# Online Combinatorial Optimization under Bandit Feedback 

MOHAMMAD SADEGH TALEBI MAZRAEH SHAHI

Licentiate Thesis
Stockholm, Sweden 2016

KTH Royal Institute of Technology<br>School of Electrical Engineering Department of Automatic Control

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie licentiatesexamen i elektro och systemteknik fredagen den 5 februari 2016 klockan 10.00 i E3, Kungliga Tekniska högskolan, Osquarsbacke 14, Stockholm.
© Mohammad Sadegh Talebi Mazraeh Shahi, February 2016
Tryck: Universitetsservice US AB


#### Abstract

Multi-Armed Bandits (MAB) constitute the most fundamental model for sequential decision making problems with an exploration vs. exploitation trade-off. In such problems, the decision maker selects an arm in each round and observes a realization of the corresponding unknown reward distribution. Each decision is based on past decisions and observed rewards. The objective is to maximize the expected cumulative reward over some time horizon by balancing exploitation (arms with higher observed rewards should be selected often) and exploration (all arms should be explored to learn their average rewards). Equivalently, the performance of a decision rule or algorithm can be measured through its expected regret, defined as the gap between the expected reward achieved by the algorithm and that achieved by an oracle algorithm always selecting the best arm.

This thesis investigates stochastic and adversarial combinatorial MAB problems, where each arm is a collection of several basic actions taken from a set of $d$ elements, in a way that the set of arms has a certain combinatorial structure. Examples of such sets include the set of fixed-size subsets, matchings, spanning trees, paths, etc. These problems are specific forms of online linear optimization, where the decision space is a subset of $d$-dimensional hypercube. Due to the combinatorial nature, the number of arms generically grows exponentially with $d$. Hence, treating arms as independent and applying classical sequential arm selection policies would yield a prohibitive regret. It may then be crucial to exploit the combinatorial structure of the problem to design efficient arm selection algorithms.

As the first contribution of this thesis, in Chapter 3 we investigate combinatorial MABs in the stochastic setting and with Bernoulli rewards. We derive asymptotic (i.e., when the time horizon grows large) lower bounds on the regret of any algorithm under bandit and semi-bandit feedback. The proposed lower bounds are problem-specific and tight in the sense that there exists an algorithm that achieves these regret bounds. Our derivation leverages some theoretical results in adaptive control of Markov chains. Under semi-bandit feedback, we further discuss the scaling of the proposed lower bound with the dimension of the underlying combinatorial structure. For the case of semi-bandit feedback, we propose ESCB, an algorithm that efficiently exploits the structure of the problem and provide a finite-time analysis of its regret. ESCB has better performance guarantees than existing algorithms, and significantly outperforms these algorithms in practice.

In the fourth chapter, we consider stochastic combinatorial MAB problems where the underlying combinatorial structure is a matroid. Specializing the results of Chapter 3 to matroids, we provide explicit regret lower bounds for this class of problems. For the case of semi-bandit feedback, we propose KL-OSM, a computationally efficient greedy-based algorithm that exploits the matroid structure. Through a finite-time analysis, we prove that the regret upper bound of KL-OSM matches the proposed lower bound, thus making it the first asymptotically optimal algorithm for this class of problems. Numerical experiments validate that KL-OSM outperforms state-of-the-art algorithms in practice, as well.


In the fifth chapter, we investigate the online shortest-path routing problem which is an instance of combinatorial MABs with geometric rewards. 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 (semi-bandit or bandit). For each case, we derive the asymptotic regret lower bound. These bounds help us to understand the performance improvements we can expect when (i) taking routing decisions at each hop rather than at the source only, and (ii) observing per-link delays rather than end-to-end path delays. In particular, we show that (i) is of no use while (ii) can have a spectacular impact. For source routing under semi-bandit feedback, we then propose two algorithms with a trade-off between computational complexity and performance. The regret upper bounds of these algorithms improve over those of the existing algorithms, and they significantly outperform state-of-the-art algorithms in numerical experiments.

Finally, we discuss combinatorial MABs in the adversarial setting and under bandit feedback. We concentrate on the case where arms have the same number of basic actions but are otherwise arbitrary. We propose Combexp, an algorithm that has the same regret scaling as state-of-the-art algorithms. Furthermore, we show that CombEXP admits lower computational complexity for some combinatorial problems.

## Acknowledgements

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 your continuous support, inspiration, and motivation, I would not have come this far. Thank you for teaching me to think clearly and independently, and to present concisely. I am also very grateful to my co-advisor, Prof. Mikael Johansson, for fruitful discussions and invaluable advices during my research.

I am deeply grateful to Dr. Richard Combes, whose mentoring and collaboration proved invaluable during my research. I have the good fortune to be able to work with you and I owe you a lot. I am also very grateful to Dr. Marc Lelarge for his kind helps and collaborations.

I would like to thank my co-author Dr. Zhenhua Zou, with whom I had fruitful discussions and collaborations. My appreciation also goes to all the present and past colleagues in the Department of Automatic Control, and in particular to Stefan, Themis, Euhanna, Hamid, Burak, and Hossein for uncountable discussions we had on various technical topics. I acknowledge Arda for his patience and valuable helps in programming and Demia for her help in proofreading this thesis. I owe special thanks to the group of Persians, Euhanna, Afrooz, Hossein, Hamid, Kaveh, and Reza. My thank also goes to my officemates Pedro, Olle, Riccardo, Demia, Patricio, Kuo-Yun, and Antonio, for making the room C:628 a wonderful office. I also want to thank the administrators Karin, Anneli, Hanna, Gerd, and Silvia for all the support to make the department run smoothly.

I am deeply grateful for the support and love I received from my parents, Esmaeel and Batool, and my sisters, Saeedeh, Zohreh, Azadeh, and Fatemeh. I am also thankful to my parents-in-law for their support and encouragements.

Finally, I must thank my love and wonderful wife, Zahra. I am deeply indebted to you for your love, support, and understanding. Thank you for making me happy to be with you!
M. Sadegh Talebi

Stockholm, January 2016.

## Contents

Contents ..... vi
1 Introduction ..... 1
1.1 Problem Statement and Objectives ..... 2
1.1.1 Online Combinatorial Optimization under Bandit Feedback ..... 2
1.1.2 Objectives ..... 3
1.2 Motivating Examples ..... 4
1.2.1 Dynamic Spectrum Access ..... 4
1.2.2 Shortest-Path Routing ..... 7
1.3 Thesis Outline and Contributions ..... 8
2 Background ..... 11
2.1 Stochastic MAB ..... 11
2.1.1 Lower Bounds on the Regret ..... 12
2.1.2 Algorithms ..... 15
2.2 Adversarial MAB ..... 18
2.A Proof of Theorem 12.2 ..... 19
3 Stochastic Combinatorial MABs: Bernoulli Rewards ..... 21
3.1 Contributions and Related Work ..... 21
3.1.1 Related Work ..... 22
3.2 Model and Objectives ..... 23
3.3 Regret Lower Bounds ..... 24
3.3.1 Semi-bandit Feedback ..... 24
3.3.2 Bandit Feedback ..... 26
3.4 Algorithms ..... 28
3.4.1 Indexes ..... 28
3.4.2 Index Computation ..... 29
3.4.3 The ESCB Algorithm ..... 30
3.4.4 Epoch-ESCB: An Algorithm with Lower Computational Com- plexity ..... 30
3.5 Numerical Experiments ..... 31
3.6 Summary ..... 34
3.A Proof of Theorem 3.1 ..... 35
3.B Proof of Theorem 13.2 ..... 36
3.C Proof of Corollary 3.1 ..... 39
3.D Proof of Theorem 3.3 ..... 39
3.E Proof of Theorem 3.4 ..... 40
3.F Proof of Theorem 3.6 ..... 41
3.G Proof of Theorem 3.5 ..... 44
4 Stochastic Matroid MABs ..... 47
4.1 Motivation and Contributions ..... 48
4.1.1 Contributions ..... 48
4.1.2 Related Work ..... 49
4.2 Matroid Structure ..... 49
4.2.1 Weighted Matroids ..... 51
4.3 Model and Objectives ..... 52
4.4 Regret Lower Bound ..... 52
4.4.1 Semi-Bandit Feedback ..... 53
4.4.2 Bandit Feedback ..... 54
4.5 The KL-0SM Algorithm ..... 54
4.5.1 Implementation ..... 55
4.6 Numerical Experiments ..... 55
4.7 Summary ..... 57
4.A Proof of Theorem 4.2 ..... 58
4.B Proof of Theorem 4.3 ..... 60
4.C Proof of Theorem 4.4 ..... 61
4.D Proof of Supporting Lemmas ..... 65
5 Stochastic Combinatorial MABs: Geometric Rewards ..... 67
5.1 Contributions and Related Work ..... 67
5.1.1 Related Work ..... 69
5.2 Model and Objectives ..... 70
5.2.1 Network Model ..... 70
5.2.2 Objectives and Feedback ..... 70
5.3 Regret Lower Bounds ..... 71
5.3.1 Source Routing with Semi-Bandit (Per-Link) Feedback ..... 72
5.3.2 Source Routing with Bandit Feedback ..... 73
5.3.3 Hop-by-hop Routing ..... 74
5.3.4 Line Networks ..... 75
5.4 Algorithms ..... 75
5.4.1 Path Indexes ..... 76
5.4.2 The GeoCombUCB Algorithm ..... 77
5.4.3 Improved Regret Bound for The KL-SR Algorithm ..... 77
5.4.4 Implementation ..... 78
5.5 Numerical Experiments ..... 81
5.6 Summary ..... 81
5.A Proof of Theorem 15.2 ..... 83
5.B Proof of Theorem 15.4 ..... 83
5.C Proof of Theorem 15.5 ..... 84
5.D Proof of Theorem 15.6 ..... 88
5.E Proof of Proposition 5.1 ..... 92
5.F Regret Upper Bound of The CUCB Algorithm ..... 92
6 Adversarial Combinatorial MABs ..... 95
6.1 Contributions and Related Work ..... 95
6.1.1 Related Work ..... 96
6.2 Model and Objectives ..... 98
6.2.1 The OSMD Algorithm ..... 98
6.3 The CombEXP Algorithm ..... 100
6.3.1 Examples ..... 102
6.3.2 Implementation ..... 103
6.4 Summary ..... 105
6.A Proof of Theorem 16.1 ..... 105
6.B Proof of Proposition 6.1 ..... 109
6.C Proof of Theorem 16.2 ..... 111
7 Conclusions and Future Work ..... 113
7.1 Conclusions ..... 113
7.2 Future Work ..... 114
A Concentration Inequalities ..... 117
B Properties of the KL-Divergence ..... 121
B. 1 Definition ..... 121
B. 2 Properties ..... 122
B.2.1 Non-negativity ..... 122
B.2.2 Convexity ..... 122
B.2.3 Chain Rule ..... 123
B. 3 The KL-Divergence between Two Bernoulli Distributions ..... 123
Bibliography ..... 125

## Chapter 1

## Introduction

This thesis investigates online combinatorial optimization problems in stochastic and adversarial settings under bandit feedback. The term bandit feedback signifies the relation of this class of problems to the multi-armed bandit (henceforth, MAB) problems, which constitute an important class of sequential optimization problems. MAB problems were introduced in the seminal paper by Robbins [1] in 1952 to study the sequential design of experiments ${ }^{1}$. In the classical MAB problem, a decision maker has to repeatedly select an arm from a finite set of arms. After selecting an arm, she observes a realization of the corresponding unknown reward distribution. Each decision is based on past decisions and observed rewards. The objective is to maximize the expected cumulative reward over some time horizon by balancing 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 performance of the decision maker (or algorithm) can be measured through its expected regret, defined as the gap between the expected reward achieved by the algorithm and that achieved by an oracle algorithm always selecting the best arm.

This setup, often referred to as classical stochastic MAB problem in the literature, was studied by Lai and Robbins [3], where they proved a lower bound scaling as $\Omega(\log (T))$ on the regret after $T$ rounds. The constant in the lower bound scales as $K$, the number of arms. This lower bound indeed provides a fundamental performance limit that no policy can beat. Lai and Robbins also constructed policies for certain distributions and showed that they asymptotically achieve the lower bound. Namely, these policies have regret upper bounds growing at most as $\mathcal{O}(\log (T))$ where the constant is the same as in the lower bound.

Four decades after the seminal work of Robbins [1, Auer et al. [4] introduced the adversarial (non-stochastic) MAB problem, in which no statistical assumption on the generation of the rewards is made. Indeed the reward sequences are arbitrary, as if they were generated by an adversary. Auer et al. established a lower bound growing at least as $\Omega(\sqrt{T})$ on the regret in which the constant scales as $\sqrt{K}$. Later,

[^0]it was shown in [5] that a regret upper bound scaling as $\sqrt{T K \log (K)}$ is achievable in this case. Namely, the regret upper bound of the proposed algorithm in [5] matches the lower bound of [4] up to a logarithmic factor.

Since its introduction, the MAB framework has served as the best model for studying the tradeoff between exploration and exploitation in sequential decision making in both adversarial and stochastic settings. MAB problems have found applications in many fields as diverse as sequential clinical trials [6], communication systems [7, 8], economics [9, 10], recommendation systems [11, 12], learning to rank [13], to name a few.

### 1.1 Problem Statement and Objectives

### 1.1.1 Online Combinatorial Optimization under Bandit Feedback

The focus of this thesis is on online combinatorial problems with linear objectives. Such combinatorial problems are specific forms of online linear optimization as studied in, e.g., 14, 15, 16, 17, where the decision space is a subset of $d$-dimensional hypercube $\{0,1\}^{d}$. In the literature, this problem is also known as combinatorial bandit or combinatorial $M A B^{2}$. In the adversarial setting, it is sometimes referred to as the combinatorial prediction game.

The setup may be concretely described as follows: The set of arms $\mathcal{M}$ is an arbitrary subset of $\{0,1\}^{d}$, such that each of its elements $M$ is a subset of at most $m$ basic actions taken from $[d]=\{1, \ldots, d\}$. Arm $M$ is identified with a binary column vector $\left(M_{1}, \ldots, M_{d}\right)^{\top}$. In each round $n \geq 1$, a decision maker selects an $\operatorname{arm} M \in \mathcal{M}$ and receives a reward $M^{\top} X(n)=\sum_{i=1}^{d} M_{i} X_{i}(n)$. The reward vector $X(n) \in \mathbb{R}_{+}^{d}$ is unknown. After selecting an arm $M$ 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 $i \in\{1, \ldots, d\}$, the component $X_{i}(n)$ of the reward vector is revealed if and only if $M_{i}=1$.
(ii) Bandit feedback under which only the reward $M^{\top} X(n)$ is revealed.

Based on the feedback received up to round $n-1$, the decision maker selects an arm for the next round $n$, and she aims to maximize her cumulative expected reward over a given time horizon consisting of $T$ rounds.

We consider two instances of combinatorial bandit problems, depending on how the sequence of reward vectors is generated. In the stochastic setting, for all $i \in$ $\{1, \ldots, d\},\left(X_{i}(n)\right)_{n \geq 1}$ are i.i.d. with unknown distribution. The reward sequences may be arbitrarily correlated across basic actions. In the adversarial setting, the

[^1]sequence of vectors $X(n)$ is arbitrarily selected from $[0,1]^{d}$ by an adversary at the beginning of the experiment ${ }^{4}$.

The objective is to identify a policy, amongst all feasible policies, maximizing the cumulative expected reward over $T$ rounds. The expectation is here taken with respect to randomness in the rewards (in the stochastic setting) and possible randomization in the policy. Equivalently, we aim at designing a policy that minimizes regret, where the regret of policy $\pi$ is defined by:

$$
R^{\pi}(T)=\max _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{n=1}^{T} X^{M}(n)\right]-\mathbb{E}\left[\sum_{n=1}^{T} X^{M^{\pi}(n)}(n)\right]
$$

The notion of regret quantifies the performance loss due to the need for learning the average rewards of the various arms.

### 1.1.2 Objectives

In combinatorial MAB problems, one could apply classical sequential arm selection policies, developed in e.g. [19, 5], as if arms would yield independent rewards. Such policies would have a regret asymptotically scaling as $|\mathcal{M}| \log (T)$ and $\sqrt{|\mathcal{M}| T}$ in the stochastic and adversarial settings, respectively. However, since the number of arms $|\mathcal{M}|$ grows exponentially with $d$, the number of basic actions, treating arms as independent would lead to a prohibitive regret. In contrast to classical MAB studied by Lai and Robbins [3] 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 which have a regret scaling as $C \log (T)$ where $C$ is much smaller than $|\mathcal{M}|$. Similarly, in the adversarial setting, exploiting the underlying combinatorial structure allows us to design policies with a regret scaling as $\sqrt{C T}$ with $C$ being much smaller than $|\mathcal{M}|$.

The sought objectives in this thesis may be formalized as follows:

- Stochastic setting: We seek two primary objectives in the stochastic settings. Firstly, we would like to study the asymptotic (namely when $T$ grows large) regret lower bounds for policies with bandit and semi-bandit feedbacks. Such lower bounds provide fundamental performance limits that no policy can beat. Correlations significantly complicate the derivation and the expression of the regret lower bounds. To derive such bounds, we use the techniques presented in Graves and Lai [20] to study the adaptive control of Markov chains with unknown transition probabilities. Study of such lower bounds is of great importance as they provide insights into the design of arm selection policies

[^2]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.

- Adversarial setting: Our focus in the adversarial setting is on the case where all arms consist of the same number $m$ of basic actions in the sense that $\|M\|_{1}=m, \forall M \in \mathcal{M}$. The set of arms is otherwise arbitrary. For this case, lower bounds on the regret have been established for both bandit and semi-bandit feedbacks by Audibert et al. [21. The state-of-the-art policies, however, either achieve a sub-optimal regret or are complicated to implement. Precisely speaking, the best existing algorithms for the case of bandit feedback suffer from both drawbacks: Firstly, their regret upper bounds are off the lower bound by a factor of $\sqrt{m}$. Secondly, they are complicated to implement and may suffer from precision issues which may in turn result in a cumulative time complexity that is super-linear in $T$. Our aim is to propose arm selection policies with reduced computational complexity while attaining at most the same regret as that of state-of-the-art policies.


### 1.2 Motivating Examples

Combinatorial MAB problems 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.

### 1.2.1 Dynamic Spectrum Access

As the first motivating example, we consider a dynamic spectrum access scenario as studied in [22]. 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 $i \in[L]=\{1, \ldots, L\}$. Each link can use one of the $K$ available radio channels indexed by $j \in[K]$. Interference is represented as an interference graph $G=(V, E)^{5}$ where vertices are links and edges indicate interference among links. More precisely, we have $\left(i, i^{\prime}\right) \in E$ if links $i$ and $i^{\prime}$ interfere, i.e., these links cannot be simultaneously active. A spectrum allocation is represented as an allocation matrix $M \in\{0,1\}^{M \times K}$, where $M_{i j}=1$ if and only if user $i$ transmitter uses channel $j$. Allocation $M$ is feasible if (i) for all $i$,

[^3]the corresponding transmitter uses at most one channel, i.e., $\sum_{j \in[K]} M_{i j} \leq 1$; (ii) two interfering links cannot be active on the same channel, i.e., for all $i, i^{\prime} \in[L]$, $\left(i, i^{\prime}\right) \in E$ implies for all $j \in[K], M_{i j} M_{i^{\prime} j}=0^{6}$. Let $\mathcal{M}$ be the set of all feasible allocation matrices. In the following we denote by $\mathcal{F}=\left\{\mathcal{F}_{\ell}, \ell \in[f]\right\}$ 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 $\mathcal{F}_{\ell}$. An example of an interference graph along with a feasible allocation is shown in Figure 1.1


Figure 1.1: Spectrum allocation in a wireless system with 5 links and 3 channels: (a) interference graph (b) an example of a feasible allocation.

