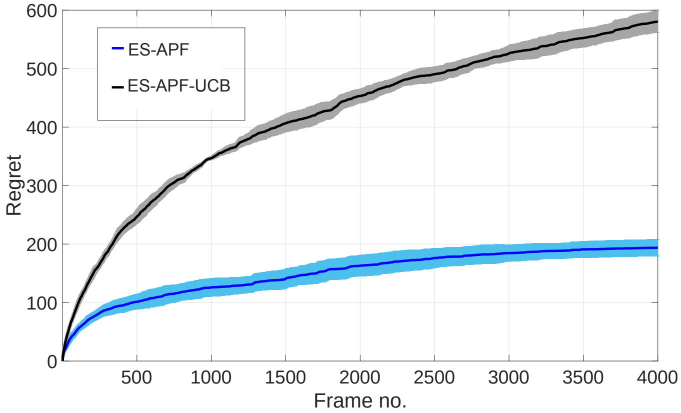


Figure 5.3: Regret of various algorithms averaged over 50 runs; 95% confidence intervals

Finally we mention that `ES-APF` has a polynomial time complexity per frame as one can compute the index for various task-server pairs efficiently, and that computation of $A(n)$ involves solving $\mathsf{PF}_1(b(n))$ and a matroid intersection problem, both of which can be solved in polynomial time.

A Numerical Example

We briefly illustrate the performance under our proposed algorithm through a simple numerical experiment. We consider a scenario comprising $m = 10$ tasks and $s = 3$ servers, where θ is sampled uniformly at random from $[0, 1]^{10\times 3}$. Figure 5.3 depicts the regret, averaged over 50 independent experiments, against the number of frames for a case where $\theta_{\min} = 0.06$ and where $(f^{\mathsf{pf}} - f^{\mathsf{apf}})/f^{\mathsf{pf}} = 1.4 \times 10^{-3}$.

We also compare the performance under `ES-APF` to a variant of `ES-APF` where confidence bounds on success probabilities are defined using the classical `UCB` index, which we refer to as `ES-APF-UCB`. Figure 5.3 shows that, as expected, `ES-APF` significantly outperforms `ES-APF-UCB`. The superior behavior of `ES-APF` over `ES-APF-UCB` have been confirmed for other set of parameters in our experiments as well.

5.7 Summary

We investigated a generic sequential resource allocation under semi-bandit feedback, where a decision maker wishes to quickly learn and converge to the PF allocation, or an approximate PF allocation. We presented a notion of approximate PF allocation, referred to as APF, which allows us to study the problem within the framework of combinatorial MAB. We derived a lower bound on the regret for learning APF allocation and presented an index policy for learning APF allocation enjoying a regret of order $\mathcal{O}(m^3\theta_{\min}^{-1}\Delta_{\min}^{-1}\log(T))$.

5.A Proof of Lemma 5.1

To prove the lemma, we first present a useful property of the optimal solution to problem $\mathsf{PF}_1(\theta)$. Throughout this section, the underlying parameter is θ. We therefore omit the dependence of various quantities on θ.

Lemma 5.4. *For all $j \in [s]$:*

$$S_{ij}^{\mathsf{pf}} = 1 \quad \textit{iff} \quad \frac{\theta_{ij}}{\gamma_i^{\mathsf{pf}}} = \max_{\ell} \frac{\theta_{\ell j}}{\gamma_\ell^{\mathsf{pf}}}.$$

Proof. Introduce the Lagrangian for problem $\mathsf{PF}_1(\theta)$:

$$L(z,\mu,\nu) = \sum_{i\in[m]} \log\Big(\sum_{j\in[s]} \theta_{ij}z_{ij}\Big) + \sum_{i\in[m]}\sum_{j\in[s]} \nu_{ij}z_{ij} - \sum_{j\in[s]} \mu_j\Big(\sum_{i\in[m]} z_{ij} - 1\Big).$$

KKT conditions at the optimal solution z^{pf} satisfy, for all i and j,

$$\begin{aligned} &\text{(i)} && \frac{\theta_{ij}}{\sum_{k\in[s]}\theta_{ik}z_{ik}^{\mathsf{pf}}} - \mu_j + \nu_{ij} = 0,\\ &\text{(ii)} && \nu_{ij}z_{ij}^{\mathsf{pf}} = 0,\\ &\text{(iii)} && \nu_{ij} \ge 0. \end{aligned}$$

Let $(i,j) \in [m]\times[s]$. If $S_{ij}^{\mathsf{pf}} = 1$, then $z_{ij}^{\mathsf{pf}} > 0$ so that (ii) implies $\nu_{ij} = 0$. Hence, by (i) we have $\theta_{ij}/\gamma_i^{\mathsf{pf}} = \mu_j$. Moreover, note that (i) and (iii) together imply that $\mu_j = \max_\ell \theta_{\ell j}/\gamma_\ell^{\mathsf{pf}}$. Therefore, $S_{ij}^{\mathsf{pf}} = 1$ implies $\theta_{ij}/\gamma_i^{\mathsf{pf}} = \max_\ell \theta_{\ell j}/\gamma_\ell^{\mathsf{pf}}$.

Now if $\theta_{ij}/\gamma_i^{\mathsf{pf}} = \max_\ell \theta_{\ell j}/\gamma_\ell^{\mathsf{pf}}$, one necessarily has $\nu_{ij} = 0$, which further implies $z_{ij}^{\mathsf{pf}} > 0$, and thus $S_{ij}^{\mathsf{pf}} = 1$. This completes the proof. □

Next we prove the lemma.

Proof of Lemma 5.1. For any $A \in \mathcal{A}$, we have that

$$\begin{aligned} f(z^{\mathsf{pf}},\theta) - f(z(A),\theta) &= \sum_i \log\gamma_i^{\mathsf{pf}} - \sum_{i,j} A_{ij}\log(\theta_{ij}/a_j)\\ &= \sum_j a_j\log a_j - \sum_{i,j} A_{ij}\log(\theta_{ij}/\gamma_i^{\mathsf{pf}})\\ &\ge \sum_j a_j\log a_j - \sum_j a_j\log(w_j^{\mathsf{pf}}), \end{aligned} \tag{5.6}$$

where the last inequality follows from the definition of w_j^{pf}. Furthermore, we deduce from Lemma 5.4 that inequality (5.6) holds with equality only if $A_{ij} = 1$ implies $S_{ij}^{\mathsf{pf}} = 1$. Recalling the definition of $\mathcal{A}$, the latter happens if $A \le S^{\mathsf{pf}}$.[2]

It then follows that

$$\begin{aligned} f(z^{\mathsf{pf}}, \theta) - \max_{A \in \mathcal{A}} f(z(A), \theta) &= f(z^{\mathsf{pf}}, \theta) - \max_{A \in \mathcal{A} : A \le S^{\mathsf{pf}}} f(z(A), \theta) \\ &= \min_{A \in \mathcal{A} : A \le S^{\mathsf{pf}}} \sum_j a_j \log(a_j / w_j^{\mathsf{pf}}). \end{aligned}$$

Noting that the constraint $A \le S^{\mathsf{pf}}$ further implies $a_j \le \sum_{i \in [m]} S_{ij}^{\mathsf{pf}}$ for $j \in [s]$, we get

$$\begin{aligned} f(z^{\mathsf{pf}}, \theta) - \max_{A \in \mathcal{A}} f(z(A), \theta) = \min_{w \in \mathbb{N}^s} \quad & \sum_{j \in [s]} w_j \log(w_j / w_j^{\mathsf{pf}}) \\ \text{subject to:} \quad & \sum_{j \in [s]} w_j = m, \\ & w_j \le \sum_{i \in [m]} S_{ij}^{\mathsf{pf}}, \quad \forall j \in [s], \end{aligned}$$

and the claim of the lemma follows directly. □

5.B Proof of Lemma 5.2

To prove the lemma, it suffices to show that any $A \in \mathcal{A}$, which is feasible for problem (5.4), is a basis for the intersection of two matroids. To verify this claim, we consider a bipartite graph $G = (U \cup S, E)$, where U and S respectively denote the set of tasks and servers, and where E denotes the set of task-server pairs in the support of z^{pf}. Fix $w \in \mathbb{N}^s$. Let $\delta(v) \in E$ denote the edge incident to node $v \in U$ or $v \in S$. Define

$$\begin{aligned} \mathcal{I}_1 &= \{F \subset E : |F \cap \delta(u)| \le 1,\ u \in U\}, \\ \mathcal{I}_2 &= \{F \subset E : |F \cap \delta(v)| \le w_v,\ v \in S\}. \end{aligned}$$

Then $M_1 = (E, \mathcal{I}_1)$ and $M_2 = (E, \mathcal{I}_2)$ define two partition matroids on E with respective ranks s and m. Let $I \subset E$ and denote by A its corresponding indicator matrix, that is $A_{ij} \in \{0, 1\}$ and $A_{ij} = 1$ iff $(i, j) \in I$. If $I \in \mathcal{I}_1$ then $\sum_{j=1}^{s} A_{ij} = 1$ for all i. Moreover, if $I \in \mathcal{I}_2$ then $\sum_{i=1}^{m} A_{ij} = w_j$ for all j. Hence, for each $A \in \mathcal{A}$ and $A \le S^{\mathsf{pf}}$, there exists $I \in \mathcal{I}_1 \cap \mathcal{I}_2$, and the claim follows. □

[2] Throughout this chapter, matrix inequalities are taken component-wise.

5.C Proof of Theorem 5.1

To derive the asymptotic lower bound on the regret, we apply [39, Theorem 1] (see Chapter 2 for an overview of the framework of [39]).

Let $\pi \in \Pi_f$ be a uniformly good policy. We first derive a lower bound for a notion of regret $\mathfrak{R}'_{\pi,F}$ that corresponds to the regret incurred after $F := F^\pi(T)$ frames. Observe that letting $\ell_{\max} := \max_{A \in \mathcal{A}} \ell_A$, we have the following relation: $\frac{T}{\ell_{\max}} \le F \le T$. As we shall see later, any asymptotic lower bound on $\mathfrak{R}'_{\pi,F}$ implies an asymptotic lower bound on $\mathfrak{R}_{\pi,T}$.

We construct the following controlled Markov chain. The state-space is $\{0,1\}^{m\times s\times \ell_{\max}}$. The control set is the set of allocations $\mathcal{A}$. The parameter $\theta = (\theta_{ij})_{i\in[m],j\in[s]}$ defines transition probabilities. The parameter θ takes value in the compact set $\Theta = [\varepsilon, 1]^{m\times s}$ for ε arbitrarily close to zero. The set of control laws are stationary and each of them corresponds to an allocation. A transition in the Markov chain occurs at time slots when a new frame is started, and the transition probabilities are $p(k,l;A,\theta) = p(l;A,\theta)$, that is they do not depend on the starting state.

For any two parameter matrices θ and λ, we define the KL-divergence under allocation A as:

$$I^A(\theta,\lambda) = \sum_y p(y;A,\theta)\log\frac{p(y;A,\theta)}{p(y;A,\lambda)}.$$

Consider a frame where allocation A is selected. The average reward under control law A is $\ell_A f(A,\theta)$. The incurred regret by the end of the frame is $\ell_A\Delta_A$. Further, in the course of the frame, for any i and j server j will be allocated to task i in ℓ_A/a_j time slots. Recalling that the outcome of various servers are independent and that the trials are independent across time, by additivity of the KL-divergence for independent random variables, we obtain

$$I^A(\theta,\lambda) = \sum_{i,j}\frac{\ell_A}{a_j}A_{ij}\texttt{kl}(\theta_{ij},\lambda_{ij}). \tag{5.7}$$

Finally, by applying [39, Theorem 1], we conclude that the regret under any uniformly good policy $\pi \in \Pi_f$ satisfies

$$\liminf_{F\to\infty}\frac{\mathfrak{R}'_{\pi,F}}{\log(F)} \ge c(\theta)\,,$$

where

$$\begin{aligned}c(\theta) &= \inf\Big\{\sum_A \Delta_A\ell_A u_A : u \ge 0,\ \inf_{\lambda\in B(\theta)}\sum_{A\neq A^{\mathsf{apf}}} I^A(\theta,\lambda)u_A \ge 1\Big\}\\ &= \inf\Big\{\sum_A \Delta_A\ell_A u_A : u \ge 0,\ \inf_{\lambda\in B(\theta)}\sum_{A\neq A^{\mathsf{apf}}} \ell_A u_A\sum_{i,j}\frac{A_{ij}}{a_j}\texttt{kl}(\theta_{ij},\lambda_{ij}) \ge 1\Big\}.\end{aligned} \tag{5.8}$$

Introducing $x_A = \ell_A u_A$ for any $A \in \mathcal{A}$ gives a lower bound for $\mathfrak{R}'_{\pi,F}$.

Now observe that $\mathfrak{R}'_{\pi,F} \le \mathfrak{R}_{\pi,T} \le \mathfrak{R}'_{\pi,F} + \ell_{\max}\Delta_{\max}$, where $\Delta_{\max} := \max_{A\in\mathcal{A}} \Delta_A$. We thus get, using the relation $\frac{T}{\ell_{\max}} \le F \le T$, that

$$\frac{\mathfrak{R}'_{\pi,F}}{\log(F)} \le \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \cdot \frac{\log(T)}{\log(T/\ell_{\max})} .$$

Letting $F \to \infty$ gives $\frac{\log(T)}{\log(T/\ell_{\max})} \to 1$. Therefore, for any $\beta > 0$,

$$\liminf_{F\to\infty} \frac{\mathfrak{R}'_{\pi,F}}{\log(F)} \ge \beta \implies \liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \ge \beta.$$

Putting this together with the derived lower bound on $\mathfrak{R}'_{\pi,F}$, we conclude the proof.
□

5.D Proof of Lemma 5.3

Decompose the set of bad parameters as $B(\theta) = \bigcup_{K\neq A^{\mathsf{apf}}} B_K(\theta)$, with

$$B_K(\theta) = \Big\{\lambda \in \Theta : \lambda_{ij} = \alpha, \forall (i,j) \in A^{\mathsf{apf}},\ f(K,\lambda) > f(A^{\mathsf{apf}},\theta)\Big\}.$$

By Theorem 5.1, the regret of any uniformly good algorithm $\pi \in \Pi_f$ for this problem instance satisfies:

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \ge c(\theta),$$

where $c(\theta)$ is the optimal value of the of following:

$$\inf_{x\ge 0} \sum_{A\neq A^{\mathsf{apf}}} \Delta_A x_A \tag{5.9}$$

$$\text{subject to:} \quad \inf_{\lambda\in B_K(\theta)} \sum_{(i,j)\in K\setminus A^{\mathsf{apf}}} \mathtt{kl}(\beta,\lambda_{ij}) \sum_A \frac{A_{ij}}{a_j} x_A \ge 1, \quad \forall K \neq A^{\mathsf{apf}}.$$

The rest of the proof proceeds in two steps.

Step 1: simplifying the objective of problem (5.9). Observe that $f(A^{\mathsf{apf}},\theta) = m\log(s\alpha/m)$. It then follows that, for any $A \neq A^{\mathsf{apf}}$,

$$\begin{aligned}\Delta_A &= m\log\alpha - \Big((m - |A\setminus A^{\mathsf{apf}}|)\log\alpha + |A\setminus A^{\mathsf{apf}}|\log\beta\Big) - m\log(m/s) + \sum_j a_j \log a_j \\ &= |A\setminus A^{\mathsf{apf}}|\log(\alpha/\beta) + \sum_j a_j\log a_j - m\log(m/s).\end{aligned}$$

Introducing

$$\delta_A := \exp\Big(\frac{1}{|A \setminus A^{\mathsf{apf}}|}\big[\sum_j a_j \log a_j - m\log(m/s)\big]\Big), \quad \forall A \neq A^{\mathsf{apf}},$$

we obtain $\Delta_A = |A \setminus A^{\mathsf{apf}}|\log(\alpha\delta_A/\beta)$. Observe that for any $A \in \mathcal{A}$, $\sum_j a_j = m$, and hence the log-sum inequality implies: $\sum_j a_j \log a_j \geq m\log(m/s)$, so that $\delta_A \geq 1$ for all $A \neq A^{\mathsf{apf}}$ and thus, $\log(\delta_A)$ is well-defined. Hence,

$$\sum_{A \neq A^{\mathsf{apf}}} \Delta_A x_A = \sum_{A \neq A^{\mathsf{apf}}} |A \setminus A^{\mathsf{apf}}|\log(\alpha\delta_A/\beta)x_A. \tag{5.10}$$

Step 2: simplifying the constraints of problem (5.9). Introduce the set of allocations that differ from A^{apf} by only one task-server pair: $\mathcal{A}' := \{A : |A \setminus A^{\mathsf{apf}}| = 1\}$. It follows that for any $A \in \mathcal{A}'$,

$$\delta_A := \overline{\delta} = \Big(\frac{m}{s}+1\Big)\log\Big(\frac{m}{s}+1\Big) + \Big(\frac{m}{s}-1\Big)\log\Big(\frac{m}{s}-1\Big) - \frac{2m}{s}\log\Big(\frac{m}{s}\Big) .$$

Hence,

$$\begin{aligned} c(\theta) \geq \inf_{x \geq 0} \;& \log(\alpha\overline{\delta}/\beta) \sum_{A \neq A^{\mathsf{apf}}} x_A \\ \text{subject to: } & \inf_{\lambda \in B_K(\theta)} \sum_{(i,j) \in K \setminus A^{\mathsf{apf}}} \mathtt{kl}(\beta, \lambda_{ij}) \sum_A \frac{A_{ij}}{a_j} x_A, \quad \forall K \in \mathcal{A}^-. \end{aligned}$$

Let $K \in \mathcal{A}^-$ and $\rho > 0$. By continuity of $z \mapsto \mathtt{kl}(\beta, z)$ for $z > \beta$, we can choose $\xi > \alpha\overline{\delta}$ such that

$$|\mathtt{kl}(\beta, \xi) - \mathtt{kl}(\beta, \alpha\overline{\delta})| \leq \rho\mathtt{kl}(\beta, \alpha\overline{\delta}) .$$

Now consider parameter $\tilde{\lambda}^K$ such that $\tilde{\lambda}^K_{ij} = \xi$ if $(i,j) \in K \setminus A^{\mathsf{apf}}$, and $\tilde{\lambda}^K_{ij} = \alpha$ for $(i,j) \in A^{\mathsf{apf}}$. Otherwise, $\tilde{\lambda}^K_{ij} = \beta$. It is straightforward to check that $\tilde{\lambda}^K \in B_K(\theta)$, and thus

$$\begin{aligned} \inf_{\lambda \in B_K(\theta)} \sum_{(i,j) \in K \setminus A^{\mathsf{apf}}} \mathtt{kl}(\beta, \lambda_{ij}) \sum_A \frac{A_{ij}}{a_j} x_A &\leq \sum_{(i,j) \in K \setminus A^{\mathsf{apf}}} \mathtt{kl}(\beta, \tilde{\lambda}^K_{ij}) \sum_A \frac{A_{ij}}{a_j} x_A \\ &= \mathtt{kl}(\beta, \xi) \sum_{(i,j) \in K \setminus A^{\mathsf{apf}}} \sum_A \frac{A_{ij}}{a_j} x_A. \end{aligned} \tag{5.11}$$

Hence, defining $\varepsilon = \frac{\rho}{1+\rho}$ and recalling the definition of $\mathcal{A}^-$, we get

$$c(\theta) \geq \inf_{x \geq 0} \; \log(\alpha\overline{\delta}/\beta) \sum_{A \neq A^{\mathsf{apf}}} x_A$$

$$\text{subject to: } \sum_A \frac{A_{ij}}{a_j} x_A \geq \frac{1-\varepsilon}{\texttt{kl}(\beta, \alpha\overline{\delta})}, \quad \forall (i,j) \notin A^{\mathsf{apf}}.$$

Noting that

$$\sum_{(i,j)\notin A^{\mathsf{apf}}} \sum_A \frac{A_{ij}}{a_j} x_A = \sum_A x_A \sum_j \frac{1}{a_j} \sum_{i:(i,j)\notin A^{\mathsf{apf}}} A_{ij} \leq s \sum_A x_A \ ,$$

we further obtain

$$c(\theta) \geq \inf_{x\geq 0} \ \log(\alpha\overline{\delta}/\beta) \sum_{A\neq A^{\mathsf{apf}}} x_A$$

$$\text{subject to: } \sum_A x_A \geq \frac{m(s-1)(1-\varepsilon)}{s\texttt{kl}(\beta, \alpha\overline{\delta})},$$

which implies $c(\theta) \geq \frac{m\log(\alpha\overline{\delta}/\beta)}{\texttt{kl}(\beta,\alpha\overline{\delta})}\big(1-\frac{1}{s}\big)$. Now observe that

$$\frac{\log(\alpha\overline{\delta}/\beta)}{\texttt{kl}(\beta,\alpha\overline{\delta})} \geq \frac{\alpha\overline{\delta}(1-\alpha\overline{\delta})}{(\alpha\overline{\delta}-\beta)^2}\log(\alpha\overline{\delta}/\beta) \geq \frac{1-\alpha\overline{\delta}}{\alpha\overline{\delta}-\beta} \geq \frac{1-\alpha\overline{\delta}}{\alpha\overline{\delta}\log(\alpha\overline{\delta}/\beta)},$$

where we used $\texttt{kl}(x,y) \leq \frac{(x-y)^2}{y(1-y)}$ for all $x, y \in (0,1)$ in the first inequality, and $\log z \geq 1 - \frac{1}{z}$ for all $z \geq 1$ (see, e.g., [98]) in the second and the third inequalities. We thus get

$$c(\theta) \geq \frac{1/\overline{\delta} - \alpha}{2\beta\log(\alpha\overline{\delta}/\beta)} \geq \frac{0.15}{\beta\log(\alpha\overline{\delta}/\beta)} \ ,$$

where the last inequality follows from the observation that under the assumption $m/s \geq 5$ we have $\overline{\delta} \leq 1.2231$. Moreover, it uses the fact that $\beta \geq \frac{\alpha}{2}$ and $\alpha \leq 0.5$. Hence,

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\pi,T}}{\log(T)} \geq \frac{0.15m}{\beta\log(\alpha\overline{\delta}/\beta)}\Big(1-\frac{1}{s}\Big).$$

The proof is completed by observing that $\theta_{\min} = \beta$ and $\Delta_{\min} = \log(\alpha\overline{\delta}/\beta)$. □

5.E Proof of Theorem 5.2

We first provide Lemma 5.5, which gives an upper bound on the index function u.

Lemma 5.5. *For all $N, n \in \mathbb{N}$, and $\lambda \in (0,1]$:*

$$\log(u(n,\lambda,N)) \leq \log(\lambda) + \sqrt{\frac{2g(n)}{N\lambda}}.$$

Proof. Let $N, n \in \mathbb{N}$ and $x, \lambda \in (0,1]$ with $x \geq \lambda$. We have:

$$\log(x) - \log(\lambda) \leq \frac{x-\lambda}{\sqrt{\lambda x}} = \sqrt{\frac{N(x-\lambda)^2}{x}}\sqrt{\frac{1}{N\lambda}} \leq \sqrt{2N\mathtt{kl}(\lambda, x)}\ \frac{1}{\sqrt{N\lambda}},$$

where the first inequality follows from Lemma 5.6, stated in Appendix 5.H, and the second is due to Lemma A.3. Hence, $N\mathtt{kl}(\lambda, x) \leq g(n)$ implies:

$$\log(x) \leq \log(\lambda) + \sqrt{\frac{2g(n)}{N\lambda}}.$$

The above holds for all $x \in [\lambda, 1]$, and thus, the lemma follows by the definition of $u(n, \lambda, N)$. □

Define $\kappa = 1 - 2^{-\frac{1}{3}}$ and $\varepsilon = \frac{\kappa\Delta_{\min}}{\max(m,\Delta_{\min})}$. Observe that $\varepsilon \leq \kappa$. Furthermore, for $N \in \mathbb{N}^{m\times s}$ and $A \in \mathcal{A}$ define $r(N,A) = \left(\sum_{i,j}\frac{A_{ij}}{\sqrt{N_{ij}}}\right)^2$. To ease notation, define $r(n) = r(N(n), A(n))$.

Next we prove the Theorem.