We consider a time slotted system, where the duration of a slot corresponds to the transmission of a single packet. We denote by $X_{i j}(n)$ 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 $X_{i j}(n)$ 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 $\mathbb{E}\left[X_{i j}(n)\right]=\theta_{i j}$, and is supposed to be unknown initially. $X_{i j}(n)$ is a Bernoulli random variable of mean $\theta_{i j}$. The stochastic setting models scenarios where the radio channel conditions are stationary. In the adversarial setting, $X_{i j}(n) \in[0,1]$ 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.

[^4]If the radio conditions on each (user, channel) pair were known, the problem would reduce to the following combinatorial optimization problem:

$$
\begin{align*}
\max _{M \in \mathcal{M}} & \sum_{i \in[L], j \in[K]} X_{i j} M_{i j}  \tag{1.1}\\
\text { subject to: } & \sum_{j \in[K]} M_{i j} \leq 1, \quad \forall i \in[L], \\
& \sum_{i \in[L]} F_{\ell i} M_{i j} \leq 1, \quad \forall \ell \in[f], j \in[K], \\
& M_{i j} \in\{0,1\}, \quad \forall i \in[L], j \in[K] . \tag{1.2}
\end{align*}
$$

Problem (1.1) 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 [23].

In practice, the radio conditions on the various channels are not known a priori, and they evolve over time in an unpredictable manner. We model our sequential spectrum allocation problem as a combinatorial MAB problem. The objective is to identify a policy maximizing over a finite time horizon $T$ the expected number of packets successfully transmitted. Equivalently, we aim at designing a sequential channel allocation policy that minimizes the regret of policy $\pi$. Let $X^{M}(n)$ denote the total number of packets successfully transmitted during slot $n$ under allocation $M \in \mathcal{M}$, i.e.,

$$
X^{M}(n)=\sum_{i \in[L]} \sum_{j \in[K]} M_{i j} X_{i j}(n)
$$

Then the regret is defined as

$$
R^{\pi}(T)=\max _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{t=1}^{T} X^{M}(n)\right]-\mathbb{E}\left[\sum_{t=1}^{T} X^{M^{\pi}(n)}(n)\right]
$$

where $M^{\pi}(n)$ denotes the allocation selected under policy $\pi$ in time slot $n$. The expectation is here taken with respect to the possible randomness in the stochastic rewards (in the stochastic setting) and in the probabilistic successively selected channel allocations. The notion of regret indeed quantifies the performance loss due to the need for learning radio channel conditions. Spectrum sharing problems similar to this have been recently investigated in [8, 7, 24, [25, 26].

### 1.2.2 Shortest-Path Routing

Shortest-path routing is amongst the first instances of combinatorial MAB problems considered in the literature, e.g., in [27, 28, 29]. As our second example, we consider shortest-path routing in the stochastic setting as studied in [30, 31].

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 $i$ at time $t$ is successful. $\left(X_{i}(t)\right)_{t \geq 1}$ is a sequence of i.i.d. Bernoulli variables with initially unknown mean $\theta_{i}$. 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 $\theta_{\min }=\min _{i \in E} \theta_{i}>0$, and let $\theta=\left(\theta_{i}, i \in E\right)$ 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 $\mathcal{M} \subseteq\{0,1\}^{d}$ the set of loop-free paths from $u$ to $v$ in $G$, where each path $M \in \mathcal{M}$ is a $d$ dimensional binary vector; for any $i \in E, M_{i}=1$ if and only if $i$ belongs to $M$. Hence, for any $M \in \mathcal{M}$, the length of path $M$ is $\|M\|_{1}=\sum_{i \in E} M_{i}$.

We assume that the source is fully backlogged (i.e., it always has packets to send), and that the parameter $\theta$ is initially unknown. Packets are sent successively from $u$ to $v$ over various paths to estimate $\theta$, and in turn to learn the path $M^{\star}$, 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 end-to-end delays) before sending the next packet. If $\theta$ were known, one would choose the path $M^{\star}$ given by

$$
\begin{equation*}
M^{\star} \in \arg \min _{M \in \mathcal{M}} \sum_{i \in E} \frac{M_{i}}{\theta_{i}} \tag{1.3}
\end{equation*}
$$

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 path delay), we consider two different types of online routing policies: (i) Source routing with end-toend (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. Let $D^{M}(n)$ denote the end-to-end delay of the $n$-th packet if it is sent over path $M$. The goal is to design online routing policies that minimize the regret up to the $N$-th packet defined as:

$$
R^{\pi}(N)=\mathbb{E}\left[\sum_{n=1}^{N} D^{M^{\pi}(n)}(n)\right]-\min _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{n=1}^{N} D^{M}(n)\right]
$$

| Chapter | Combinatorial Structure | Reward |
| :---: | :---: | :---: |
| Chapter | 3 | Generic |
| Chapter | Bernoulli |  |
| Chapter | Matroid | Bernoulli |
| Chapter | Gen | Generic (with fixed cardinality) | Adversarial | Adveric |
| :--- |

Table 1.1: A summary of the chapters in this thesis
where $M^{\pi}(n)$ is the path chosen by policy $\pi$ for the transmission of the $n$-th packet. Here, the expectation is taken with respect to the random link transmission results and possible randomization in the policy $\pi$. The regret quantifies the performance loss due to the need to explore sub-optimal paths to learn the path with minimum delay.

### 1.3 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. Table 1.1 summarizes the organization of the chapters.

## Chapter 2: Background

This chapter is devoted to the overview of some results and algorithms on classical MAB problems in both stochastic and adversarial settings. In the stochastic setting, we study regret lower bounds and provide an overview of well-known algorithms for stochastic bandits. Similarly, in the adversarial setting we overview important algorithms along with their performance guarantees.

## Chapter 3: Stochastic Combinatorial MABs: Bernoulli Rewards

In chapter 3 we consider stochastic combinatorial MAB with Bernoulli rewards. We derive tight and problem-specific lower bounds on the regret of any admissible algorithm under bandit and semi-bandit feedbacks. Our derivation leverages the theory of optimal control of Markov chains with unknown transition probabilities. These constitute the first lower bounds proposed for generic combinatorial MABs in the literature. In the case of semi-bandit feedback, we further discuss 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 better performance guarantee than existing algorithms and significantly outperforms these algorithms in practice as confirmed by our numerical experiments. We then present the EPOCH-ESCB algorithm to alleviate the computational complexity of ESCB.

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.


## Chapter 4: Stochastic Matroid MABs

In Chapter 4 we consider stochastic combinatorial MABs where the underlying combinatorial structure is a matroid. We provide asymptotic regret lower bounds, which are specialization of the lower bounds in Chapter 3 to the case of matroids. In contrast to the lower bounds of Chapter 3, the results of Chapter 4 are explicit. To the best of our knowledge, this is the first explicit performance limit for the problem considered. In the case of semi-bandit feedback, we propose KL-OSM, which is a computationally efficient algorithm working based on the greedy algorithm. Hence, KL-OSM is capable of exploiting the matroid structure. Through a finite-time analysis, we prove that the regret upper bound of KL-OSM matches the proposed lower bound, and hence it is asymptotically optimal. This algorithm constitutes the first optimal algorithm for matroid bandits in the literature.

The chapter is based on the following work:

- M. Sadegh Talebi and Alexandre Proutiere, "An Optimal Algorithm for Stochastic Matroid Bandit Optimization," submitted to International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2016.


## Chapter 5: Stochastic Combinatorial MABs: Geometric Rewards

In Chapter 5, we study combinatorial MAB problems with geometrically distributed rewards, which is motivated by the shortest-path routing as discussed in Section 1.2.2. 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 that has to be satisfied by any online routing policy. 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 [30]. 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.

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," submitted to IEEE Transaction on Automatic Control.


## Chapter 6: Adversarial Combinatorial MABs

Chapter 6 investigates adversarial combinatorial MAB problems under bandit feedback. The focus of that chapter is on the case where all arms consist of the same number of basic actions. We propose the CombEXP algorithm, an OSMD-type algorithm, and provide a finite-time analysis of its regret. As our analysis shows, CombEXP has the same regret scaling as state-of-the-art algorithms. Furthermore, we present an analysis of the computational complexity of CombEXP showing that it has lower computational complexity than state-of-the-art algorithms for some problems of interest. The presented computational complexity analysis extends in an straightforward manner to class of OSMD-type algorithms and hence could be of independent interest.

The chapter is based on the following publication:

- Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge, "Combinatorial Bandits Revisited," in Advances in Neural Information Processing Systems 28 (NIPS), 2015.


## Chapter 7: Conclusions and Future Work

This chapter draws some conclusions and provides some directions for the future work.

## Appendices

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

## Chapter 2

## Background

In this section we give an overview of the various results for stochastic and adversarial MABs.

### 2.1 Stochastic MAB

The multi-armed bandit (MAB) problem was introduced in the seminal paper by Robbins [1 to study the sequential design of experiments. The first bandit algorithm, however, dates back to a paper by Thompson [2] in 1933. In this section we give an overview of lower bounds for stochastic MAB. It is followed by an overview of various algorithms developed.

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 $\left(X_{i}(n)\right)_{n \geq 1}$. For any $i$, the sequence of rewards $\left(X_{i}(n)\right)_{n \geq 1}$ is drawn i.i.d. from a parametric distribution $\nu\left(\theta_{i}\right)$, 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 $\pi$ is a sequence of random variables $I^{\pi}(1), I^{\pi}(2), \ldots$ all taking values from $[K]$ such that $\left\{I^{\pi}(n)=i\right\} \in \mathcal{F}_{n}$ for all $i \in[K]$ and $n \geq$ 0 . Let $\Pi$ be the set of all feasible policies. The objective is to identify a policy in $\Pi$ maximizing the cumulative expected reward over a finite time horizon $T$. The expectation is here taken with respect to 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:

$$
R^{\pi}(T)=\max _{i \in[K]} \mathbb{E}\left[\sum_{n=1}^{T} X_{i}(n)\right]-\mathbb{E}\left[\sum_{n=1}^{T} X_{I^{\pi}(n)}(n)\right]
$$

For any $i \in[K]$ introduce $\Delta_{i}=\max _{j \in[K]} \mu\left(\theta_{j}\right)-\mu\left(\theta_{i}\right)$. Moreover, let $t_{i}^{\pi}(n)$ denote the number of times arm $i$ is selected up to round $n$ under policy $\pi$, i.e.,
$t_{i}^{\pi}(n)=\sum_{s=1}^{n} \mathbb{1}\left\{I^{\pi}(s)=i\right\}$. Then, the regret $R^{\pi}(T)$ can be decomposed as follows:

$$
R^{\pi}(T)=\sum_{i \in[K]} \Delta_{i} \mathbb{E}\left[t_{i}^{\pi}(T)\right]
$$

### 2.1.1 Lower Bounds on the Regret

In this subsection we present lower bounds on the regret for stochastic MAB problems. The first lower bound was proposed by Lai and Robbins in their seminal paper [3]. 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 define the concept of uniformly good rules.

Definition 2.1 ([3). A policy is uniformly good if for all $\theta \in \Theta$, the regret satisfies $R^{\pi}(T)=o\left(T^{\alpha}\right)$ for any $\alpha>0$.

Let $i^{\star}$ be an optimal arm, namely $\mu\left(\theta_{i^{\star}}\right)=\max _{i \in[K]} \mu\left(\theta_{i}\right)$. 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 $\pi$ satisfies:

$$
\liminf _{T \rightarrow \infty} \frac{\mathbb{E}\left[t_{i}^{\pi}(T)\right]}{\log (T)} \geq \frac{1}{\operatorname{KL}\left(\nu\left(\theta_{i}\right), \nu\left(\theta_{i^{\star}}\right)\right)},
$$

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

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq \sum_{i: \Delta_{i}>0} \frac{\Delta_{i}}{\operatorname{KL}\left(\theta_{i}, \theta_{i^{\star}}\right)}
$$

This result indeed defines the asymptotic optimality criterion: an algorithm $\pi$ is said to be asymptotically optimal if the following holds for any $\theta \in \Theta$ :

$$
\limsup _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \leq \sum_{i: \Delta_{i}>0} \frac{\Delta_{i}}{\operatorname{KL}\left(\theta_{i}, \theta_{i^{\star}}\right)}
$$

Lai and Robbins' lower bound was generalized in subsequent works, e.g., [33, 34, 35]. Extension to multiple play, i.e. the case where multiple arms are pulled at the same time, is addressed by Anantharam et al. [33, 34. Let us assume that arms are enumerated such that $\mu\left(\theta_{1}\right) \geq \mu\left(\theta_{2}\right) \geq \cdots>\mu\left(\theta_{m+1}\right) \geq \cdots \geq \mu\left(\theta_{K}\right)$ and that at each round, $m$ arms are played. Anantharam et al. [33] show that the regret of any uniformly good rule $\pi$ satisfies:

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \leq \sum_{i=m+1}^{K} \frac{\mu\left(\theta_{m}\right)-\mu\left(\theta_{i}\right)}{\mathrm{KL}\left(\theta_{i}, \theta_{m}\right)} .
$$

[^5]Furthermore, [34 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 [35] to distributions that rely on multiple parameters, and by Graves and Lai [20] to a more general framework of adaptive control of Markov chains.

## Regret Lower Bound for Adaptive Control of Markov Chains

Graves and Lai [20] 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 compact metric space. The framework of Graves and Lai generalizes those of Lai and Robbins [3], Anantharam et al. [34], and Burnetas and Katehakis [35], and plays a pivotal role in 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 $\left(X_{n}\right)_{n \geq 0}$ on a finite state space $\mathcal{S}$ with a control set $U$. The transition probabilities given control $u \in U$ are parameterized by $\theta$ taking values in a compact metric space $\Theta$ : The probability to move from state $x$ to state $y$ given the control $u$ and the parameter $\theta$ is $p(x, y ; u, \theta)$. The parameter $\theta$ is not known. The decision maker is provided with a finite set of stationary control laws $G=\left\{g_{1}, \ldots, g_{K}\right\}$, 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 $\pi_{\theta}^{g}$. 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 $\theta$, an optimal control law is optimal if its expected reward is

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

Letting

$$
J(\theta)=\left\{j \in[K]: \mu_{\theta}\left(g_{j}\right)=\mu_{\theta}^{\star}\right\}
$$

the set of optimal stationary control laws is $\left\{g_{j}, j \in J(\theta)\right\}$. Of course when the optimal stationary law is unique, $J(\theta)$ is a singleton. 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 $T$. The performance of a decision scheme can be quantified through the notion of regret which compares the expected reward to
that obtained by always applying the optimal control law:

$$
R(T)=T \mu_{\theta}^{\star}-\mathbb{E}\left[\sum_{n=1}^{T} r\left(X_{n}, u_{n}\right)\right]=\sum_{g \in G: \mu_{\theta}(g)<\mu_{\theta}^{\star}}\left(\mu_{\theta}^{\star}-\mu_{\theta}(g)\right) \mathbb{E}\left[t_{g}(T)\right]
$$

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 KullbackLeibler 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 $\Theta$ into $L$ subsets $\left\{\Theta_{j}, j \in[L]\right\}$, such that for any $\theta \in \Theta_{j}, g_{j}$ is the stationary control law, i.e.,

$$
\Theta_{j}=\left\{\theta \in \Theta: \mu_{\theta}\left(g_{j}\right)=\max _{g \in G} \mu_{\theta}(g)\right\}
$$

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

$$
B(\theta)=\left\{\lambda \in \Theta: \lambda \notin \bigcup_{j \in J(\theta)} \Theta_{j} \text { and } I^{g_{j}}(\theta, \lambda)=0, \forall j \in J(\theta)\right\}
$$

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

An adaptive control rule $\phi$ is a sequence of random variables $I(1), I(2), \ldots$ that belong to $G$ such that $\{I(n)=g\} \in \mathcal{F}_{n}$ for all $g \in G$ and $n \geq 0$. An adaptive control rule $\phi$ is said to be uniformly good if for all $\theta \in \Theta$, we have that $R(T)=\mathcal{O}(\log (T))$ 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 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 ([20, Theorem 1]). For every $\theta \in \Theta$ and for any uniformly good algorithm $\phi$,

$$
\liminf _{T \rightarrow \infty} \frac{R(T)}{\log (T)} \geq c(\theta)
$$

where

$$
c(\theta)=\inf \left\{\sum_{j \notin J(\theta)} x_{j}\left(\mu^{\star}-\mu\left(g_{j}\right)\right): x_{j} \in \mathbb{R}_{+}, \inf _{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} x_{j} I^{g_{j}}(\theta, \lambda) \geq 1\right\}
$$

```
Algorithm 2.1 Index policy using index \(\xi\)
    for \(n \geq 1\) do
        Select \(\operatorname{arm} I(n) \in \arg \max _{i \in[K]} \xi_{i}(n)\).
        Observe the rewards, and update \(t_{i}(n)\) and \(\hat{\theta}_{i}(n), \forall i \in[K]\).
    end for
```

We remark that $c(\theta)$ is the optimal value of a linear semi-infinite program (LSIP) [36]. 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 $M$ should be of the order of $x_{M}^{\star} \log (T)$ where $x_{M}^{\star}$ 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 require to solve, in each round, a linear semi-infinite program which might be computationally expensive.

### 2.1.2 Algorithms

In this section we present 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 $\xi$.

An index policy relies on constructing an upper confidence bound for the expected reward of each $\operatorname{arm}^{3}$ in a way that $\mu_{i} \in\left[\hat{\mu}_{i}(n)-\delta_{i}(n), \hat{\mu}_{i}(n)+\delta_{i}(n)\right]$ with high probability, where $\hat{\mu}_{i}$ is the empirical average reward of arm $i$. 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 [37] for the MABs in Bayesian setting. For non-Bayesian stochastic MAB problems, the first index policy was introduced by Lai and Robbins [3]. This policy constitutes the first

[^6]asymptotically optimal algorithm for the classic MAB problem. Lai and Robbins' algorithm was very complicated; hence it motivated developments of simpler index policies; see, e.g., [38, 19, 39, 40, 41. Agrawal [38 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 [19]. It wasn't until the paper by Auer et al. 19] 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

$$
b_{i}(n)=\hat{\mu}_{i}(n)+\sqrt{\frac{\alpha \log (n)}{t_{i}(n)}} .
$$

Originally, Auer et al. chose $\alpha=2$. 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 2.2 ( 19 ). The regret under $\pi=$ UCB1 satisfies

$$
R^{\pi}(T) \leq 6 \sum_{i>1} \frac{\log (T)}{\Delta_{i}}+\frac{K \pi^{2}}{6}+\sum_{i>1} \frac{4}{\Delta_{i}} .
$$

We provide a proof of this result in the appendix. Observe that UCB1 achieves a sub-optimal regret in view of Lai and Robbins' lower bound since $\mathrm{kl}\left(\theta_{i}, \theta_{1}\right)>2 \Delta_{i}^{2}$.

The KL-UCB Algorithm [39]. The KL-UCB algorithm is an optimal algorithm for stochastic MABs with bounded rewards proposed by Garivier and Cappé [39]. KL-UCB relies on the following index:

$$
b_{i}(n)=\max \left\{q \in \Theta: t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), q\right) \leq \log (n)+c \log (\log (n))\right\} .
$$

The following theorem provides the regret bound of KL-UCB.
Theorem 2.3 ([39]). The regret under $\pi=K L-U C B$ satisfies

$$
R^{\pi}(T) \leq(1+\varepsilon) \sum_{i>1} \frac{\Delta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{1}\right)} \log (T)+C_{1} \log (\log (T))+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}
$$

[^7]where $C_{1}$ is a positive constant and where $C_{2}(\varepsilon)$ and $\beta(\varepsilon)$ denote positive functions of $\varepsilon$. Hence,
$$
\limsup _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \leq \sum_{i>1} \frac{\Delta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{1}\right)}
$$

It is noted that the regret upper bound of KL-UCB matches the lower bound of Burnetas and Katehakis 35].

Using Variance Estimates: The UCB-V Algorithm [40]. Incorporating variance estimates into index function allows to have superior algorithms. One such index policy is UCB-Tuned [19], for which no theoretical guarantee is proposed. Audibert et al. [40] proposed UCB-V (UCB 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 $n$ :

$$
V_{i}(n)=\frac{1}{t_{i}(n)} \sum_{n=1}^{t_{i}(n)}\left(X_{i}(n)-\hat{\theta}_{i}(n)\right)^{2}
$$

Let $\sigma_{i}^{2}$ denote the variance of arm $i$. It is shown that UCB-V achieves the following regret upper bound [40]:

$$
R^{\mathrm{UCB}-\mathrm{V}}(T) \leq 10\left(\sum_{i>1} \frac{\sigma_{i}^{2}}{\Delta_{i}}+2\right) \log (T)
$$

## The Thompson Sampling Algorithm

Thompson Sampling was proposed by Thompson [2] in 1933. It wasn't until very recently, however, that its regret analysis was presented by Agrawal and Goyal [42, 43] and Kaufmann et al. [44.

In contrast to previously described index policies, Thompson Sampling belongs to the family of randomized probability matching algorithms and selects an arm based on posterior samples. The underlying idea in Thompson Sampling is to assume a prior distribution on the parameters of the reward distribution of every arm. Then at any time step, Thompson Sampling plays an arm according to its posterior probability of being the best arm. Algorithm 2.2 presents the pseudo-code of Thompson Sampling for the case of Bernoulli distributed rewards, for which the appropriate prior distribution is the Beta distribution (see, e.g., 43] for details).

The first regret analysis for Thompson Sampling was proposed by Agrawal and Goyal [42]. Later, Kaufmann et al. [44] improved this regret analysis and proved

```
Algorithm 2.2 Thompson Sampling
    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 \(\operatorname{Beta}\left(S_{i}+1, F_{i}+1\right)\).
        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 asymptotic optimality of Thompson Sampling for classical stochastic MABs. Optimality of Thompson Sampling was also addressed by Agrawal and Goyal 43 with a different regret analysis. In the following theorem, we provide the regret upper bound for Thompson Sampling with Beta priors.

Theorem 2.4 ([43, Theorem 1]). The regret under Thompson Sampling using Beta Priors satisfies:

$$
R(T) \leq(1+\varepsilon) \sum_{i>1} \frac{\Delta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{1}\right)} \log (T)+C\left(\varepsilon, \theta_{1}, \ldots, \theta_{K}\right),
$$

where $C\left(\varepsilon, \theta_{1}, \ldots, \theta_{K}\right)$ is a problem-dependent constant independent of $T$. In particular, $C\left(\varepsilon, \theta_{1}, \ldots, \theta_{K}\right)=\mathcal{O}\left(K \varepsilon^{-2}\right)$.

### 2.2 Adversarial MAB

In this section we consider another variant of MAB problems called adversarial, also known as non-stochastic, MAB problem introduced by Auer et al. 4. The term non-stochastic comes from the fact the no statistical assumption on the reward sequence is made. Indeed, the rewards are arbitrary as if they were generated by an adversary.

In their paper [4, Auer et al. propose Exp3 (EXPonential-weight algorithm for EXPloration and EXPlotation) for the non-stochastic MAB problem. Algorithm 2.3 presents the pseudo-code of Exp3. The design of EXP3 is indeed inspired by the Hedge algorithm of Freund and Schapire [45], which itself is based on the multiplicative weight algorithm ${ }^{5}$ of [47] and aggregating strategies of Vovk [48].

Auer et al. establish a regret of $\mathcal{O}\left(T^{2 / 3}\right)$ for Exp3 and a lower bound of $\Omega(\sqrt{T})$ for the adversarial MAB problem. It wasn't until the final version of their work in [5] that they show Exp3 achieves $\mathcal{O}(\sqrt{K T \log (K)})$ and hence is optimal up to a logarithmic factor.

[^8]```
Algorithm 2.3 Exp3 [5]
    Initialization: Set \(w_{i}(1)=1\) for \(i=1, \ldots, K\).
    for \(n \geq 1\) do
        For \(i=1, \ldots, K\) set
\[
p_{i}(n)=(1-\gamma) \frac{w_{i}(n)}{\sum_{j=1}^{K} w_{j}(n)}+\frac{\gamma}{K} .
\]
```

Choose arm $I(n)$ from the probability distribution $p(n)$. Play arm $I(n)$ and observe the reward $X_{I(n)}(n)$. Compute the estimate reward vector $\tilde{X}_{i}(n)=\frac{X_{i}(n)}{p_{i}(n)} \mathbb{1}\{I(n)=i\}$ For $i=1, \ldots, K$ set $w_{i}(n+1)=w_{i}(n) e^{\gamma \tilde{X}_{i}(n) / K}$.
end for

The following theorem provides the regret guarantee for the expected regret of Exp3.

Theorem 2.5 ([5] Theorem 3.1]). For any $T \geq 1$, assume that Exp3 is run with parameter $\gamma=\min \left(1, \sqrt{\frac{K \log (K)}{(e-1) T}}\right)$. Then

$$
R(T) \leq 2 \sqrt{e-1} \sqrt{T K \log (K)}
$$

Several variants of the Exp3 algorithm have also been proposed in [5]. One of the variants is Exp3.p, which attains a regret guarantee that holds with highprobability. We note that Theorem 2.5 implies that Exp3 has $\mathcal{O}(\sqrt{K T \log (K)})$ regret. The following theorem from [5] establishes a lower bound on the regret signifying the tightness of the result of Theorem 2.5

Theorem 2.6 ([5, Theorem 5.1]). We have that:

$$
\inf \sup R(T) \geq \frac{1}{20} \sqrt{K T}
$$

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 is at least $\frac{1}{20} \sqrt{K T}$. Therefore, Exp3 achieves a regret which is optimal up to a logarithmic factor. Finally, we mention that Audibert and Bubeck [49] propose an algorithm, called MOSS, whose upper bound scales as $\mathcal{O}(\sqrt{K T})$ which matches the abovementioned lower bound.

## 2.A Proof of Theorem 2.2

Proof. Let $T>0$. First we derive an upper bound for the expected number of plays of any sub-optimal arm $i$ up to round $T$. Fix $i>1$ and consider round $n$ when arm
$i$ is selected. Either the index of the best arm underestimates the average reward of this arm, or the average reward of arm $i$ is badly estimated, or the number of times arm $i$ is played is not sufficient.

For any $n \geq 1$, introduce the following events: $A_{n}=\left\{b_{1}(n)<\mu_{1}\right\}$ and $B_{n, i}=$ $\left\{I(n)=i, \hat{\mu}_{i}(n)-\mu_{i} \geq \Delta_{i} / 2\right\}$. Note that

$$
\{I(n)=i\} \subset A_{n} \cup B_{n, i} \cup\left(\{I(n)=i\} \cap \overline{A_{n}} \cap \overline{B_{n, i}}\right)
$$

Define $\ell_{i}=\frac{6}{\Delta_{i}^{2}} \log (T)$. Next we show that

$$
\{I(n)=i\} \cap \overline{A_{n}} \cap \overline{B_{n, i}} \subseteq\left\{t_{i}(n)<\ell_{i}\right\}
$$

Consider $n$ such that $\overline{A_{n}}$ and $\overline{B_{n, i}}$ occur and arm $i$ is selected. It then follows that

$$
\mu_{i}+\frac{\Delta_{i}}{2}+\sqrt{\frac{3 \log (T)}{2 t_{i}(n)}}>b_{i}(n) \geq b_{1}(n) \geq \mu_{1}
$$

which further implies that $t_{i}(n)<\frac{6 \log (T)}{\Delta_{i}^{2}}$. Hence, $\{I(n)=i\} \subset A_{n} \cup B_{n, i} \cup\left\{t_{i}(n)<\right.$ $\left.\ell_{i}\right\}$, and consequently

$$
\mathbb{E}\left[t_{i}(T)\right]=\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\{I(n)=i\}\right] \leq \ell_{i}+\sum_{n=1}^{T}\left(\mathbb{P}\left[A_{n}\right]+\mathbb{P}\left[B_{n, i}\right]\right)
$$

Note that $\sum_{n \geq 1} \mathbb{P}\left[B_{n, i}\right] \leq 4 / \Delta_{i}^{2}$ on the account of Corollary A. 2 Moreover,

$$
\begin{aligned}
\mathbb{P}\left[A_{n}\right] & =\mathbb{P}\left[\mu_{1}-\hat{\mu}_{1}(n)>\sqrt{1.5 \log (n) / t_{i}(n)}\right] \\
& =\sum_{s=1}^{n} \mathbb{P}\left[\mu_{1}-\hat{\mu}_{1, s}>\sqrt{1.5 \log (n) / s}, t_{i}(n)=s\right] \\
& \leq \sum_{s=1}^{n} \mathbb{P}\left[\mu_{1}-\hat{\mu}_{1, s}>\sqrt{1.5 \log (n) / s}\right] \leq n e^{-3 \log (n)}=\frac{1}{n^{2}}
\end{aligned}
$$

where the last line follows from Hoeffding inequality (Theorem A.2. Here $\hat{\mu}_{1, s}$ denotes the empirical average reward of arm 1 when it is sampled $s$ times. Thus $\sum_{n=1}^{T} \mathbb{P}\left[A_{n}\right] \leq \sum_{n \geq 1} n^{-2}=\pi^{2} / 6$. Putting these together, we obtain

$$
\mathbb{E}\left[t_{i}(T)\right] \leq \frac{6}{\Delta_{i}^{2}} \log (T)+\frac{\pi^{2}}{6}+\frac{4}{\Delta_{i}^{2}}
$$

Finally,

$$
R(T)=\sum_{i>1} \Delta_{i} \mathbb{E}\left[t_{i}(T)\right] \leq \sum_{i>1} \frac{6}{\Delta_{i}} \log (T)+\sum_{i>1} \frac{4}{\Delta_{i}}+\frac{K \pi^{2}}{6}
$$

which concludes the proof.

## Chapter 3

## Stochastic Combinatorial MABs: Bernoulli Rewards

This chapter, which constitutes the core part of this thesis, investigates combinatorial MABs with Bernoulli rewards. We first derive lower bounds on the regret to determine the fundamental performance limit of these problems under bandit and semi-bandit feedback. Our derivation leverages the theory of optimal control of Markov chains as studied by Graves and Lai [20]. In contrast to Lai and Robbins' lower bound, which has a closed-form expression, the derivation and the expression of regret lower bounds for generic combinatorial bandit problems are complicated due to correlation between the rewards of the various actions. We then propose ESCB, an algorithm working under semi-bandit feedback, and provide upper bounds on its regret. This upper bound constitutes the best regret guarantee that has been proposed in the literature for the problem considered.

This chapter is based on the publications [50] and [22]. It is organized as follows: Section 3.1 outlines contributions of the chapter and provides an overview of related works. 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 3.3. 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
in $m$ and $d$ of the lower bounds is unfortunately not explicit. For semi-bandit feedback, we further provide a simplified lower bound (Theorem 3.2) 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 $\mathcal{O}\left(\frac{\sqrt{m} d}{\Delta_{\min }} \log (T)\right)$ (Theorem 3.6, where $\Delta_{\text {min }}$ 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 [39, 19]. 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 semibandit feedback. Most of these contributions focused on specific combinatorial structures, e.g., fixed-size subsets [33, 51], matroids [52, 53], or permutations [26, [54. Generic combinatorial problems were investigated in [55], [56], and 57]. Gai et al. [55] propose LLR, a variant of the UCB algorithm which assigns index to basic actions. Gai et al. 55] establish a loose regret bound of $\mathcal{O}\left(\frac{m^{3} d \Delta_{\max }}{\Delta_{\min }} \log (T)\right)$ for LLR, where $\Delta_{\text {max }}$ denotes the expected reward difference between the best and the worst arm. Chen et al. 56] 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, CUCB, is a variant of UCB that assigns index to basic actions. For linear combinatorial problems, CUCB achieves a regret $\mathcal{O}\left(\frac{m^{2} d}{\Delta_{\text {min }}} \log (T)\right)$, which improves over the regret bound of LLR by a factor of $m \Delta_{\max } / \Delta_{\min }$. For linear combinatorial problems, Kveton et al. [57] improve the regret upper bound of $\operatorname{CUCB}^{1}$ to $\mathcal{O}\left(\frac{m d}{\Delta_{\min }} \log (T)\right)$. However, the constant in the leading term of this regret bound is fairly large. They also derive another regret bound scaling as $\mathcal{O}\left(\frac{m^{4 / 3} d}{\Delta_{\text {min }}} \log (T)\right)$ with better constants ${ }^{2}$. Our algorithms improve over LLR and CUCB by a multiplicative factor of (at least) $\sqrt{m}$. The performance guarantees of these algorithms are presented in Table 3.1.

In spite of specific lower bound examples, regret lower bounds, which 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 $\mathcal{O}\left(\frac{m d}{\Delta_{\text {min }}} \log (T)\right)$ cannot be beaten, Kveton et al. 57] artificially create an instance of the 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 $\mathcal{O}\left(\frac{\sqrt{m} d}{\Delta_{\min }} \log (T)\right)$ since we assume independence among the rewards of various basic actions.

[^9]| Algorithm | Regret |
| :---: | :---: |
| LLR 55] | $\mathcal{O}\left(\frac{m^{3} d \Delta_{\max }}{\Delta_{\min }^{2}} \log (T)\right)$ |
| CUCB [56] | $\mathcal{O}\left(\frac{m^{2} d}{\Delta_{\min }} \log (T)\right)$ |
| CombUCB1 (CUCB) 57] | $\mathcal{O}\left(\frac{m^{4 / 3} d}{\Delta_{\min }} \log (T)\right)$ |
| CombUCB1 (CUCB) [57] | $\mathcal{O}\left(\frac{m d}{\Delta_{\min }} \log (T)\right)$ |
| ESCB (Theorem 3.6 | $\mathcal{O}\left(\frac{\sqrt{m d}}{\Delta_{\min }} \log (T)\right)$ |

Table 3.1: Regret upper bounds for stochastic combinatorial bandits under semibandit feedback.

Linear combinatorial MABs may be viewed as linear optimization over a polyhedral set. Dani et al. [17] consider stochastic linear optimization over compact and convex sets under bandit feedback. They propose algorithms with regret bounds which scale as $\mathcal{O}\left(\log ^{3}(T)\right)$ and hold with high probability. We stress, however, that Dani et al. [17] assume that $\mathcal{M}$ is full rank and therefore, their algorithms are not applicable to all classes of $\mathcal{M}$.

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. [58, most of existing works focused on specific problems, e.g., fixed-size subsets [34] and permutations [59, 25].

### 3.2 Model and Objectives

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

The set of arms $\mathcal{M}$ is an arbitrary subset of $\{0,1\}^{d}$, such that each of its elements $M$ has at most $m$ basic actions. Arm $M$ is identified with a binary column vector $\left(M_{1}, \ldots, M_{d}\right)^{\top}$, and we have $\|M\|_{1} \leq m, \forall M \in \mathcal{M}$. At the beginning of each round $n$, an algorithm or policy $\pi$, selects an $\operatorname{arm} M^{\pi}(n) \in \mathcal{M}$ based on the arms chosen in previous rounds and their observed rewards. The reward of arm $M^{\pi}(n)$ selected in round $n$ is $X^{M^{\pi}(n)}(n)=\sum_{i \in[d]} M_{i}^{\pi}(n) X_{i}(n)=M^{\pi}(n)^{\top} X(n)$.

We consider both semi-bandit and bandit feedbacks. Under semi-bandit feedback and policy $\pi$, at the end of round $n$, the outcome of basic actions $X_{i}(n)$ for all $i \in M^{\pi}(n)$ are revealed to the decision maker, whereas under bandit feedback, $M^{\pi}(n)^{\top} X(n)$ only can be observed. Let $\Pi_{\mathrm{s}}$ and $\Pi_{\mathrm{b}}$ be respectively the set of all feasible policies with semi-bandit and bandit feedback. The objective is to identify a policy in $\Pi_{\mathrm{s}}$ and $\Pi_{\mathrm{b}}$ maximizing the cumulative expected reward over a finite time
horizon $T$. The expectation is here taken with respect to 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 $\pi$ is defined by:

$$
R^{\pi}(T)=\max _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{n=1}^{T} X^{M}(n)\right]-\mathbb{E}\left[\sum_{n=1}^{T} X^{M^{\pi}(n)}(n)\right]
$$

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

### 3.3 Regret Lower Bounds

### 3.3.1 Semi-bandit Feedback

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

$$
B_{\mathrm{s}}(\theta)=\left\{\lambda \in \Theta: \lambda_{i}=\theta_{i}, \forall i \in M^{\star}(\theta), \mu^{\star}(\lambda)>\mu^{\star}(\theta)\right\} .
$$

Let $\operatorname{kl}(u, v)$ be the Kullback-Leibler divergence between Bernoulli distributions of respective means $u$ and $v$, i.e., $\mathrm{kl}(u, v)=u \log (u / v)+(1-u) \log ((1-u) /(1-v))$. We derive a regret lower bound valid for any uniformly good algorithm in $\Pi_{\mathrm{s}}$. The proof of this result relies on a general result on controlled Markov chains due to Graves and Lai [20], which was described in Chapter 2

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

$$
\begin{equation*}
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{s}}(\theta) \tag{3.1}
\end{equation*}
$$

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

$$
\begin{align*}
\inf _{x \geq 0} & \sum_{M \in \mathcal{M}} x_{M} \Delta_{M}  \tag{3.2}\\
\text { subject to: } & \sum_{i=1}^{d} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{M \in \mathcal{M}} x_{M} M \geq 1, \quad \forall \lambda \in B_{\mathrm{s}}(\theta) . \tag{3.3}
\end{align*}
$$

Observe first that optimization problem (3.10) is a semi-infinite linear program [36] which can be solved for any fixed $\theta$, 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 (3.10 has the following interpretation: Assume that (3.10) has a unique solution $x^{\star}$. Then any uniformly good algorithm
must select action $M$ at least $x_{M}^{\star} \log (T)$ times over the $T$ first rounds. From [20], we know that there exists an algorithm which is asymptotically optimal, so that 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 3.10 $T$ 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. Further remark that if $\mathcal{M}$ is the set of singletons (classical bandit), Theorem 3.1 reduces to the Lai and Robbins' bound [3] and if $\mathcal{M}$ is the set of fixed-size subsets (bandit with multiple plays), Theorem 3.1 reduces to the lower bound derived in [33]. Finally, 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.