Proof of Theorem 5.2. Let $T \geq 1$ and denote by $F(T)$ the number of frames initiated by the algorithm up to time T (note that $F(T)$ is a bounded stopping time with $F(T) \leq T$). For any $n \geq 1$, the regret incurred in frame n is $\ell_{A(n)}\Delta_{A(n)}$. Hence,

$$\mathfrak{R}_{\pi,T} \leq \mathbb{E}\Big[\sum_{n=1}^{F(T)} \Delta_{A(n)}\ell_{A(n)}\mathbb{I}\{A(n) \neq A^{\mathsf{apf}}\}\Big].$$

For any frame n, define the following events:

$$B_{n,i,j} = \big\{b_{ij}(n) < \theta_{ij}\big\}, \quad B_n = \bigcup_{(i,j)\in A^{\mathsf{apf}}} B_{n,i,j},$$
$$C_{n,i,j} = \big\{A_{ij}(n) = 1,\ |\hat\theta_{ij}(n) - \theta_{ij}| \geq \varepsilon\theta_{ij}\big\}, \quad C_n = \bigcup_{i,j} C_{n,i,j}.$$

Moreover, for any time t, let n_t denote the frame to which t belongs, and define

$$D_t = \Big\{\Delta_{A(n_t)} \leq (1-\kappa)^{-3/2}\theta_{\min}^{-1/2}\sqrt{2g(T)r(n_t)}\Big\}.$$

Consider a time slot t where $A(n_t) \neq A^{\mathsf{apf}}$. We show that $A(n_t) \neq A^{\mathsf{apf}}$ implies: $t \in B_{n_t} \cup C_{n_t} \cup D_t$. First observe that

$$f(A(n_t), b(n_t)) = \max_{A\in\mathcal{A}} f(A, b(n_t)) \geq f(A^{\mathsf{apf}}, b(n_t)).$$

Assume that $B_{n_t} \cup C_{n_t}$ does not occur. One the one hand, $b_{ij}(n_t) \geq \theta_{ij}$ for all $(i,j) \in A^{\mathsf{apf}}$. Hence, $f(A^{\mathsf{apf}}, b(n_t)) \geq f(A^{\mathsf{apf}}, \theta)$ since $f(\cdot, \theta)$ is increasing in θ. Since C_{n_t} does not occur, for all i and j, we have

$$(1-\kappa)\theta_{\min} \leq \hat{\theta}_{ij}(n_t) \leq (1+\varepsilon)\theta_{ij},$$

where we used $\varepsilon \leq \kappa$. Using the above results together with Lemma 5.5, we deduce

$$\begin{aligned}
&\big\{A(n_t) \neq A^{\mathsf{apf}},\ \overline{B}_{n_t},\ \overline{C}_{n_t}\big\} \subset \big\{f(A(n_t), b(n_t)) \geq f^{\mathsf{apf}},\ \overline{C}_{n_t}\big\} \\
&\subset \left\{\sum_{i,j} A_{ij}(n_t)\sqrt{\frac{2g(T)}{N_{ij}(n_t)\hat{\theta}_{ij}(n_t)}} + f(A(n_t), \hat{\theta}(n_t)) \geq f^{\mathsf{apf}},\ \overline{C}_{n_t}\right\} \\
&\subset \left\{\sqrt{\frac{2g(T)}{(1-\kappa)\theta_{\min}}}\sum_{i,j}\frac{A_{ij}(n_t)}{\sqrt{N_{ij}(n_t)}} + f(A(n_t), (1+\varepsilon)\theta) \geq f^{\mathsf{apf}}\right\} \\
&\subset \left\{(1-\kappa)^{-1/2}\theta_{\min}^{-1/2}\sqrt{2g(T)r(n_t)} + f(A(n_t), \theta) + \kappa\Delta_{A(n_t)} \geq f^{\mathsf{apf}}\right\} \qquad (5.12) \\
&\subset \left\{\Delta_{A(n_t)} \leq (1-\kappa)^{-3/2}\theta_{\min}^{-1/2}\sqrt{2g(T)r(n_t)}\right\},
\end{aligned}$$

where (5.12) follows from that fact that for any $A \in \mathcal{A}$, we have

$$f(A, (1+\varepsilon)\theta) - f(A, \theta) = \sum_{i,j} A_{ij}\log(1+\varepsilon) \leq m\varepsilon = \frac{m\kappa\Delta_{\min}}{\max(m, \Delta_{\min})} \leq \kappa\Delta_A,$$

where we used $\log(z+1) \leq z$ for all $z > -1$. Hence, $A(n_t) \neq A^{\mathsf{apf}}$ implies: $t \in B_{n_t} \cup C_{n_t} \cup D_t$.

Hence, the regret $\mathfrak{R}_{\pi,T}$ is upper bounded by:

$$\begin{aligned}
\mathfrak{R}_{\pi,T} &\leq \mathbb{E}\Big[\sum_{n=1}^{F(T)} \Delta_{A(n)}\ell_{A(n)}\mathbb{I}\{A(n) \neq A^{\mathsf{apf}}\}\Big] \\
&\leq \mathbb{E}\Big[\sum_{n=1}^{F(T)} \ell_{A(n)}\Delta_{A(n)}(\mathbb{I}\{B_n\} + \mathbb{I}\{C_n\})\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}\Delta_{A(n_t)}\mathbb{I}\{D_t\}\Big] \\
&\leq (\max_A \ell_A\Delta_A)\sum_{n=1}^{T}(\mathbb{P}(B_n) + \mathbb{P}(C_n)) + \mathbb{E}\Big[\sum_{t=1}^{T}\Delta_{A(n_t)}\mathbb{I}\{D_t\}\Big],
\end{aligned}$$

where we used $F(T) \leq T$. We will prove the following inequalities:

$$\begin{aligned}
&\sum_{n=1}^{T}\mathbb{P}(B_n) \leq 8m, \\
&\sum_{n=1}^{T}\mathbb{P}(C_n) \leq 2\varepsilon^{-2}\sum_{i,j}\theta_{ij}^{-2},
\end{aligned}$$

$$\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{1}\{D_t\}\Big] \leq 360 m^2 g(T)(\theta_{\min}\Delta_{\min})^{-1}.$$

Hence as announced:

$$\mathfrak{R}_{\pi,T} \leq \frac{360 m^3 g(T)}{\theta_{\min}\Delta_{\min}} + (\max_A \ell_A \Delta_A)\Big(8m + 2\varepsilon^{-2}\sum_{i,j}\theta_{ij}^{-2}\Big).$$

Set B_n. By Theorem B.5 ([43, Theorem 10]), we have $\mathbb{P}(B_{n,i,j}) \leq \lceil g(n)\log(n)\rceil e^{1-g(n)}$. Using a union bound and since $\sum_{i,j} A_{ij}^{\mathsf{apf}} = m$,

$$\sum_{n=1}^{T} \mathbb{P}(B_n) \leq \sum_{(i,j)\in A^{\mathsf{apf}}} \sum_{n=1}^{T} \mathbb{P}(B_{n,i,j}) \leq 8m. \tag{5.13}$$

Set C_n. Define $\tau_{ij}(n) = \sum_{n'=1}^{n} \mathbb{I}\{C_{n',i,j}\}$. Since $C_{n',i,j}$ implies $A_{ij}(n') = 1$, we have $N_{ij}(n) \geq \tau_{ij}(n)$. Applying Theorem B.7 ([70, Lemma B.1]), we have $\sum_{n=1}^{T} \mathbb{P}(C_{n,i,j}) \leq 2(\varepsilon\theta_{ij})^{-2}$. A union bound yields:

$$\sum_{n=1}^{T} \mathbb{P}(C_n) \leq 2\varepsilon^{-2}\sum_{i,j}\theta_{ij}^{-2}. \tag{5.14}$$

Set D_t. To derive an upper bound on the regret incurred due to event D_t, we borrow some techniques from [62]. Define $V = 2m^2 g(T)(1-\kappa)^{-3}\theta_{\min}^{-1}$. Similarly to the proof of [62, Theorem 5], consider $\alpha, \beta > 0$, and for $l \in \mathbb{N}$ define $\alpha_l = \left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^2 \alpha^l$ and $\beta_l = \beta^l$. Introduce set $S_{l,t}$ and event $G_{l,t}$:

$$\begin{aligned} S_{l,t} &= \{(i,j) \in A(n_t),\ N_{ij}(n_t) \leq V\alpha_l \Delta_{A(n_t)}^{-2}\}, \\ G_{l,t} &= \{|S_{l,t}| \geq \beta_l m\} \cap \{|S_{k,t}| < \beta_k m,\ k = 1,\dots,l-1\}. \end{aligned}$$

If $\overline{\cup_{l\geq 1} G_{l,t}} = \{|S_{l,t}| < m\beta_l, l \geq 1\}$ occurs, then:

$$\begin{aligned} \sum_{l\geq 1} \frac{|S_{l-1,t}| - |S_{l,t}|}{\sqrt{\alpha_l}} &= \frac{|S_{0,t}|}{\sqrt{\alpha_1}} + \sum_{l\geq 1} |S_{l,t}|\Big(\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_l}}\Big) \\ &< \frac{m\beta_0}{\sqrt{\alpha_1}} + \sum_{l\geq 1} m\beta_l\Big(\frac{1}{\sqrt{\alpha_{\ell+1}}} - \frac{1}{\sqrt{\alpha_l}}\Big) \\ &= m\sum_{l\geq 1} \frac{\beta_l - \beta_{l-1}}{\sqrt{\alpha_l}} \leq m, \end{aligned}$$

since $\frac{1}{\sqrt{\alpha_{l+1}}} - \frac{1}{\sqrt{\alpha_l}} \geq 0$. Observe that

$$|\{(i,j) : N_{ij}(n_t) \in V\Delta_{A(n_t)}^{-2}[\alpha_l, \alpha_{l-1}]\}| = |S_{l-1,t}| - |S_{l,t}|.$$

Hence,

$$\sqrt{r(n_t)} \leq \sum_{l \geq 1} \frac{|S_{l-1,t}| - |S_{l,t}|}{\sqrt{\alpha_l}} \frac{\Delta_{A(n_t)}}{\sqrt{V}} < m \frac{\Delta_{A(n_t)}}{\sqrt{V}}.$$

Hence $\Delta^2_{A(n_t)} > Vm^{-2} r(n_t)$, and D_t does not occur. Therefore, $D_t \subset \cup_{l \geq 1} G_{l,t}$ and:

$$\sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\{D_t\} \leq \sum_{t=1}^{T} \sum_{l \geq 1} \Delta_{A(n_t)} \mathbb{I}\{G_{l,t}\}.$$

We further decompose $G_{l,t}$ as:

$$G_{i,j,l,t} = G_{l,t} \cap \big\{(i,j) \in A(n_t),\ N_{ij}(n_t) \leq V \alpha_l \Delta^{-2}_{A(n_t)}\big\}.$$

Observe that:

$$\mathbb{I}\{G_{l,t}\} \leq \frac{|S_{l,t}|}{m \beta_l} \mathbb{I}\{G_{l,t}\} = \frac{1}{m \beta_l} \sum_{i,j} \mathbb{I}\{G_{i,j,l,t}\}.$$

Hence,

$$\begin{aligned}
\sum_{i,j} \sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\{G_{i,j,l,t}\} &\leq \sum_{i,j} \sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\Big\{A_{ij}(n_t) = 1,\ N_{ij}(n_t) \leq \frac{V \alpha_l}{\Delta^2_{A(n_t)}}\Big\} \\
&\leq \sum_{i,j} \sum_{t=1}^{T} \sum_{\tau=1}^{V \alpha_l \Delta^{-2}_{\min}} \mathbb{I}\{A_{ij}(n_t) = 1,\ N_{ij}(n_t) = \tau\} \sqrt{\frac{V \alpha_l}{\tau}} \\
&\leq m^2 \sum_{\tau=1}^{V \alpha_l \Delta^{-2}_{\min}} \sqrt{\frac{V \alpha_l}{\tau}},
\end{aligned}$$

where in the second line we used $\sum_{i,j} \sum_{t=1}^{T} \mathbb{I}\{A_{ij}(n_t) = 1,\ N_{ij}(n_t) = \tau\} \leq m^2$. Using the inequality $\sum_{t=1}^{T} t^{-\frac{1}{2}} \leq 1 + \int_1^T t^{-\frac{1}{2}} dt \leq 2\sqrt{T}$ yields

$$\sum_{i,j} \sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\{G_{i,j,l,t}\} \leq \frac{2m^2 V \alpha_l}{\Delta_{\min}},$$

so that

$$\sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\{D_t\} \leq \sum_{i,j} \sum_{t=1}^{T} \sum_{l \geq 1} \Delta_{A(n_t)} \frac{1}{m \beta_l} \mathbb{I}\{G_{i,j,l,t}\} \leq \frac{2mV}{\Delta_{\min}} \sum_{l \geq 1} \frac{\alpha_l}{\beta_l}.$$

By choosing $\alpha = 0.15$ and $\beta = 0.24$ so that $\sum_{l\geq 1} \frac{\alpha_l}{\beta_l} \leq 45$, we obtain

$$\sum_{t=1}^{T} \Delta_{A(n_t)} \mathbb{I}\{D_t\} \leq \frac{90mV}{\Delta_{\min}} \; . \tag{5.15}$$

Combining (5.13), (5.14), and (5.15) yields the desired result and concludes the proof. □

5.F Dynamic Programming for Solving Problem (5.4)

Consider the following problem:

$$\begin{aligned} \min_{x\in\mathbb{N}^s} \quad & \sum_{j=1}^{s} x_j \log(x_j/\alpha_j) \qquad (5.16) \\ \text{subject to:} \quad & \sum_{j=1}^{s} x_j = m \; , \\ & x_j \leq c_j, \; \forall j. \end{aligned}$$

The above problem can be solved using dynamic programming. To this end, for any $j \in [s]$ and $u \in \mathbb{N}$, we introduce

$$\delta_j(u) = (u+1)\log\Big(\frac{u+1}{\alpha_j}\Big) - u\log\Big(\frac{u}{\alpha_j}\Big) \; .$$

Algorithm 5.3 describes the pseudo-code of dynamic programming for solving problem (5.16).

Algorithm 5.3 Dynamic Programming for Problem (5.16)

```
Set $X(k, j, w) = 1$ for all $j, k, w \in [s]$.
for $j = 1..s$ do
    $X(k, j, s+1) \leftarrow X(k, j, s) + \mathbb{I}\{k = j\}, \;\; \forall k$
    $V(j, s+1) \leftarrow V(j, s) + \delta_j(X(j, j, s))\mathbb{I}\{k = j\}, \;\; \forall k$
end for
for $j = 1..s$ do
    for $w = s+2..m$ do
        Let $k_j \in \mathrm{argmin}_{j:X(k,j,w-1)\leq c_k} \delta_k(X(k, j, w-1))$.
        $X(k, j, w) \leftarrow X(k, j, w-1) + \mathbb{I}\{k = k_j\}$
        $V(j, w) \leftarrow V(j, w-1) + \delta_{k_j}(X(k_j, j, w-1))\mathbb{I}\{k = k_j\}$
    end for
end for
```

5.G Decomposition to Assignments

We present a simple algorithm to decompose a given allocation matrix A to a minimal sequence of assignments $(M(k))_{k\in\Lambda}$ satisfying:

$$M(k) \in \mathcal{M}, \qquad \forall k \in \Lambda, \tag{5.17}$$

$$\frac{1}{|\Lambda|}\sum_{k\in\Lambda} M_{ij}(k) = z_{ij}(A), \quad \forall i, \forall j. \tag{5.18}$$

Here, Λ denotes the set of time slots in a given frame where allocation A is chosen. In particular, $|\Lambda| = \ell_A$.

Algorithm 5.4 Decomposition of Allocation A

```
$c \leftarrow \ell_A$
Let matrix $B$ with $B_{ij} = cA_{ij}/a_j$ for all $i$ and $j$.
$k \leftarrow 0$
while $B \neq \mathbf{0}$ do
  $M(k) \leftarrow \mathbf{0}$
  for $j = 1..s$ do
    Let $i_0 \in \operatorname{argmin}_{i:B_{ij}>0} B_{ij}$.
    $M_{i_0 j}(k) \leftarrow 1$
    $B_{i_0 j} \leftarrow B_{i_0 j} - 1$
  end for
  $k \leftarrow k + 1$
end while
Output $(M(k))_k$.
```

A simple procedure to make a sequence $(M(k))_k$ satisfying (5.17)-(5.18) is provided in Algorithm 5.4. Observe that for any j, it holds that $\sum_i M_{ij}(k) = M_{i_0 j}(k) = 1$. Moreover, by the design of the algorithm, $M_{ij}(k) = 1$ implies $B_{ij} > 0$. Recalling that $B_{ij} > 0$ iff $A_{ij} = 1$, and that $A \in \mathcal{A}$, we deduce $\sum_j M_{ij}(k) \leq 1$ so that $M(k) \in \mathcal{M}$. Moreover, for any i and j, by the design of the algorithm, task-server pair (i, j) will be chosen B_{ij} times. Hence,

$$\frac{1}{c}\sum_{k=1}^{c} M_{ij}(k) = \frac{B_{ij}}{c} = \frac{A_{ij}}{a_j} = z_{ij}(A),$$

so that constraint (5.18) is satisfied.

5.H Technical Lemmas

Lemma 5.6. *For all $x \geq 1$, we have $\log x \leq \frac{x-1}{\sqrt{x}}$.*

Proof. Consider the function $u(x) = (x-1)/\sqrt{x} - \log x$ defined for $x \geq 1$. Clearly, f is continuous in $[1, +\infty)$ and $u(1) = 0$. Furthermore,

$$u'(x) = \frac{2x - x + 1}{2x\sqrt{x}} - \frac{1}{x} = \frac{(\sqrt{x}-1)^2}{2x\sqrt{x}} \geq 0\,, \quad \forall x > 1.$$

Hence, $u(x) \geq 0$ for all $x \geq 1$ and the claim of the lemma follows. □

Chapter 6

Variance-Aware Regret Bounds for Undiscounted RL

In this chapter[1], we consider Reinforcement Learning (RL) in an unknown and discrete Markov Decision Process (MDP) under the average-reward criterion, when the decision maker interacts with the system in a single stream of observations, starting from an initial state without any reset. More formally, let $M = (\mathcal{S}, \mathcal{A}, \nu, P)$ denote an MDP where $\mathcal{S}$ is a finite set of states and $\mathcal{A}$ is a finite set of actions available at any state, with respective cardinalities S and A. ν and P denote the reward function and transition kernel, respectively. The game goes as follows: The decision maker starts in some state $s_1 \in \mathcal{S}$ at time $t = 1$. At each time step $t \in \mathbb{N}$, the decision maker chooses one action $a \in \mathcal{A}$ in her current state $s \in \mathcal{S}$ based on her past decisions and observations. When executing action a in state s, she receives a random reward r drawn independently from distribution $\nu(s, a)$ with support $[0, 1]$ and mean $\mu(s, a)$. The state then transits to a next state $s' \in \mathcal{S}$ sampled with probability $p(s'|s, a)$, and a new decision step begins. As the transition probabilities and reward functions are unknown, the decision maker has to learn them by trying different actions and recording the realized rewards and state transitions. For background material on RL and MDPs, we refer to Chapter 2.

The performance of the decision maker can be quantified through the notion of regret, which compares her collected reward to that obtained by an oracle always following an optimal policy, where a policy is a mapping from states to actions. More formally, let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ denote a possibly stochastic policy. We further introduce the notation $p(s'|s, \pi(s)) = \mathbb{E}_{Z\sim\pi(s)}[p(s'|s, Z)]$, and $P_\pi f$ to denote the function $s \mapsto \sum_{s'\in\mathcal{S}} p(s'|s, \pi(s))f(s')$. Likewise, let $\mu_\pi(s) = \mathbb{E}_{Z\sim\pi(s)}[\mu(s, Z)]$ denote the mean reward after choosing action $\pi(s)$ in step s.

We can now introduce the notion of regret. Given a learning algorithm $\mathbb{A}$, consider the following quantity that compares the cumulative reward after T steps

[1] Part of the research presented in this chapter was carried out while the author was a visiting PhD student at SequeL team in INRIA - Lille Nord Europe. The main part of this chapter is based on the work [99].

obtained by an optimal algorithm (denoted by $\star$) to that obtained by $\mathbb{A}$:

$$\mathrm{Reg}_{\mathbb{A},T} := \sum_{t=1}^{T} r(s_t^\star, \star(s_t^\star)) - \sum_{t=1}^{T} r(s_t, a_t)\,,$$

where $a_t = \mathbb{A}(s_t, (\{s_{t'}, a_{t'}, r_{t'}\})_{t'<t})$ and $s_t^\star \sim p(\cdot|s_t^\star, \star(s_t^\star))$ with $s_1^\star = s_1$ is a sequence of states generated by the optimal strategy, and finally $r(s,a) \sim \nu(s,a)$.

By an application of Azuma-Hoeffding's inequality for bounded martingales, it is immediate to show that with probability higher than $1-\delta$,

$$\begin{aligned}\mathrm{Reg}_{\mathbb{A},T} &\le \sum_{t=1}^{T}\Big(P_\star^{t-1}\mu_\star - r(s_t,a_t)\Big) + \sqrt{2T\log(1/\delta)}\\ &= \sum_{t=1}^{T}(P_\star^{t-1} - \overline{P}_\star)\mu_\star + \left[Tg^\star - \sum_{t=1}^{T} r(s_t,a_t)\right] + \sqrt{2T\log(1/\delta)}\,.\end{aligned}$$

where for any policy π, $\overline{P}_\pi = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} P_\pi^{t-1}$.

Thus, following [25], it makes sense to focus on the control of the middle term in brackets only, which we now call the *effective regret*:

$$\mathrm{Regret}_{\mathbb{A},T} := Tg^\star - \sum_{t=1}^{T} r(s_t,a_t)\ .$$

6.1 Motivation and Contributions

To date, several algorithms have been proposed in order to minimize the regret based on the "*optimism in the face of uncertainty*" principle, coming from the literature on stochastic MABs (see [16]). Algorithms designed based on this principle typically maintain confidence bounds on the unknown reward and transition distributions, and choose an optimistic model that leads to the highest average long-term reward. One of the first algorithms based on this principle for MDPs is Burnetas-Katehakis [22], which is shown to be asymptotically optimal; we refer to Chapter 2 for the description of this algorithm. Burnetas-Katehakis uses the KL-divergence to define confidence bounds for transition probabilities. Subsequent studies by [100], [51], [25], and [101] propose algorithms that maintain confidence bounds on transition kernel defined by L_1 or total variation norm. The use of L_1 norm, instead of KL-divergence, allows one to describe the uncertainty of the transition kernel by a *polytope*, which in turn brings computational advantages and ease in the regret analysis. On the other hand, such polytopic models are typically known to provide poor representations of underlying uncertainties; for a thorough discussion on this matter, we refer to the literature on the robust control of MDPs with uncertain transition kernels, e.g., [23], and more appropriately to [24]. Indeed, as argued in [24], optimistic models designed by L_1 norm suffer from two shortcomings:

(i) The L_1 optimistic model could lead to *inconsistent* models by assigning a zero mass to an already observed element.

(ii) Due to polytopic shape of L_1-induced confidence bounds, the maximizer of a linear optimization over L_1 ball could significantly vary for a small change in the value function, thus resulting in sub-optimal exploration.

These shortcomings are discussed in [24], and we refer to pages 120–121 there for relevant illustrations.

Both of these shortcomings are avoided by using the Kullback-Leibler (KL) divergence and the properties of corresponding KL-ball. In [24], the authors introduce the KL-Ucrl algorithm that modifies Ucrl2 [25] by replacing L_1 norms with KL-divergences in order to define the confidence bound on transition probabilities. Further, they provide an efficient way to carry out linear optimization over the KL-ball, which is necessary in each iteration of the Extended Value Iteration. Despite these favorable properties and the strictly superior performance in numerical experiments (even for very short time horizons), the best known regret bound for KL-Ucrl matches that of Ucrl2. Hence, from a theoretical perspective, the potential gain of use of KL-divergence has remained largely unexplored.

6.1.1 Contributions of the Chapter

The main objective in this chapter is to investigate the benefits of KL-based confidence bounds for regret minimization in RL. In particular, we study KL-Ucrl and provide a new high-probability regret bound for it, scaling as

$$\widetilde{\mathcal{O}}\Big(\sqrt{S\textstyle\sum_{s,a}\mathbf{V}^\star_{s,a}T} + D\sqrt{T}\Big),$$

for ergodic MDPs with S states, A actions, and diameter D. Here, $\mathbf{V}^\star_{s,a} := \mathbb{V}_{p(\cdot|s,a)}(b^\star)$ denotes the variance of the optimal bias function $b^\star$ of the true (unknown) MDP with respect to next state distribution under state-action (s,a). This bound improves over the best previous bound $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for KL-Ucrl [24] as $\sqrt{\mathbf{V}^\star_{s,a}} \le D$. Interestingly, in several examples $\sqrt{\mathbf{V}^\star_{s,a}} \ll D$ and actually $\sqrt{\mathbf{V}^\star_{s,a}}$ is comparable to $\sqrt{D}$. Our numerical experiments on typical MDPs further confirm that $\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}}$ could be much smaller than $DS\sqrt{A}$. To prove the above regret bound, we provide novel concentration inequalities inspired by the transportation method that relate the so-called transportation cost under two discrete probability measures to the KL-divergence between the two measures and the associated variances. These concentration inequalities enable to decouple the concentration properties of the transition kernel from the specific structure of the involved bias (or value) functions. To the best of our knowledge, these inequalities are new and could be independently interesting.

Leveraging these inequalities also enables us to simplify the (implicit) regret bound of the Burnetas-Katehakis algorithm and obtain an explicit one asymptotically growing as

$$\mathcal{O}\Big(\Big[\sum_{s,a}\frac{\mathbf{V}^\star_{s,a}}{\varphi(s,a)}+SA\Psi\Big]\log(T)\Big),$$

where Ψ denotes the span[2] of bias function and $\varphi(s,a)$ is a notion of gap between the average rewards of an optimal and of a sub-optimal action a in state s. To the best of our knowledge, these regret bounds are the first to provide such insights into the benefit of KL-divergence over L_1 norm for RL.