## A Simplified Lower Bound

We now 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 $\theta$, we say that a set $\mathcal{H} \subset \mathcal{M} \backslash M^{\star}$ has property $P(\theta)$ iff, for $\operatorname{all}\left(M, M^{\prime}\right) \in \mathcal{H}^{2}, M \neq M^{\prime}$ we have $\left(M \backslash M^{\star}\right) \cap\left(M^{\prime} \backslash M^{\star}\right)=\emptyset$.

Theorem 3.2. Let $\mathcal{H}$ be a maximal (inclusion-wise) subset of $\mathcal{M}$ satisfying the property $P(\theta)$. Define $\beta(\theta)=\min _{M \neq M^{\star}} \frac{\Delta^{M}}{\left|M \backslash M^{\star}\right|}$. Then:

$$
c_{\mathrm{s}}(\theta) \geq \sum_{M \in \mathcal{H}} \frac{\beta(\theta)}{\max _{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \theta_{j}\right)} .
$$

Corollary 3.1. Let $\theta \in[a, 1]^{d}$ for some constant $a>0$ and $\mathcal{M}$ be such that each arm $M \in \mathcal{M}, M \neq M^{\star}$ has at most $k$ sub-optimal basic actions. Then: $c_{\mathrm{s}}(\theta)=\Omega(|\mathcal{H}| / k)$.

Theorem 3.2 provides explicit regret lower bound and Corollary 3.1 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, $|\mathcal{H}|$ is proportional to $d-m$, which implies that in these cases one cannot obtain a regret smaller than $\mathcal{O}\left((d-m) \Delta_{\min }^{-1} \log (T)\right)$. 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 $\mathcal{O}\left(d \sqrt{m} \Delta_{\min }^{-1} \log (T)\right)$, which is acceptable since typically, $\sqrt{m}$ is much smaller than $d$.

Next we examine Theorem 3.2 for some concrete classes of $\mathcal{M}$.

Matchings. In the first example, we assume that $\mathcal{M}$ is the set of perfect matchings in the complete bipartite graph $\mathcal{K}_{m, m}$, with $|\mathcal{M}|=m$ ! and $d=m^{2}$. A maximal


Figure 3.1: Matchings in $\mathcal{K}_{4,4}$ : (a) The optimal matching $M^{\star}$, (b)-(g) Elements of $\mathcal{H}$.

(a) $M^{\star}$

(b)

(c)

(d)

(e)

(f)

(g)

Figure 3.2: Spanning trees in $\mathcal{K}_{5}$ : (a) The optimal spanning tree $M^{\star}$, (b)-(g) Elements of $\mathcal{H}$.
subset $\mathcal{H}$ of $\mathcal{M}$ satisfying property $P(\theta)$ can be constructed by adding all matchings that differ from the optimal matching by only two edges, see Figure 3.1 for illustration in the case of $m=4$. Here $|\mathcal{H}|=\binom{m}{2}$ and thus, $|\mathcal{H}|$ scales as $d-m$.

Spanning trees. Consider the problem of finding the minimum spanning tree in a complete graph $\mathcal{K}_{N}$. This corresponds to letting $\mathcal{M}$ be the set of all spanning trees in $\mathcal{K}_{N}$, where $|\mathcal{M}|=N^{N-2}$ (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 $\mathcal{M}$ satisfying property $P(\theta)$ can be constructed by composing all spanning trees that differ from the optimal tree by one edge only, see Figure 3.2. In this case, $\mathcal{H}$ has $d-m=\frac{(N-1)(N-2)}{2}$ elements.

### 3.3.2 Bandit Feedback

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

$$
\begin{equation*}
\psi_{\theta}^{M}(k)=\sum_{A \subseteq M,|A|=k} \prod_{i \in A} \theta_{i} \prod_{i \in M \backslash A}\left(1-\theta_{i}\right) . \tag{3.4}
\end{equation*}
$$

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

$$
\begin{equation*}
I^{M}(\theta, \lambda)=\sum_{k=0}^{m} \psi_{\theta}^{M}(k) \log \frac{\psi_{\theta}^{M}(k)}{\psi_{\lambda}^{M}(k)} . \tag{3.5}
\end{equation*}
$$

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

$$
B_{\mathrm{b}}(\theta)=\left\{\lambda \in \Theta:\left\{\lambda_{i}, i \in M^{\star}\right\}=\left\{\theta_{i}, i \in M^{\star}\right\}, \mu^{\star}(\lambda)>\mu^{\star}(\theta)\right\} .
$$

It is important to observe that in the definition of $B_{\mathrm{b}}(\theta)$, the equality $\left\{\lambda_{i}, i \in\right.$ $\left.M^{\star}\right\}=\left\{\theta_{i}, i \in M^{\star}\right\}$ is a set equality, i.e., order does not matter (e.g., if $M^{\star}=$ $(0,1,1,0)^{\top}$, the equality means that either $\theta_{2}=\lambda_{2} ; \theta_{3}=\lambda_{3}$ or $\left.\theta_{2}=\lambda_{3} ; \theta_{3}=\lambda_{2}\right)$. 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 parameters in the case of bandit feedback contains that of semi-bandit feedback, i.e., $B_{\mathrm{s}}(\theta) \subset B_{\mathrm{b}}(\theta)$.

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 3.3. For all $\theta \in \Theta$, for any uniformly good policy $\pi \in \Pi_{\mathrm{b}}$,

$$
\begin{equation*}
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{b}}(\theta) \tag{3.6}
\end{equation*}
$$

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

$$
\begin{align*}
\inf _{x \geq 0} & \sum_{M \in \mathcal{M}} x_{M} \Delta_{M}  \tag{3.7}\\
\text { subject to: } & \sum_{M \in \mathcal{M}} x_{M} I^{M}(\theta, \lambda) \geq 1, \quad \forall \lambda \in B_{\mathrm{b}}(\theta) . \tag{3.8}
\end{align*}
$$

The variables $x_{M}, M \in \mathcal{M}$ 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 3.1. 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., $\Pi_{\mathrm{b}} \subset \Pi_{\mathrm{s}}$ ).

### 3.4 Algorithms

Next we present ESCB, an algorithm for the case of semi-bandit feedback that relies on arm indexes as in UCB1 [19] and KL-UCB [39].

### 3.4.1 Indexes

ESCB relies on arm indexes. In general, an index of arm $M$ in round $n$, say $\xi_{M}(n)$, should be defined so that $\xi_{M}(n) \geq M^{\top} \theta$ with high probability. Then as for UCB1 and KL-UCB, applying the principle of "optimism in face of uncertainty", a natural way to devise algorithms based on indexes is to select in each round the arm with the highest index. Under a given algorithm, at time $n$, we define $t_{i}(n)=\sum_{s=1}^{n} M_{i}(s)$ the number of times basic action $i$ has been sampled. The empirical mean reward of action $i$ is then defined as $\hat{\theta}_{i}(n)=\left(1 / t_{i}(n)\right) \sum_{s=1}^{n} X_{i}(s) M_{i}(s)$ if $t_{i}(n)>0$ and $\hat{\theta}_{i}(n)=0$, otherwise. We define the corresponding vectors $t(n)=\left(t_{i}(n)\right)_{i \in[d]}$ and $\hat{\theta}(n)=\left(\hat{\theta}_{i}(n)\right)_{i \in[d]}$.

The indexes we propose are functions of the round $n$ and of $\hat{\theta}(n)$. Our first index for $\operatorname{arm} M$, referred to as $b_{M}(n, \hat{\theta}(n))$ or $b_{M}(n)$ for short, is an extension of KL-UCB index. Let $f(n)=\log (n)+4 m \log (\log (n)) . b_{M}(n, \hat{\theta}(n))$ is the optimal value of the following optimization problem:

$$
\begin{align*}
\max _{q \in \Theta} & M^{\top} q  \tag{3.9}\\
\text { subject to: } & \sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), q_{i}\right) \leq f(n) .
\end{align*}
$$

As we show later, $b_{M}(n)$ may be computed efficiently using a line search procedure similar to that used to determine KL-UCB index.

Our second index $c_{M}(n, \hat{\theta}(n))$ or $c_{M}(n)$ for short is a generalization of the UCB1 and UCB-Tuned indexes:

$$
c_{M}(n)=M^{\top} \hat{\theta}(n)+\sqrt{\frac{f(n)}{2}\left(\sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}\right)}
$$

Note that, in the classical bandit problems with independent arms, i.e., when $m=1$, $b_{M}$ reduces to the KL-UCB index (which yields an asymptotically optimal algorithm) and $c_{M}$ reduces to the UCB-Tuned index [19]. The next theorem provides generic properties of our indexes. An important consequence of these properties is that the expected number of times where $b_{M^{\star}}(n, \hat{\theta}(n))$ or $c_{M^{\star}}(n, \hat{\theta}(n))$ underestimate $\mu^{\star}(\theta)$ is finite, as stated in the corollary below.

Theorem 3.4. (i) For all $n \geq 1, M \in \mathcal{M}$ and $\tau \in[0,1]^{d}$, we have $b_{M}(n, \tau) \leq$ $c_{M}(n, \tau)$. (ii) There exists $C_{m}>0$ depending on $m$ only such that, for all $M \in \mathcal{M}$ and $n \geq 2$ :

$$
\mathbb{P}\left[b_{M}(n, \hat{\theta}(n)) \leq M^{\top} \theta\right] \leq C_{m} n^{-1}(\log (n))^{-2}
$$

Corollary 3.2. We have:

$$
\sum_{n \geq 1} \mathbb{P}\left[b_{M^{\star}}(n, \hat{\theta}(n)) \leq \mu^{\star}\right] \leq 1+C_{m} \sum_{n \geq 2} n^{-1}(\log (n))^{-2}<\infty .
$$

Statement (i) in the above theorem is obtained combining Pinsker and CauchySchwarz inequalities. The proof of statement (ii) is based on a concentration inequality on sums of empirical KL-divergences proven in [60] (see Appendix A). It enables to control the fluctuations of multivariate empirical distributions for exponential families. It should also be observed that indexes $b_{M}(n)$ and $c_{M}(n)$ 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. $b_{M}(n)$ can also be extended to the case where reward distributions are not Bernoulli but lie in an exponential family (e.g. Gaussian, Exponential, Gamma, etc.), replacing kl by a suitably chosen divergence measure. A close look at $c_{M}(n)$ reveals that the indexes proposed in [56], [57], and [55] are too conservative to be optimal in our setting: there the "confidence bonus" $\sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}$ was replaced by (at least) $m \sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}$. Note that [56], [57] assumed that the various basic actions are arbitrarily correlated, while we assume independence among basic actions.

### 3.4.2 Index Computation

While the index $c_{M}$ is explicit, $b_{M}$ is defined as the optimal value of an optimization problem. We show that it may be computed by a simple line search. For $\lambda \geq 0$, $w \in[0,1]$ and $v \in \mathbb{N}$, define:

$$
g(\lambda, w, v)=\frac{1}{2}\left(1-\lambda v+\sqrt{(1-\lambda v)^{2}+4 w v \lambda}\right) .
$$

Fix $n, M, \hat{\theta}(n)$ and $t(n)$. Define $I=\left\{i: M_{i}=1, \hat{\theta}_{i}(n) \neq 1\right\}$, and for $\lambda>0$, define:

$$
F(\lambda)=\sum_{i \in I} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), g\left(\lambda, \hat{\theta}_{i}(n), t_{i}(n)\right)\right) .
$$

Theorem 3.5. If $I=\emptyset, b_{M}(n)=\|M\|_{1}$. Otherwise:
(i) $\lambda \mapsto F(\lambda)$ is strictly increasing, and $F\left(\mathbb{R}^{+}\right)=\mathbb{R}^{+}$.
(ii) Define $\lambda^{\star}$ as the unique solution to $F(\lambda)=f(n)$. Then

$$
b_{M}(n)=\|M\|_{1}-|I|+\sum_{i \in I} g\left(\lambda^{\star}, \hat{\theta}_{i}(n), t_{i}(n)\right) .
$$

Theorem 3.5 shows that $b_{M}$ can be computed using a line search procedure such as bisection, as this computation amounts to solving the nonlinear equation $F(\lambda)=f(n)$, where $F$ is a strictly increasing function. The proof of Theorem 3.5 follows from KKT conditions and the convexity of the KL-divergence (see Appendix B).

```
Algorithm 3.1 ESCB
    for \(n \geq 1\) do
        Select arm \(M(n) \in \arg \max _{M \in \mathcal{M}} \xi_{M}(n)\).
        Observe the rewards, and update \(t_{i}(n)\) and \(\hat{\theta}_{i}(n), \forall i \in M(n)\).
    end for
```


### 3.4.3 The ESCB Algorithm

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 $\xi_{M}$ : ESCB-1 when $\xi_{M}=b_{M}$ and ESCB-2 if $\xi_{M}=c_{M}$. 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 3.4 (i)). The following theorem provides a finite-time analysis of our ESCB algorithms. The proof of this theorem borrows some ideas from the proof of [57] Theorem 3].

Theorem 3.6. The regret under algorithms $\pi \in\{$ ESCB-1, ESCB-2\} satisfies for any time horizon $T$ :

$$
R^{\pi}(T) \leq \frac{16 d \sqrt{m}}{\Delta_{\min }}(\log (T)+4 m \log (\log (T)))+\frac{4 d m^{3}}{\Delta_{\min }^{2}}+C_{m}^{\prime}
$$

where $C_{m}^{\prime} \geq 0$ does not depend on $\theta$, d, and $T$. As a consequence $R^{\pi}(T)=$ $\mathcal{O}\left(d \sqrt{m} \Delta_{\min }^{-1} \log (T)\right)$ when $T \rightarrow \infty$.

ESCB with time horizon $T$ has a complexity of $\mathcal{O}(|\mathcal{M}| T)$ as neither $b_{M}$ nor $c_{M}$ can be written as $M^{\top} y$ for some vector $y \in \mathbb{R}^{d}$. Assuming that the offline (static) combinatorial problem is solvable in $\mathcal{O}(V(\mathcal{M}))$ time, the complexity of CUCB in 56] and [57] after $T$ rounds is $\mathcal{O}(V(\mathcal{M}) T)$. Thus, if the offline problem is efficiently implementable, i.e., $V(\mathcal{M})=\mathcal{O}(\operatorname{poly}(d))$, CUCB is efficient, whereas ESCB is not since $|\mathcal{M}|$ may generically have exponentially many elements. Next, we provide an extension to ESCB, which we may call Epoch-ESCB, that attains almost the same regret as ESCB while enjoying much lower computational complexity.

### 3.4.4 Epoch-ESCB: An Algorithm with Lower Computational Complexity

Еросн-ESCB algorithm works in epochs of varying lengths. Epoch $k$ comprises rounds $\left\{N_{k}, \ldots, N_{k+1}-1\right\}$, 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., $N_{k+1}=N_{k}+2\left(N_{k}-N_{k-1}\right)$. Otherwise, it will be set to 1 . In contrast to ESCB, Еросн-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 Еросн-ESCB is presented in Algorithm 3.2

```
Algorithm 3.2 Epoch-ESCB
    Initialization: Set \(k=1\) and \(N_{0}=N_{1}=1\).
    for \(n \geq 1\) do
        Compute \(L(n) \in \arg \max _{M \in \mathcal{M} M} M^{\top} \hat{\theta}(n)\).
        if \(n=N_{k}\) then
            Select \(\operatorname{arm} M(n) \in \arg \max _{M \in \mathcal{M}} \xi_{M}(n)\).
            if \(M(n)=L(n)\) then
                Set \(N_{k+1}=N_{k}+2\left(N_{k}-N_{k-1}\right)\).
            else
                Set \(N_{k+1}=N_{k}+1\).
            end if
            Increment \(k\).
        else
            Select \(\operatorname{arm} M(n)=L(n)\).
        end if
        Observe the rewards, and update \(t_{i}(n)\) and \(\hat{\theta}_{i}(n), \forall i \in M(n)\).
    end for
```

We assess the performance of Еросн-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 Eросн-ESCB after $T$ rounds will be $\mathcal{O}(V(\mathcal{M}) T+\log (T)|\mathcal{M}|)$. Compared to CUCB, the complexity is penalized by $|\mathcal{M}| \log (T)$, which may become dominated by the term $V(\mathcal{M}) T$ 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 $\mathcal{M}$. When implementing ESCB, 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 $\mathcal{K}_{5,5}$, which corresponds to $d=5^{2}=25$ and $m=5$. We also set $\theta$ such that $\theta_{i}=a$ if $i \in M^{\star}$, and $\theta_{i}=b$ otherwise, with $0<b<a<1$. In this case the lower bound in Theorem 3.1 becomes $c_{\mathrm{s}}(\theta)=\frac{m(m-1)(a-b)}{2 \mathrm{kl}(b, a)}$.

Figure 3.3 (a)-(b) depicts the regret of various algorithms for the case of $a=0.7$ and $b=0.5$. The curves in Figure 3.3(a) are shown with a $95 \%$ confidence interval. We observe that ESCB-1 has the lowest regret. Moreover, ESCB-2 significantly


Figure 3.3: Regret of various algorithms for matchings with $a=0.7$ and $b=0.5$.
outperforms CUCB and LLR, and its regret is close to that of ESCB-1. Moreover, we observe that the regret of Еросн-ESCB is quite close to that of ESCB-2.

Figures 3.4 (a)-(b) presents the regret of various algorithms for the case of $a=$ 0.95 and $b=0.3$. 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 the most of the basic actions were close to $1 / 2$, 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 $1 / 2$, 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, similar to previous experiment, Еросн-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


Figure 3.4: Regret of various algorithms for matchings with $a=0.95$ and $b=0.3$.
in Figure 3.5 (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 $\log (n) / n$ at round $n$. Since the number of times Epoch-ESCB computes the index $c_{M}$ is equal to the number of epochs, these curves suggest that the computational complexity of index computations in EPосн-ESCB after $n$ rounds scales as $|\mathcal{M}| \log (n)$.

## Experiment 2: Spanning Trees

In the second experiment, we consider spanning trees problem described in Section 3.3.1 for the case of $N=5$. In this case, we have $d=\binom{5}{2}=10, m=4$, and $|\mathcal{M}|=5^{3}=125$. We generate parameter $\theta$ uniformly at random from $[0,1]^{10}$. Figure 3.6 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


Figure 3.5: Number of epochs in Еросн-ESCB vs. time for Experiment 1 and 2 (\%95 confidence interval).
and LLR.

### 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 the 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 Еросн-ESCB which has lower computational complexity than ESCB. The regret analysis of EPосн-ESCB is much more complicated than that of ESCB, and hence is let for future work.


Figure 3.6: Regret of various algorithms for spanning trees with $N=5$ and $\Delta_{\min }=$ 0.54 .

## 3.A Proof of Theorem 3.1

Proof. To derive regret lower bounds, we apply the techniques used by Graves and Lai [20] to investigate efficient adaptive decision rules in controlled Markov chains ${ }^{3}$.

The parameter $\theta$ takes values in $[0,1]^{d}$. The Markov chain has values in $\mathcal{S}=$ $\{0,1\}^{d}$. The set of controls corresponds to the set of feasible actions $\mathcal{M}$, and the set of control laws is also $\mathcal{M}$. These laws are constant, in the sense that the control applied by control law $M \in \mathcal{M}$ does not depend on the state of the Markov chain, and corresponds to selecting action $M$. The transition probabilities are given as follows: for all $x, y \in \mathcal{S}$,

$$
p(x, y ; M, \theta)=p(y ; M, \theta)=\prod_{i \in[d]} p_{i}\left(y_{i} ; M, \theta\right),
$$

[^10]where for all $i \in[d]$, if $M_{i}=0, p_{i}(0 ; M, \theta)=1$, and if $M_{i}=1, p_{i}\left(y_{i} ; M, \theta\right)=$ $\theta_{i}^{y_{i}}\left(1-\theta_{i}\right)^{1-y_{i}}$. Finally, the reward $r(y, M)$ is defined by $r(y, M)=M^{\top} y$. 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 [20].

We can now apply Theorem 1 in [20]. Note that the KL number under action $M$ is

$$
I^{M}(\theta, \lambda)=\sum_{i \in[d]} M_{i} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right)
$$

From [20, Theorem 1], we conclude that for any uniformly good rule $\pi$,

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{s}}(\theta)
$$

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

$$
\begin{array}{cc}
\inf _{x \geq 0} & \sum_{M \neq M^{\star}} x_{M} \Delta_{M}, \\
\text { subject to: } & \inf _{\lambda \in B_{\mathrm{s}}(\theta)} \sum_{Q \neq M^{\star}} x_{Q} I^{Q}(\theta, \lambda) \geq 1 . \tag{3.11}
\end{array}
$$

The result is obtained by observing that $B_{\mathrm{s}}(\theta)=\bigcup_{M \neq M^{\star}} B_{\mathrm{s}, M}(\theta)$, where

$$
B_{\mathrm{s}, M}(\theta)=\left\{\lambda \in \Theta: \lambda_{i}=\theta_{i}, \forall i \in M^{\star}, \mu^{\star}(\theta)<\mu_{M}(\lambda)\right\} .
$$

## 3.B Proof of Theorem 3.2

Proof. The proof proceeds in three steps. In the subsequent analysis, given the optimization problem $P$, we use $\operatorname{val}(\mathrm{P})$ to denote its optimal value.

Step 1. In this step, first we introduce an equivalent formulation for problem (3.10) above by simplifying its constraints. We show that constraint 3.11) is equivalent to:

$$
\inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q \in \mathcal{M}} Q_{i} x_{Q} \geq 1, \quad \forall M \neq M^{\star} .
$$

Observe that:

$$
\sum_{Q \neq M^{\star}} x_{Q} I^{Q}(\theta, \lambda)=\sum_{Q \neq M^{\star}} x_{Q} \sum_{i \in[d]} Q_{i} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right)=\sum_{i \in[d]} \operatorname{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q \neq M^{\star}} Q_{i} x_{Q}
$$

Fix $M \neq M^{\star}$. In view of the definition of $B_{\mathrm{s}, M}(\theta)$, we can find $\lambda \in B_{\mathrm{s}, M}(\theta)$ such that $\lambda_{i}=\theta_{i}, \forall i \in([d] \backslash M) \cup M^{\star}$. Thus, for the r.h.s. of the $M$-th constraint in (3.11), we get:

$$
\begin{aligned}
\inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{Q \neq M^{\star}} x_{Q} I^{Q}(\theta, \lambda) & =\inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{i \in[d]} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q \neq M^{\star}} Q_{i} x_{Q} \\
& =\inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q} Q_{i} x_{Q}
\end{aligned}
$$

and therefore problem 3.10 can be equivalently written as:

$$
\begin{align*}
c_{\mathrm{s}}(\theta) & =\inf _{x \geq 0}  \tag{3.12}\\
& \sum_{M \neq M^{\star}} \Delta_{M} x_{M},  \tag{3.13}\\
\text { subject to: } & \inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q} Q_{i} x_{Q} \geq 1, \quad \forall M \neq M^{\star} .
\end{align*}
$$

Next, we formulate an LP whose value gives a lower bound for $c_{\mathrm{s}}(\theta)$. Define $\hat{\lambda}(M)=\left(\hat{\lambda}_{i}(M), i \in[d]\right)$ with

$$
\hat{\lambda}_{i}(M)= \begin{cases}\frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \theta_{j} & \text { if } i \in M \backslash M^{\star} \\ \theta_{i} & \text { otherwise }\end{cases}
$$

Clearly $\hat{\lambda}(M) \in B_{\mathrm{s}, M}(\theta)$, and therefore:

$$
\inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q} Q_{i} x_{Q} \leq \sum_{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \hat{\lambda}_{i}(M)\right) \sum_{Q} Q_{i} x_{Q}
$$

Then, we can write:

$$
\begin{align*}
c_{\mathrm{s}}(\theta) & \geq \inf _{x \geq 0} \sum_{M \neq M^{\star}} \Delta_{M} x_{M}  \tag{3.14}\\
& \text { subject to: } \sum_{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \hat{\lambda}_{i}(M)\right) \sum_{Q} Q_{i} x_{Q} \geq 1, \quad \forall M \neq M^{\star} . \tag{3.15}
\end{align*}
$$

For any $M \neq M^{\star}$ introduce: $g_{M}=\max _{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \hat{\lambda}_{i}(M)\right)$. Now we form P1 as follows:

$$
\begin{array}{ll}
\text { P1: } \quad \inf _{x \geq 0} \sum_{M \neq M^{\star}} \Delta_{M} x_{M} \\
\text { subject to: } & \sum_{i \in M \backslash M^{\star}} \sum_{Q} Q_{i} x_{Q} \geq \frac{1}{g_{M}}, \quad \forall M \neq M^{\star} . \tag{3.17}
\end{array}
$$

Observe that $c_{\mathrm{s}}(\theta) \geq \operatorname{val}(\mathrm{P} 1)$ since the feasible set of problem $(3.14)$ is contained in that of P1.

Step 2. In this step, we formulate an LP to give a lower bound for val(P1). To $\overline{\text { this end }}$, for any sub-optimal basic action $i \in[d]$, we define $z_{i}=\sum_{M} M_{i} x_{M}$. Further, we let $z=\left(z_{i}, i \in[d]\right)$. Next, we represent the objective of P1 in terms of $z$, and give a lower bound for it as follows:

$$
\begin{aligned}
\sum_{M \neq M^{\star}} \Delta_{M} x_{M} & =\sum_{M \neq M^{\star}} x_{M} \sum_{i \in M \backslash M^{\star}} \frac{\Delta_{M}}{\left|M \backslash M^{\star}\right|} \\
& =\sum_{M \neq M^{\star}} x_{M} \sum_{i \in[d] \backslash M^{\star}} \frac{\Delta_{M}}{\left|M \backslash M^{\star}\right|} M_{i} \\
& \geq \min _{M \neq M^{\star}} \frac{\Delta_{M}}{\left|M \backslash M^{\star}\right|} \cdot \sum_{i \in[d] \backslash M^{\star}} \sum_{M^{\prime} \neq M^{\star}} M_{i}^{\prime} x_{M^{\prime}} \\
& =\min _{M \neq M^{\star}} \frac{\Delta_{M}}{\left|M \backslash M^{\star}\right|} \cdot \sum_{i \in[d] \backslash M^{\star}} z_{i} \\
& =\beta(\theta) \sum_{i \in[d] \backslash M^{\star}} z_{i} .
\end{aligned}
$$

Then, defining

$$
\begin{aligned}
& \text { P2: } \quad \inf _{z \geq 0} \beta(\theta) \sum_{i \in[d] \backslash M^{\star}} z_{i} \\
& \text { subject to: } \sum_{i \in M \backslash M^{\star}} z_{i} \geq \frac{1}{g_{M}}, \quad \forall M \neq M^{\star},
\end{aligned}
$$

yields: $\operatorname{val}(\mathrm{P} 1) \geq \operatorname{val}(\mathrm{P} 2)$.
Step 3. Introduce a set $\mathcal{H}$ satisfying property $P(\theta)$ as stated in Section 4. Now define

$$
\mathcal{Z}=\left\{z \in \mathbb{R}_{+}^{d}: \sum_{i \in M \backslash M^{\star}} z_{i} \geq \frac{1}{g_{M}}, \quad \forall M \in \mathcal{H}\right\}
$$

and

$$
\text { P3: } \inf _{z \in \mathcal{Z}} \beta(\theta) \sum_{i \in[d] \backslash M^{\star}} z_{i} .
$$

Observe that $\operatorname{val}(\mathrm{P} 2) \geq \operatorname{val}(\mathrm{P} 3)$ since the feasible set of P 2 is contained in $\mathcal{Z}$. The definition of $\mathcal{H}$ implies that $\sum_{i \in[d] \backslash M^{\star}} z_{i}=\sum_{M \in \mathcal{H}} \sum_{i \in M \backslash M^{\star}} z_{i}$. It then follows that

$$
\operatorname{val}(\mathrm{P} 3)=\sum_{M \in \mathcal{H}} \frac{\beta(\theta)}{g_{M}}
$$

$$
\begin{aligned}
& \geq \sum_{M \in \mathcal{H}} \frac{\beta(\theta)}{\max _{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \hat{\lambda}_{i}(M)\right)} \\
& =\sum_{M \in \mathcal{H}} \frac{\beta(\theta)}{\max _{i \in M \backslash M^{\star}} \operatorname{kl}\left(\theta_{i}, \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M^{\prime}} \theta_{j}\right)} .
\end{aligned}
$$

The proof is completed by observing that: $c_{\mathrm{s}}(\theta) \geq \operatorname{val}(\mathrm{P} 1) \geq \operatorname{val}(\mathrm{P} 2) \geq \operatorname{val}(\mathrm{P} 3)$.

## 3.C Proof of Corollary 3.1

Proof. Fix $M \neq M^{\star}$. For any $i \in M \backslash M^{\star}$, we have:

$$
\begin{aligned}
\mathrm{kl}\left(\theta_{i}, \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \theta_{j}\right) & \leq \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \mathrm{kl}\left(\theta_{i}, \theta_{j}\right) \quad(\text { By convexity of } \mathrm{kl}) \\
& \leq \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \frac{\left(\theta_{i}-\theta_{j}\right)^{2}}{\theta_{j}\left(1-\theta_{j}\right)} \\
& \leq \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \frac{\left(1-\theta_{j}\right)^{2}}{\theta_{j}\left(1-\theta_{j}\right)} \\
& \leq \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M}\left(\frac{1}{\theta_{j}}-1\right) \\
& \leq \frac{1}{\min _{j \in M^{\star} \backslash M} \theta_{j}}-1 \\
& \leq \frac{1}{a}-1
\end{aligned}
$$

where the second inequality follows from the inequality $\operatorname{kl}(p, q) \leq \frac{(p-q)^{2}}{q(1-q)}$ for all $(p, q) \in[0,1]^{2}$. Moreover, we have that

$$
\beta(\theta)=\min _{M \neq M^{\star}} \frac{\Delta^{M}}{\left|M \backslash M^{\star}\right|} \geq \frac{\Delta_{\min }}{\max _{M}\left|M \backslash M^{\star}\right|}=\frac{\Delta_{\min }}{k}
$$

Applying Theorem 3.2, we get:

$$
c_{\mathrm{s}}(\theta) \geq \sum_{M \in \mathcal{H}} \frac{\beta(\theta)}{\max _{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \frac{1}{\left|M \backslash M^{\star}\right|} \sum_{j \in M^{\star} \backslash M} \theta_{j}\right)} \geq \frac{a \Delta_{\min }}{k(1-a)}|\mathcal{H}|
$$

which gives the required lower bound and completes the proof.

## 3.D Proof of Theorem 3.3

Proof. The parameter $\theta$ takes values in $[0,1]^{d}$. The Markov chain has values in $\mathcal{S}=\{0, \ldots, m\}$. The set of controls corresponds to the set of arms $\mathcal{M}$, and the set
of control laws is also $\mathcal{M}$. The probability that the reward under arm $M$ is equal to $k$ is then $\psi_{\theta}^{M}(k)$ defined in (3.4), and so:

$$
p\left(k^{\prime}, k ; M, \theta\right)=\psi_{\theta}^{M}(k), \quad \forall k, k^{\prime} \in \mathcal{S} .
$$

From [20, Theorem 1], we conclude that for any uniformly good rule $\pi$,

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{b}}(\theta)
$$

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

$$
\begin{align*}
c_{\mathrm{b}}(\theta)= & \inf _{x \geq 0}  \tag{3.18}\\
\text { subject to: } & \sum_{M \neq M^{\star}} x_{M} \Delta_{M},  \tag{3.19}\\
& \sum_{\lambda \in B_{\mathrm{b}}(\theta)} x_{Q \neq M^{\star}} I^{Q}(\theta, \lambda) \geq 1,
\end{align*}
$$

where $I^{Q}(\theta, \lambda)$ is defined in 3.5. This concludes the proof.

## 3.E Proof of Theorem 3.4

## Proof. First statement:

Consider $q \in \Theta$, and apply Cauchy-Schwarz inequality:

$$
\begin{aligned}
M^{\top}(q-\hat{\theta}(n)) & =\sum_{i=1}^{d} \sqrt{t_{i}(n)}\left(q_{i}-\hat{\theta}_{i}(n)\right) \frac{M_{i}}{\sqrt{t_{i}(n)}} \\
& \leq \sqrt{\sum_{i=1}^{d} M_{i} t_{i}(n)\left(q_{i}-\hat{\theta}_{i}(n)\right)^{2}} \sqrt{\sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}} .
\end{aligned}
$$

By Pinsker's inequality, for all $(p, q) \in[0,1]^{2}$ we have $2(p-q)^{2} \leq \mathrm{kl}(p, q)$ so that:

$$
M^{\top}(q-\hat{\theta}(n)) \leq \sqrt{\frac{1}{2} \sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), q_{i}\right)} \sqrt{\sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}} .
$$

Hence, $\sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), q_{i}\right) \leq f(n)$ implies:

$$
M^{\top} q=M^{\top} \hat{\theta}(n)+M^{\top}(q-\hat{\theta}(n)) \leq M^{\top} \hat{\theta}(n)+\sqrt{\frac{f(n)}{2} \sum_{i=1}^{d} \frac{M_{i}}{t_{i}(n)}}=c_{M}(n)
$$

so that, by definition of $b_{M}(n)$, we have $b_{M}(n) \leq c_{M}(n)$.

## Second statement:

If $\sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right) \leq f(n)$, then by definition of $b_{M}(n)$ we have $b_{M}(n) \geq$ $M^{\top} \theta$. Therefore, using Lemma A.4 there exists $C_{m}$ such that for all $n \geq 2$ we have:

$$
\mathbb{P}\left[b_{M}(n)<M^{\top} \theta\right] \leq \mathbb{P}\left[\sum_{i \in M} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f(n)\right] \leq C_{m} n^{-1}(\log (n))^{-2}
$$

which concludes the proof.

## 3.F Proof of Theorem 3.6

Proof. To prove Theorem 3.6, we borrow some ideas from the proof of [57, Theorem 3].

For any $n \in \mathbb{N}, s \in \mathbb{R}^{d}$, and $M \in \mathcal{M}$ define $h_{n, s, M}=\sqrt{\frac{f(n)}{2} \sum_{i=1}^{d} \frac{M_{i}}{s_{i}}}$, and introduce the following events:

$$
\begin{aligned}
G_{n} & =\left\{\sum_{i \in M^{\star}} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f(n)\right\}, \\
H_{i, n} & =\left\{M_{i}(n)=1,\left|\hat{\theta}_{i}(n)-\theta_{i}\right| \geq m^{-1} \Delta_{\min } / 2\right\}, \quad H_{n}=\cup_{i=1}^{d} H_{i, n} \\
F_{n} & =\left\{\Delta_{M(n)} \leq 2 h_{T, t(n), M(n)}\right\} .
\end{aligned}
$$

Then the regret can be bounded as:

$$
\begin{aligned}
R^{\pi}(T) & =\mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)}\right] \leq \mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)}\left(\mathbb{1}\left\{G_{n}\right\}+\mathbb{1}\left\{H_{n}\right\}\right)\right]+\mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{\overline{G_{n}}, \overline{H_{n}}\right\}\right] \\
& \leq m \mathbb{E}\left[\sum_{n=1}^{T}\left(\mathbb{1}\left\{G_{n}\right\}+\mathbb{1}\left\{H_{n}\right\}\right)\right]+\mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{\overline{G_{n}}, \overline{H_{n}}\right\}\right],
\end{aligned}
$$

since $\Delta_{M(n)} \leq m$.
Next we show that for any $n$ such that $M(n) \neq M^{\star}$, it holds that $\overline{G_{n} \cup H_{n}} \subset F_{n}$. Recall that $c_{M}(n) \geq b_{M}(n)$ for any $M$ and $n$ (Theorem 3.4). Moreover, if $\overline{G_{n}}$ holds, we have $\sum_{i \in M^{\star}} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right) \leq f(n)$, which by definition of $b_{M}$ implies: $b_{M^{\star}}(n) \geq M^{\star}{ }^{\top} \theta$. Hence we have:

$$
\begin{aligned}
\mathbb{1}\left\{\overline{G_{n}}, \overline{H_{n}}, M(n) \neq M^{\star}\right\} & =\mathbb{1}\left\{\overline{G_{n}}, \overline{H_{n}}, \xi_{M(n)}(n) \geq \xi_{M^{\star}}(n)\right\} \\
& \leq \mathbb{1}\left\{\overline{H_{n}}, c_{M(n)}(n) \geq M^{\star \top} \theta\right\} \\
& =\mathbb{1}\left\{\overline{H_{n}}, M(n)^{\top} \hat{\theta}(n)+h_{n, t(n), M(n)} \geq M^{\star^{\top}} \theta\right\} \\
& \leq \mathbb{1}\left\{M(n)^{\top} \theta+\Delta_{M(n)} / 2+h_{n, t(n), M(n)} \geq M^{\star^{\top}} \theta\right\} \\
& =\mathbb{1}\left\{2 h_{n, t(n), M(n)} \geq \Delta_{M(n)}\right\} \\
& \leq \mathbb{1}\left\{2 h_{T, t(n), M(n)} \geq \Delta_{M(n)}\right\}
\end{aligned}
$$

$$
=\mathbb{1}\left\{F_{n}\right\}
$$

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

Hence, the regret is upper bounded by:

$$
R^{\pi}(T) \leq m \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{G_{n}\right\}\right]+m \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{H_{n}\right\}\right]+\mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right]
$$

We will prove the following inequalities: (i) $\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{G_{n}\right\}\right] \leq m^{-1} C_{m}^{\prime}$, with $C_{m}^{\prime} \geq 0$ independent of $\theta, d$, and $T$, (ii) $\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{H_{n}\right\}\right] \leq 4 d m^{2} \Delta_{\text {min }}^{-2}$, and (iii) $\mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right] \leq 16 d \sqrt{m} \Delta_{\text {min }}^{-1} f(T)$.