In order to justify these results, we revisit existing lower bounds on the regret for the considered setup to make appear their dependence on the aforementioned variance terms. Specifically, building on the minimax regret lower bound of Jaksch et al. [25], we provide an alternative lower bound of order $\Omega(\sqrt{SA\mathbf{V}_{\max}T})$, where $\mathbf{V}_{\max} := \max_{s,a}\mathbf{V}^\star_{s,a}$. In view of this lower bound, our regret upper bound for KL-Ucrl can be improved by only a factor $\sqrt{S}$. Further, we study a family of ergodic MDPs and show that an application of Burnetas-Katehakis lower bound [22] to them leads to an explicit regret lower bound growing at least as $\Omega\Big(\sum_{s,a}\frac{\mathbf{V}^\star_{s,a}}{\varphi(s,a)}\log(T)\Big)$ as T grows large.

6.2 Related Work

The study of RL in MDPs under average-reward criterion dates back to the seminal papers by Graves and Lai [39], and Burnetas and Katehakis [22]. This line of research was further followed over the last decade by several studies including [100, 25, 102, 24, 103, 104].

In the sequel, we briefly discuss the contribution of these papers. Under some reasonable assumption, Burnetas and Katehakis [22] provide an MDP-dependent asymptotic lower bound of the form $c_{\text{bk}}\log(T)$ on the regret for the class of ergodic MDPs, where c_{bk} is an implicit MDP-dependent constant (see Theorem 2.6 for a precise definition). Furthermore, they propose an index policy, which we refer to as Burnetas-Katehakis (Algorithm 2.3), that achieves this regret bound asymptotically. Tewari and Bartlett [100] study RL for the same class of MDPs and propose OLP (Optimistic Linear Programming), which is quite similar to Burnetas-Katehakis except that instead of the KL-divergence, it uses L_1 norm to define confidence bounds. OLP is computationally simpler than Burnetas-Katehakis and enjoys a regret bound scaling as $\mathcal{O}\left(\frac{SA\Psi^2}{\Phi}\log(T)\right)$, where $\Phi = \min_{(s,a)}\varphi(s,a)$. We note that both OLP and Burnetas-Katehakis rely on the knowledge of average reward functions and satisfy regret bounds that grow asymptotically logarithmically in T. Furthermore, these asymptotic bounds hide an additive term that could be exponential in the number of state S (cf. [25, p. 1566], [22]).

Auer and Ortner [51] study the same problem for the larger class of unichain MDPs but under less restrictive assumptions. They propose Ucrl whose regret

[2] The span of function f that takes values in a set $\mathcal{X}$ is defined as $\mathbb{S}(f) = \max_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} f(x)$.

scales at most as $\mathcal{O}\left(\frac{S^5 A T_M \kappa_M^2}{\Delta^2}\log(T)\right)$ uniformly over time, where T_M denotes the mixing time of MDP (see Definition 2.5) and κ_M is a constant defined in terms of hitting times of stationary policies in the true MDP. Jaksch et al. [25, 102] propose Ucrl2 as a generalization of Ucrl for the class of communicating MDPs (see Algorithm 2.4). Ucrl2 achieves a regret of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ with high probability uniformly over time. They also provide a problem-dependent regret bound for Ucrl2 growing as $\widetilde{\mathcal{O}}\left(\frac{D^2S^2A}{\Delta}\log(T)\right)$, where Δ denotes the smallest gap between the average reward of an optimal policy and of a sub-optimal policy. Inspired by Ucrl2, Filippi et al. [24] propose KL-Ucrl, which enjoys the same performance guarantee as Ucrl2, though numerically it shows a superior performance. KL-Ucrl has a similar design as Ucrl2 except that it maintains confidence intervals based on the KL-divergence instead of L_1 norm. For the class of communicating MDPs, Jaksch et al. [25] also establish a non-asymptotic minimax lower bound on the expected regret scaling as $\Omega(\sqrt{DSAT})$. Thus, the regret bounds of Ucrl2 and KL-Ucrl are far from the optimal scaling, at most, by a factor $\sqrt{DS}$.

Bartlett and Tewari [101] address the larger class of weakly communicating MDPs, under the assumption that reward functions are known and that an upper bound D' on the span of the bias function is given. Their proposed algorithm, Regal, is inspired by Ucrl2 but also uses the idea of regularization. It attains a $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$ regret with high probability. It is however still an open problem to incorporate the imposed assumptions into an implementable algorithm. Ortner [105, 106] studies learning in unknown MDPs with deterministic transitions. Despite the similarity of this setting to the classical MAB, one cannot directly use corresponding algorithms since the number of policies grows exponentially in S. Ortner [105, 106] presents UCycle, which is an adaptation of Ucrl2 to the case of deterministic transitions and achieves a regret of $\mathcal{O}(\frac{SA}{\Delta}\log(T))$.

There are two recent studies [107, 103] that present algorithms for RL under average-reward criterion based on posterior sampling. Under the assumption of known reward function and *known time horizon*, the algorithm of Agrawal and Jia [103] enjoys a regret scaling as $\widetilde{\mathcal{O}}\left(D\sqrt{SAT} + DS^{7/4}A^{3/4}T^{1/4}\right)$. In particular, for $T \ge S^5A$, this bound grows as $\widetilde{\mathcal{O}}\left(D\sqrt{SAT}\right)$, which constitutes the best known regret upper bound for learning in communicating MDPs and has tight dependencies on S and A. The TSDE algorithm by Ouyang et al. [107] achieves a regret growing as $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$ for the class of weakly communicating MDPs, where D' is a given bound on the span of the bias function. We refer to Tables 6.1 and 6.2 for a summary of these results.

All the papers cited above consider RL in MDPs with finite state-space. Undiscounted RL in continuous state-space is recently investigated in a few studies, e.g., by Ortner and Ryabko [108], Ortner [109], Lakshmanan et al. [110]. Regret bounds reported in these works hold under various assumptions on the structure of reward and transition functions. Ortner and Ryabko [108] investigate undiscounted RL with continuous state-space under a fairly general setting in which only smooth-

Algorithm	**Setting**	**Regret**
Burnetas-Katehakis [22]	ergodic, known rewards	$\mathcal{O}(c_{\mathrm{bk}} \log(T))$ – asymptotically
Burnetas-Katehakis [22] (Theorem 6.2)	ergodic, known rewards	$\mathcal{O}\Big(\big(\frac{SA\mathbf{V}_{\max}}{\Phi} + DSA\big)\log(T)\Big)$ – asymptotically
OLP [100]	ergodic, known rewards	$\mathcal{O}\left(\frac{D^2SA}{\Phi}\log(T)\right)$ – asymptotically
Ucrl [51]	unichain	$\mathcal{O}\left(\frac{S^5AT_M\kappa}{\Delta^2}\log(T)\right)$
Ucrl2 [102, 25]	communicating	$\mathcal{O}\left(\frac{D^2S^2A}{\Delta}\log(T)\right)$
KL-Ucrl [24]	communicating	$\mathcal{O}\left(\frac{D^2S^2A}{\Delta}\log(T)\right)$
Lower Bound [22]	ergodic (generic), known rewards	$\Omega(c_{\mathrm{bk}} \log(T))$ – asymptotically
Lower Bound (Proposition 6.1)	ergodic (specific instance)	$\Omega\Big(\sum_{s,a} \frac{\mathbf{V}^\star_{s,a}}{\varphi(s,a)}\log(T)\Big)$ – asymptotically

Table 6.1: Comparison of various problem-dependent regret bounds for RL under average reward criterion

ness assumptions on rewards and transition probabilities are made. Assuming α-Hölder continuity of reward functions and transition probabilities, they present a Ucrl2-style algorithm called UCCRl, which combines state aggregation and the optimistic principle. For the case $\mathcal{S} = [0,1]^d$ and under the assumption that the Hölder parameter is known to the decision maker, UCCRl achieves a regret of order $\widetilde{\mathcal{O}}\left(T^{(2d+1)/(2d+2\alpha)}\right)$ with high probability. Lakshmanan et al. [110] investigate RL in the same setup as in [106], but further assume that transition probabilities are κ-time smoothly differentiable. Employing kernel density estimation techniques, they propose UCCRl-KD achieving a regret growing as $\widetilde{\mathcal{O}}\left(T^{\frac{\beta+\alpha\beta+2\alpha}{\beta+2\alpha\beta+2\alpha}}\right)$ with high probability, where $\beta := \alpha + \kappa$.

We finally mention that some studies consider regret minimization in MDPs in the *episodic* setting, where the length of each episode is fixed and known; see, e.g., [111], [112], and [113]. Although these problems bear some similarities to the average-reward setting, the techniques developed in these paper strongly rely on the fixed length of the episode, which is usually considered to be small, and do not directly carry over to the problem considered in this chapter.

6.3 The KL-Ucrl Algorithm

The KL-Ucrl algorithm [24, 114] is a model-based algorithm inspired by Ucrl2 [25]. To present the algorithm, we first describe how it defines, at each given time t, the set of plausible MDPs based on the observation available at time t. To

Algorithm	**Setting**	**Regret**
Ucrl2 [102, 25]	communicating	$\widetilde{\mathcal{O}}\left(DS\sqrt{AT}\right)$
KL-Ucrl [24]	communicating	$\widetilde{\mathcal{O}}\left(DS\sqrt{AT}\right)$
[103]	communicating, known rewards	$\widetilde{\mathcal{O}}\left(D\sqrt{SAT}\right), \quad T \geq S^5 A$
Regal [101]	weakly communicating, known rewards	$\widetilde{\mathcal{O}}\left(D'S\sqrt{AT}\right)$
TSDE [107]	weakly communicating	$\widetilde{\mathcal{O}}\left(D'S\sqrt{AT}\right)$
Minimax Lower Bound [102, 25]	communicating	$\Omega\left(\sqrt{DSAT}\right)$

Table 6.2: Comparison of various problem-independent regret bounds for RL under average reward criterion

this end, we introduce the following notations. Under a given algorithm and for a state-action pair (s,a), let $N_t(s,a)$ denote the number of visits, up to time t, to (s,a): $N_t(s,a) = \sum_{t'=0}^{t-1} \mathbb{I}\{s_{t'} = s, a_{t'} = a\}$. Then, let $N_t(s,a)^+ = \max\{N_t(s,a), 1\}$. Similarly, $N_t(s,a,s')$ denotes the number of visits to (s,a), up to time t, followed by a visit to state s': $N_t(s,a,s') = \sum_{t'=0}^{t-1} \mathbb{I}\{s_{t'} = s, a_{t'} = a, s_{t'+1} = s'\}$. We introduce the empirical estimates of transition probabilities and rewards:

$$\hat{\mu}_t(s,a) = \frac{\sum_{t'=0}^{t-1} r_t \mathbb{I}\{s_{t'} = s, a_{t'} = a\}}{N_t(s,a)^+}, \quad \hat{p}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a)^+}\,.$$

KL-Ucrl, as an optimistic model-based algorithm, considers the set $\mathcal{M}_t$ as a collection of all MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', P')$, whose transition kernels and reward functions satisfy:

$$\begin{aligned} \texttt{KL}(\hat{p}_t(\cdot|s,a), p'(\cdot|s,a)) &\leq C_p/N_t(s,a)\,, & (6.1)\\ |\hat{\mu}_t(s,a) - \mu'(s,a)| &\leq \sqrt{C_\mu/N_t(s,a)}\,, & (6.2) \end{aligned}$$

where $C_p = S\left(B + \log(G)(1+1/G)\right)$, with $B = \log(2eS^2A\log(T)/\delta)$ and $G = B + 1/\log(T)$, and $C_\mu = \log(4SA\log(T)/\delta)/1.99$. Here $\delta \in (0,1]$ is an input parameter of the algorithm. Importantly, as proven in [24, Proposition 1], with probability at least $1-2\delta$, the true MDP M belongs to the set $\mathcal{M}_t$ uniformly over all time steps $t \leq T$.

Similarly to Ucrl2, KL-Ucrl (Algorithm 6.1) proceeds in episodes of varying lengths. We index an episode by $k \in \mathbb{N}$. The starting time of the k-th episode is denoted t_k, and by a slight abuse of notation, let $N_k := N_{t_k}$, $\mathcal{M}_k := \mathcal{M}_{t_k}$, $\hat{\mu}_k = \hat{\mu}_{t_k}$, and $\hat{p}_k := \hat{p}_{t_k}$. At $t = t_k$, the algorithm forms the set of plausible MDPs $\mathcal{M}_k$ based on the observations gathered so far. It then defines an extended MDP $M_{\texttt{ext},k} = (S, A \times \mathcal{M}_k, \mu_{\texttt{ext}}, P_{\texttt{ext}})$, where for an extended action $a_{\texttt{ext}} = (a, M')$, it defines $\mu_{\texttt{ext}}(s, a_{\texttt{ext}}) = \mu'(s,a)$ and $p_{\texttt{ext}}(s'|s, a_{\texttt{ext}}) = p'(s'|s,a)$. Then, a $\frac{1}{\sqrt{t_k}}$-optimal extended policy $\pi_{\texttt{ext},k}$ is computed in the form $\pi_{\texttt{ext},k}(s) = (\widetilde{M}_k, \tilde{\pi}_k(s))$, in the sense that it satisfies

$$\tilde{g}_k := g_{\tilde{\pi}_k}(\tilde{M}_k) \geq \max_{M' \in \mathcal{M}_k, \pi} g_\pi(M') - \frac{1}{\sqrt{t_k}} .$$

Here $g_\pi(M)$ denotes the gain of policy π in MDP M. $\tilde{M}_k$ and $\tilde{\pi}_k$ are respectively called the optimistic MDP and the optimistic policy. Finally, an episode stops at the first step $t = t_{k+1}$ when number of local counts $v_{k,t}(s,a) = \sum_{t'=t_k}^{t} \mathbb{I}\{s_{t'} = s, a_{t'} = a\}$ exceeds $N_{t_k}(s,a)$ for some (s,a). We denote with some abuse $v_k = v_{k,t_{k+1}}$.

Remark 6.1. *The value $1/\sqrt{t_k}$ is a parameter of extended value iteration and is only here for computational reasons: with sufficient computational power, it could be replaced with* 0*.*

Algorithm 6.1 KL-Ucrl [24], with input parameter $\delta \in (0,1]$

```
Initialize: For all $(s,a)$, set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. Set $t = 1$, $k = 1$, and observe initial state $s_1$
for episodes $k \geq 1$ do
  Set $t_k = t$
  Set $N_k(s,a) = N_{k-1}(s,a) + v_{k-1}(s,a)$ for all $(s,a)$
  Find an $\frac{1}{\sqrt{t_k}}$-optimal policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ using Extended Value Iteration
  while $v_k(s_t,a_t) \geq N_k(s_t,a_t)$ do
    Play action $a_t = \tilde{\pi}_k(s_t)$, and observe the next state $s_{t+1}$ and reward $r_t$
    Update $N_k(s,a,x)$ and $v_{t+1}(s,a)$ for all actions $a$ and states $s,x$
  end while
end for
```

6.4 Variance-Aware Regret Lower Bounds

In order to motivate the dependence of the regret on the local variance of the bias function (namely, w.r.t. transition laws), we revisit some regret lower bounds for our setup in the literature. We begin with the following minimax lower bound for communicating MDPs that makes appear this scaling:

Theorem 6.1. *There exists an MDP M with S states and A actions with $S, A \geq 10$, such that the expected regret under any algorithm $\mathbb{A}$ after $T \geq DSA$ steps for any initial state satisfies*

$$\mathbb{E}[\text{Regret}_{\mathbb{A},T}] \geq 0.0123\sqrt{\mathbf{V}_{\max} SAT} ,$$

where $\mathbf{V}_{\max} := \max_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star)$.

Let us recall that Jaksch et al. [25] present a minimax lower bound on the regret that scales as $\Omega(\sqrt{DSAT})$. Their lower bound is derived by considering a family of *hard-to-learn* MDPs and using similar techniques used in the derivation of minimax lower bound in MAB, as studied in [21]. To prove the above theorem, we also consider the same MDP instances as in [25] and leverage their techniques. We show however that choosing slightly different transition probabilities for the problem instance leads to a lower bound scaling as $\Omega(\sqrt{\mathbf{V}_{\max}SAT})$, which does not depend on the diameter.

We also remark that for the considered problem instance, easy calculations show that for any state-action pair (s,a), the variance of bias function satisfies $c_1\sqrt{D} \le \mathbb{V}_{p(\cdot|s,a)}(b^\star) \le c_2 D$ for some constants c_1 and c_2. Hence, the lower bound in Theorem 6.1 can serve as an alternative minimax lower bound without any dependence on the diameter.

Next we consider a family of *hard-to-learn* ergodic MDPs and examine the lower bound of Burnetas and Katehakis [22] for them (for an overview on this lower bound, we refer to Section 2.3.1 in Chapter 2).

Proposition 6.1. *There exists a family $\mathfrak{C}$ of ergodic MDPs, such that for any $M \in \mathfrak{C}$, the expected regret under any admissible algorithm $\mathbb{A}$ starting from any initial state in M satisfies*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}]}{\log(T)} \ge \sum_{(s,a)\in\mathcal{C}_M} \frac{\mathbb{V}_{p(\cdot|s,a)}(b^\star)}{\varphi(s,a)} \,.$$

To prove the above proposition, we present a family of MDPs whose design is inspired by the ones in [25] (see also the proof of Theorem 6.1).

Remark 6.2. *Theorem 6.1 and Proposition 6.1 suggest that the local variance of bias function can serve as a hardness metric for regret in the considered setup.*

6.5 Concentration Inequalities and the Kullback-Leibler Divergence

Before providing variance-aware regret bounds, let us discuss some important tools that we use for the regret analysis. We believe that these results could also be of independent interest beyond RL.

Let us first recall a powerful result known as the transportation lemma; see, e.g., [115, Lemma 4.18]:

Lemma 6.1 (Transportation Lemma). *For any function f, let us introduce $\varphi_f : \lambda \mapsto \log \mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f]))]$. Whenever φ_f is defined on some possibly unbounded interval I containing 0, define its dual $\varphi_{\star,f}(x) = \sup_{\lambda\in I}(\lambda x - \varphi_f(\lambda))$. Then it holds*

$$\begin{aligned}
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \quad &\le \quad \varphi_{+,f}^{-1}(\mathtt{KL}(Q,P)) \\
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \quad &\ge \quad \varphi_{-,f}^{-1}(\mathtt{KL}(Q,P))
\end{aligned}$$

where

$$\begin{aligned}\varphi_{+,f}^{-1}(t) &= \inf\{x \geq 0 : \varphi_{\star,f}(x) > t\}\\ \varphi_{-,f}^{-1}(t) &= \sup\{x \leq 0 : \varphi_{\star,f}(x) > t\}\,.\end{aligned}$$

For the sake of completeness, we provide the proof of this lemma in Appendix B. This result is especially interesting when Q is the empirical version of P built from n i.i.d. observations, since in that case it enables to *decouple* the concentration properties of the distribution from the specific structure of the considered function. Further, it shows that controlling the KL divergence between Q and P induces a concentration result valid for all (nice enough) functions f, which is especially useful when we do not know in advance the function f we want to handle (such as bias function $b^\star$).

The quantities $\varphi_{+,f}^{-1}$, $\varphi_{-,f}^{-1}$ may look complicated. When $f(X)$ (where $X \sim P$) is Gaussian, they coincide with $t \mapsto \pm\sqrt{2\mathbb{V}_P(f)t}$. Although controlling them in general is challenging, for bounded functions a Bernstein-type relaxation can be derived that uses the variance $\mathbb{V}_P(f)$ and the span $\mathbb{S}(f)$. Let us recall that the span of function f that takes values in a set $\mathcal{X}$ is defined as $\mathbb{S}(f) = \max_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} f(x)$.

Corollary 6.1 (Bernstein Transportation). *For any f such that $\mathbb{V}_P(f)$ and $\mathbb{S}(f)$ are finite,*

$$\begin{aligned}\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leq \sqrt{2\mathbb{V}_P(f)\mathtt{KL}(Q,P)} + \frac{2}{3}\mathbb{S}(f)\mathtt{KL}(Q,P)\,,\\ \forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leq \sqrt{2\mathbb{V}_P(f)\mathtt{KL}(Q,P)}\,.\end{aligned}$$

We also provide below another variation of this result that is especially useful when the bounds of Corollary 6.1 cannot be handled. This result, to the best of our knowledge, is new.

Lemma 6.2 (Transportation Method II). *Let $P \in \mathcal{P}(\mathcal{X})$ be a probability distribution on a finite alphabet $\mathcal{X}$. Then, for any real-valued function f defined on $\mathcal{X}$, it holds that*

$$\forall P \ll Q, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2\mathtt{KL}(P,Q)} + \mathbb{S}(f)\mathtt{KL}(P,Q),$$

$$\textit{where} \quad \mathcal{V}_{P,Q}(f) \overset{\text{def}}{=} \sum_{x\in\mathcal{X}:P(x)\geq Q(x)} P(x)(f(x) - \mathbb{E}_P[f])^2\,.$$

Next we discuss the implication of these results in our regret analysis. When P is the transition law under a state-action pair (s,a) and Q is its empirical estimates up to time t, i.e., $Q = \hat{p}_t(\cdot|s,a)$ and $P = p(\cdot|s,a)$, Corollary 6.1 can be used to decouple $\mathbb{E}_Q[f] - \mathbb{E}_P[f]$ from specific structure of f. In particular, if f is some optimal value function (or bias function)[3], using the observation that $\mathtt{KL}(Q,P) = \widetilde{\mathcal{O}}(N_t^{-1})$, we

[3] Note that for communicating MDPs, the span of the optimal bias function is bounded by diameter D, so that use of Corollary 6.1 is allowed.

derive an upper bound for $\mathbb{E}_Q[f] - \mathbb{E}_P[f]$, whose first order terms makes appear the variance of f. This would result in a term scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}T}\Big)$ in our regret bound, where $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic terms.

Now, for the case when $Q = \hat{p}_t(\cdot|s,a)$ and $P = \tilde{p}_t(\cdot|s,a)$ is the optimistic transition law at time t, the second inequality in Corollary 6.1 allows us to bound $\mathbb{E}_P[f] - \mathbb{E}_Q[f]$ by the variance of f under law $\tilde{p}(\cdot|s,a)$, which itself is controlled by the variance of f under the true law $p(\cdot|s,a)$. This approach would lead to a term scaling as $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}T} + DS^2T^{1/4}\Big)$. We can remove the term scaling as $\widetilde{\mathcal{O}}(T^{1/4})$ in our regret analysis by resorting to Lemma 6.2 instead, in combination with the following property of the operator $\mathcal{V}$:

Lemma 6.3. *Consider two distributions $P, Q \in \mathcal{P}(\mathcal{X})$ with $|\mathcal{X}| \geq 2$. Then, for any real-valued function f defined on $\mathcal{X}$, it holds that*

$$\begin{aligned} (i) &\quad \mathcal{V}_{P,Q}(f) \leq \mathbb{V}_P(f)\,, \\ (ii) &\quad \sqrt{\mathcal{V}_{P,Q}(f)} \leq \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q,P)}\,. \end{aligned}$$

6.6 Variance-Aware Regret Upper Bounds

In this section, we present regret upper bounds for Burnetas-Katehakis and KL-Ucrl that leverage the results presented in the previous section. Let $\Psi := \mathbb{S}(b^\star)$ denote the span of the bias function, and for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ define $\mathbf{V}^\star_{s,a} := \mathbb{V}_{p(\cdot|s,a)}(b^\star)$ as the variance of the bias function under law $p(\cdot|s,a)$.

We begin with revisiting the regret bound of Burnetas-Katehakis. The following corollary to [22, Theorem 1] (see also Theorem 2.8) gives an explicit variance-aware regret bound for Burnetas-Katehakis valid for any ergodic MDP:

Corollary 6.2 (Variance-Aware Regret Bound for Burnetas-Katehakis)**.** *For any ergodic MDP M and any initial state in M, the regret under algorithm $\mathbb{A} =$ Burnetas-Katehakis satisfies*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}]}{\log(T)} \leq 4\sum_{(s,a)\in\mathcal{C}_M} \frac{\mathbf{V}^\star_{s,a}}{\varphi(s,a)} + 4SA\Psi\,.$$

To the best of our knowledge, the bound reported in the above corollary is the best explicit bound for Burnetas-Katehakis. The gap-independent term (scaling as $\mathcal{O}(SA\Psi\log(T))$) could be an artefact of the proof, and can probably be removed with a more careful analysis. We further remark that the OLP algorithm of Tewari and Bartlett [100], which relies on L_1-based confidence bounds, attains a regret bound asymptotically growing as $\mathcal{O}\Big(\frac{SA\Psi^2}{\Phi}\log(T)\Big)$. Corollary 6.2 improves over this bound since $\mathbf{V}^\star_{s,a} \leq \Psi^2$. As we shall see later, the leading constant $\sum_{s,a}\mathbf{V}^\star_{s,a}$ could be much smaller than $SA\Psi^2$ as confirmed by some illustrative examples.

Next we provide a refined regret bound for a variant of KL-Ucrl for the class of ergodic MDPs. Let $\tilde{\star}_k$ denote the optimal policy in the extended MDP $\mathcal{M}_k$, whose gain $\tilde{g}_{\tilde{\star}_k}$ satisfies $\tilde{g}_{\tilde{\star}_k} = \max_{M' \in \mathcal{M}_k, \pi} g_\pi(M')$. We consider a variant of KL-Ucrl, which computes, in every episode k, a policy $\tilde{\pi}_k$ satisfying: $\max_s |\tilde{b}_k(s) - \tilde{b}_{\tilde{\star}_k}(s)| \le \frac{1}{\sqrt{t_k}}$, and $\tilde{g}_k \ge \tilde{g}_{\tilde{\star}_k} - \frac{1}{\sqrt{t_k}}$.[4] We have:

Theorem 6.2 (Variance-Aware Regret Bound for KL-Ucrl). *With probability at least $1-6\delta$, the regret under the variant of* KL-Ucrl *described above for any initial state satisfies*

$$\begin{aligned}\mathrm{Regret}_{\texttt{KL-Ucrl},T} \le{}& \Big(31\sqrt{S\textstyle\sum_{s,a}\mathbf{V}^\star_{s,a}} + 35S\sqrt{A} + \sqrt{2}D + 1\Big)\sqrt{T\log(\log(T)/\delta)} \\ &+ \widetilde{\mathcal{O}}\Big(SA(T_M SA + D + S^{3/2})\log(T)\Big)\,,\end{aligned}$$

where $\widetilde{\mathcal{O}}$ hides terms scaling as $\mathrm{polylog}(\log(T)/\delta)$.

Remark 6.3. *If the cardinality of the set $\mathcal{S}^+_{s,a} := \{s' : p(s'|s,a) > 0\}$ for state-action (s,a) is known, then one can use the following improved confidence bound for the pair (s,a) (instead of (6.1)):*

$$N_t(s,a)\texttt{KL}(\hat{p}_t(\cdot|s,a), p'(\cdot|s,a)) \le C_p^{s,a}\,, \tag{6.3}$$

where $C_p^{s,a} = \frac{|\mathcal{S}^+_{s,a}|}{S}C_p$ (see, e.g., [114, Proposition 4.1] for the corresponding concentration result). As a result, if $|\mathcal{S}^+_{s,a}|$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ is known, it is then straightforward to show that the corresponding variant of KL-Ucrl*, which relies on (6.3), achives a regret growing as $\widetilde{\mathcal{O}}\big(\sqrt{\sum_{s,a}|\mathcal{S}^+_{s,a}|\mathbf{V}^\star_{s,a}T} + D\sqrt{T}\big)$.*

The regret bound provided in the aforementioned remark is of particular importance in the case of *sparse MDPs*, where most states transit to only a few next-states under various actions. We would like to stress that to get an improvement of a similar flavour for Ucrl2, to the best of our knowledge, one has to know the sets $\mathcal{S}^+_{s,a}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ rather than their cardinalities.

Illustrative numerical experiments. In order to better highlight the magnitude of the main terms in Theorem 6.2 when compared to other existing results, and strengthen the discussion, we consider a standard class of environments for which we compute them explicitly.

For the sake of illustration, we consider the *RiverSwim* MDP, introduced in [53], as our benchmark environment. In order to satisfy ergodicity, here we consider a slightly modified version of the original *RiverSwim* (see Figure 6.1). Furthermore,

[4] We study such a variant to facilitate the analysis and presentation of the proof. This variant of KL-Ucrl may be computationally less efficient than Algorithm 6.1. We stress however that, in view of the number of episodes (growing as $SA\log(T)$) as well as Remark 6.1, with sufficient computational power such an algorithm could be practical.

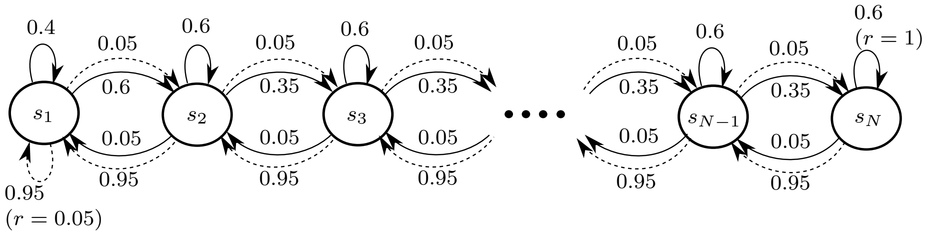


Figure 6.1: The N-state Ergodic *RiverSwim* MDP

S	Ψ	$\max_{s,a}\mathbf{V}^\star_{s,a}$	$\Psi\sqrt{SA}$	$\sqrt{\sum_{s,a}\mathbf{V}^\star_{s,a}}$
6	6.3	0.6322	21.9	1.8
12	14.9	0.6327	72.9	2.8
20	26.3	0.6327	166.4	3.7
40	54.9	0.6327	490.9	5.3
70	97.7	0.6327	1156.5	7.1
100	140.6	0.6327	1988.3	8.5

Table 6.3: Comparison of span and variance for *Ergodic RiverSwim* with various number of states.

to convey more intuition about the potential gains, we consider varying number of states. The benefits of KL-Ucrl have already been studied experimentally in [24], and we compute in Table 6.3 features that we believe explain the reason behind this. In particular, it is apparent that while $\Psi\sqrt{SA} \leq D\sqrt{SA}$ grows very large with S, $\mathbf{V}^\star_{s,a}$ is about constant and very small on all tested environments. Further, even on this simple environment, we see that $\sqrt{\sum_{s,a}\mathbf{V}^\star_{s,a}}$ is an order or magnitude smaller than $\Psi\sqrt{SA}$. We believe that these computations highlight the fact that the regret bound of Theorem 6.2 captures a massive improvement over the initial analysis of KL-Ucrl in [24], and over alternative algorithms such as Ucrl2.

6.7 Summary

In this chapter, we provided variance-aware regret bounds for Burnetas-Katehakis and KL-Ucrl, and also revisited existing regret lower bounds, in order to make appear the local variance of the bias function of the MDP. Computations of these terms in some illustrative enviroments show that reported upper bound for KL-Ucrl may improve an order of magnitude over the existing ones (as observed experimentally in [114]), thus highlighting the fact that trading the diameter of the MDP to the local variance of the bias function may result in huge improvements.

6.A Derivation of Minimax Regret Lower Bound

The proof of Theorem 6.1 mainly relies on the problem instance for the derivation of the minimax lower bound in Jaksch et al. [25] and related arguments there. For the sake of completeness, we first recall this family of MDPs and then compute the variance of the corresponding bias function.

To get there, we first consider the two-state MDP M' shown in Figure 6.2, where there are two states $\{s_0, s_1\}$, each having $A' = \lfloor \frac{A-1}{2} \rfloor$ actions. We consider deterministic rewards defined as $r(s_0, a) = 0$ and $r(s_1, a) = 1$ for all $a \in \mathcal{A}$. The decision maker knows the rewards but not the transition probabilities. Let $\delta := \frac{4}{D}$, where D is the diameter for which we derive the lower bound. Under any action a, $p(s_0|s_1, a) = \delta$. In state s_0, there is a unique optimal action $a^\star$, which will be referred to as "good" action. For any $a \neq a^\star$, we have $p(s_1|s_0, a) = \delta$ whereas $p(s_1|s_0, a^\star) = \delta + \varepsilon$ for some $\varepsilon \in (0, \frac{\delta}{2})$ that will be determined later. Note that the diameter D' of M' satisfies: $D' = \frac{1}{\delta} = \frac{D}{4}$.

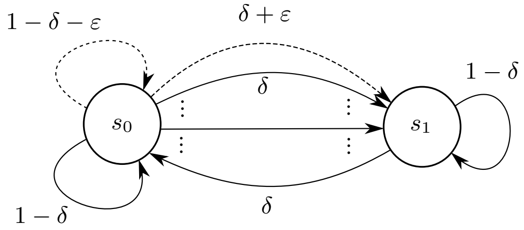


Figure 6.2: The MDP M' for lower bound [25]

We consider $\delta \in (0, \frac{1}{3})$.[5] After straightforward calculations, one finds that the average reward in M' is given by

$$g^\star = \frac{1/\delta}{1/\delta + 1/(\delta+\varepsilon)} = \frac{\delta+\varepsilon}{2\delta+\varepsilon} \,.$$

Furthermore, from Bellman's optimality equations, we obtain

$$b^\star(s_0) + \frac{\delta+\varepsilon}{2\delta+\varepsilon} = (\delta+\varepsilon)b^\star(s_1) + (1-\delta-\varepsilon)b^\star(s_0) \,,$$

thus giving $\Psi := \mathbb{S}(b^\star) = b^\star(s_1) - b^\star(s_0) = \frac{1}{2\delta+\varepsilon}$. Consider $a \neq a^\star$ and let $p = p(\cdot|s_0, a)$. It follows that:

$$\mathbb{E}_p[b^\star] = \delta b^\star(s_1) + (1-\delta)b^\star(s_0) = b^\star(s_0) + \delta\Psi \,,$$
$$\mathbb{V}_p(b^\star) = \delta(b^\star(s_1) - \mathbb{E}_p[b^\star])^2 + (1-\delta)(b^\star(s_0) - \mathbb{E}_p[b^\star])^2 = \delta(1-\delta)\Psi^2 \,.$$

Similarly, we obtain

$$\mathbb{V}_{p(\cdot|s_0,a^\star)}(b^\star) = (\delta+\varepsilon)(1-\delta-\varepsilon)\Psi^2 \,,$$

[5] The case of $\delta > 1/3$ can be handled similarly to the analysis of [25].

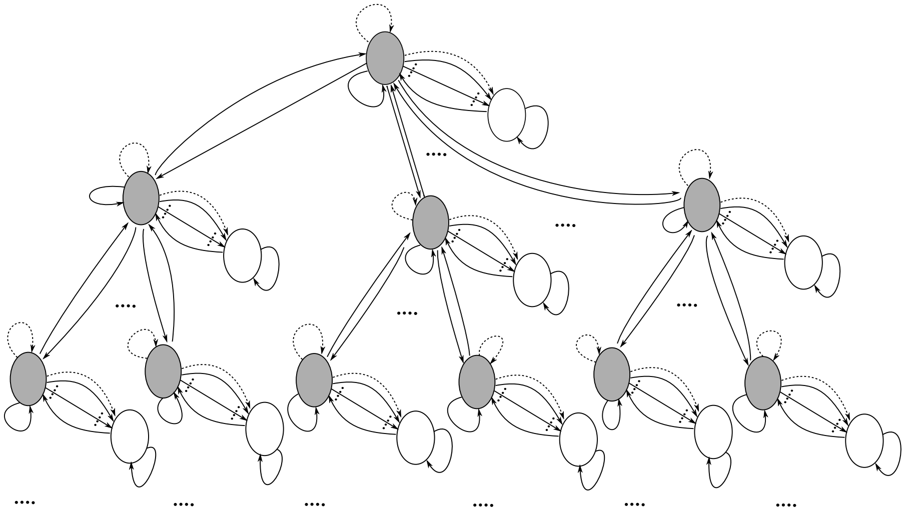


Figure 6.3: The composite MDP M [25]

$$\mathbb{V}_{p(\cdot|s_1,a)}(b^\star) = \delta(1-\delta)\Psi^2 \,.$$

Hence, using the facts that $x \mapsto x(1-x)$ is increasing for $x \in [0, \frac{1}{2}]$ and $\varepsilon + \delta \le \frac{1}{2}$, we obtain

$$\mathbf{V}_{\max} := \max_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star) = (\delta+\varepsilon)(1-\delta-\varepsilon)\Psi^2 \,.$$

The Composite MDP

We now build a composite MDP M as considered in [25], as a concatenation of $k := \lfloor S/2 \rfloor$ copies of M' in the form of an A'-ary tree, where only one copy contains the good action $a^\star$ (see Figure 6.3). To this end, we first add $A'+1$ additional actions, so that M has at most A actions per state. For any state s_0, one of these new actions connects s_0 to the root, and the rest connect s_0 to the leaves. Whereas for any state s_1, all new actions make a transition to the same state s_1. By construction, the diameter of the composite MDP M does not exceed $2(D/4 + \log_{A'} k)$, so that MDP M has $2\lfloor S/2 \rfloor \le S$ states, $\lfloor \frac{A'-1}{2} \rfloor + \lfloor \frac{A'-1}{2} \rfloor + 1 \le A$ actions, and a diameter less than D.

6.A.1 Proof of Theorem 6.1

To derive the claimed result, we derive a lower bound on the regret for the composite MDP presented above. Our analysis is largely built on the techniques used in the proof of [25, Theorem 5]. We also closely follow the notations used in [25].

Let us assume, as in the proof of [25, Theorem 5], that all states s_0 are identified so that M is equivalent to an MDP M' with kA' actions (note that following the same argument as in [25], despite the same maximal average reward, learning in M' is easier than in M, and so any regret lower bound for M' implies a lower bound in M, too). Note that by construction of M, it holds that $\mathbf{V}_{\max}$ in M equals $\mathbf{V}_{\max}$ in M'. Denote by $(s_0^\star, a^\star)$ the *good copy*, i.e., the one containing good action $a^\star$. We assume that $a^\star$ is chosen uniformly at random among all actions $\{1, \dots, k\} \times \{1, \dots, A'\}$. Let $\mathbb{E}_\star[\cdot]$ and $\mathbb{E}_{\text{unif}}[\cdot]$ respectively denote the expectation with respect to the random choice of $(s_0^\star, a^\star)$ and the expectation when there is no good action. Furthermore, let $\mathbb{E}_a[\cdot]$ denote the expectation conditioned on $a = a^\star$, and introduce N_1, N_0, and $N_0^\star$ as the respective number of visits to s_1, s_0, and $(s_0, a^\star)$.

The proof proceeds in the same steps as in the proof of [25, Theorem 5] up to equation (36) there, where it is shown that under the assumption that the initial state is s_0,

$$\begin{aligned}\mathbb{E}_a[N_1] &\le \mathbb{E}_a[N_0 - N_0^\star] + (\delta + \varepsilon)D'\mathbb{E}_a[N_0^\star] \le T - \mathbb{E}_{\text{unif}}[N_1] + \varepsilon D'\mathbb{E}_a[N_0^\star]\,, \\ \mathbb{E}_{\text{unif}}[N_1] &\ge \frac{T - D'}{2}\,,\end{aligned}$$

so that the accumulated reward $R_{\mathbb{A},T}$ by the algorithm $\mathbb{A}$ in M' up to time step T satisfies

$$\mathbb{E}_a[R_{\mathbb{A},T}] \le \mathbb{E}_a[N_1] \le \frac{T + D'}{2} + \varepsilon D'\mathbb{E}_a[N_0^\star].$$

The following lemma, which is a straightforward modification to [25, Lemma 13], enables us to control $\mathbb{E}_a[N_0^\star]$:

Lemma 6.4. *Let $f : \{s_0, s_1\}^{T+1} \mapsto [0, B]$ be any function defined on any trajectory $\boldsymbol{s}_{T+1} = (s_t)_{1\le t\le T+1}$ in M'. Then, for any $\delta \in [0, \frac{1}{3}]$, $\varepsilon \in (0, 1 - 2\delta)$, and $a \in \{1, \dots, kA'\}$,*

$$\mathbb{E}_a[f(\boldsymbol{s})] \le \mathbb{E}_{\text{unif}}[f(\boldsymbol{s})] + \varepsilon B\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2(\delta + \varepsilon)(1 - \delta - \varepsilon)}}\,.$$

Noting that $N_0^\star$ is a function of $\boldsymbol{s}_{T+1}$ satisfying $N_0^\star \in [0, T]$, by Lemma 6.4 we deduce that

$$\begin{aligned}\mathbb{E}_a[N_0^\star] &\le \mathbb{E}_{\text{unif}}[N_0^\star] + \varepsilon T\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2(\delta + \varepsilon)(1 - \delta - \varepsilon)}} \\ &= \mathbb{E}_{\text{unif}}[N_0^\star] + \varepsilon \Psi T\sqrt{\frac{\log(2)\mathbb{E}_{\text{unif}}[N_0^\star]}{2\mathbf{V}_{\max}}}\,,\end{aligned}$$

where we used $\sqrt{\mathbf{V}_{\max}} = \Psi\sqrt{(\delta + \varepsilon)(1 - \varepsilon - \delta)}$. As shown in the proof of [25, Theorem 5], $\sum_{a=1}^{kA'} \mathbb{E}_{\text{unif}}[N_0^\star] \le (T + D')/2$ and $\sum_{a=1}^{kA'} \sqrt{\mathbb{E}_{\text{unif}}[N_0^\star]} \le \sqrt{kA'(T + D')/2}$, so that we finally get, using the relation $\mathbb{E}_\star[R_{\mathbb{A},T}] = \frac{1}{kA'}\sum_{a=1}^{kA'}\mathbb{E}_a[R_{\mathbb{A},T}]$,

$$\begin{aligned}\mathbb{E}_\star[\mathrm{Regret}_{\mathbb{A},T,M'}] &= \frac{\delta+\varepsilon}{2\delta+\varepsilon}T - \mathbb{E}_\star[R_{\mathbb{A},T}] \\ &\ge \frac{\delta+\varepsilon}{2\delta+\varepsilon}T - \frac{T}{2} - \frac{\varepsilon D'T}{2kA'} - \frac{\varepsilon D'^2}{2kA'} \\ &\quad - \frac{\varepsilon^2\Psi D'T}{kA'}\sqrt{\frac{\log(2)kA'T}{4\mathbf{V}_{\max}}} - \frac{\varepsilon^2\Psi D'T}{kA'}\sqrt{\frac{\log(2)kA'D'}{4\mathbf{V}_{\max}}} - \frac{D'}{2} \\ &\ge \frac{\varepsilon T}{4\delta+2\varepsilon} - \frac{\varepsilon D'}{2kA'}(T+D') - \frac{0.42\varepsilon^2\Psi D'T}{\sqrt{kA'\mathbf{V}_{\max}}}(\sqrt{T}+\sqrt{D'}) - \frac{D'}{2}\ .\end{aligned}$$

Noting that the assumption $T \ge DSA$ implies $T \ge 16D'kA'$, we deduce that

$$\begin{aligned}\mathbb{E}_\star[\mathrm{Regret}_{\mathbb{A},T,M'}] &\ge \frac{\varepsilon T}{4\delta+2\varepsilon} - \frac{\varepsilon D'T}{2kA'}\Big(1+\frac{1}{16kA'}\Big) \\ &\quad - \frac{0.42\varepsilon^2\Psi D'T\sqrt{T}}{\sqrt{kA'\mathbf{V}_{\max}}}\Big(1+\frac{1}{4\sqrt{kA'}}\Big) - \frac{D'}{2}\ .\end{aligned}$$

Note that the first term in the right-hand side satisfies

$$\frac{\varepsilon T}{4\delta+2\varepsilon} = \frac{\varepsilon T\Psi}{2} \ge \frac{5\varepsilon\mathbf{V}_{\max}T}{6}\ ,$$

since

$$\frac{\Psi}{\mathbf{V}_{\max}} = \frac{2\delta+\varepsilon}{(\delta+\varepsilon)(1-\delta-\varepsilon)} \ge 1 + \frac{\delta}{\delta+\varepsilon} > \frac{5}{3}\ ,$$

where we use $\varepsilon \le \delta/2$ in the last step. Hence, we get

$$\begin{aligned}\mathbb{E}_\star[\mathrm{Regret}_{\mathbb{A},T,M'}] &\ge \frac{5}{6}\varepsilon\mathbf{V}_{\max}T - \frac{\varepsilon D'T}{2kA'}\Big(1+\frac{1}{16kA'}\Big) \\ &\quad - \frac{0.42\varepsilon^2\Psi D'T\sqrt{T}}{\sqrt{kA'\mathbf{V}_{\max}}}\Big(1+\frac{1}{4\sqrt{kA'}}\Big) - \frac{D'}{2}\ .\end{aligned}$$

In particular, setting $\varepsilon = c\sqrt{\frac{kA'}{\mathbf{V}_{\max}T}}$ for some c (which will be determined later) yields

$$\begin{aligned}\mathbb{E}_\star[\mathrm{Regret}_{\mathbb{A},T,M'}] &\ge \frac{5}{6}c\sqrt{kA'\mathbf{V}_{\max}T} - \sqrt{kA'\mathbf{V}_{\max}T}\left(\frac{cD'}{2kA'\mathbf{V}_{\max}}\Big(1+\frac{1}{16kA'}\Big)\right) \\ &\quad - \sqrt{kA'\mathbf{V}_{\max}T}\left(\frac{0.42c^2}{kA'}\frac{D'\Psi}{\mathbf{V}_{\max}^2}\Big(1+\frac{1}{4\sqrt{kA'}}\Big)\right) - \frac{D'}{2}\ .\end{aligned}$$

To simplify the above bound, note that

$$\frac{D'}{\mathbf{V}_{\max}} \le \frac{(2\delta+\varepsilon)^2}{\delta(\delta+\varepsilon)(1-\delta-\varepsilon)} \le 2\Big(\frac{2\delta+\varepsilon}{\delta}\Big)^2 \le 12.5\ , \tag{6.4}$$

where we used $1-\varepsilon-\delta \geq \frac{1}{2}$ since $\varepsilon \leq \frac{\delta}{2}$. Moreover,

$$\begin{aligned}\frac{D'\Psi}{\mathbf{V}_{\max}^2} &= \frac{D'\Psi}{\Psi^4(\delta+\varepsilon)^2(1-\delta-\varepsilon)^2}\\ &= \frac{(2\delta+\varepsilon)^3}{\delta(\delta+\varepsilon)^2(1-\delta-\varepsilon)^2} \leq 4\Big(\frac{2\delta+\varepsilon}{\delta}\Big)^3 \leq 62.5\,.\end{aligned}$$

Putting these together with the fact that

$$\frac{D'}{2} \leq \frac{\sqrt{D'}}{2}\sqrt{\frac{T}{16kA'}} \leq \frac{\sqrt{12.5/16}}{2}\sqrt{\frac{\mathbf{V}_{\max}T}{kA'}} \leq 0.45\sqrt{\frac{\mathbf{V}_{\max}T}{kA'}}\,,$$

which follows from (6.4), we deduce that

$$\begin{aligned}\mathbb{E}_\star[\mathrm{Regret}_{\mathbb{A},T,M'}] \geq \sqrt{kA'\mathbf{V}_{\max}T}\Big(&\frac{5c}{6} - \frac{12.5c}{2kA'} - \frac{12.5c}{32(kA')^2}\\ &- \frac{26.25c^2}{kA'} - \frac{6.6c^2}{(kA')^{3/2}} - \frac{0.45}{kA'}\Big),\end{aligned}$$

Taking $c = 0.132$ and using the facts $k = \lfloor\frac{S}{2}\rfloor \geq 5$ and $A' = \lfloor\frac{A-1}{2}\rfloor \geq 4$ yield the announced result. This completes the proof provided that we show that this choice of c satisfies $\varepsilon \leq \frac{\delta}{2}$. To this end, observe that by the assumption $T \geq DSA \geq \frac{16kA'}{\delta}$, it follows that

$$\begin{aligned}\varepsilon = 0.132\sqrt{\frac{kA'}{\mathbf{V}_{\max}T}} &\leq \frac{0.132}{4}\sqrt{\frac{\delta}{\mathbf{V}_{\max}}}\\ &\leq \frac{0.132}{4}\sqrt{\frac{\delta(2\delta+\varepsilon)^2}{(\delta+\varepsilon)(1-\delta-\varepsilon)}} \leq 0.047(2\delta+\varepsilon)\,,\end{aligned}$$

so that $\varepsilon \leq 0.1\delta$. This concludes the proof. □

6.A.2 Proof of Lemma 6.4

The lemma follows by a slight modification of the proof of [25, Lemma 13]. We recall that according to equations (49)-(51) in [25],

$$\mathbb{E}_a[f(\boldsymbol{s})] - \mathbb{E}_{\mathrm{unif}}[f(\boldsymbol{s})] \leq \frac{B}{2}\sqrt{2\log(2)\mathtt{KL}(\mathbb{P}_{\mathrm{unif}},\mathbb{P}_a)}\,, \tag{6.5}$$

where

$$\begin{aligned}\mathtt{KL}(\mathbb{P}_{\mathrm{unif}},\mathbb{P}_a) &= \sum_{t=1}^T \mathtt{KL}(\mathbb{P}_{\mathrm{unif}}(s_{t+1}|\boldsymbol{s}^t),\mathbb{P}_a(s_{t+1}|\boldsymbol{s}^t))\\ &= \sum_{t=1}^T \mathbb{P}_{\mathrm{unif}}(s_t = s_0, a_t = a)\left(\delta\log\Big(\frac{\delta}{\delta+\varepsilon}\Big) + (1-\delta)\log\Big(\frac{1-\delta}{1-\delta-\varepsilon}\Big)\right)\,.\end{aligned}$$

Now using the inequality $\mathtt{kl}(a,b) \le \frac{(a-b)^2}{b(1-b)}$ valid for all $a,b\in(0,1)$ (instead of [25, Lemma 20]) and noting that $\mathbb{E}_{\text{unif}}[N_0^\star] = \sum_{t=1}^T \mathbb{P}_{\text{unif}}(s_t = s_0, a_t = a)$, we obtain

$$\mathtt{KL}(\mathbb{P}_{\text{unif}}, \mathbb{P}_a) = \mathtt{kl}(\delta, \delta+\varepsilon)\mathbb{E}_{\text{unif}}[N_0^\star] \le \frac{\varepsilon^2}{(1-\delta)(1-\delta-\varepsilon)}\mathbb{E}_{\text{unif}}[N_0^\star] \; .$$

Plugging this into (6.5) completes the proof. □

6.B Proof of Proposition 6.1

To prove the proposition, we present a family of ergodic MDPs for which the application of Burnetas and Katehakis' lower bound yields the desired result.