Hence as announced:

$$
R^{\pi}(T) \leq 16 d \sqrt{m} \Delta_{\min }^{-1} f(T)+4 d m^{3} \Delta_{\min }^{-2}+C_{m}^{\prime}
$$

Inequality (i): An application of Lemma A.4 gives

$$
\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{G_{n}\right\}\right] & =\sum_{n=1}^{T} \mathbb{P}\left[\sum_{i \in M^{\star}} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f(n)\right] \\
& \leq 1+\sum_{n \geq 2} C_{m} n^{-1}(\log (n))^{-2} \equiv m^{-1} C_{m}^{\prime}<\infty .
\end{aligned}
$$

Inequality (ii): Fix $i$ and $n$. Define $s=\sum_{n^{\prime}=1}^{n} \mathbb{1}\left\{H_{n^{\prime}, i}\right\}$. Observe that $H_{n^{\prime}, i}$ implies $M_{i}\left(n^{\prime}\right)=1$, hence $t_{i}(n) \geq s$. Therefore, applying [61, Lemma B.1], we have that $\sum_{n=1}^{T} \mathbb{P}\left[H_{n, i}\right] \leq 4 m^{2} \Delta_{\text {min }}^{-2}$. Using the union bound: $\sum_{n=1}^{T} \mathbb{P}\left[H_{n}\right] \leq 4 d m^{2} \Delta_{\min }^{-2}$.

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

$$
\begin{aligned}
& S_{n}=\left\{i \in M(n): t_{i}(n) \leq 4 m f(T) \Delta_{M(n)}^{-2}\right\} \\
& A_{n}=\left\{\left|S_{n}\right| \geq \ell\right\} \\
& B_{n}=\left\{\left|S_{n}\right|<\ell, \quad\left[\exists i \in M(n): t_{i}(n) \leq 4 \ell f(T) \Delta_{M(n)}^{-2}\right]\right\}
\end{aligned}
$$

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

$$
\Delta_{M(n)} \leq 2 h_{T, t(n), M(n)}=2 \sqrt{\frac{f(T)}{2}} \sqrt{\sum_{i \in[d] \backslash S_{n}} \frac{M_{i}(n)}{t_{i}(n)}+\sum_{i \in S_{n}} \frac{M_{i}(n)}{t_{i}(n)}}
$$

$$
\begin{equation*}
<2 \sqrt{\frac{f(T)}{2}} \sqrt{m \frac{\Delta_{M(n)}^{2}}{4 m f(T)}+\left|S_{n}\right| \frac{\Delta_{M(n)}^{2}}{4 \ell f(T)}}<\Delta_{M(n)} \tag{3.20}
\end{equation*}
$$

where the last inequality uses the observation that $\overline{A_{n} \cup B_{n}}$ implies $\left|S_{n}\right|<\ell$. Clearly, 3.20 is a contradiction. Thus $F_{n} \subset\left(A_{n} \cup B_{n}\right)$ and consequently:

$$
\begin{equation*}
\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\} \leq \sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{A_{n}\right\}+\sum_{n=1}^{T} \Delta_{M(n)} \mathbb{1}\left\{B_{n}\right\} . \tag{3.21}
\end{equation*}
$$

To further bound the r.h.s. of the above, we introduce the following events for any $i$ :

$$
\begin{aligned}
& A_{i, n}=A_{n} \cap\left\{i \in M(n), t_{i}(n) \leq 4 m f(T) \Delta_{M(n)}^{-2}\right\}, \\
& B_{i, n}=B_{n} \cap\left\{i \in M(n), t_{i}(n) \leq 4 \ell f(T) \Delta_{M(n)}^{-2}\right\}
\end{aligned}
$$

It is noted that:

$$
\sum_{i \in[d]} \mathbb{1}\left\{A_{i, n}\right\}=\mathbb{1}\left\{A_{n}\right\} \sum_{i \in[d]} \mathbb{1}\left\{i \in S_{n}\right\}=\left|S_{n}\right| \mathbb{1}\left\{A_{n}\right\} \geq \ell \mathbb{1}\left\{A_{n}\right\}
$$

and hence: $\mathbb{1}\left\{A_{n}\right\} \leq \frac{1}{\ell} \sum_{i \in[d]} \mathbb{1}\left\{A_{i, n}\right\}$. Moreover $\mathbb{1}\left\{B_{n}\right\} \leq \sum_{i \in[d]} \mathbb{1}\left\{B_{i, n}\right\}$. 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 (3.21), we have

$$
\begin{aligned}
\sum_{n=1}^{T} \Delta_{M(n)} & \mathbb{1}
\end{aligned} \begin{aligned}
& \left.F_{n}\right\} \leq \sum_{n=1}^{T} \sum_{i=1}^{d} \frac{\Delta_{M(n)}}{\ell} \mathbb{1}\left\{A_{i, n}\right\}+\sum_{n=1}^{T} \sum_{i=1}^{d} \Delta_{M(n)} \mathbb{1}\left\{B_{i, n}\right\} \\
& =\sum_{n=1}^{T} \sum_{i=1}^{d} \frac{\Delta_{M(n)}}{\ell} \mathbb{1}\left\{A_{i, n}, M(n) \neq M^{\star}\right\} \\
& +\sum_{n=1}^{T} \sum_{i=1}^{d} \Delta_{M(n)} \mathbb{1}\left\{B_{i, n}, M(n) \neq M^{\star}\right\} \\
& \leq \sum_{n=1}^{T} \sum_{i=1}^{d} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{\ell} \mathbb{1}_{1}\left\{A_{i, n}, M(n)=k\right\} \\
& +\sum_{n=1}^{T} \sum_{i=1}^{d} \sum_{k \in\left[K_{i}\right]} \Delta^{i, k} \mathbb{1}\left\{B_{i, n}, M(n)=k\right\} \\
& \leq \sum_{i=1}^{d} \sum_{n=1}^{T} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{\ell} \mathbb{1}^{T}\left\{i \in M(n), t_{i}(n) \leq 4 m f(T)\left(\Delta^{i, k}\right)^{-2}, M(n)=k\right\} \\
& +\sum_{i=1}^{d} \sum_{n=1}^{T} \sum_{k \in\left[K_{i}\right]} \Delta^{i, k} \mathbb{1}\left\{i \in M(n), t_{i}(n) \leq 4 \ell f(T)\left(\Delta^{i, k}\right)^{-2}, M(n)=k\right\}
\end{aligned}
$$

$$
\leq \frac{8 d f(T)}{\Delta_{\min }}\left(\frac{m}{\ell}+\ell\right)
$$

where the last inequality follows from Lemma 3.1 which is proven next. The proof is completed by setting $\ell=\sqrt{m}$.

Lemma 3.1. Let $C>0$ be a constant independent of $n$. Then for any $i$ such that $K_{i} \geq 1$ :

$$
\sum_{n=1}^{T} \sum_{k=1}^{K_{i}} \mathbb{1}\left\{i \in M(n), t_{i}(n) \leq C\left(\Delta^{i, k}\right)^{-2}, M(n)=k\right\} \Delta^{i, k} \leq \frac{2 C}{\Delta_{\min }}
$$

Proof. We have:

$$
\begin{aligned}
\sum_{n=1}^{T} & \sum_{k=1}^{K_{i}} \mathbb{1}\left\{i \in M(n), t_{i}(n) \leq C\left(\Delta^{i, k}\right)^{-2}, M(n)=k\right\} \Delta^{i, k} \\
& =\sum_{n=1}^{T} \sum_{k=1}^{K_{i}} \sum_{j=1}^{k} \mathbb{1}\left\{i \in M(n), t_{i}(n) \in\left(C\left(\Delta^{i, j-1}\right)^{-2}, C\left(\Delta^{i, j}\right)^{-2}\right], M(n)=k\right\} \Delta^{i, k} \\
& \leq \sum_{n=1}^{T} \sum_{k=1}^{K_{i}} \sum_{j=1}^{k} \mathbb{1}\left\{i \in M(n), t_{i}(n) \in\left(C\left(\Delta^{i, j-1}\right)^{-2}, C\left(\Delta^{i, j}\right)^{-2}\right], M(n)=k\right\} \Delta^{i, j} \\
& \leq \sum_{n=1}^{T} \sum_{k=1}^{K_{i}} \sum_{j=1}^{K_{i}} \mathbb{1}\left\{i \in M(n), t_{i}(n) \in\left(C\left(\Delta^{i, j-1}\right)^{-2}, C\left(\Delta^{i, j}\right)^{-2}\right], M(n)=k\right\} \Delta^{i, j} \\
& \leq \sum_{n=1}^{T} \sum_{j=1}^{K_{i}} \mathbb{1}\left\{i \in M(n), t_{i}(n) \in\left(C\left(\Delta^{i, j-1}\right)^{-2}, C\left(\Delta^{i, j}\right)^{-2}\right], M(n) \neq M^{\star}\right\} \Delta^{i, j} \\
& \leq \frac{C}{\Delta^{i, 1}}+\sum_{j=2}^{K_{i}} C\left(\left(\Delta^{i, j}\right)^{-2}-\left(\Delta^{i, j-1}\right)^{-2}\right) \Delta^{i, j} \\
& \leq \frac{C}{\Delta^{i, 1}}+\int_{\Delta^{i, K_{i}}}^{\Delta^{i, 2}} C x^{-2} \mathrm{~d} x \leq \frac{2 C}{\Delta^{i, K_{i}}} \leq \frac{2 C}{\Delta_{\min }}
\end{aligned}
$$

which completes the proof.

## 3.G Proof of Theorem 3.5

Proof. We recall the following facts about the KL-divergence kl , for all $p \in[0,1]$ :
(i) $q \mapsto \operatorname{kl}(p, q)$ is strictly convex on $[0,1]$ and attains its minimum at $p$, with $\operatorname{kl}(p, p)=0$.
(ii) Its derivative with respect to the second parameter $q \mapsto \mathrm{kl}^{\prime}(p, q)=\frac{q-p}{q(1-q)}$ is strictly increasing on $(p, 1)$.
(iii) For $p<1$, we have $\mathrm{kl}(p, q) \underset{q \rightarrow 1^{-}}{\rightarrow} \infty$ and $\mathrm{kl}^{\prime}(p, q) \underset{q \rightarrow 1^{-}}{\rightarrow} \infty$.

Consider $M$ and $n$ fixed throughout the proof. Define $I=\left\{i \in M: \hat{\theta}_{i}(n) \neq 1\right\}$. Consider $q^{\star} \in \Theta$ the optimal solution of optimization problem:

$$
\begin{aligned}
\max _{q \in \Theta} & M^{\top} q \\
\text { subject to: } & \sum_{i \in M} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), q_{i}\right) \leq f(n),
\end{aligned}
$$

so that $b_{M}(n)=M^{\top} q^{\star}$. Consider $i \notin M$, then $M^{\top} q$ does not depend on $q_{i}$ and from (i) we get $q_{i}=\hat{\theta}_{i}(n)$. Now consider $i \in M$. From (i) we get that $1 \geq q_{i}^{\star} \geq \hat{\theta}_{i}(n)$. Hence $q_{i}^{\star}=1$ if $\hat{\theta}_{i}(n)=1$. If $I$ is empty, then $q_{i}^{\star}=1$ for all $i \in M$, so that $b_{M}(n)=\|M\|_{1}$.

Consider the case where $I \neq \emptyset$. From (iii) and the fact that $\sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), q_{i}^{\star}\right)<$ $\infty$ we get $\hat{\theta}_{i}(n) \leq q_{i}^{\star}<1$. From the Karush-Kuhn-Tucker (KKT) conditions, there exists $\lambda^{\star}>0$ such that for all $i \in I$ :

$$
1=\lambda^{\star} t_{i}(n) \mathrm{kl}^{\prime}\left(\hat{\theta}_{i}(n), q_{i}^{\star}\right)
$$

For $\lambda>0$ define $\hat{\theta}_{i}(n) \leq \bar{q}_{i}(\lambda)<1$ as a solution to the equation:

$$
1=\lambda t_{i}(n) \operatorname{kl}^{\prime}\left(\hat{\theta}_{i}(n), \bar{q}_{i}(\lambda)\right)
$$

From (i) we have that $\lambda \mapsto \bar{q}_{i}(\lambda)$ is uniquely defined, strictly decreasing, and $\hat{\theta}_{i}(n)<\bar{q}_{i}(\lambda)<1$. From (iii) we get that $\bar{q}_{i}\left(\mathbb{R}^{+}\right)=\left[\hat{\theta}_{i}(n), 1\right]$. Define the function:

$$
F(\lambda)=\sum_{i \in I} t_{i}(n) \operatorname{kl}\left(\hat{\theta}(n), \bar{q}_{i}(\lambda)\right)
$$

From the reasoning below, $F$ is well defined, strictly increasing, and $F\left(\mathbb{R}^{+}\right)=$ $\mathbb{R}^{+}$. Therefore, $\lambda^{\star}$ is the unique solution to $F\left(\lambda^{\star}\right)=f(n)$, and $q_{i}^{\star}=\bar{q}_{i}\left(\lambda^{\star}\right)$. Furthermore, replacing $\mathrm{kl}^{\prime}$ by its expression we obtain the quadratic equation:

$$
\bar{q}_{i}(\lambda)^{2}+\bar{q}_{i}(\lambda)\left(\lambda t_{i}(n)-1\right)-\lambda t_{i}(n) \hat{\theta}_{i}(n)=0 .
$$

Solving for $\bar{q}_{i}(\lambda)$, we obtain that $\bar{q}_{i}(\lambda)=g\left(\lambda, \hat{\theta}_{i}(n), t_{i}(n)\right)$, which concludes the proof.

## Chapter 4

## Stochastic Matroid Bandits

In this chapter we study stochastic combinatorial MAB problems where the underlying combinatorial structure is a matroid. Given a set of basic actions $E$ (called ground set), a matroid is a pair $(E, \mathcal{I})$ with some $\mathcal{I} \subset 2^{E}$ such that $\mathcal{I}$ is an independence system (i.e., it is closed under subset) and satisfies the so-called augmentation property (see Definition 4.1 for a precise definition). This sub-class of combinatorial MABs, often referred to as matroid bandits [52], considers weighted matroids where each element of $E$ is assigned a weight (its average reward). Each arm is then a basis (i.e., an inclusion-wise maximal element of $\mathcal{I}$ ) of the matroid. The weight of various basic actions are fixed and a priori unknown. The decision maker aims at learning the maximum weight basis by sequentially selecting various arms. Hence, at each round she faces a linear optimization problem under a matroid constraint.

Linear optimization over matroid bases is a sub-class of matroid optimization problems. These latter problems have been investigated extensively, e.g. in 62, 63, 64, and are of special interests in the area of combinatorial optimization both theoretically and practically: Firstly, because matroid structures occur naturally in many problems with practical applications. Secondly, optimization over matroids is relatively easy. In particular, linear optimization over matroid bases is proven to be greedily solvable. More precisely, a well-known result in combinatorial optimization states that an independence system is a matroid if and only if the Greedy algorithm, described in Section 4.2 leads to a maximum weight basis; see, e.g., [62].

Matroid theory brings a two-fold advantage in the corresponding bandit optimization problems: Firstly, it is possible to devise computationally efficient algorithms that, in most cases, select arms greedily. Secondly, the corresponding regret analysis is usually more tractable. Despite such advantage, lack of optimal algorithms for matroid bandits in the literature is evident. Here we provide a sequential arm selection algorithm, KL-OSM, and prove its asymptotic optimality.

This chapter is based on the joint work [65]. Here is an organization of this chapter: Section 4.1 discusses the motivation for bandit optimization over matroids through more examples and outlines contributions of the chapter. Section 4.2 gives an overview of matroid definitions and Section 4.3 describes the model and formulates the problem. In Section 4.4 we present regret lower bounds for the case of
bandit and semi-bandit feedbacks. In Section 4.5, we present KL-OSM, our proposed algorithm for matroid bandits under semi-bandit feedback, and provide a finitetime analysis of its regret. Section 4.6 presents our numerical examples. Finally, Section 4.7 summarizes the chapter. All proofs are presented in the appendix.

### 4.1 Motivation and Contributions

Matroid structures occur naturally in many problems with practical applications ranging from bidding in ad exchange [66], product search [67], task assignment in crowdsourcing [68, leader selection in multi-agent systems [69, 70], and a variety of engineering applications. Hence, matroid constraints are quite natural for combinatorial problems that arise in these applications. For example, assume that the elements of ground set $E$ are categorized into $L$ disjoint categories. A natural requirement for some applications is to force to choose at most one element from each category. In the context of product search, each category might be a specific brand, whereas for news aggregation, a category may correspond to a news domain. We are interested in finding a subset $M \subset E$ while maximizing the total reward such that at most one element from each category belongs to $M$. This constraint forms a partition matroid constraint. Another natural type of constraint is to have cardinality constraint on the set $M$, which is related to a uniform matroid. Another notable instance of matroid constraints appears in the problem of finding the minimum-weight spanning tree in a graph, which arises in various engineering disciplines. The selection of leaders in leader-follower multi-agent systems is yet another engineering application in which matroid constraints arise.

### 4.1.1 Contributions

We make the following contributions for matroid bandits:
(a) We derive asymptotic (as the time horizon $T$ grows large) lower bounds on the regret, satisfied by any algorithm (Theorem 4.2 and Theorem 4.3). The proposed lower bounds are tight and problem-dependent. Similarly to the lower bounds in previous chapters, we leverage the theory of optimal control of Markov chains with unknown transition probabilities. However, the lower bound here for the case of semi-bandit feedback is explicit. To the best of our knowledge, this is the first time that such an explicit fundamental performance limit is presented for matroid bandits.
(b) We propose KL-OSM (KL-based Efficient Sampling for Matroid), which is a greedy-based index policy that maintains a KL-UCB index for each basic action. Hence, it is provably computationally efficient assuming access to an independence oracle (see Section 4.2 for a precise definition). Through a finite-time analysis (Theorem 4.4), we show that KL-OSM attains a regret (asymptotically) growing as the proposed lower bound in Theorem 4.2 Hence, it is asymptotically optimal. To our best knowledge, this is the first optimal algorithm for this class of combinatorial

MABs. Moreover, the regret upper bound of KL-OSM (Theorem 4.4) beats that of existing algorithms. Numerical experiments for some specific matroid problems show that KL-OSM significantly outperforms existing algorithms in practice, as well.