Consider MDP M shown in Figure 6.4, which has $N+1$ states and A actions per state. Let $\delta\in(0,\frac{1}{3})$ and $\varepsilon\in(0,\frac{\delta}{2})$ so that $\delta+\varepsilon\le\frac{1}{2}$. There is only one state (denoted by s_{N+1}) that gives a non-zero reward: $r(s_{N+1},a)=1$ for all a. MDP M has quite similar design to MDP M' in Figure 6.2: It has the same transition probabilities and reward functions in all states $s\in\mathcal{S}\setminus\{s_{N+1}\}$. In state s_{N+1} and under any action a (shown in blue), the system makes a transition to state $s\neq s_{N+1}$ with probability $\frac{\delta}{N}$, and stays in s_{N+1} with probability $1-\delta$. Note that MDP M is ergodic since any state, under any policy, is reachable with positive probability from any other state.

It is easy to check that the average reward for this MDP is the same as that of M' shown in Figure 6.2, that is $g^\star = \frac{\delta+\varepsilon}{2\delta+\varepsilon}$. Moreover, for any state $s\in\mathcal{S}\setminus\{s_{N+1}\}$ Bellman's optimality equation reads

$$g^\star + b^\star(s) = (\varepsilon+\delta)b^\star(s_{N+1}) + (1-\varepsilon-\delta)b^\star(s) \; ,$$

so that $\Psi'(s) := b^\star(s_{N+1}) - b^\star(s) = \frac{1}{2\delta+\varepsilon}$. It then follows, after algebraic calculations similar to the ones for MDP M' in Figure 6.2, that

$$\mathbb{V}_{p(\cdot|s,a)}(b^\star) = \delta(1-\delta)\Psi'(s)^2 \; ,$$

and that $\varphi(s,a) = \varepsilon\Psi'(s)$ for $a\neq a^\star$.

Now consider a sub-optimal action $a\neq a^\star$ in state s. We have

$$\begin{aligned}\mathcal{K}(s,a) &= \inf\Big\{\mathtt{KL}(p(\cdot|s,a),q) : q\in\Lambda(s,a)\Big\}\\ &= \inf\Big\{\mathtt{kl}(\delta,x) : x>\delta+\varepsilon\Big\} = \mathtt{kl}(\delta,\delta+\varepsilon) \; ,\end{aligned}$$

which, using the inequality $\mathtt{kl}(x,y)\le\frac{(x-y)^2}{y(1-y)}$ valid for all $x,y\in(0,1)$, gives

$$\begin{aligned}\frac{\varphi(s,a)}{\mathcal{K}(s,a)} = \frac{\varepsilon\Psi'(s)}{\mathtt{kl}(\delta,\delta+\varepsilon)} &\ge \frac{(\varepsilon+\delta)(1-\varepsilon-\delta)\Psi'(s)^2}{\varepsilon\Psi'(s)}\\ &\ge \frac{\delta(1-\delta)\Psi'(s)^2}{\varepsilon\Psi'(s)} = \frac{\mathbb{V}_{p(\cdot|s,a)}(b^\star)}{\varphi(s,a)} \; ,\end{aligned}$$

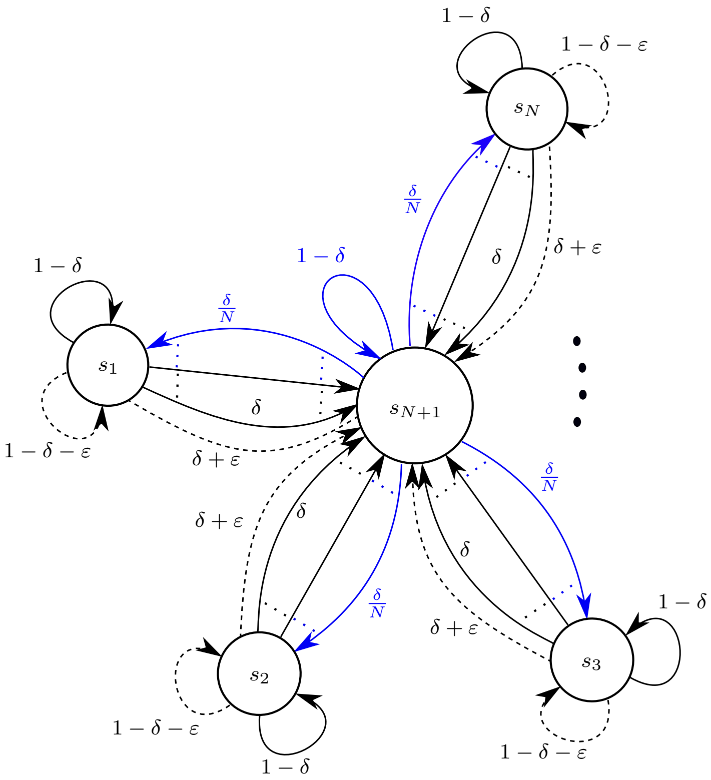


Figure 6.4: MDP M for proof of Proposition 6.1

where we used that the mapping $x \mapsto x(1-x)$ is increasing for $x \in [0, \frac{1}{2}]$.

Finally, applying [22, Theorem 2] yields the desired result:

$$c_{\mathrm{bk}}(M) = \sum_{(s,a)\in\mathcal{C}_M} \frac{\varphi(s,a)}{\mathcal{K}(s,a)} \geq \sum_{s\neq s_{N+1}} \sum_{a\neq a^\star} \frac{\mathbb{V}_{p(\cdot|s,a)}(b^\star)}{\varphi(s,a)} ,$$

and completes the proof. □

6.C Proof of Corollary 6.1

By a standard Bernstein argument (see for instance [115, Section 2.8]), it holds

$$\begin{aligned} \forall \lambda \in [0, 3/\mathbb{S}(f)), \quad \varphi_f(\lambda) \quad &\leq \quad \frac{\mathbb{V}_P(f)}{2} \frac{\lambda^2}{1-\frac{\mathbb{S}(f)\lambda}{3}} , \\ \forall x \geq 0, \quad \varphi_{\star,f}(x) \quad &\geq \quad \frac{x^2}{2(\mathbb{V}_P(f)+\frac{\mathbb{S}(f)}{3}x)} . \end{aligned}$$

Then, a direct computation (solving for x in $\varphi_{\star,f}(x) = t$) shows that

$$\begin{aligned} \varphi^{-1}_{+,f}(t) &\le \frac{\mathbb{S}(f)}{3}t + \sqrt{2t\mathbb{V}_P(f) + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2} \le \sqrt{2t\mathbb{V}_P[f]} + \frac{2}{3}t\mathbb{S}(f), \\ \varphi^{-1}_{-,f}(t) &\ge \frac{\mathbb{S}(f)}{3}t - \sqrt{2t\mathbb{V}_P(f) + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2} \ge -\sqrt{2t\mathbb{V}_P[f]}\,, \end{aligned}$$

where we used that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$. Combining these bounds, we get

$$\begin{aligned} \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\le \sqrt{2\mathbb{V}_P(f)\mathtt{KL}(Q,P)} + \frac{2}{3}\mathbb{S}(f)\mathtt{KL}(Q,P)\,, \\ \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\le \sqrt{2\mathbb{V}_P(f)\mathtt{KL}(Q,P)}\,. \end{aligned}$$

□

6.D Proof of Lemma 6.2

If $\mathbb{E}_Q[f] \le \mathbb{E}_P[f]$, then the result holds trivially. We thus assume that $\mathbb{E}_Q[f] > \mathbb{E}_P[f]$. It is straightforward to verify that

$$\begin{aligned} \mathbb{E}_Q[f] - \mathbb{E}_P[f] = &\sum_{x:Q(x)\ge P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) \\ &+ \sum_{x:Q(x)<P(x)} (f(x) - \mathbb{E}_P[f])(Q(x) - P(x)) \\ &+ \sum_{x:P(x)>Q(x)} (\mathbb{E}_P[f] - \mathbb{E}_Q[f])(Q(x) - P(x))\,. \end{aligned} \tag{6.6}$$

The first term in the right-hand side of (6.6) is upper bounded as

$$\begin{aligned} &\sum_{x:Q(x)\ge P(x)} (f(x) - \mathbb{E}_Q[f])(Q(x) - P(x)) \\ &\quad = \sum_{x:Q(x)\ge P(x)} \sqrt{Q(x)}(f(x) - \mathbb{E}_Q[f])\frac{Q(x)-P(x)}{\sqrt{Q(x)}} \\ &\quad \overset{(a)}{\le} \sqrt{\sum_{x:Q(x)\ge P(x)} Q(x)(f(x) - \mathbb{E}_Q[f])^2}\sqrt{\sum_{x:Q(x)\ge P(x)} \frac{(Q(x)-P(x))^2}{Q(x)}} \\ &\quad \overset{(b)}{\le} \sqrt{\mathcal{V}_{Q,P}(f)}\sqrt{2\mathtt{KL}(P,Q)}\,, \end{aligned} \tag{6.7}$$

where (a) follows from Cauchy-Schwarz inequality and (b) follows from Lemma A.7.

Similarly, the second term in (6.6) satisfies

$$\begin{aligned}\sum_{x:Q(x)<P(x)} &(f(x)-\mathbb{E}_P[f])(Q(x)-P(x)) \\ &= \sum_{x:Q(x)<P(x)} \sqrt{P(x)}(f(x)-\mathbb{E}_P[f])\frac{Q(x)-P(x)}{\sqrt{P(x)}} \\ &\le \sqrt{\mathcal{V}_{P,Q}(f)}\sqrt{2\mathtt{KL}(P,Q)}\ . \qquad (6.8)\end{aligned}$$

Finally, we bound the last term in (6.6):

$$\begin{aligned}(\mathbb{E}_P[f]-\mathbb{E}_Q[f])\sum_{x:P(x)>Q(x)}(Q(x)-P(x)) &\overset{(a)}{=} \frac{1}{2}(\mathbb{E}_Q[f]-\mathbb{E}_P[f])\|P-Q\|_1 \\ &\le \frac{1}{2}\mathbb{S}(f)\|P-Q\|_1^2 \\ &\overset{(b)}{\le} \mathbb{S}(f)\mathtt{KL}(P,Q)\ , \qquad (6.9)\end{aligned}$$

where (a) follows from the fact that for any pair of distributions U and V on the same alphabet $\mathcal{X}$, it holds that $\sum_{x\in\mathcal{X}}|U(x)-V(x)| = 2\sum_{x:U(x)\ge V(x)}(U(x)-V(x))$, and where (b) follows from Pinsker's inequality.

The proof is concluded by combining (6.7), (6.8), and (6.9). □

6.E Proof of Lemma 6.3

Statement (i) is a direct consequence of the definition of $\mathcal{V}_{P,Q}$. We next prove statement (ii).

Observe that Lemma A.7 implies that for all $x\in\mathcal{X}$

$$|P(x)-Q(x)| \le \sqrt{2\max(P(x),Q(x))\mathtt{KL}(Q,P)}\ .$$

Hence,

$$\begin{aligned}\mathcal{V}_{P,Q}(f) &= \sum_{x:P(x)\ge Q(x)} P(x)(f(x)-\mathbb{E}_P[f])^2 \\ &\le \sum_{x:P(x)\ge Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 \\ &\quad + \sqrt{2\mathtt{KL}(Q,P)}\sum_{x:P(x)\ge Q(x)}\sqrt{P(x)}(f(x)-\mathbb{E}_P[f])^2\ . \qquad (6.10)\end{aligned}$$

The first term in the right-hand side of (6.10) is bounded as follows:

$$\sum_{x:P(x)\ge Q(x)} Q(x)(f(x)-\mathbb{E}_P[f])^2 \le 2\sum_{x:P(x)\ge Q(x)} Q(x)(f(x)-\mathbb{E}_Q[f])^2$$

$$
\begin{aligned}
&+ 2(\mathbb{E}_Q[f] - \mathbb{E}_P[f])^2 \\
&\le 2\mathbb{V}_Q(f) + 2(\mathbb{E}_Q[f] - \mathbb{E}_P[f])^2 \,.
\end{aligned}
$$

Note that

$$(\mathbb{E}_Q[f] - \mathbb{E}_P[f])^2 \le \mathbb{S}(f)^2 \|P - Q\|_1^2 \le 2\mathbb{S}(f)^2 \mathtt{KL}(Q, P) \,,$$

which further gives

$$\sum_{x: P(x) \ge Q(x)} Q(x)(f(x) - \mathbb{E}_P[f])^2 \le 2\mathbb{V}_Q(f) + 4\mathbb{S}(f)^2 \mathtt{KL}(Q, P) \,.$$

Now we consider the second term in (6.10). First observe that

$$
\begin{aligned}
&\sum_{x: P(x) \ge Q(x)} \sqrt{P(x)}(f(x) - \mathbb{E}_P[f])^2 \\
&\qquad \le \sqrt{\sum_{x: P(x) \ge Q(x)} P(x)(f(x) - \mathbb{E}_P[f])^2} \sqrt{\sum_x (f(x) - \mathbb{E}_P[f])^2} \\
&\qquad \le \sqrt{\mathcal{V}_{P,Q}(f)}\mathbb{S}(f)\sqrt{|\mathcal{X}|} \,,
\end{aligned}
$$

thanks to Cauchy-Schwarz inequality. Hence, the second term in (6.10) is upper bounded by

$$\mathbb{S}(f)\sqrt{2|\mathcal{X}|\mathcal{V}_{P,Q}(f)\mathtt{KL}(Q, P)} \,.$$

Combining the previous bounds together, we get

$$\mathcal{V}_{P,Q}(f) \le 2\mathbb{V}_Q(f) + 4\mathbb{S}(f)^2 \mathtt{KL}(Q, P) + \mathbb{S}(f)\sqrt{2|\mathcal{X}|\mathcal{V}_{P,Q}(f)\mathtt{KL}(Q, P)} \,,$$

which leads to

$$\Big(\sqrt{\mathcal{V}_{P,Q}(f)} - \mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q, P)/2}\Big)^2 \le 2\mathbb{V}_Q(f) + \mathbb{S}(f)^2(|\mathcal{X}|/2 + 4)\mathtt{KL}(Q, P) \,,$$

so that using the inequality $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$, we finally obtain

$$
\begin{aligned}
\sqrt{\mathcal{V}_{Q,P}(f)} &\le \sqrt{2\mathbb{V}_Q(f) + \mathbb{S}(f)^2(|\mathcal{X}|/2 + 4)\mathtt{KL}(Q, P)} + \mathbb{S}(f)\sqrt{|\mathcal{X}|\mathtt{KL}(Q, P)/2} \\
&\le \sqrt{2\mathbb{V}_Q(f)} + \mathbb{S}(f)(\sqrt{2|\mathcal{X}|} + 2)\sqrt{\mathtt{KL}(Q, P)} \,.
\end{aligned}
$$

The proof is completed by observing that $\sqrt{2|\mathcal{X}|} + 2 \le 3\sqrt{|\mathcal{X}|}$ for $|\mathcal{X}| \ge 2$. □

6.F Proof of Corollary 6.2

Let M be an ergodic MDP. Recall that by [22, Theorem 1], the regret under $\mathbb{A} =$ Burnetas-Katehakis for MDP M satisfies:

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Regret}_{\mathbb{A},T}]}{\log(T)} \le c_{\mathrm{bk}}(M) := \sum_{(s,a)\in\mathcal{C}_M} \frac{\varphi(s,a)}{\mathtt{KL}(p(\cdot|s,a), q^\star_{s,a})} ,$$

where $\mathcal{C}_M$ denotes the set of critical state-action pairs in M (see Chapter 2), and where for any $(s,a)\in\mathcal{C}_M$ we define

$$q^\star_{s,a} \in \arg\min_{q\in\Lambda(s,a)} \mathtt{KL}(p(\cdot|s,a), q) .$$

Recall that $q^\star_{s,a}\in\Lambda(s,a)$ implies $\mu(s,a) + {q^\star_{s,a}}^\top b^\star > g^\star + b^\star(s)$ so that

$$\varphi(s,a) = g^\star + b^\star(s) - \mu(s,a) - p(\cdot|s,a)^\top b^\star < (q^\star_{s,a} - p(\cdot|s,a))^\top b^\star .$$

Hence, we get for any $(s,a)\in\mathcal{C}_M$,

$$\frac{\varphi(s,a)}{\mathtt{KL}(p(\cdot|s,a), q^\star_{s,a})} \le \frac{\big[(q^\star_{s,a} - p(\cdot|s,a))^\top b^\star\big]^2}{\varphi(s,a)\mathtt{KL}(p(\cdot|s,a), q^\star_{s,a})} \le \frac{2\mathbb{V}_{q^\star_{s,a}}(b^\star)}{\varphi(s,a)} , \qquad (6.11)$$

thanks to Corollary 6.1 in the last inequality.

To derive an explicit upper bound for $\varphi(s,a)/\mathtt{KL}(p(\cdot|s,a), q^\star_{s,a})$ (and in turn for $c_{\mathrm{bk}}(M)$), we bound $\mathbb{V}_{q^\star_{s,a}}(b^\star)$. To this end, fix state-action pair $(s,a)\in\mathcal{C}_M$. To ease notation, define the short-hands $P = p(\cdot|s,a)$ and $Q = q^\star_{s,a}$. We have that

$$\begin{aligned}
\mathbb{V}_Q(b^\star) &= \sum_x Q(x)(b^\star(x) - \mathbb{E}_Q[b^\star])^2 \\
&= \sum_x P(x)(b^\star(x) - \mathbb{E}_Q[b^\star])^2 + \sum_x (Q(x)-P(x))(b^\star(x) - \mathbb{E}_Q[b^\star])^2 \\
&\le 2\sum_x P(x)(b^\star(x) - \mathbb{E}_P[b^\star])^2 + 2\underbrace{(\mathbb{E}_P[b^\star] - \mathbb{E}_Q[b^\star])^2}_{G_1} \\
&\quad + \underbrace{\sum_x (Q(x)-P(x))(b^\star(x) - \mathbb{E}_Q[b^\star])^2}_{G_2} .
\end{aligned}$$

Defining $\Psi := \mathbb{S}(b^\star)$, we observe that

$$G_1 \le \Psi^2\|P-Q\|_1^2 \le 2\Psi^2\mathtt{KL}(P,Q) ,$$

thanks to Pinsker's inequality. Moreover, applying Cauchy-Schwarz inequality and Lemma A.7 gives

$$G_2 \le \sum_{x:Q(x)\ge P(x)} \frac{Q(x)-P(x)}{\sqrt{Q(x)}}\sqrt{Q(x)}(b^\star(x) - \mathbb{E}_Q[b^\star])^2$$

$$\leq \sqrt{\sum_{x:Q(x)\geq P(x)} (Q(x)-P(x))^2/Q(x)}\sqrt{\sum_{x:Q(x)\geq P(x)} Q(x)(b^\star(x)-\mathbb{E}_Q[b^\star])^4}$$
$$\leq \Psi\sqrt{2\mathbb{V}_Q(b^\star)\mathtt{KL}(P,Q)}\ .$$

Combining these bounds, we deduce

$$\mathbb{V}_Q(b^\star) \leq 2\mathbb{V}_P(b^\star) + \Psi\sqrt{2\mathbb{V}_Q(b^\star)\mathtt{KL}(P,Q)} + 4\Psi^2\mathtt{KL}(P,Q)\ ,$$

or equivalently

$$\Big(\sqrt{\mathbb{V}_Q(b^\star)} - \Psi\sqrt{\mathtt{KL}(P,Q)/2}\Big)^2 \leq 2\mathbb{V}_P(b^\star) + \frac{9}{2}\Psi^2\mathtt{KL}(P,Q)\ .$$

We thus get

$$\sqrt{\mathbb{V}_Q(b^\star)} \leq \sqrt{2\mathbb{V}_P(b^\star)} + 2\Psi\sqrt{2\mathtt{KL}(P,Q)}\ ,$$

and $\mathbb{V}_Q(b^\star) \leq 4\mathbb{V}_P(b^\star) + 16\Psi^2\mathtt{KL}(P,Q)$. Define the short-hand $\varphi := \varphi(s,a)$. Now combining the above bound with (6.11), we deduce

$$\frac{\varphi}{\mathtt{KL}(P,Q)} \leq \frac{2\mathbb{V}_Q(b^\star)}{\varphi} \leq \frac{4\mathbb{V}_P(b^\star) + 16\Psi^2\mathtt{KL}(P,Q)}{\varphi}\ .$$

Defining $X = \varphi/\mathtt{KL}(P,Q)$, $A = 4\mathbb{V}_P(b^\star)/\varphi$, and $B = 16\Psi^2$, the above inequality reads $X \leq A + B/X$. Solving for X yields $X \leq A/2 + \sqrt{B + A^2/4}$, and so

$$\frac{\varphi}{\mathtt{KL}(P,Q)} \leq \frac{2\mathbb{V}_P(b^\star)}{\varphi} + \sqrt{16\Psi^2 + \frac{4\mathbb{V}_P(b^\star)^2}{\varphi^2}}$$
$$\leq \frac{4\mathbb{V}_P(b^\star)}{\varphi} + 4\Psi\ ,$$

thus giving

$$c_{\mathrm{bk}}(M) \leq 4\sum_{(s,a)\in\mathcal{C}_M} \frac{\mathbb{V}_{p(\cdot|s,a)}(b^\star)}{\varphi(s,a)} + 4SA\Psi$$

and completing the proof. □

6.G Proof of Theorem 6.2

In this section, we provide the regret analysis of KL-Ucrl. We will try to follow the notations used in the proof of [25, Theorem 2].

We first recall the following result indicating that the true model belongs to the set of plausible MDPs with high probability. Recall that for $\delta \in (0,1]$ and $T > 1$,

$$C_\mu := C_\mu(T,\delta) = \log(4SA\log(T)/\delta)/1.99$$

$$C_p := C_p(T,\delta) = S\left(B + \log(G)(1+1/G)\right) ,$$

where $B := B(T,\delta) = \log(2eS^2A\log(T)/\delta)$ and $G := G(T,\delta) = B + 1/\log(T)$. Moreover, observe that $C_p(T,\delta) \le 4SB$.

Lemma 6.5 ([24, Proposition 1])**.** *For all $T \ge 1$ and $\delta > 0$, and for any pair (s,a) it holds that*

$$\mathbb{P}\Big(\forall t \le T,\ |\hat{\mu}_t(s,a) - \mu(s,a)| \le \sqrt{C_\mu/N_t(s,a)}\Big) \ge 1 - \frac{\delta}{SA}$$
$$\mathbb{P}\Big(\forall t \le T,\ N_t(s,a)\mathtt{KL}(\hat{p}_t(s,a), p(\cdot|s,a)) \le C_p\Big) \ge 1 - \frac{\delta}{SA} ,$$

In particular, $\mathbb{P}(\forall t \le T,\ M \in \mathcal{M}_t) \ge 1 - 2\delta$.

Next we prove the theorem.

Proof of Theorem 6.2. Let $T \ge 1$ and $\delta \in (0,1)$. Fix algorithm $\mathbb{A} = \texttt{KL-Ucrl}$. Denote by $m(T)$ the number of episodes started by KL-Ucrl up to time step T (hence, $1 \le k \le m(T)$).

Applying Azuma-Hoeffding inequality (see Theorem B.3), as in the proof of [24, Theorem 1], we deduce that

$$\mathrm{Regret}_{\mathbb{A},T} = Tg^\star - \sum_{t=1}^{T} r(s_t,a_t) \le \sum_{s,a} N_T(s,a)(g^\star - \mu(s,a)) + \sqrt{\tfrac{1}{2}T\log(1/\delta)} ,$$

with probability at least $1-\delta$. The regret up to time T can be decomposed as the sum of the regret incurred in various episodes. Let Δ_k denote the regret in episode k:

$$\Delta_k := \sum_{s,a} v_k(s,a)(g^\star - \mu(s,a)) .$$

Therefore, Lemma 6.5 implies that with probability at least $1-3\delta$,

$$\mathrm{Regret}_{\mathbb{A},T} \le \sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} + \sqrt{\tfrac{1}{2}T\log(1/\delta)} .$$

Consider an episode $k \ge 1$ such that $M \in \mathcal{M}_k$. The pair (s,a) is considered as sufficiently sampled if its number of observations satisfies $N_k(s,a) \ge \ell_{s,a}$, where

$$\ell_{s,a} = \ell_{s,a}(T,\delta) := \max\Big\{\frac{128SB\max(\Psi^2,1)}{\varphi(s,a)^2},\ 32SB\Big(\frac{\log(D)}{\log(1/\gamma)}\Big)^2\Big\}, \quad \forall s,a,$$

where γ denotes the contraction factor of the mapping induced by the transition probability matrix $P_\star$ of the optimal policy (γ can be determined as a function of elements of $P_\star$).

Consider the case where all state-action pairs are sufficiently sampled (we analyse the case where some pairs are under-sampled (i.e., not sufficiently sampled) at the end of the proof). We have

$$|\tilde{\mu}_k(s,a) - \mu(s,a)| \le |\tilde{\mu}_k(s,a) - \hat{\mu}_k(s,a)| + |\hat{\mu}_k(s,a) - \mu(s,a)| \le 2\sqrt{\frac{C_\mu}{N_k(s,a)^+}} \,.$$

Hence,

$$\begin{aligned}\Delta_k &= \sum_{s,a} v_k(s,a)(g^\star - \tilde{\mu}_k(s,a)) + \sum_{s,a} v_k(s,a)(\tilde{\mu}_k(s,a) - \mu(s,a)) \\ &\le \sum_{s,a} v_k(s,a)(g^\star - \tilde{\mu}_k(s,a)) + 2\sqrt{C_\mu}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \,.\end{aligned}$$

Let $\tilde{\mu}_k$ and $\widetilde{P}_k$ respectively denote the reward vector and transition probability matrix induced by policy $\tilde{\pi}_k$ on $\widetilde{M}_k$, i.e., $\tilde{\mu}_k := (\tilde{\mu}_k(s,\tilde{\pi}_k(s)))_s$, $\widetilde{P}_k := \big(\tilde{p}_k(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$. By Bellman's optimality equation, $\tilde{g}_k - \tilde{\mu}_k(s,a) = (\widetilde{P}_k - I)\tilde{b}_k$. Hence, defining $v_k = (v_k(s,\tilde{\pi}_k(s))_s$ yields

$$\Delta_k \le v_k(\widetilde{P}_k - I)\tilde{b}_k + (g^\star - \tilde{g}_k)v_k\mathbf{1} + 2\sqrt{C_\mu}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \,.$$

We use the following decomposition for the first term in the right-hand side of the above inequality:

$$v_k(\widetilde{P}_k - I)\tilde{b}_k = \underbrace{v_k(\widetilde{P}_k - P_k)b^\star}_{F_1(k)} + \underbrace{v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star)}_{F_2(k)} + \underbrace{v_k(P_k - I)\tilde{b}_k}_{F_3(k)} \,.$$

Let $c = 1 + \sqrt{2}$. The following two lemmas provide upper bounds for $F_1(k)$ and $F_2(k)$:

Lemma 6.6. *For all $k \in \mathbb{N}$ such that $M \in \mathcal{M}_k$, with probability at least $1-\delta$, it holds that*

$$F_1(k) \le (4 + 6\sqrt{2})\sqrt{SB}\sum_{s,a} v_k(s,a)\sqrt{\frac{\mathbf{V}^\star_{s,a}}{N_k(s,a)^+}} + 63\Psi S^{3/2}B^{3/2}\sum_{s,a}\frac{v_k(s,a)}{N_k(s,a)^+} \,.$$

Lemma 6.7. *For all $k \in \mathbb{N}$ such that $M \in \mathcal{M}_k$, it holds that*

$$F_2(k) + (g^\star - \tilde{g}_k)v_k\mathbf{1} \le \big(2\sqrt{32SB} + 1\big)\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \,.$$

Analysis of Term F_3. Now we bound the term $\sum_{k=1}^{m(T)} F_3(k)$. To this end, similarly to the proof of [25, Theorem 2] and [24, Theorem 1], we define the martingale difference sequence $(Z_t)_{t\ge 1}$, where $Z_t = (p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}})\tilde{b}_{k(t)}\mathbb{I}\{M \in \mathcal{M}_{k(t)}\}$ for $t \in \{t_k, t_{k+1} - 1\}$, where $k(t)$ denotes the episode containing t. Note that for all t, $|Z_t| \le 2D$. Now applying Azuma-Hoeffding inequality (see Theorem B.3), we deduce that with probability at least $1 - \delta$

$$\begin{aligned}\sum_{k=1}^{m(T)} F_3(k) &\le \sum_{t=1}^{T} Z_t + 2m(T)D \\ &\le D\sqrt{2T\log(1/\delta)} + 2DSA\log_2\left(\tfrac{8T}{SA}\right) .\end{aligned}$$

The regret due to under-sampled state-action pairs. To analyze the under-sampled regime, where some state-action pair is not sufficiently sampled, we borrow some techniques from [51]. For any state-action pair (s, a), let $L_{s,a}$ denote the set of indexes of episodes in which (s, a) is chosen and yet (s, a) is under-sampled; namely $k \in L_{s,a}$ if $\tilde{\pi}_k(s) = a$ and $N_k(s, a) \le \ell_{s,a}$. Furthermore, let $\tau_k(s, a)$ denote the length of such an episode.

Consider an episode $k \in L_{s,a}$. By Markov's inequality, with probability at least $\frac{1}{2}$, it takes at most $2T_M$ to reach state s from any state s' in k, where T_M denotes the mixing time of M. Let us divide episode k into $\lfloor \frac{\tau_k(s,a)}{2T_M} \rfloor$ sub-episodes, each with length greater than $2T_M$. It then follows that in each sub-episode, (s, a) is visited with probability at least $\frac{1}{2}$.

Using Hoeffding's inequality, if we consider n such sub-episodes, with probability at least $1 - \frac{\delta}{SA}$,

$$N(s, a) > n/2 - \sqrt{n\log(SA/\delta)}.$$

Now we find n that implies $N(s, a) < \ell_{s,a}$. Noting that $x \mapsto \frac{x}{2} - \sqrt{\alpha x}$ is increasing for $x \ge \alpha$, we have that for $n > 10\max(\ell_{s,a}, \log(SA/\delta))$,

$$\begin{aligned}n/2 - \sqrt{n\log(SA/\delta)} &> 5\max(\ell_{s,a}, \log(SA/\delta)) - \sqrt{10\max(\ell_{s,a}, \log(SA/\delta))\log(SA/\delta)} \\ &> \max(\ell_{s,a}, \log(SA/\delta)) .\end{aligned}$$

Hence, with probability at least $1 - \frac{\delta}{SA}$, it holds that

$$\sum_{k\in L_{s,a}} \left\lfloor \frac{\tau_k(s, a)}{2T_M} \right\rfloor \le 10\max(\ell_{s,a}, \log(SA/\delta)) .$$

Hence, the regret due to under-sampled state-action pairs can be upper bounded by

$$\sum_{s,a}\sum_{k\in L_{s,a}} \tau_k(s, a) \le 20T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M \sum_{s,a} |L_{s,a}|$$

$$\leq 20T_M \sum_{s,a} \max(\ell_{s,a}, \log(SA/\delta)) + 2T_M S^2 A^2 \log_2\left(\tfrac{8T}{SA}\right) ,$$

with probability at least $1-\delta$. Here we used that $|L_{s,a}| \leq m(T)$.

Now applying Lemmas 6.6 and 6.7 together with the above bounds, and using the fact $C_\mu \leq B/1.99$, we deduce that with probability at least $1-3\delta$

$$\begin{aligned}\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} &\leq (4+6\sqrt{2})\sqrt{SB}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\sqrt{\mathbf{V}^\star_{s,a}}\\ &+ (2\sqrt{32SB}+3\sqrt{B}+1)\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\\ &+ 63\Psi S^{3/2}B^{3/2}\sum_{s,a}\frac{v_k(s,a)}{N_k(s,a)^+}\\ &+ D\sqrt{2T\log(1/\delta)} + 2DSA\log_2\left(\tfrac{8T}{SA}\right)\\ &+ 20T_M\sum_{s,a}\max(\ell_{s,a},\log(SA/\delta)) + 2T_M S^2A^2\log_2\left(\tfrac{8T}{SA}\right) .\end{aligned}$$

To simplify the above bound, we will use Lemmas 6.9, 6.10, and 6.11 together with Jensen's inequality:

$$\sum_{k=1}^{m(T)}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} \leq c\sum_{s,a}\sqrt{N_T(s,a)} \leq c\sqrt{SAT} ,$$
$$\sum_{k=1}^{m(T)}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}}\sqrt{\mathbf{V}^\star_{s,a}} \leq c\sum_{s,a}\sqrt{\mathbf{V}^\star_{s,a}N_T(s,a)} \leq c\sqrt{T\textstyle\sum_{s,a}\mathbf{V}^\star_{s,a}} ,$$
$$\sum_{k=1}^{m(T)}\sum_{s,a}\frac{v_k(s,a)}{N_k(s,a)^+} \leq 2\sum_{s,a}\log(N_T(s,a)) + SA \leq 2SA\log\left(\tfrac{T}{SA}\right) + SA .$$

Putting everything together, we deduce that with probability at least $1-6\delta$,

$$\begin{aligned}\mathrm{Regret}_{\mathbb{A},T} &\leq \sum_{k=1}^{m(T)}\Delta_k\mathbb{I}\{M\in\mathcal{M}_k\} + \sqrt{\tfrac{1}{2}T\log(1/\delta)}\\ &\leq 31\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}TB} + 35S\sqrt{ATB} + (\sqrt{2}D+1)\sqrt{T\log(1/\delta)}\\ &+ 126S^{5/2}AB^{5/2}\log\left(\tfrac{T}{SA}\right) + 2DSA\log_2\left(\tfrac{8T}{SA}\right)\\ &+ 20T_M\sum_{s,a}\max(\ell_{s,a},\log(SA/\delta)) + 2T_MS^2A^2\log_2\left(\tfrac{8T}{SA}\right) + 63S^{5/2}A .\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Regret}_{\mathbb{A},T} &\le 31\sqrt{S\sum_{s,a}\mathbf{V}^\star_{s,a}TB} + 35S\sqrt{ATB} + (\sqrt{2}D+1)\sqrt{T\log(1/\delta)}\\ &\quad + \widetilde{\mathcal{O}}\Big(SA(T_MSA + D + S^{3/2})\log(T)\Big)\ .\end{aligned}$$

Noting that $B = \mathcal{O}(\log(\log(T)/\delta))$ gives the desired scaling and completes the proof. □

6.G.1 Proof of Lemma 6.6

We have

$$F_1(k) = \underbrace{v_k(\widehat{P}_k - P_k)b^\star}_{G_1} + \underbrace{v_k(\widetilde{P}_k - \widehat{P}_k)b^\star}_{G_2}$$

Next we provide upper bounds for G_1 and G_2.

Term G_1. We have

$$\begin{aligned}G_1 &= \sum_s v_k(s,\tilde{\pi}_k(s))\sum_{s'} b^\star(s')\big(\hat{p}_k(s'|s,\tilde{\pi}_k(s)) - p(s'|s,\tilde{\pi}_k(s))\big)\\ &\le \sum_{s,a} v_k(s,a)\sum_{s'} b^\star(s')\big(\hat{p}_k(s'|s,a) - p(s'|s,a)\big)\ .\end{aligned}$$

Fix $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Define the short-hands $p = p(\cdot|s,a)$, $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, and $N_k^+ = N_k(s,a)^+$. Applying Corollary 6.1 (the first statement) and using the fact that $M\in\mathcal{M}_k$ give:

$$\begin{aligned}\sum_{s'} b^\star(s')(\hat{p}_k(s') - p(s')) &\le \sqrt{2\mathbf{V}^\star_{s,a}\mathtt{KL}(\hat{p}_k,p)} + \frac{2}{3}\Psi\mathtt{KL}(\hat{p}_k,p)\\ &\le \sqrt{8S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{8\Psi SB}{3N_k^+}\ .\end{aligned}$$

Therefore,

$$G_1 \le \sqrt{8SB}\sum_{s,a} v_k(s,a)\sqrt{\mathbf{V}^\star_{s,a}/N_k(s,a)^+} + \frac{8}{3}\Psi SB\sum_{s,a} v_k(s,a)/N_k(s,a)^+\ .$$

Term G_2. We have

$$G_2 \le \sum_{s,a} v_k(s,a)\sum_{s'} b^\star(s')\big(\tilde{p}_k(s'|s,a) - \hat{p}_k(s'|s,a)\big)\ .$$

Fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Define the short-hands $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, $\tilde{p}_k = \tilde{p}_k(\cdot|s,a)$, and $N_k^+ = N_k(s,a)^+$. An application of Lemma 6.2 and Lemma 6.3 gives

$$\begin{aligned}\sum_{s'} b^\star(s')(\tilde{p}_k(s') - \hat{p}_k(s')) &\le \left(\sqrt{\mathcal{V}_{\tilde{p}_k,\hat{p}_k}(b^\star)} + \sqrt{\mathcal{V}_{\hat{p}_k,\tilde{p}_k}(b^\star)}\right)\sqrt{2\mathtt{KL}(\hat{p}_k,\tilde{p}_k)} + \Psi\mathtt{KL}(\hat{p}_k,\tilde{p}_k)\\ &\le c\sqrt{2\mathbb{V}_{\hat{p}_k}(b^\star)\mathtt{KL}(\hat{p}_k,\tilde{p}_k)} + \Psi(1+3\sqrt{2S})\mathtt{KL}(\hat{p}_k,\tilde{p}_k)\ ,\end{aligned}$$

where $c = 1+\sqrt{2}$.

Note that when $M \in \mathcal{M}_k$, an application of Lemma 6.8, stated below, implies that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{s'} b^\star(s')(\tilde{p}_k(s') - \hat{p}_k(s')) &\le 4c\sqrt{S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{\Psi S^{3/2}B^{3/2}}{N_k^+}(12c\sqrt{2} + 12\sqrt{2} + 4/\sqrt{S})\\ &\le 4c\sqrt{S\mathbf{V}^\star_{s,a}B/N_k^+} + \frac{61\Psi S^{3/2}B^{3/2}}{N_k^+}\ ,\end{aligned}$$

where we used that $S \ge 2$. Multiplying by $v_k(s,a)$ and summing over s,a yields

$$G_2 \le 4c\sqrt{SB}\sum_{s,a} v_k(s,a)\sqrt{\mathbf{V}^\star_{s,a}/N_k(s,a)^+} + 61\Psi S^{3/2}B^{3/2}\sum_{s,a} v_k(s,a)/N_k(s,a)^+\ .$$

The lemma follows by combing the bounds on G_1 and G_2. □

Lemma 6.8. *For any episode $k \ge 1$ such that $M \in \mathcal{M}_k$, it holds that for any pair (s,a),*

$$\sqrt{\mathbb{V}_{\hat{p}_k(\cdot|s,a)}(f)} \le \sqrt{2\mathbb{V}_{p(\cdot|s,a)}(f)} + \frac{6S\mathbb{S}(f)B}{\sqrt{N_k(s,a)}}$$

with probability at least $1-\delta$.

Proof. Let $\delta \in (0,1)$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$. Consider an episode $k \ge 1$ such that $M \in \mathcal{M}_k$, and define $\hat{p}_k = \hat{p}_k(\cdot|s,a)$, $p = p(\cdot|s,a)$, and $N_k = N_k(s,a)$. Observe that by a Bernstein-like inequality [113, Lemma F.2] (see Theorem B.4), we have: for all $s' \in \mathcal{S}$, with probability at least $1-\delta$,

$$\hat{p}_k(s') - p(s') \le \sqrt{\frac{2p(s')C_b}{N_k}} + \frac{2C_b}{N_k}\ ,$$

with $C_b = C_b(t,\delta) := \log(3\log(\max(e,t))/\delta)$. It then follows that with probability at least $1-\delta$,

$$\mathbb{V}_{\hat{p}_k}(f) = \sum_{s'} \hat{p}_k(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2$$

$$\leq \sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2 + \sqrt{\frac{2C_b}{N_k}} \sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2$$
$$+ \frac{2C_b}{N_k} \sum_{s'} (f(s') - \mathbb{E}_{\hat{p}_k}[f])^2$$
$$\leq \underbrace{\sum_{s'} p(s')(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_1} + \sqrt{\frac{2C_b}{N_k}} \underbrace{\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_{\hat{p}_k}[f])^2}_{Z_2} + \frac{2C_b S \mathbb{S}(f)^2}{N_k} . \tag{6.12}$$

Next we bound Z_1 and Z_2. Observe that

$$Z_1 \leq 2 \sum_{s'} p(s')(f(s') - \mathbb{E}_p[f])^2 + 2(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2$$
$$\leq 2\mathbb{V}_p(f) + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) ,$$

where the last inequality follows from

$$(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \leq \mathbb{S}(f)^2 \|p - \hat{p}_k\|_1^2 \leq 2\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) . \tag{6.13}$$

For Z_2 we have

$$Z_2 \leq 2 \sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_p[f])^2 + 2(\mathbb{E}_p[f] - \mathbb{E}_{\hat{p}_k}[f])^2 \sum_{s'} \sqrt{p(s')} .$$

Now, using Cauchy-Schwarz inequality

$$\sum_{s'} \sqrt{p(s')}(f(s') - \mathbb{E}_p[f])^2 \leq \sqrt{\sum_{s'} p(s')(f(s') - \mathbb{E}_p[f])^2 \sum_{s'} (f(s') - \mathbb{E}_p[f])^2}$$
$$\leq \sqrt{S\mathbb{V}_p(f)} \mathbb{S}(f) ,$$

so that using (6.13), we deduce that

$$Z_2 \leq 2\mathbb{S}(f)\sqrt{S\mathbb{V}_p(f)} + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p) \sum_{s'} \sqrt{p(s')}$$
$$\leq 2\mathbb{S}(f)\sqrt{S\mathbb{V}_p(f)} + 4\mathbb{S}(f)^2 \mathtt{KL}(\hat{p}_k, p)\sqrt{S} ,$$

where the last inequality follows from Jensen's inequality:

$$\sum_{s'} \sqrt{p(s')} = \sum_{s'} p(s') \sqrt{\frac{1}{p(s')}} \leq \sum_{s'} \sqrt{\frac{p(s')}{p(s')}} = \sqrt{S} .$$

Putting together, we deduce that with probability at least $1 - \delta$,

$$\mathbb{V}_{\hat{p}_k}(f) \leq 2\mathbb{V}_p(f) + 2\mathbb{S}(f)\sqrt{\frac{2SC_b}{N_k}} \Big(\sqrt{\mathbb{V}_p(f)} + 2\mathbb{S}(f)\mathtt{KL}(\hat{p}_k, p) \Big)$$

$$+\,\mathbb{S}(f)^2\Big(4\mathtt{KL}(\hat{p}_k,p)+\frac{2SC_b}{N_k}\Big)\,.$$

Noting that $M\in\mathcal{M}_k$, we obtain

$$\begin{aligned}\mathbb{V}_{\hat{p}_k}(f)&\le 2\mathbb{V}_p(f)+\mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)C_b}{N_k}}+4\mathbb{S}(f)^2\frac{\sqrt{2SC_b}C_p}{N_k^{3/2}}+\frac{(4C_p+2SC_b)\mathbb{S}(f)^2}{N_k}\\&\le 2\mathbb{V}_p(f)+\mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)C_b}{N_k}}+\frac{S\mathbb{S}(f)^2}{N_k}(16B\sqrt{2SC_b}+16B+2C_b)\\&\le 2\mathbb{V}_p(f)+\mathbb{S}(f)\sqrt{\frac{8S\mathbb{V}_p(f)B}{N_k}}+\frac{36S^{3/2}B^{3/2}\mathbb{S}(f)^2}{N_k}\,,\end{aligned}$$

with probability at least $1-\delta$, where we used $C_p=4SB$, $C_b\le B$, and $S\ge 2$. The proof is concluded by observing that

$$\begin{aligned}\sqrt{\mathbb{V}_{\hat{p}_k}(f)}&\le\sqrt{2\mathbb{V}_p(f)}+\mathbb{S}(f)\sqrt{\frac{SB}{N_k}}+6\mathbb{S}(f)B\sqrt{\frac{S^{3/2}}{N_k}}\\&\le\sqrt{2\mathbb{V}_p(f)}+\frac{6S\mathbb{S}(f)B}{\sqrt{N_k}}\,,\end{aligned}$$

with probability at least $1-\delta$. □

6.G.2 Proof of Lemma 6.7

Let $k\ge 1$ be the index of an episode such that $M\in\mathcal{M}_k$. Let $\tilde{\star}:=\tilde{\star}_k$ denote the optimal policy in $\mathcal{M}_k$. The proof proceeds in three steps.

Step 1 We remark that by definition of the bias functions, it holds that

$$\begin{aligned}\tilde{b}_k-b^\star&=(g^\star-\tilde{g}_k)\mathbf{1}+\tilde{\mu}_k+\widetilde{P}_kb^\star-\mu_\star-P_\star b^\star+\widetilde{P}_k(\tilde{b}_k-b^\star)\\&\le(\tilde{g}_{\tilde{\star}}-\tilde{g}_k)\mathbf{1}+\tilde{\mu}_k-\mu_k+(\widetilde{P}_k-P_k)b^\star+\widetilde{P}_k(\tilde{b}_k-b^\star)-\varphi_k\,,\end{aligned}$$

where we define $\varphi_k(s):=\varphi(s,\tilde{\pi}_k(s))$ for all s. Defining

$$\xi_k(s)=2\sqrt{C_\mu/N_k(s,\tilde{\pi}_k(s))^+},\qquad\zeta_k(s)=\Psi\sqrt{32SB/N_k(s,\tilde{\pi}_k(s))^+}\,,$$

we obtain the following bound:

$$\tilde{b}_k-b^\star\le\frac{1}{\sqrt{t_k}}\mathbf{1}+\xi_k+\zeta_k-\varphi_k+\widetilde{P}_k(\tilde{b}_k-b^\star)\,.$$

It is straightforward to check that the assumption $N_k(s,\tilde{\pi}_k(s))\ge\ell_{s,\tilde{\pi}_k(s)}$ for all s implies

$$\tilde{b}_k-b^\star\le\widetilde{P}_k(\tilde{b}_k-b^\star)\,.\qquad(6.14)$$

Note also that $\varphi(s,\tilde{\pi}_k(s)) \geq 0$ since $\star$ is $b^\star$-improving.

On the other hand, it holds that

$$\begin{aligned} b^\star - \tilde{b}_{\tilde\star} &= (\tilde{g}_{\tilde\star} - g^\star)\mathbf{1} + \mu_\star + P_\star b^\star - \tilde{\mu}_{\tilde\star} - \widetilde{P}_{\tilde\star}\tilde{b}_{\tilde\star} \\ &\leq (\tilde{g}_{\tilde\star} - g^\star)\mathbf{1} + \mu_\star + P_\star b^\star - \mu_\star - P_\star\tilde{b}_{\tilde\star} \\ &= (\tilde{g}_{\tilde\star} - g^\star)\mathbf{1} + P_\star(b^\star - \tilde{b}_{\tilde\star})\,. \end{aligned}$$

Noting $P_\star\mathbf{1} = \mathbf{1}$, and since all entries of $P_\star$ are non-negative, we thus get for all $J \in \mathbb{N}$

$$b^\star - \tilde{b}_{\tilde\star} \leq J(\tilde{g}_{\tilde\star} - g^\star)\mathbf{1} + P_\star^J(b^\star - \tilde{b}_{\tilde\star})\,.$$

Step 2 Let us now introduce $\mathcal{S}_s^+ = \{x \in \mathcal{S} : \widetilde{P}_k(s,x) > P_k(s,x)\}$ as well as its complementary set $\mathcal{S}_s^- = \mathcal{S} \setminus \mathcal{S}_s^+$. Using (6.14), $\tilde{b}_k - b^\star \leq 0$ so that

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) &= \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}} (\widetilde{P}_k(s,x) - P_k(s,x))(\tilde{b}_k(x) - b^\star(x)) \\ &\leq \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} \underbrace{(P_k(s,x) - \widetilde{P}_k(s,x))}_{\geq 0}(b^\star(x) - \tilde{b}_k(x))\,. \end{aligned}$$

We thus obtain

$$\begin{aligned} v_k(\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star) &\leq \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(b^\star(x) - \tilde{b}_{\tilde\star}(x)) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde\star}(x) - \tilde{b}_k(x)) \\ &\leq \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde\star})](x) \\ &\quad + \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde\star}(x) - \tilde{b}_k(x)) \\ &\quad - J\sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))(g^\star - \tilde{g}_{\tilde\star})\,. \end{aligned} \tag{6.15}$$

We thus get

$$\begin{aligned} &\sum_s v_k(s,\tilde{\pi}_k(s))\Big((\widetilde{P}_k - P_k)(\tilde{b}_k - b^\star)(s) + g^\star - \tilde{g}_{\tilde\star}\Big) \\ &\qquad \leq \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde\star})](x) + \eta_k \\ &\qquad + \sum_s v_k(s,\tilde{\pi}_k(s))\Big[1 - J\sum_{x\in\mathcal{S}_s^-} (P_k(s,x) - \widetilde{P}_k(s,x))\Big](g^\star - \tilde{g}_{\tilde\star})\,, \end{aligned} \tag{6.16}$$

where $\eta_k := \sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))(\tilde{b}_{\tilde{\star}}(x) - \tilde{b}_k(x))$ is controlled by the error of computing $\tilde{b}_k$ in episode k. In particular, for the considered variant of the algorithm,

$$\begin{aligned}\eta_k &\leq \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \frac{1}{\sqrt{t_k}}\\ &\leq \sqrt{32SB}\sum_s \frac{v_k(s,\tilde{\pi}_k(s))}{N_k(s,\tilde{\pi}_k(s))^+}\,,\end{aligned}$$

where we used $t_k \geq N_k(s,\tilde{\pi}_k(s))$ for all s.