### 4.1.2 Related Work

Despite tractability of linear optimization over matroid structures, corresponding combinatorial MABs have not been well addressed so far. The underlying structure in bandits with multiple plays (i.e., when $\mathcal{M}$ is the set of fixed-size subsets) is a uniform matroid, and hence [33, 34, 51] have indeed addressed specific matroid bandits. MABs with generic matroid structures were investigated in 52 , 53 . The proposed algorithm, called OMM, is a UCB-type policy relying on the Greedy algorithm. OMM achieves a regret scaling at most as $\mathcal{O}\left(\frac{d-m}{\Delta_{\min }} \log (T)\right)$. The dependence of this bound on $(d, m)$ is tight and cannot be improved. We remark that all these works addressed problems with semi-bandit feedback. Furthermore, none of these studies provide regret lower bound. Only in [52] through a specific problem instance, the authors show that the regret upper bound of OMM is order-optimal.

### 4.2 Matroid Structure

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, 23].

Definition 4.1. Let $E$ be a finite set and $\mathcal{I} \subset 2^{E}$. The pair $G=(E, \mathcal{I})$ 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 $Y \in \mathcal{I}$.
(iii) If $X, Y \in \mathcal{I}$ with $|X|>|Y|$, then there is some element $\ell \in X \backslash Y$ 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 4.1 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 $G$. In other words, if $X \in \mathcal{I}$ is a basis for $G$, then $X \cup\{\ell\} \notin \mathcal{I}$ for all $\ell \in E \backslash X$.

Proposition 4.1 ([71). Let $G=(E, \mathcal{I})$ be a matroid. Then
(i) all bases of $G$ have the same cardinality,
(ii) for all bases $X, Y$ of $G$, if $\ell \in X \backslash Y$, then there exists $k \in Y \backslash X$ such that $(X \backslash \ell) \cup\{k\}$ is a basis for $G .{ }^{1}$

[^11](iii) for all bases $X, Y$ of $G$, if $\ell \in X \backslash Y$ then there exists $k \in Y \backslash X$ such that $(Y \backslash k) \cup\{\ell\}$ is a basis for $G$.

The common cardinality of bases of $G$ is referred to as rank of $G$. For any $X \in \mathcal{I}$, we let $A(X)$ denote the set of elements of the ground set that can augment $X$ such that the resulting set remains independent:

$$
A(X)=\{\ell: \ell \notin X, X \cup\{\ell\} \in \mathcal{I}\}
$$

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 $U_{m, d}$.

Partition matroid. Let $E$ be a finite set. Assume that $\left\{E_{i}\right\}_{i \in[l]}$ is a partition of $E$, i.e., $E_{i}, i \in[l]$ are disjoint sets and $\cup_{i \in[l]} E_{i}=E$. For some given parameters $k_{1}, \ldots, k_{l}$, define

$$
\mathcal{I}=\left\{X \subseteq E:\left|X \cap E_{i}\right| \leq k_{i}, \forall i \in[l]\right\}
$$

Then $(E, \mathcal{I})$ is a partition matroid of rank $\sum_{i \in[l]} k_{i} .{ }^{2}$

Linear matroid. Let $\mathbb{F}$ be a field and $E \subset \mathbb{F}^{k}$ be a finite set of vectors. Let

$$
\mathcal{I}=\{H \subseteq E: H \text { is linearly independent over } \mathbb{F}\}
$$

Then $G=(E, \mathcal{I})$ is a linear matroid.

Graphic matroid. Given an undirected graph $\mathcal{G}=(V, H)$ (that may contain loops), define

$$
\mathcal{I}=\{F \subseteq H:(V, F) \text { is a forest }\}
$$

Then, it can be shown that $G(\mathcal{G})=(H, \mathcal{I})$ is a matroid, referred to as graphic matroid. Every spanning forest of the graph $G$ is indeed a basis for matroid $G(\mathcal{G})$.

[^12]
### 4.2.1 Weighted Matroids

For any $\ell \in E$, let $w_{\ell}$ denote the weight assigned to $\ell$. Finding a maximum-weight independent set of $G$ is then defined as finding a set $X \in \mathcal{I}$ which maximizes $\sum_{\ell \in X} w_{\ell}:$

$$
\begin{equation*}
\max _{X \in \mathcal{I}} \sum_{\ell \in X} w_{\ell} . \tag{4.1}
\end{equation*}
$$

It is noted that the optimal solution to this problem is necessarily a basis of $G$. The above problem can be solved efficiently by the Greedy algorithm [71], which works based on the following greedy principle: At each step, add an element (that is not chosen so far) with the largest weight so that the resulting set remains independent. The pseudo-code of Greedy is shown in Algorithm 4.1

```
Algorithm 4.1 Greedy [71]
    Sort weights \(w_{i}, i \in E\). Denote the new ordering by a bijection \(k: E \rightarrow E\) :
                \(w_{k(1)} \geq w_{k(2)} \geq \cdots \geq w_{k(d)}\).
    \(X \leftarrow \emptyset\)
    for \(i=1, \ldots, d\) do
        if \(k(i) \in A(X)\) then
            \(X \leftarrow X \cup\{i\}\)
        end if
    end for
```

As a matter of fact, Greedy leads to an optimal solution of problem 4.1) only if $\mathcal{M}$ is the set of bases of a matroid, as stated in the following theorem which is due to Edmonds 62].

Theorem 4.1 ([23, Theorem 40.1]). Let $\mathcal{I}$ be a nonempty collection of subsets of a set $E$, closed under subsets. Then the pair $(E, \mathcal{I})$ is a matroid if and only if for each weight function $w: E \rightarrow \mathbb{R}_{+}$the Greedy algorithm leads to a set $I \in \mathcal{I}$ of maximum weight.

Let us assume that testing whether a given subset of the ground set $E$ is independent would take $\mathcal{O}(h(d))$ time for some function $h$. Noting that sorting can be carried out in $\mathcal{O}(d \log (d))$ time, we observe that the time complexity of Greedy is $\mathcal{O}(d \log (d)+d h(d))$. In some computational models, it is assumed that an algorithm has access to an independence oracle, that is a routine returning whether $X \in \mathcal{I}$ or not for any given $X \subset E$. Under the independence oracle model, the Greedy algorithm has a time complexity of $\mathcal{O}(d \log (d))$. In words, a maximumweight independent set of a matroid can be found in strongly polynomial time in the independence oracle model ([23, Corollary 40.1]). It is also noted that the independence oracle allows computing the rank of any $X \in E$.

### 4.3 Model and Objectives

Consider a finite set of basic actions $E=\{1, \ldots, d\}$ and a matroid $G=(E, \mathcal{I})$, of rank $m$. We consider a combinatorial MAB problem, where each arm $M$ is a basis of $G$. We let $\mathcal{M}$ denote the set of arms, i.e., the collection of all bases of $G$. Each arm $M$ is identified with a binary column vector $\left(M_{1}, \ldots, M_{d}\right)^{\top}$, and we have $\|M\|_{1}=m, \forall M \in \mathcal{M}$ since $G$ is of rank $m$. Time proceeds in rounds. For $i \in E$, $X_{i}(n)$ denotes the random reward of basic action $i$ in round $n$. For each $i$, the sequence $\left(X_{i}(n)\right)_{n \geq 1}$ is i.i.d. with Bernoulli distribution of mean $\theta_{i}$. The reward sequences across various basic actions may be correlated as considered in 56, 57. We denote by $\theta=\left(\theta_{1}, \ldots, \theta_{d}\right)^{\top} \in \Theta=[0,1]^{d}$ the vector of unknown expected rewards of the various basic actions.

At the beginning of each round $n$, an algorithm or policy $\pi$, selects an arm $M^{\pi}(n) \in \mathcal{M}$ based on the arms chosen in previous rounds and their observed rewards. The reward of arm $M^{\pi}(n)$ selected in round $n$ is

$$
X^{M^{\pi}(n)}(n)=\sum_{i \in E} M_{i}^{\pi}(n) X_{i}(n)=M^{\pi}(n)^{\top} X(n)
$$

Let $\Pi_{s}$ and $\Pi_{\mathrm{b}}$ be the set of all feasible policies with semi-bandit and bandit feedback, respectively. The objective is to identify a policy $\pi$, which maximizes the cumulative expected reward over a finite time horizon $T$. Here the expectation is understood with respect to 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 $\pi$ is defined by:

$$
R^{\pi}(T)=\max _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{n=1}^{T} X^{M}(n)\right]-\mathbb{E}\left[\sum_{n=1}^{T} X^{M^{\pi}(n)}(n)\right]
$$

Finally, we denote by $\mu_{M}(\theta)=M^{\top} \theta$ the expected reward of arm $M$, and let $M^{\star}(\theta) \in \mathcal{M}$, or $M^{\star}$ for short, be any arm with maximum expected reward:

$$
M^{\star}(\theta) \in \arg \max _{M \in \mathcal{M}} \mu_{M}(\theta)
$$

To simplify the presentation in subsequent analysis, we assume that the elements of $\theta$ are distinct, and hence the optimal arm $M^{\star}$ is unique. We further define: $\mu^{\star}(\theta)=M^{\star}(\theta)^{\top} \theta, \quad$ and $\Delta_{\min }=\min _{M \neq M^{\star}} \Delta_{M}$, where $\Delta_{M}=\mu^{\star}(\theta)-\mu_{M}(\theta)$.

### 4.4 Regret Lower Bound

In this section, we present regret lower bounds under the assumption that the reward sequences across basic actions are independent. Our derivations rely on the results presented in Chapter 3 .


Figure 4.1: 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 $\mathcal{K}_{6}=\{1,2,5\}$.

### 4.4.1 Semi-Bandit Feedback

In order to present a regret lower bound for any uniformly good policy in $\Pi_{s}$, we introduce mapping $\sigma_{\theta}: E \backslash M^{\star} \rightarrow M^{\star}$ with

$$
\sigma_{\theta}(i)=\underset{j \in \mathcal{K}_{i}}{\operatorname{argmin}} \theta_{j}, \quad \forall i \in E \backslash M^{\star}
$$

where $\mathcal{K}_{i}=\left\{\ell \in M^{\star}: i \in A\left(M^{\star} \backslash \ell\right)\right\}$. For brevity, we will refer to $\sigma_{\theta}$ by $\sigma$. Figure 4.1 shows an example of $\mathcal{K}_{i}$ for the case of graphic matroids.

By Proposition 4.1, we have that $\mathcal{K}_{i} \neq \emptyset$ for any $i \notin M^{\star}$. Moreover, for any $i \notin M^{\star}$, 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 $\theta$ comprises distinct elements. Consider $M^{\prime}=\left(M^{\star} \backslash \ell\right) \cup\{i\}$. Then, by Proposition 4.1. $M^{\prime} \in \mathcal{M}$. Moreover,

$$
\mu_{M^{\prime}}(\theta)-\mu^{\star}(\theta)=\sum_{k \in M^{\prime}} \theta_{k}-\sum_{k \in M^{\star}} \theta_{k}=\theta_{i}-\theta_{\ell}>0
$$

which contradicts the optimality of $M^{\star}$. Hence, $\theta_{\ell}>\theta_{i}$ for any $\ell \in \mathcal{K}_{i}$.
The next theorem provides a regret lower bound for the policies in $\Pi_{s}$, which may be viewed as the specialization of Theorem 3.1 for the case of matroids.

Theorem 4.2. For all $\theta \in \Theta$ and for any uniformly good algorithm $\pi \in \Pi_{\mathrm{s}}$,

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq \sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}
$$

Remark 4.1. When the underlying matroid is a uniform matroid, the problem reduces to MAB with multiple plays as studied in [33, 51]. Assume that actions are enumerated such that $\theta_{1} \geq \theta_{2} \geq \ldots \theta_{m}>\cdots \geq \theta_{d}$. Then $M^{\star}=\{1,2, \ldots, m\}$ and $\sigma(i)=m$ for all $i \notin M^{\star}$. Hence, the regret lower bound of Theorem 4.2 reduces to the lower bound of Anantharam et al. [33].

For the case of semi-bandit feedback, a specific lower bound example for the case of partition matroid is presented in Kveton et al. [52] to support the claim that regret scaling of $\mathcal{O}\left(\frac{d-m}{\Delta_{\min }} \log (T)\right)$ is tight. Our result is consistent with their result. Moreover, contrary to their lower bound, ours presented in Theorem 4.2 is problem-dependent and tight, i.e. it holds for any parameter $\theta$ and any matroid $G$, and cannot be improved.

### 4.4.2 Bandit Feedback

Now we provide a lower bound on the regret of any uniformly good policy in $\Pi_{b}$.
Theorem 4.3. For all $\theta \in \Theta$ and for any uniformly goods algorithm $\pi \in \Pi_{\mathrm{b}}$,

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq \sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\max _{M: i \in M} I^{M}\left(\theta, \zeta^{i}\right)}
$$

where $\zeta^{i}$ is a vector of parameters defined as $\zeta_{j}^{i}=\theta_{j}$ if $j \neq i$, and $\zeta_{i}^{i}=\theta_{\sigma(i)}$.
The above theorem is indeed an specialization of Theorem 3.3 for the case of matroids.

### 4.5 The KL-OSM Algorithm

Next we present KL-OSM, which is a natural extension of KL-UCB [39] to matroid bandits. The necessary notations are collected as follows: At time $n$, we define $t_{i}(n)=$ $\sum_{s=1}^{n} M_{i}(s)$ the number of times basic action $i$ has been sampled. At time $n$, we define the empirical mean reward of action $i$ as $\hat{\theta}_{i}(n)=\left(1 / t_{i}(n)\right) \sum_{s=1}^{n} X_{i}(s) M_{i}(s)$ if $t_{i}(n)>0$ and $\hat{\theta}_{i}(n)=0$ otherwise.

Our algorithm is an index policy relying on KL-UCB index [39] maintained for each basic action. More precisely, the index of basic action $i$ in round $n$ is denoted by $\omega_{i}(n)$ and defined as:

$$
\omega_{i}(n)=\max \left\{q \in\left[\hat{\theta}_{i}(n), 1\right]: t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), q\right) \leq f(n)\right\},
$$

with $f(n)=\log (n)+3 \log (\log (n))$.
In each round $n \geq 1$, the KL-OSM algorithm simply consists in computing indexes $\omega_{i}(n)$ for all $i$ and then selecting an arm $M(n)$ by solving

$$
M(n) \in \arg \max _{M \in \mathcal{M}} \sum_{i \in M} \omega_{i}(n)
$$

using the Greedy algorithm. The pseudo-code of KL-OSM is given in Algorithm 4.2
The following theorem gives an upper bound on the regret of KL-OSM.

```
Algorithm 4.2 KL-OSM
    for \(n \geq 1\) do
        Select \(M(n) \in \operatorname{argmax}_{M \in \mathcal{M}} \sum_{i \in M} \omega_{i}(n)\) using Greedy.
        Play \(M(n)\), observe the rewards, and update \(t_{i}(n)\) and \(\hat{\theta}_{i}(n), \forall i \in M(n)\).
    end for
```

Theorem 4.4. For any $\varepsilon>0$, there exists positive constants $C_{1}, C_{2}(\varepsilon)$, and $\beta(\varepsilon)$ such that the regret under algorithm $\pi=$ KL-OSM satisfies:

$$
R^{\pi}(T) \leq \sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}(1+\varepsilon) \log (T)+(d-m)\left(C_{1} \log (\log (T))+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}\right)
$$

Hence,

$$
\limsup _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \leq \sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}
$$

Comparing the regret bound of Theorem 4.4 with that of Theorem 4.2 we observe that for the case of Bernoulli rewards, KL-OSM is asymptotically optimal. Next we compare KL-OSM and OMM [52] in terms of their regret upper bounds. OMM achieves a regret upper-bounded by

$$
R(T) \leq \sum_{i \in E \backslash M^{\star}} \frac{16}{\Delta_{\min , i}} \log (T)+\mathcal{O}(1)
$$

where for any sub-optimal $i: \Delta_{\min , i}=\min _{j \in E \backslash M^{\star}}\left|\theta_{i}-\theta_{j}\right|$. Note that by Pinsker's inequality, $\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right) \geq 2\left(\theta_{i}-\theta_{\sigma(i)}\right)^{2} \geq 2 \Delta_{\min , i}^{2}$. Hence, the regret upper bound for KL-OSM is better than that of OMM. The numerical experiment in the next section also shows that KL-OSM outperforms OMM in practice and in some cases the difference is significant.

### 4.5.1 Implementation

The KL-OSM algorithm finds a basis with the maximum index using Greedy, whose time complexity under independence oracle model is $\mathcal{O}(d \log (d))$. We also remark that the computation of index $\omega_{i}(n)$ amounts to finding the roots of a strictly convex and increasing function in one variable (since $z \mapsto \mathrm{kl}(p, z)$ is an increasing function for $z \geq p$ ). Hence, $\omega_{i}(n)$ can be computed straightforwardly by a simple line search such as bisection. Therefore, the time complexity of KL-OSM after $T$ rounds is $\mathcal{O}(d T \log (d))$.

### 4.6 Numerical Experiments

We briefly illustrate the performance under the KL-OSM algorithm for the case of spanning trees in the complete graph $K_{5}$. In this case, there are $d=10$ basic actions


Figure 4.2: Regret of various algorithms for spanning trees
and by Cayley's formula, there are $5^{3}$ spanning trees or arms. Depending on the way the parameter $\theta$ is chosen, we consider two scenarios. In the first scenario, parameter $\theta$ is chosen such that $\theta_{i}=0.8$ if $i \in M^{\star}$ and $\theta_{i}=0.6$ otherwise, whereas in the second one, $\theta$ is drawn uniformly at random from $[0,1]^{10}$.

Figures 4.2 and 4.3 present the regret vs. time horizon under KL-OSM and OMM for the two scenarios. In these figures, the curve in red represents the lower bound of Theorem 4.2. We observe that in both scenarios, KL-OSM significantly outperforms OMM. The curves in Figures 4.2 (b) and $4.3(\mathrm{~b})$ show that the slope of the regret of KL-OSM is becoming identical to that the 'lower bound' curve when the number of rounds grows large. In words, these results imply that the regret of KL-OSM is growing at the same rate of the 'lower bound' curve in the long run, thus verifying the asymptotic optimality of KL-OSM.


Figure 4.3: Regret of various algorithms for spanning trees

### 4.7 Summary

In this chapter we investigated combinatorial bandits where arms are bases of a matroid. We provided explicit regret lower bounds under semi-bandit and bandit feedbacks. These results were specializations of Theorem 3.1 and Theorem 3.3 for the case of matroids. In the case of semi-bandit feedback, we presented KL-OSM and provided its finite-time regret analysis. Our analysis shows that KL-OSM is an asymptotically optimal algorithm as its regret upper bound matches the proposed lower bound. We also provided numerical experiments to validate superiority of KL-OSM over existing algorithms in practice.