Step 3 It remains to choose J. To this end, we remark that the mapping induced by $P_\star$ is a contractive mapping, namely there exists some $\gamma < 1$ such that for any function f,

$$\mathbb{S}(P_\star f) \leq \gamma \mathbb{S}(f)\,.$$

Let us choose $J \geq \frac{\log(D)}{\log(1/\gamma)}$, so that with a simple upper bound, it comes

$$\begin{aligned}&\sum_s v_k(s,\tilde{\pi}_k(s)) \sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))[P_\star^J(b^\star - \tilde{b}_{\tilde{\star}})](x)\\ &\qquad\leq \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \frac{\mathbb{S}(P_\star^J(b^\star - \tilde{b}_{\tilde{\star}}^\star))}{2}\\ &\qquad\leq \sum_s v_k(s,\tilde{\pi}_k(s))\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 D e^{-\log(D)}\\ &\qquad\leq \sum_s v_k(s,\tilde{\pi}_k(s))\sqrt{\frac{32SB}{N_k(s,\tilde{\pi}_k(s))^+}}\,.\end{aligned}$$

In the sequel, we take $J = \frac{\log(D)}{\log(1/\gamma)}$. This enables us to control the first two terms in (6.16) and it remains to control the term

$$\sum_s v_k(s,\tilde{\pi}_k(s))\left[1 - J\sum_{x\in\mathcal{S}_s^-}(P_k(s,x) - \widetilde{P}_k(s,x))\right](g^\star - \tilde{g}_{\tilde{\star}})\,.$$

In particular we would like to ensure that the bracket is non-negative, since in that case, it is multiplied by a term that is negative. To this end, we note that the term in brackets is lower bounded by

$$1 - J\|p(\cdot|s,\tilde{\pi}_k(s)) - \tilde{p}_k(\cdot|s,\tilde{\pi}_k(s))\|_1 \geq 1 - \frac{\log(D)}{\log(1/\gamma)}\sqrt{\frac{32SB}{N_k(s,\tilde{\pi}_k(s))^+}}\,,$$

and is thus guaranteed to be non-negative since

$$N_k(s,\tilde\pi_k(s)) \ge \ell_{s,\tilde\pi_k(s)} \ge 32SB\Big(\frac{\log(D)}{\log(1/\gamma)}\Big)^2 .$$

Putting together, we finally have shown that

$$\begin{aligned} v_k(\widetilde P_k - P_k)(\tilde b_k - b^\star) + v_k(g^\star - \tilde g_k)\mathbf{1} &\le v_k(\widetilde P_k - P_k)(\tilde b_k - b^\star) + v_k(g^\star - \tilde g_{\tilde\star})\mathbf{1} + \frac{1}{\sqrt{t_k}} v_k\mathbf{1} \\ &\le \big(2\sqrt{32SB}+1\big)\sum_s \frac{v_k(s,\tilde\pi_k(s))}{\sqrt{N_k(s,\tilde\pi_k(s))^+}} \\ &\le \big(2\sqrt{32SB}+1\big)\sum_{s,a} \frac{v_k(s,a)}{\sqrt{N_k(s,a)^+}} , \end{aligned}$$

which completes the proof.

□

6.H Technical Lemmas

Lemma 6.9 ([25, Lemma 19]). *Consider a sequence $(z_k)_{1\le k\le n}$ with $0 \le z_k \le Z_{k-1} := \max\big\{1, \sum_{i=1}^{k-1} z_i\big\}$ for $k \ge 1$ and $Z_0 \ge 1$. Then,*

$$\sum_{k=1}^n \frac{z_k}{\sqrt{Z_{k-1}}} \le (\sqrt{2}+1)\sqrt{Z_n} .$$

Lemma 6.10. *Consider a sequence $(z_k)_{1\le k\le n}$ with $0 \le z_k \le Z_{k-1} := \max\big\{1, \sum_{i=1}^{k-1} z_i\big\}$ for $k \ge 1$ and $Z_0 = z_1$. Then,*

$$\sum_{k=1}^n \frac{z_k}{Z_{k-1}} \le 2\log(Z_n) + 1 .$$

Proof. We prove the lemma by induction over n. For $n = 1$, we have $z_1/Z_0 = 1$. Since $Z_1 = \max\{1, z_1\}$, it holds that $z_1/Z_0 \le 2\log(Z_1) + 1$.

Now consider $n > 1$. By the induction hypothesis, it holds that $\sum_{k=1}^{n-1} z_k/Z_{k-1} \le 2\log(Z_{n-1}) + 1$. Now it follows from the facts $z_n = Z_n - Z_{n-1}$ and $Z_{n-1} \le Z_n \le 2Z_{n-1}$ for $n \ge 2$, that

$$\begin{aligned} \sum_{k=1}^n \frac{z_k}{Z_{k-1}} &\le 2\log(Z_{n-1}) + \frac{z_n}{Z_{n-1}} + 1 \\ &\le 2\log(Z_{n-1}) + 2\frac{Z_n - Z_{n-1}}{Z_n} + 1 \\ &= 2\log(Z_{n-1}) + 2\Big(1 - \frac{1}{Z_n/Z_{n-1}}\Big) + 1 \le 2\log(Z_n) + 1 , \end{aligned}$$

where the last inequality follows from $\log(x) \geq 1 - \frac{1}{x}$ valid for all $x \geq 1$ (see, e.g., [98]). This concludes the proof. □

Lemma 6.11. *Let $\alpha_i, \ldots, \alpha_d$ be non-negative numbers and $T \geq 1$, and denote by V the optimal value of the following problem:*

$$\begin{aligned} \max_x \quad & \sum_{i=1}^d \sqrt{\alpha_i x_i} \\ \text{subject to :} \quad & \sum_{i=1}^d x_i = T \ , \\ & x_i \geq 0, \quad i = 1, \ldots, d \ . \end{aligned}$$

Then, $V = \sqrt{T \sum_{i=1}^d \alpha_i}$

Proof. Introduce the partial Lagrangian

$$L(x, \lambda) = \sum_{i=1}^d \sqrt{\alpha_i x_i} + \lambda\Big(T - \sum_{i=1}^d x_i\Big) \ .$$

Writing KKT conditions, we observe that the optimal point $x_i^\star, i = 1, \ldots, d$ satisfies

$$\frac{\alpha_i}{2\sqrt{x_i^\star}} - \lambda = 0, \quad \forall i \text{ and } \quad \sum_{i=1}^d x_i^\star - T = 0 \, .$$

Hence, we obtain $x_i^\star = \alpha_i / (4\lambda^2)$ (note that this choice of $x_i^\star$ satisfies the inequality constraints too). Plugging this into the equality constraint, it follows that $\lambda = \sqrt{\frac{1}{4T} \sum_{j=1}^d \alpha_j}$, thus giving $x_i^\star = \alpha_i T / \sum_{j=1}^d \alpha_j$. Therefore,

$$V = \sum_{i=1}^d \sqrt{\alpha_i x_i^\star} = \sum_{i=1}^d \frac{\alpha_i}{\sum_{j=1}^d \alpha_j} \sqrt{T \sum_{j=1}^d \alpha_j} = \sqrt{T \sum_{j=1}^d \alpha_j},$$

which completes the proof. □

Chapter 7

Conclusions and Future Work

This chapter concludes the thesis by summarizing the main results and proposing some directions for future research.

7.1 Conclusions

In Chapter 3, we investigated stochastic combinatorial MABs with Bernoulli rewards. Leveraging the theory of adaptive control of Markov chains with unknown transition probabilities, we derived tight and problem-specific lower bounds on the regret under bandit and semi-bandit feedback. These bounds are unfortunately implicit (more precisely, they are optimal values of semi-infinite linear programs). We then investigated how these lower bounds scale with the dimension of the set of arms $\mathcal{A}$ for some problems of interest. We proposed the ESCB algorithm for the case of semi-bandit feedback and showed that its regret grows at most as $\mathcal{O}(\sqrt{m}d\Delta_{\min}^{-1}\log(T))$ after T rounds, where d denotes the number of basic actions (i.e., the dimension of $\mathcal{A}$), and m denotes the maximal number of basic actions per arm. ESCB improves over the state-of-the-art algorithms proposed for combinatorial MABs in the literature. ESCB is unfortunately computationally expensive. To alleviate its computational complexity, we proposed Epoch-ESCB and assessed its performance numerically.

In Chapter 4, we studied stochastic online shortest-path routing, which was formulated as a stochastic combinatorial MAB problem with geometrically distributed rewards. Three types of routing policies were considered that include source routing with bandit feedback, source routing with semi-bandit feedback, and hop-by-hop routing. We presented regret lower bounds for each type of routing. Our derivations showed that the regret lower bounds for source routing policies with semi-bandit feedback and that for hop-by-hop routing policies are identical, indicating that taking routing decisions hop by hop does not bring any advantage. On the contrary, the regret lower bounds for source routing policies with bandit and semi-bandit feedback can be significantly different, illustrating the importance of having semi-bandit feedback. In the case of semi-bandit feedback, we proposed two source routing policies, namely GeoCombUCB-1 and GeoCombUCB-2, which attain a regret scaling as

$\mathcal{O}(\sqrt{m}d\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$ after transmitting N packets. Furthermore, we provided an improved regret bound for KL-SR [33] growing as $\mathcal{O}(md\Delta_{\min}^{-1}\theta_{\min}^{-2}\log(N))$. These routing policies strike an interesting trade-off between computational complexity and performance, and exhibit better regret upper bounds than state-of-the-art algorithms.

In Chapter 5, we investigated a generic sequential resource allocation under semi-bandit feedback, where a decision maker wishes to quickly learn and converge to an approximate Proportionally Fair (PF) allocation, referred to as APF allocation. The materials presented in that chapter provide a precise description of the system model as well as a notion of regret that captures the rate at which the allocation chosen by the decision maker can converge towards the APF allocation. We derived an asymptotic problem-specific regret lower bound for a class of policies for this problem and further showed that it scales as $\Omega(m\theta_{\min}^{-1}\Delta_{\min}^{-1}\log(T))$ for specific instances. We also presented an optimistic algorithm for learning APF allocation, which we called ES-APF, enjoying a regret bound of order $\mathcal{O}(m^3\theta_{\min}^{-1}\Delta_{\min}^{-1}\log(T))$.

Chapter 6 concerns RL in MDPs under average-reward criterion. We revisited some existing lower bounds in the literature and provided alternative presentations, which make appear the local variance of the bias function of the MDP. Furthermore, we revisited the regret analysis of KL-Ucrl [24] and showed that the leading term $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ obtained for the regret of KL-Ucrl in [24] can be reduced to $\widetilde{\mathcal{O}}\Big(\sqrt{S\sum_{s,a}\mathbf{V}_{s,a}^\star T} + D\sqrt{T}\Big)$. Computations of these regret bounds in some illustrative MDP showed that the reported upper bound may improve an order of magnitude over the existing ones (as observed experimentally in [114]), thus highlighting the fact that trading the diameter of the MDP to the local variance of the bias function may result in huge improvements. Our regret analysis relies on novel transportation concentration inequalities presented in that chapter, which could be of independent interest in the performance analysis of RL in MDPs in various setups.

7.2 Future Work

There are several directions to extend the work carried out in this thesis. Some of them are outlined next.

Analysis of TS for combinatorial MAB problems. One intriguing direction for future research is to analyze the performance of TS for the combinatorial MAB problems considered. Despite its popularity in the MAB literature, TS is seldom studied for combinatorial problems except for the recent work of Komiyama et al. [56], which concerns the very simple setting of fixed-size subsets. Regret analysis of TS for generic combinatorial structures proves quite challenging. Nonetheless, it is a promising direction since (i) if the offline problem is polynomial-time solvable, efficient implementations for TS might exist (because arm selection can be cast as the same linear combinatorial problem as the offline problem), and (ii) in empirical evaluations TS exhibits superior performance over existing algorithms.

Efficient algorithms for problems that admit efficient oracles. The current implementation of ESCB presented in Chapter 3 (as well as that of GeoCombUCB in Chapter 4) is computationally expensive as it requires $\mathcal{O}(|\mathcal{A}|)$ computations in each round and as $|\mathcal{A}|$ could well grow exponentially with the number of basic actions d. A promising future work is to present a wiser way to implement ESCB efficiently for problems that admit efficient oracles. To determine whether or not such an implementation exists will provide a further insight into the trade-off between computational complexity and performance (in terms of regret) of online combinatorial problems. This trade-off is certainly hard to characterize and yet of extreme importance. To the best of our knowledge, existing literature does not provide rigorous results about such a trade-off. For the case of shortest-path routing (as an instance of a problem admitting an efficient oracle), our proposed algorithms in Chapter 4 provide a first insight, up to our knowledge, into such a trade-off.

Non-linear reward functions. In this thesis we mostly concentrated on combinatorial problems with modular objective functions. Nonetheless, a lot of interesting applications may be cast as combinatorial MABs whose average reward function is non-linear. An interesting direction to continue this work is to devise algorithms for these cases. A particular case of interest is submodular reward functions under matroid constraints. There are numerous applications of combinatorial problems that fall within this framework, which include bidding in ad exchange [116], product search [117], leader selection in leader-follower multi-agent systems [118], coverage problem, and influence maximization [119]. Despite some recent studies (see, e.g., [120]), there are only few results for stochastic MAB problems with submodular reward functions in the stochastic setting, though these problems have received more attention in the adversarial setting [121, 116].

Stochastic combinatorial MABs under bandit feedback. Stochastic combinatorial MABs under bandit feedback have seldom been studied, though the problem is very well investigated in the adversarial setting.

Although the setting of semi-bandit feedback is strongly motivated by several applications involving multiple agents, that of bandit feedback does arguably make more sense for some other applications. A notable instance is shortest-path routing scenarios in which the decision maker has only access to the end-to-end (bandit) feedback rather than per-link (semi-bandit) feedback. Therefore, the need for devising arm selection algorithms for bandit feedback is evident from a practical standpoint. Although this task could be much more complicated than that of semi-bandit feedback, we conjecture that for the case of matroids, it might be relatively straightforward due to the unimodality of these structures.

Learning PF allocations. An interesting future direction relevant to the resource allocation problem studied in Chapter 5 is to devise an algorithm for learning the PF allocation with analytical performance guarantees. A natural candidate to accommodate this task is an algorithm that relies on the `KL-UCB` index (for each task-server pair) and consists in computing the optimistic PF allocation, namely the solution to the PF problem parameterized by the indexes. It is however much more difficult to carry out the corresponding regret analysis due to random nature of the received feedback. As a future work, we plan to derive a finite-time regret bound for such an algorithm as well as a regret lower bound for learning such allocations.

Minimax-optimal regret bounds for undiscounted RL. In view of the minimax lower bound presented in Chapter 6, there is a possibility to remove a factor $\sqrt{S}$ from the regret upper bound of `KL-Ucrl`. In the simpler setting of episodic RL with *known* episode horizon H, recently a few studies have shown that by taking advantage of this knowledge, it is possible to devise minimax-optimal RL algorithms, namely algorithms whose regret grows as $\widetilde{\mathcal{O}}(\sqrt{HSAT})$; see, e.g., [112]. We note, however that such techniques do not apply straightforwardly to RL with average reward criterion, which was the setup of interest in Chapter 6. Nonetheless, we believe that combining techniques of such studies with the tools developed in this thesis is a promising research direction.

Exploiting structure in reinforcement learning. Recalling the regret bounds in Chapter 6, we have seen that state-of-the-art algorithms typically incur a problem-independent (resp. problem-independent) regret bound, whose leading term scales as $S\sqrt{T}$ (resp. $S^2\log(T)$), where only dependencies on S and T are shown. Moreover, in view of existing lower bounds, a regret of order $\sqrt{ST}$ (or problem-dependent regret of $S\log(T)$) cannot be beaten by any admissible algorithm. An implication of these bounds is that existing algorithms would work well for when the size of state-space is not very large. On the other hand, the majority of nowadays applications admit a huge state-space, and often, endowed with some structural properties. In order to successfully apply RL algorithms to such real-world applications, it is therefore crucial to exploit the structural properties of the underlying problems. Using the structural properties could result in the reduction of the corresponding state-space, which would in turn bring huge potential benefits.

To be more specific, identifying the problem structure would result in a reduced model with orders of magnitude smaller state-space (and possibly actions). Now devising RL algorithms on this reduced model, one can therefore hope for incurring much lower regret than that could be obtained by an structure-oblivious algorithm. Exploiting problem structure in RL is definitely a challenging task that, to the best of our knowledge, is not rigorously investigated in the literature and existing heuristic-based algorithms for discounted RL capable of doing so often fail to provide performance guarantees in terms of regret or sample complexity.

A crucial step in the design of an structure-aware algorithm for RL is *state aggregation*. There are some structure-aware algorithms for RL in MDPs with continuous state-space, which consider a fairly broad structural properties, such as Lipschitz or Hölder continuity of reward functions or transition kernel; see, e.g., [108, 109, 110] and also refer to the discussion in Section 6.2. We remark, however, that in practical scenarios we are often interested in more specific structural properties than the one characterized by a globally determined, for instance, Lipschtiz constant. To summarize, we believe that it would be very promising to devise structure-aware algorithms for RL with analytical performance guarantees.

Appendix A

Properties of the KL-Divergence

In this appendix we briefly overview some of the properties of the Kullback-Leibler (KL) divergence, which prove instrumental throughout this thesis. The KL-divergence, originally introduced by Kullback and Leibler in [122], defines a distance measure between two distributions. It has been given other names such as *KL information number*, *relative entropy*, and *information divergence*. The KL-divergence is a special case of a larger class of functions referred to as f-divergence; see, e.g., [96] for a through treatment.

The majority of the results presented here can be found in, e.g., [123] and [96].

A.1 Definition

Let F and G be two distributions on the same set $\mathcal{X}$ with $G \ll F$, i.e., G is absolutely continuous with respect to F. Then, the KL-divergence between F and G is defined as

$$\texttt{KL}(F, G) = \mathbb{E}_F \left[\log \frac{F(\mathrm{d}x)}{G(\mathrm{d}x)} \right] = \int_{\mathcal{X}} \log \frac{F(\mathrm{d}x)}{G(\mathrm{d}x)} F(\mathrm{d}x),$$

where $F(\mathrm{d}x)/G(\mathrm{d}x)$ denotes the Radon-Nikodym derivative of F with respect to G. $\texttt{KL}(F, G)$ may be derived using densities as well: Let $m(\mathrm{d}x)$ be an appropriate measure. Then,

$$\texttt{KL}(F, G) = \int_{\mathcal{X}} \log \frac{f(x)}{g(x)} f(x) m(\mathrm{d}x).$$

We remark that the above expression does not depend on the choice of $m(\mathrm{d}x)$. It is also noted that if G is not absolutely continuous with respect to F, then $\texttt{KL}(F, G) = \infty$. In the discrete case, namely when F and G are probability vectors (and $\mathcal{X}$ is a finite set), the above definition reads

$$\texttt{KL}(F, G) = \sum_{x \in \mathcal{X}} F(x) \log \frac{F(x)}{G(x)},$$

with the usual convention that $p \log \frac{p}{q}$ is defined to be 0 if $p = 0$ and $+\infty$ if $p > q = 0$. In what follows, we mainly concern the KL-divergence between two discrete distributions.

First observe that the KL-divergence is always non-negative: $\mathtt{KL}(F, G) \geq 0$ with equality if $F(x) = G(x), \forall x \in \mathcal{X}$.

The next result, referred to as the chain rule for the KL-divergence, may prove useful when working with the KL-divergence of joint probability distributions. A consequence of this result is that the KL-divergence is additive for independent random variables.

Theorem A.1 (Chain Rule)**.** *For two random variables $x, y \in \mathcal{X}$ we have:*

$$\mathtt{KL}(F(x,y), G(x,y)) = \mathtt{KL}(F(x), G(x)) + \mathtt{KL}(F(y|x), G(y|x)),$$

where $\mathtt{KL}(F(y|x), G(y|x)) = \mathbb{E}_x[\log(F(y|x)/G(y|x))]$.

The KL-divergence between two Bernoulli distributions with respective parameters p and q, denoted by $\mathtt{kl}(p, q)$, is:

$$\mathtt{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

The function $\mathtt{kl}$ is sometimes referred to as the *binary relative entropy.* Some of the properties of $\mathtt{kl}(p, q)$ are summarized in the following lemma.

Lemma A.1 ([54])**.** *The mapping $q \mapsto \mathtt{kl}(p, q)$ satisfies the following properties for all $p \in [0, 1]$:*
(i) It is strictly convex on $[0, 1]$ and attains its minimum at p with $\mathtt{kl}(p, p) = 0$.
(ii) Its derivative with respect to the second parameter $q \mapsto \mathtt{kl}'(p, q) = \frac{q-p}{q(1-q)}$ is strictly increasing on $(p, 1)$.
(iii) For $p < 1$, we have $\mathtt{kl}(p, q) \underset{q \to 1^-}{\to} \infty$ and $\mathtt{kl}'(p, q) \underset{q \to 1^-}{\to} \infty$.

The following lemma provides upper and lower bounds for the KL-divergence of Bernoulli distributions.

Lemma A.2. *For any $p, q \in [0, 1]$, it holds that*

$$2(p-q)^2 \leq \mathtt{kl}(p,q) \leq \frac{(p-q)^2}{q(1-q)}.$$

The lower bound in the above lemma is referred to as Pinsker's inequality. The following lemma presents a local version of Pinsker's inequality:

Lemma A.3 ([124, Lemma 2])**.** *For $0 \leq u < v \leq 1$ we have:*

$$\mathtt{kl}(u,v) \geq \frac{1}{2v}(u-v)^2.$$

We also provide the following lemma relating the KL-divergence between two geometric distributions to that of corresponding Bernoulli distributions.

Lemma A.4 ([75, Lemma 3]). *For any $u, v \in (0, 1]$, we have:*

$$\mathtt{KLG}(u, v) = \frac{\mathtt{kl}(u, v)}{u}.$$

Proof. We have:

$$\begin{aligned}
\mathtt{KLG}(u, v) =& \sum_{i=1}^{\infty} u(1-u)^{i-1} \log \frac{u(1-u)^{i-1}}{v(1-v)^{i-1}} \\
=& \sum_{i=1}^{\infty} u(1-u)^{i-1} \log \frac{u}{v} + \sum_{i=1}^{\infty} (i-1) u(1-u)^{i-1} \log \frac{1-u}{1-v} \\
=& \log \frac{u}{v} + \frac{1-u}{u} \log \frac{1-u}{1-v} = \frac{\mathtt{kl}(u, v)}{u}.
\end{aligned}$$

□

We now turn back to the KL-divergence of discrete probability distributions and study some of its properties below.

Lemma A.5 ([125, Lemma A.5]). *Let P and Q be two probability distributions on a finite alphabet $\mathcal{X}$. Then,*

$$\mathtt{KL}(P, Q) \le \sum_{x \in \mathcal{X}} \mathtt{kl}(P(x), Q(x)) \; .$$

Lemma A.6 (Pinsker's Inequality). *For any probability distributions P and Q on a finite alphabet $\mathcal{X}$, it holds that*

$$\mathtt{KL}(P, Q) \ge \frac{1}{2} \|P - Q\|_1^2 \; .$$

The following lemma provides a local version of Pinsker's inequality for two probability vectors, which can be seen as the extension of Lemma A.3 ([124, Lemma 2]) for the case of discrete probability measures. To the best of our knowledge, this lemma is new.

Lemma A.7. *For any probability distributions P and Q on a finite alphabet $\mathcal{X}$, it holds that*

$$\mathtt{KL}(P, Q) \ge \frac{1}{2} \sum_{x : P(x) \ne Q(x)} \frac{(P(x) - Q(x))^2}{\max(P(x), Q(x))} \; .$$

Proof. The first and second derivatives of $\mathtt{KL}$ satisfy:

$$\frac{\partial}{\partial P(x)}\mathtt{KL}(P,Q) = 1 + \log\frac{P(x)}{Q(x)}, \quad \forall x \in \mathcal{X},$$
$$\frac{\partial^2}{\partial P(x)\partial P(y)}\mathtt{KL}(P,Q) = \frac{\mathbb{I}\{x=y\}}{P(x)}, \quad \forall x, y \in \mathcal{X}.$$

By Taylor's Theorem, there exists a probability vector Ξ, where $\Xi = tP+(1-t)Q$ for some $t \in (0,1)$, so that

$$\begin{aligned}
\mathtt{KL}(P,Q) &= \mathtt{KL}(Q,Q) + \sum_x (P(x)-Q(x))\frac{\partial}{\partial P}\mathtt{KL}(Q,Q) \\
&+ \frac{1}{2}\sum_{x,y}(P(x)-Q(x))(P(y)-Q(y))\frac{\partial^2}{\partial P(x)\partial P(y)}\mathtt{KL}(\Xi,Q) \\
&= \sum_x (P(x)-Q(x)) + \sum_x \frac{(P(x)-Q(x))^2}{2\Xi(x)} \\
&\geq \sum_{x:P(x)\neq Q(x)} \frac{(P(x)-Q(x))^2}{2\max(P(x),Q(x))}\ ,
\end{aligned}$$

thus concluding the proof. □

Appendix B

Concentration Inequalities

This appendix is devoted to the overview of some important concentration inequalities used in various chapters of this thesis.

B.1 Bounded Random Variables

We begin with stating the celebrated Chernoff-Hoeffding bounds:

Theorem B.1 (Chernoff-Hoeffding Bound)**.** *Let $X_1, \dots, X_n$ be 0-1 independent random variables with $\mathbb{E}[X_i] = p_i$. Let $Y = \frac{1}{n}\sum_{t=1}^n X_t$ and $\mu = \mathbb{E}[Y] = \frac{1}{n}\sum_{t=1}^n p_i$. Then for all $0 < \lambda < 1 - \mu$,*

$$\mathbb{P}(X \geq \mu + \lambda) \leq e^{-n\mathtt{kl}(\mu+\lambda,\mu)},$$

and for all $0 < \lambda < \mu$,

$$\mathbb{P}(X \leq \mu - \lambda) \leq e^{-n\mathtt{kl}(\mu-\lambda,\mu)}.$$

Theorem B.2 (Chernoff-Hoeffding Bound)**.** *Let $X_1, \dots, X_n$ be random variables with common ranges $[0, 1]$ and such that $\mathbb{E}[X_t|X_1 \dots, X_{t-1}] = \mu$. Let $S_n = \sum_{t=1}^n X_t$. Then for all $a \geq 0$:*

$$\mathbb{P}(S_n \geq n\mu + a) \leq e^{-2a^2/n},$$
$$\mathbb{P}(S_n \leq n\mu - a) \leq e^{-2a^2/n}.$$

The next theorem states Azuma-Hoeffding inequality for bounded martingale difference sequences:

Theorem B.3 (Azuma-Hoeffding Inequality [126])**.** *Let $(X_t)_{1\leq t\leq n}$ be a martingale difference sequence with $|X_i| \leq c$ for all i. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$:*

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \geq \varepsilon\Big) \leq e^{-\frac{\varepsilon^2}{2nc^2}} .$$

The following results give the concentration for self-normalized form of bounded random variables.

Theorem B.4 ([113, Lemma F.2])**.** *Let $(X_n)_{1\leq n\leq t}$ be a sequence of Bernoulli random variables with mean $\mu \in [0,1]$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists n \colon \hat{\mu}_n - \mu \geq \sqrt{\frac{2\mu}{n}h(n,\delta)} + \frac{h(n,\delta)}{n}\right) \leq \delta \ ,$$

where $h(n,\delta) := 2\mathrm{llnp}(n) + \log(3/\delta)$.

Theorem B.5 ([43, Theorem 10])**.** *Let $(X_t)_{t\geq 1}$ be a sequence of independent random variables bounded in $[0,1]$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with common expectation $\mu = \mathbb{E}[X_t]$. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields of $\mathcal{F}$ such that for each t, $\sigma(X_1, \dots, X_t) \subset \mathcal{F}_t$ and for $s > t$, X_s is independent from $\mathcal{F}_t$. Consider a previsible sequence $(\varepsilon_t)_{t\geq 1}$ of Bernoulli variables (for all $t > 0$, ε_t is $\mathcal{F}_{t-1}$-measurable). Let $\delta > 0$ and for every $t \in [n]$ let*

$$S(t) = \sum_{s=1}^{t} \varepsilon_s X_s, \quad N(t) = \sum_{s=1}^{t} \varepsilon_s, \quad \hat{\mu}(t) = \frac{S(t)}{N(t)},$$

$$u(n) = \max\{q > \hat{\mu}(n) : N(n)\mathtt{kl}(\hat{\mu}(n), q) \leq \delta\}.$$

Then: $\mathbb{P}(u(n) < \mu) \leq \lceil \delta \log(n) \rceil e^{-(\delta+1)}$.

The following theorem is a generalization of Theorem B.5 and gives a concentration inequality on sums of empirical KL-divergences.

Theorem B.6 ([68, Theorem 2])**.** *For all $\delta \geq K+1$ and $n \in \mathbb{N}$ we have:*

$$\mathbb{P}\Big(\sum_{i=1}^{K} N_i(n)\mathtt{kl}(\hat{\mu}_i(n), \mu_i) \geq \delta\Big) \leq \left(\frac{\lceil \delta \log(n) \rceil \delta}{K}\right)^K e^{-\delta(K+1)}.$$

In particular, we have:

Corollary B.1. *There exists a constant C_K that only depends on K, such that for all $n \geq 2$ we have:*

$$\mathbb{P}\Big(\sum_{i=1}^{K} N_i(n)\mathtt{kl}(\hat{\mu}_i(n), \mu_i) \geq \log(n) + 4K\log(\log(n))\Big) \leq C_K n^{-1}(\log(n))^{-2}.$$

The following lemma proves useful in the proof of various regret bounds throughout the thesis. It states that if a set of instants Λ can be decomposed into a family of singletons such that the arm i is drawn sufficiently many times, then the number of times in Λ (in expectations) at which the empirical average reward of i is badly estimated is finite.

Theorem B.7 ([70, Theorem B.1]). *Let $i \in \{1, \dots, K\}$ and $\delta > 0$. Define $\mathcal{F}_n$ the σ-algebra generated by $(X_i(t))_{1 \le t \le n, 1 \le i \le K}$. Let $\Lambda \subset \mathbb{N}$ be a (random) set of instants. Assume that there exists a sequence of (random) sets $(\Lambda(s))_{s \ge 1}$ such that (i) $\Lambda \subset \cup_{s \ge 1} \Lambda(s)$, (ii) for all $s \ge 1$ and all $n \in \Lambda(s)$, $N_i(n) \ge \varepsilon s$, (iii) $|\Lambda(s)| \le 1$, and (iv) the event $n \in \Lambda(s)$ is $\mathcal{F}_n$-measurable. Then, for all $\delta > 0$:*

$$\mathbb{E}[\sum_{n \ge 1} \mathbb{I}\{n \in \Lambda,\ |\hat{\mu}_i(n) - \mu_i| \ge \delta\}] \le \frac{1}{\varepsilon \delta^2}.$$

The proof of the above lemma leverages a concentration inequality proposed in [70][1]. A consequence of the above lemma is the following corollary which states that the expected number of times at which basic action i is sampled and the empirical average reward of i exceeds the true mean reward of i by some threshold is finite. Note that this result holds irrespective of how arm i is chosen. To present the corollary we let $A_i(n)$ denote the event of sampling basic action $i \in \{1, \dots, K\}$ at round n.

Corollary B.2. *For all $i \in \{1, \dots, K\}$ and all $\delta > 0$:*

$$\mathbb{E}[\sum_{n \ge 1} \mathbb{I}\{A_i(n),\ |\hat{\mu}_i(n) - \mu_i| \ge \delta\}] \le \frac{1}{\delta^2}.$$

Proof. Let $\Lambda = \{n : \mathbb{I}\{A_i(n)\} = 1\}$. Observe that for each $s \in \mathbb{N}$, there exists at most one time index $\varphi_s \in \mathbb{N}$ such that $N_i(\varphi_s) = s$ and $\varphi_s \in \Lambda$, since $N_i(n) = N_i(n-1) + 1$ for all $n \in \Lambda$. The set Λ is included in $\cup_{s \ge 1} \{\varphi_s\}$. The announced result is then a direct consequence of Lemma B.7 with $\varepsilon = 1$. □

B.2 Discrete Probability Distributions

In the following theorems, we provide concentration for the L_1 deviation of empirical distributions of discrete probability measures.

Theorem B.8 (L_1-Deviation Bound for Empirical Distribution, [127]). *Let P be a probability distribution on the finite alphabet $\mathcal{X}$. Let $(X_n)_{n \ge 1}$ be a set of i.i.d. samples distributed according to P, and $\hat{P}_n$ be the corresponding empirical estimation of P. Define $\pi_P := \max_{X \subseteq \mathcal{X}} \min(\mathbb{P}_P(X), 1 - \mathbb{P}_P(X))$, where $\mathbb{P}_P(X)$ is the probability of X under P. Furthermore, define*

$$\varphi(z) := \frac{1}{1-2p} \log \frac{1-p}{p}, \quad \forall p \in [0, 1/2)$$

and by convention define $\varphi(1/2) = 2$. Then with probability at least $1 - \delta$, it holds that

$$\|\hat{P}_n - P\|_1 \le \sqrt{\frac{2}{n\varphi(\pi_P)} \log \frac{2^{|\mathcal{X}|} - 2}{\delta}} \le \sqrt{\frac{2|\mathcal{X}|}{n} \log \frac{2}{\delta}}\ .$$

[1] We note that a slightly worse bound can be obtained from [48, Lemma 3].

Theorem B.9 (Uniform L_1-Deviation Bound for Empirical Distribution, [113, Lemma F.3])**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. categorical variables on a finite alphabet $\mathcal{X}$ with distribution P. Let $\hat{P}_n$ be the corresponding empirical distribution. Then for all $\delta \in (0,1]$*

$$\mathbb{P}\left(\exists n : \|\hat{P}_n - P\|_1 \geq \sqrt{\frac{4}{n}\Big(2\mathrm{llnp}(n) + \log\frac{3(2^{|\mathcal{X}|}-2)}{\delta}\Big)}\right) \leq \delta\ ,$$

where $\mathrm{llnp}(n) := \log(\log(\max(n,e)))$.

Theorem B.10 ([125, Lemma A.5])**.** *Let P be a probability distribution defined on a finite alphabet $\mathcal{X}$. Let $\hat{P}_t$ be the empirical estimation of P using the samples drawn from P up to time t. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\Big(N_t \mathtt{KL}(\widehat{P}_t, P) > \varepsilon\Big) \leq 2e(\varepsilon\log(t) + |\mathcal{X}|)e^{-\frac{\varepsilon}{|\mathcal{X}|}}\ .$$

B.2.1 Transportation Lemma

We restate the following theorem from Chapter 6, known as the transportation lemma; see, e.g., [115, Lemma 4.18]:

Theorem B.11 (Transportation Lemma)**.** *For any function f, let us introduce $\varphi_f : \lambda \mapsto \log\mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f]))]$. Whenever φ_f is defined on some possibly unbounded interval I containing 0, define its dual $\varphi_{\star,f}(x) = \sup_{\lambda\in I}(\lambda x - \varphi_f(\lambda))$. Then it holds*

$$\begin{aligned}
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leq \varphi^{-1}_{+,f}(\mathtt{KL}(Q,P))\,,\\
\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\geq \varphi^{-1}_{-,f}(\mathtt{KL}(Q,P))\,,
\end{aligned}$$

where

$$\begin{aligned}
\varphi^{-1}_{+,f}(t) &= \inf\{x \geq 0 : \varphi_{\star,f}(x) > t\}\,,\\
\varphi^{-1}_{-,f}(t) &= \sup\{x \leq 0 : \varphi_{\star,f}(x) > t\}\,.
\end{aligned}$$

Proof. Let us recall the fundamental equality

$$\forall\lambda\in\mathbb{R},\quad \log\mathbb{E}_P[\exp(\lambda(X - \mathbb{E}_P[X]))] = \sup_{Q\ll P}\left[\lambda\Big(\mathbb{E}_Q[X] - \mathbb{E}_P[X]\Big) - \mathtt{KL}(Q,P)\right].$$

In particular, we obtain on the one hand that (see also [115, Lemma 2.4])

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \min_{\lambda\in\mathbb{R}_+}\frac{\varphi_f(\lambda) + \mathtt{KL}(Q,P)}{\lambda}$$

Since $\varphi_f(0) = 0$, then the right-hand side quantity is non-negative. Let us call it u. Now, we note that for any t such that $u \geq t \geq 0$, then by construction of u, it holds $\mathtt{KL}(Q, P) \geq \varphi_{\star,f}(t)$. Thus, $\{x \geq 0 : \varphi_{f,\star}(x) > \mathtt{KL}(Q, P)\} = (u, \infty)$ and hence $u = \varphi_{+,f}^{-1}(\mathtt{KL}(Q, P))$.

On the other hand, it holds

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \quad \geq \quad \max_{\lambda \in \mathbb{R}_-} \frac{\varphi_f(\lambda) + \mathtt{KL}(Q, P)}{\lambda}$$

Since $\varphi(0) = 0$, then the right hand side quantity is non-positive. Let us call it v. Now, we note that for any t such that $v \leq t \leq 0$, then by construction of v, it holds $\mathtt{KL}(Q, P) \geq \varphi_{\star,f}(t)$. Thus, $\{x \leq 0 : \varphi_{\star,f}(x) > \mathtt{KL}(Q, P)\} = (-\infty, v)$ and hence $v = \varphi_{-,f}^{-1}(\mathtt{KL}(Q, P))$. □

Bibliography

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[2] I. Szita, "Reinforcement learning in games," in *Reinforcement Learning,* Vol. 12 of *Adaptation, Learning, and Optimization*, M. Wiering and M. van Otterlo, Eds. Springer, 2012, ch. 17, pp. 539–577.

[3] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *The Annals of Statistics*, pp. 1091–1114, 1987.

[4] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," *Statistics in Medicine*, vol. 28, no. 26, pp. 3294–3315, 2009.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[6] L. Lai, H. El Gamal, H. Jiang, and H. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Transactions on Mobile Computing*, vol. 10, no. 2, pp. 239–253, 2011.

[7] R. Kleinberg and T. Leighton, "The value of knowing a demand curve: Bounds on regret for online posted-price auctions," in *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2003, pp. 594–605.

[8] M. Babaioff, S. Dughmi, R. Kleinberg, and A. Slivkins, "Dynamic pricing with limited supply," *ACM Transactions on Economics and Computation*, vol. 3, no. 1, p. 4, 2015.

[9] L. Massoulié, M. I. Ohannessian, and A. Proutière, "Greedy-bayes for targeted news dissemination," in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015, pp. 285–296.

[10] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 2010, pp. 661–670.

[11] J. Kober and J. Peters, "Reinforcement learning in robotics: A survey," in *Reinforcement Learning,* Vol. 12 of *Adaptation, Learning, and Optimization*, M. Wiering and M. van Otterlo, Eds. Springer, 2012, pp. 579–610.

[12] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: Reinforcement learning framework," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 427–435, 2004.

[13] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *The Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[14] R. S. Sutton, A. G. Barto, and R. J. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Systems*, vol. 12, no. 2, pp. 19–22, 1992.

[15] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, 2009.

[16] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[17] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, pp. 285–294, 1933.

[18] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[19] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Minimax policies for combinatorial prediction games," in *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011, pp. 107–132.

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

[21] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[22] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, vol. 22, no. 1, pp. 222–255, 1997.

[23] A. Nilim and L. El Ghaoui, "Robust control of markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[24] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and Kullback-Leibler divergence," in *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 115–122.

[25] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *The Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.

[26] M. Lelarge, A. Proutiere, and M. S. Talebi, "Spectrum bandit optimization," in *Proceedings of 2013 IEEE Information Theory Workshop (ITW)*, 2013, pp. 34–38.

[27] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency.* Springer, 2003.

[28] B. Radunovic, A. Proutiere, D. Gunawardena, and P. Key, "Dynamic channel, rate selection and scheduling for white spaces," in *Proceedings of the 7th Conference on Emerging Networking EXperiments and Technologies (CoNEXT)*, 2011.

[29] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[30] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, 2010, pp. 1–9.

[31] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, 2004, pp. 45–53.

[32] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The on-line shortest path problem under partial monitoring." *The Journal of Machine Learning Research*, vol. 8, no. Oct, 2007.

[33] Z. Zou, A. Proutiere, and M. Johansson, "Online shortest path routing: The value of information," in *Proceedings of American Control Conference (ACC)*, 2014, pp. 2142–2147.

[34] T. He, D. Goeckel, R. Raghavendra, and D. Towsley, "Endhost-based shortest path routing in dynamic networks: An online learning approach," in *Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM)*, 2013, pp. 2202–2210.

[35] E. Kaufmann, "Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources," Ph.D. dissertation, Ecole nationale supérieure des telecommunications-ENST, 2014.

[36] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays–part I: IID rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[37] ——, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays–part II: Markovian rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, 1987.

[38] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17, no. 2, pp. 122–142, 1996.

[39] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled markov chains," *SIAM Journal on Control and Optimization*, vol. 35, no. 3, pp. 715–743, 1997.

[40] A. Shapiro, "Semi-infinite programming, duality, discretization and optimality conditions," *Optimization*, vol. 58, no. 2, pp. 133–161, 2009.

[41] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

[42] R. Agrawal, "Sample mean based index policies with $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, pp. 1054–1078, 1995.

[43] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011, pp. 359–376.

[44] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[45] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, "Kullback–Leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, vol. 41, no. 3, pp. 1516–1541, 2013.

[46] O.-A. Maillard, R. Munos, and G. Stoltz, "A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences," in *Proceedings of the 24th Annual Conference On Learning Theory (COLT)*, 2011, pp. 497–514.

[47] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012, pp. 39.1–39.26.

[48] ——, "Further optimal regret bounds for Thompson sampling," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 99–107.

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

[50] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[51] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," *Advances in Neural Information Processing Systems 19 (NIPS)*, vol. 19, p. 49, 2007.

[52] O.-A. Maillard, Personal Communication.

[53] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for markov decision processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

[54] R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge, "Combinatorial bandits revisited," in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015, pp. 2107–2115.

[55] M. S. Talebi and A. Proutiere, "An optimal algorithm for stochastic matroid bandit optimization," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, 2016, pp. 548–556.

[56] J. Komiyama, J. Honda, and H. Nakagawa, "Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 1152–1161.

[57] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, "Matroid bandits: Fast combinatorial optimization with learning," in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014, pp. 420–429.

[58] B. Kveton, Z. Wen, A. Ashkan, and H. Eydgahi, "Matroid bandits: Practical large-scale combinatorial bandits," in *Proceedings of AAAI Workshop on Sequential Decision-Making with Big Data*, 2014.

[59] R. Watanabe, A. Nakamura, and M. Kudo, "An improved upper bound on the expected regret of UCB-type policies for a matching-selection bandit problem," *Operations Research Letters*, vol. 43, no. 6, pp. 558–563, 2015.

[60] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

[61] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 151–159.

[62] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 535–543.

[63] Z. Wen, B. Kveton, and A. Ashkan, "Efficient learning in large-scale combinatorial semi-bandits," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 1113–1122.

[64] R. Degenne and V. Perchet, "Combinatorial semi-bandit with known covariance," in *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016, pp. 2964–2972.

[65] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback." in *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008, pp. 355–366.

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

[67] ——, "Online learning in opportunistic spectrum access: A restless bandit approach," in *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM)*, 2011, pp. 2462–2470.

[68] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bounds and optimal algorithms," in *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pp. 975–999.

[69] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge University Press, 2004.

[70] R. Combes and A. Proutiere, “Unimodal bandits: Regret lower bounds and optimal algorithms,” *arXiv:1405.5096 [cs.LG]*, 2014. [Online]. Available: http://arxiv.org/abs/1405.5096

[71] J. G. Oxley, *Matroid theory.* Oxford University Press, 2006, vol. 3.

[72] A. György and G. Ottucsák, “Adaptive routing using expert advice,” *The Computer Journal*, vol. 49, no. 2, pp. 180–189, 2006.

[73] O. Brun, L. Wang, and E. Gelenbe, “Big data for autonomic intercontinental overlays,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 575–583, 2016.

[74] J.-Y. Audibert, S. Bubeck, and G. Lugosi, “Regret in online combinatorial optimization,” *Mathematics of Operations Research*, vol. 39, no. 1, pp. 31–45, 2014.

[75] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, “Stochastic online shortest path routing: The value of feedback,” *IEEE Transactions on Automatic Control*, to appear.

[76] A. Gopalan, S. Mannor, and Y. Mansour, “Thompson sampling for complex online problems,” in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 100–108.

[77] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory.* Cambridge University Press, 2007.

[78] K. Liu and Q. Zhao, “Adaptive shortest-path routing under unknown and stochastically varying link states,” in *Proceedings of the 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 232–237.

[79] P. Tehrani and Q. Zhao, “Distributed online learning of the shortest path under unknown random edge weights.” in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3138–3142.

[80] A. Sen and N. Balakrishnan, “Convolution of geometrics and a reliability problem,” *Statistics & Probability Letters*, vol. 43, no. 4, pp. 421–426, 1999.

[81] A. Garivier and E. Moulines, “On upper-confidence bound policies for switching bandit problems,” in *Algorithmic Learning Theory.* Springer, 2011, pp. 174–188.

[82] P. Joulani, A. György, and C. Szepesvári, “Online learning under delayed feedback,” in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 1453–1461.

[83] R. Durrett, *Probability: theory and examples.* Cambridge University Press, 2010.

[84] F. P. Kelly, A. K. Maulloo, and D. K. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[85] S.-Y. Yun and A. Proutiere, “Distributed proportional fair load balancing in heterogenous systems,” in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015, pp. 17–30.

[86] T. Bonald, L. Massoulié, A. Proutière, and J. T. Virtamo, “A queueing analysis of max-min fairness, proportional fairness and balanced fairness,” *Queueing Syst.*, vol. 53, no. 1-2, pp. 65–84, 2006.

[87] D. Tse and P. Viswanath, *Fundamentals of wireless communication.* Cambridge University Press, 2005.

[88] T. Lattimore, K. Crammer, and C. Szepesvári, “Optimal resource allocation with semi-bandit feedback,” in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014, pp. 477–486.

[89] ——, “Linear multi-resource allocation with semi-bandit feedback,” in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015, pp. 964–972.

[90] R. Johari, V. Kamble, and Y. Kanoria, “Know your customer: Multi-armed bandits with capacity constraints,” *arXiv preprint arXiv:1603.04549*, 2016.

[91] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, “Bandits with knapsacks,” in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013, pp. 207–216.

[92] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings, “Knapsack based optimal policies for budget-limited multi-armed bandits,” in *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, 2012, pp. 1134–1140.

[93] R. Combes, C. Jiang, and R. Srikant, “Bandits with budgets: Regret lower bounds and optimal algorithms,” in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015, pp. 245–257.

[94] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, “Fairness in learning: Classic and contextual bandits,” in *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

[95] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," in *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011, pp. 1035–1043.

[96] I. Csiszár and P. Shields, *Information theory and statistics: A tutorial.* Now Publishers Inc, 2004.

[97] A. Schrijver, *A course in combinatorial optimization.* TU Delft, 2000.

[98] F. Topsøe, "Some bounds for the logarithmic function," *Inequality theory and applications*, vol. 4, p. 137, 2006.

[99] M. S. Talebi and O.-A. Maillard, "Variance-aware regret bounds for undiscounted reinforcement learning in mdps," *Submitted*, 2017.

[100] A. Tewari and P. L. Bartlett, "Optimistic linear programming gives logarithmic regret for irreducible MDPs," in *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008, pp. 1505–1512.

[101] P. L. Bartlett and A. Tewari, "REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009, pp. 35–42.

[102] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Advances in Neural Information Processing Systems 22 (NIPS)*, 2009, pp. 89–96.

[103] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 1184–1194.

[104] O.-A. Maillard, T. A. Mann, and S. Mannor, "How hard is my MDP? "the distribution-norm to the rescue"," in *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014, pp. 1835–1843.

[105] R. Ortner, "Online regret bounds for markov decision processes with deterministic transitions," in *Algorithmic Learning Theory.* Springer, 2008, pp. 123–137.

[106] ——, "Online regret bounds for markov decision processes with deterministic transitions," *Theoretical Computer Science*, vol. 411, no. 29-30, pp. 2684–2695, 2010.

[107] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown markov decision processes: A thompson sampling approach," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 1333–1342.

[108] R. Ortner and D. Ryabko, "Online regret bounds for undiscounted continuous reinforcement learning," in *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012, pp. 1763–1771.

[109] R. Ortner, "Adaptive aggregation for reinforcement learning in average reward markov decision processes," *Annals of Operations Research*, vol. 208, no. 1, pp. 321–336, 2013.

[110] K. Lakshmanan, R. Ortner, and D. Ryabko, "Improved regret bounds for undiscounted continuous reinforcement learning." in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 524–532.

[111] I. Osband, D. Russo, and B. Van Roy, "(more) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013, pp. 3003–3011.

[112] M. Gheshlaghi Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 263–272.

[113] C. Dann, T. Lattimore, and E. Brunskill, "Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 5711–5721.

[114] S. Filippi, "Stratégies optimistes en apprentissage par renforcement," Ph.D. dissertation, Ecole nationale supérieure des telecommunications-ENST, 2010.

[115] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

[116] M. Streeter, D. Golovin, and A. Krause, "Online learning of assignments," in *Advances in Neural Information Processing Systems 22 (NIPS)*, 2009, pp. 1794–1802.

[117] Z. Abbassi, V. S. Mirrokni, and M. Thakur, "Diversity maximization under matroid constraints," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013, pp. 32–40.

[118] A. Clark, L. Bushnell, and R. Poovendran, "On leader selection for performance and controllability in multi-agent systems," in *Proceedings of the 51st Annual Conference on Decision and Control (CDC)*, 2012, pp. 86–93.

[119] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003, pp. 137–146.

[120] T. Lin, J. Li, and W. Chen, “Stochastic online greedy learning with semi-bandit feedbacks,” in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015, pp. 352–360.

[121] Y. Yue and C. Guestrin, “Linear submodular bandits and their application to diversified retrieval,” in *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011, pp. 2483–2491.

[122] S. Kullback, *Information theory and statistics.* Courier Corporation, 1968.

[123] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[124] A. Garivier, P. Ménard, and G. Stoltz, “Explore first, exploit next: The true shape of regret in bandit problems,” *arXiv preprint arXiv:1602.07182*, 2016.

[125] A. Garivier and F. Leonardi, “Context tree selection: A unifying view,” *Stochastic Processes and their Applications*, vol. 121, no. 11, pp. 2488–2506, 2011.

[126] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[127] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, “Inequalities for the L1 deviation of the empirical distribution,” *Hewlett-Packard Labs, Technical Report*, 2003.