## 4.A Proof of Theorem 4.2

Proof. Recalling Theorem 3.1 the regret of any uniformly good policy $\pi \in \Pi_{\mathrm{s}}$ for any $\theta \in \Theta$ satisfies

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{s}}(\theta)
$$

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

$$
\begin{align*}
\inf _{x \geq 0} & \sum_{M \in \mathcal{M}} \Delta_{M} x_{M}  \tag{4.2}\\
\text { subject to: } & \sum_{M \in \mathcal{M}} x_{M} \sum_{i \in E} M_{i} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \geq 1, \quad \forall \lambda \in B_{\mathrm{s}}(\theta),
\end{align*}
$$

with $B_{\mathrm{s}}(\theta)=\left\{\lambda \in \Theta: \lambda_{i}=\theta_{i}, \forall i \in M^{\star}(\theta), \mu^{\star}(\lambda)>\mu^{\star}(\theta)\right\}$. From the proof of Theorem 3.1, recall that problem 4.2 can be equivalently written as

$$
\begin{array}{cc}
\inf _{x \geq 0} & \sum_{M \neq M^{\star}} \Delta_{M} x_{M},  \tag{4.3}\\
\text { subject to: } & \inf _{\lambda \in B_{\mathrm{s}, M^{\prime}}(\theta)} \sum_{i \in M \backslash M^{\star}} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q \in \mathcal{M}} Q_{i} x_{Q} \geq 1, \quad \forall M \neq M^{\star},
\end{array}
$$

where, for any $M \neq M^{\star}$

$$
B_{\mathrm{s}, M}(\theta)=\left\{\lambda \in \Theta: \lambda_{i}=\theta_{i}, \forall i \in M^{\star}(\theta), \mu^{\star}(\theta)<M^{\top} \lambda\right\} .
$$

Fix $i \in E \backslash M^{\star}$. Let $M^{(i)}=\{i\} \cup M^{\star} \backslash \sigma(i)$. Proposition 4.1 implies that $M^{(i)} \in \mathcal{M}$. Figure 4.4 portrays an instance of $\left\{M^{(i)}, i \in E \backslash M^{\star}\right\}$ for the case of graphic matroids. We may simplify the l.h.s. of the constraint corresponding to arm $M^{(i)}$ as follows:

$$
\begin{aligned}
\inf _{\lambda \in B_{\mathrm{s}, M^{(i)}}(\theta)} \sum_{j \in M^{(i)} \backslash M^{\star}} \mathrm{kl}\left(\theta_{j}, \lambda_{j}\right) \sum_{Q} Q_{j} x_{Q} & =\inf _{\lambda \in B_{\mathrm{s}, M^{(i)}}(\theta)} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q} Q_{i} x_{Q} \\
& =\inf _{\lambda \in \Theta: \lambda_{i}>\theta_{\sigma(i)}} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \sum_{Q} Q_{i} x_{Q} \\
& =\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right) \sum_{Q} Q_{i} x_{Q}
\end{aligned}
$$

where the first equality follows from $M^{(i)} \backslash M^{\star}=\{i\}$. Hence, $M^{(i)}$-th constraint in problem (4.3) may be equivalently written as

$$
\sum_{Q} Q_{i} x_{Q} \geq \frac{1}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}
$$


(a) $M^{\star}$

(b)

(c)

(d)

(e)

(f)

(g)

Figure 4.4: Spanning trees in $K_{5}$ : (a) The optimal spanning tree $M^{\star},(\mathrm{b})-(\mathrm{g}) M^{(i)}$.

Let $\mathcal{M}^{-}=\mathcal{M} \backslash\left(\left\{M^{\star}\right\} \cup\left\{M^{(i)}, i \in E \backslash M^{\star}\right\}\right)$. It then follows that

$$
\begin{align*}
c_{\mathrm{s}}(\theta)=\inf _{x \geq 0} & \sum_{M \in \mathcal{M}} \Delta_{M} x_{M}  \tag{4.4}\\
\text { subject to: } & \sum_{Q \neq M^{\star}} Q_{i} x_{Q} \geq \frac{1}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}, \quad \forall i \in E \backslash M^{\star}, \\
& \inf _{\lambda \in B_{\mathrm{s}, M}(\theta)} \sum_{Q \in \mathcal{M}} x_{Q} \sum_{i \in E} Q_{i} \mathrm{kl}\left(\theta_{i}, \lambda_{i}\right) \geq 1, \quad \forall M \in \mathcal{M}^{-} .
\end{align*}
$$

Let $\tau_{M}: E \rightarrow E$ be a bijection defined as follows: If $i \in M \backslash M^{\star}$, then $\tau_{M}(i)=j$ for some $j \in \mathcal{K}_{i}$. Otherwise, $\tau_{M}(i)=i$. We have:

$$
\begin{aligned}
\Delta_{M} & =\sum_{i \in M}\left(\theta_{\tau_{M}(i)}-\theta_{i}\right)=\sum_{i \in M \backslash M^{\star}}\left(\theta_{\tau_{M}(i)}-\theta_{i}\right) \\
& =\sum_{i \in E \backslash M^{\star}} M_{i}\left(\theta_{\tau_{M}(i)}-\theta_{i}\right) \geq \sum_{i \in E \backslash M^{\star}} M_{i}\left(\theta_{\sigma(i)}-\theta_{i}\right) .
\end{aligned}
$$

Hence, introducing $z_{i}=\sum_{M} M_{i} x_{M}$ for any $i \in E \backslash M^{\star}$, we obtain:

$$
\sum_{M} x_{M} \Delta_{M} \geq \sum_{M} x_{M} \sum_{i \notin M^{\star}} M_{i}\left(\theta_{\sigma(i)}-\theta_{i}\right)=\sum_{i \notin M^{\star}}\left(\theta_{\sigma(i)}-\theta_{i}\right) z_{i} .
$$

As a result, defining

$$
\begin{aligned}
\text { P1: } & \inf _{z \geq 0} \sum_{i \in E \backslash M^{\star}}\left(\theta_{\sigma(i)}-\theta_{i}\right) z_{i} \\
\text { subject to: } & z_{i} \geq \frac{1}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}, \quad \forall i \in E \backslash M^{\star},
\end{aligned}
$$

yields: $c_{\mathrm{s}}(\theta) \geq \operatorname{val}(\mathrm{P} 1)$. The proof is completed by observing that

$$
\operatorname{val}(\mathrm{P} 1)=\sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)} .
$$

## 4.B Proof of Theorem 4.3

Proof. Recall from Theorem 3.3 that the regret of any uniformly good policy $\pi \in \Pi_{\mathrm{b}}$ for any $\theta \in \Theta$ satisfies

$$
\liminf _{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log (T)} \geq c_{\mathrm{b}}(\theta)
$$

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

$$
\begin{align*}
\inf _{x \geq 0} & \sum_{M \in \mathcal{M}} \Delta_{M} x_{M}  \tag{4.5}\\
\text { subject to: } & \sum_{M \neq M^{\star}} x_{M} I^{M}(\theta, \lambda) \geq 1, \quad \forall \lambda \in B_{\mathrm{b}}(\theta),
\end{align*}
$$

and $B_{\mathrm{b}}(\theta)$ is the set of bad parameters that cannot be distinguished from true parameter $\theta$ when selecting arm $M^{\star}(\theta)$, and for which $\operatorname{arm} M^{\star}(\theta)$ is sub-optimal:

$$
B_{\mathrm{b}}(\theta)=\left\{\lambda \in \Theta:\left\{\lambda_{i}, i \in M^{\star}\right\}=\left\{\theta_{i}, i \in M^{\star}\right\}, \mu^{\star}(\lambda)>\mu^{\star}(\theta)\right\} .
$$

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 $B_{\mathrm{b}}(\theta)$ into sets where in each set, action $i$ is better than action $\sigma(i)$ under $\lambda$. For any $i \notin M^{\star}$, define

$$
A_{i}(\theta)=\left\{\lambda:\left\{\lambda_{\ell}, \ell \in M^{\star}\right\}=\left\{\theta_{\ell}, \ell \in M^{\star}\right\}, \lambda_{i}>\theta_{\sigma(i)}\right\} .
$$

Then, $B_{\mathrm{b}}(\theta)=\bigcup_{i \notin M^{\star}} A_{i}(\theta)$ and problem 4.5 reads

$$
c_{\mathrm{b}}(\theta)=\inf _{x \geq 0} \sum_{M} x_{M} \Delta_{M}
$$

$$
\text { subject to: } \inf _{\lambda \in A_{i}(\theta)} \sum_{M \neq M^{\star}} x_{M} I^{M}(\theta, \lambda) \geq 1, \quad \forall i \notin M^{\star} .
$$

Consider $\zeta^{i}$ with $\zeta_{i}^{i}=\theta_{\sigma(i)}$ and $\zeta_{j}^{i}=\theta_{j}$ for $j \neq i$. Since $\zeta^{i} \in A_{i}(\theta)$, we have

$$
\begin{aligned}
\inf _{\lambda \in A_{i}(\theta)} \sum_{M \neq M^{\star}} x_{M} I^{M}(\theta, \lambda) & \leq \sum_{M} x_{M} I^{M}\left(\theta, \zeta^{i}\right) \\
& =\sum_{M} M_{i} x_{M} I^{M}\left(\theta, \zeta^{i}\right) \\
& \leq \max _{M: i \in M} I^{M}\left(\theta, \zeta^{i}\right) \sum_{M} M_{i} x_{M} .
\end{aligned}
$$

Hence, problem 4.6 is lower bounded as follows:

$$
\begin{equation*}
c_{\mathrm{b}}(\theta) \geq \inf _{x \geq 0} \sum_{M} x_{M} \Delta_{M} \tag{4.7}
\end{equation*}
$$

$$
\text { subject to: } \max _{M: i \in M} I^{M}\left(\theta, \zeta^{i}\right) \sum_{M} M_{i} x_{M} \geq 1, \quad \forall i \notin M^{\star} .
$$

Recall from the proof of Theorem 4.2 that $\sum_{M} x_{M} \Delta_{M} \geq \sum_{i \in E \backslash M^{*}}\left(\theta_{\sigma(i)}-\theta_{i}\right) z_{i}$. Hence, problem (4.7) is further lower bounded as

$$
\begin{aligned}
c_{\mathrm{b}}(\theta) \geq \inf _{z \geq 0} & \sum_{i \in E \backslash M^{\star}}\left(\theta_{\sigma(i)}-\theta_{i}\right) z_{i} \\
\text { subject to: } & z_{i} \geq \frac{1}{\max _{M: i \in M} I^{M}\left(\theta, \zeta^{i}\right)}, \quad \forall i \notin M^{\star},
\end{aligned}
$$

which further gives

$$
c_{\mathrm{b}}(\theta) \geq \sum_{i \in E \backslash M^{\star}} \frac{\theta_{\sigma(i)}-\theta_{i}}{\max _{M: i \in M} I^{M}\left(\theta, \zeta^{i}\right)}
$$

and concludes the proof.

## 4.C Proof of Theorem 4.4

Proof. Let $T>0$. Consider round $n$ where $M(n) \neq M^{\star}$ is selected by the algorithm $\pi=$ KL-OSM. Then, there exists a bijection $\tau_{n}: M(n) \rightarrow M^{\star}$ such that: $\tau_{n}(i)=i$ if $i \in M^{\star} \cap M(n)$. Otherwise, $\tau_{n}(i)=j$ for some $j \in \mathcal{K}_{i}$. The bijection $\tau_{n}$ simply maps the sub-optimal basic actions of $M(n)$ to the corresponding ones in $M^{\star}$ that are not chosen by the algorithm at round $n$. It then follows that for any $i \in E$ :

$$
\mathbb{1}\left\{M_{i}(n)=1\right\}=\sum_{j \in \mathcal{K}_{i}} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j\right\}
$$

and that $\sum_{j \in \mathcal{K}_{i}} \mathbb{1}\left\{\tau_{n}(i)=j\right\} \leq 1$ since $\tau_{n}$ is a bijection. An example of bijection $\tau_{n}$ for the case of graphic matroids is shown in Figure 4.5 It should be noted that $\tau_{n}$ may not be unique.

For any $i, j \in E$, define $\Delta_{j, i}=\theta_{j}-\theta_{i}$. Then, the regret under policy $\pi=$ KL-OSM is upper bounded as:

$$
\begin{aligned}
R^{\pi}(T) & \leq \mathbb{E}\left[\sum_{n=1}^{T} \Delta_{M(n)}\right]=\mathbb{E}\left[\sum_{n=1}^{T} \sum_{i \in E \backslash M^{\star}} \Delta_{\tau_{n}(i), i} \mathbb{1}\left\{M_{i}(n)=1\right\}\right] \\
& =\mathbb{E}\left[\sum_{i \in E \backslash M^{\star}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j\right\}\right] .
\end{aligned}
$$

Let $i \in E \backslash M^{\star}$. We use the following decomposition:

$$
\mathbb{1}\left\{M_{i}(n)=1, \omega_{i}(n) \geq \omega_{\tau_{n}(i)}(n)\right\} \leq \mathbb{1}\left\{\omega_{\tau_{n}(i)}(n)<\theta_{\tau_{n}(i)}\right\}
$$



Figure 4.5: An example of bijection $\tau_{n}$ for the case of graphic matroids. In this case: $\tau_{n}(1)=1, \tau_{n}(3)=2, \tau_{n}(4)=4, \tau_{n}(6)=5$.

$$
+\mathbb{1}\left\{M_{i}(n)=1, \omega_{i}(n) \geq \theta_{\tau_{n}(i)}\right\}
$$

Hence,

$$
\begin{aligned}
\sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j\right\} & \leq \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{\tau_{n}(i)=j, \omega_{j}(n)<\theta_{j}\right\} \\
& +\sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{i \in E \backslash M^{*}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j\right\}\right. \\
& \leq \mathbb{E}\left[\sum_{i \in E \backslash M^{*}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \mathbb{1}\left\{\tau_{n}(i)=j, \omega_{j}(n)<\theta_{j}\right\}\right] \\
& +\mathbb{E}\left[\sum_{i \in E \backslash M^{*}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}\right],
\end{aligned}
$$

since $\Delta_{j, i} \leq 1$.
We prove that there exist positive constants $C_{1}, C_{2}(\varepsilon)$, and $\beta(\varepsilon)$ such that

$$
\begin{equation*}
\mathbb{E}\left[\sum_{i \in E \backslash M^{\star}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \mathbb{1}\left\{\tau_{n}(i)=j, \omega_{j}(n)<\theta_{j}\right\}\right] \leq(d-m) C_{1} \log (\log (T)), \tag{4.8}
\end{equation*}
$$

$\mathbb{E}\left[\sum_{i \in E \backslash M^{\star}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}\right]$

$$
\begin{equation*}
\leq(1+\varepsilon) \frac{\theta_{\sigma(i)}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)} \log (T)+(d-m) \frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}} . \tag{4.9}
\end{equation*}
$$

Hence, we get the announced result:

$$
\begin{aligned}
R^{\pi}(T) & \leq \mathbb{E}\left[\sum_{i \in E \backslash M^{\star}} \sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j\right\}\right] \\
& \leq \sum_{i \in E \backslash M^{\star}}(1+\varepsilon) \log (T) \frac{\theta_{\sigma(i)}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}+(d-m)\left(\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}+C_{1} \log (\log (T))\right) .
\end{aligned}
$$

## Inequality (4.8):

Fix $j \in \mathcal{K}_{i}$. By the concentration inequality in [39, Theorem 10], we have

$$
\mathbb{P}\left[\omega_{j}(n)<\theta_{j}\right] \leq\lceil f(n) \log (n)\rceil e^{1-f(n)},
$$

and hence following the same steps as in the proof of [39, Theorem 2], we observe that there exists constant $C_{1} \leq 7$ such that $\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{\omega_{j}(n)<\theta_{j}\right\}\right] \leq$ $C_{1} \log (\log (T))$. It then follows that

$$
\sum_{j \in \mathcal{K}_{i}} \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{\tau_{n}(i)=j, \omega_{j}(n)<\theta_{j}\right\}\right] \leq C_{1}(\log (\log (T)))
$$

since $\tau_{n}$ for any $n$ is a bijection. As a result:

$$
\sum_{i \notin M^{\star}} \sum_{j \in \mathcal{K}_{i}} \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left\{\tau_{n}(i)=j, \omega_{j}(n)<\theta_{j}\right\}\right] \leq(d-m) C_{1}(\log (\log (T)))
$$

## Inequality (4.9):

For $x, y \in[0,1]$, introduce $\mathrm{kl}^{+}(x, y)=\operatorname{kl}(x, y) \mathbb{1}\{x<y\}$. Fix $j \in \mathcal{K}_{i}$. Observe that the event $\omega_{i}(n) \geq \theta_{j}$ implies that $\mathrm{kl}^{+}\left(\hat{\theta}_{i}(n), \theta_{j}\right) \leq \mathrm{kl}\left(\hat{\theta}_{i}(n), \omega_{i}(n)\right)=f(n) / t_{i}(n)$.

We let $\hat{\theta}_{i, s}$ denote the empirical average of rewards of action $i$ when it is selected $s$ times. Hence we obtain:

$$
\begin{aligned}
& \sum_{n=1}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\} \\
& \leq \sum_{n=1}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n) \mathrm{kl}^{+}\left(\hat{\theta}_{i}(n), \theta_{j}\right) \leq f(n)\right\} \\
& =\sum_{n=1}^{T} \sum_{s=1}^{n} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s, s \mathrm{kl}^{+}\left(\hat{\theta}_{i, s}, \theta_{j}\right) \leq f(n)\right\} \\
& \leq \sum_{n=1}^{T} \sum_{s=1}^{n} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s, s \mathrm{kl}^{+}\left(\hat{\theta}_{i, s}, \theta_{j}\right) \leq f(T)\right\}
\end{aligned}
$$

$$
=\sum_{s=1}^{T} \mathbb{1}\left\{s \mathrm{kl}^{+}\left(\hat{\theta}_{i, s}, \theta_{j}\right) \leq f(T)\right\} \sum_{n=s}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s\right\} .
$$

We therefore get

$$
\begin{align*}
& \mathbb{E}\left[\sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}\right] \\
& \quad \leq \mathbb{E}\left[\sum_{j \in \mathcal{K}_{i}} \sum_{s=1}^{T} \Delta_{j, i} \mathbb{1}\left\{s \mathrm{kl}^{+}\left(\hat{\theta}_{i, s}, \theta_{j}\right) \leq f(T)\right\} \sum_{n=s}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s\right\}\right] . \tag{4.10}
\end{align*}
$$

From [39, Lemma 8], we have that for any $j \in \mathcal{K}_{i}$ :

$$
\mathbb{E}\left[\sum_{s=1}^{T} \Delta_{j, i} \mathbb{1}\left\{s \mathrm{kl}^{+}\left(\hat{\theta}_{i, s}, \theta_{j}\right) \leq f(T)\right\}\right] \leq \frac{(1+\varepsilon) \Delta_{j, i} \log (T)}{\operatorname{kl}\left(\theta_{i}, \theta_{j}\right)}+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}
$$

Combining this with 4.10, we get

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}\right] \\
& \quad \leq \sum_{j \in \mathcal{K}_{i}}\left(\frac{(1+\varepsilon) \Delta_{j, i} \log (T)}{\mathrm{kl}\left(\theta_{i}, \theta_{j}\right)}+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}\right) \\
& \quad \times \mathbb{E}\left[\sum_{n=s}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s\right\}\right] \\
& \quad \leq\left(\max _{j \in \mathcal{K}_{i}} \frac{(1+\varepsilon) \Delta_{j, i} \log (T)}{\mathrm{kl}\left(\theta_{i}, \theta_{j}\right)}+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}\right) \\
& \quad \times \mathbb{E}\left[\sum_{j \in \mathcal{K}_{i}} \sum_{n=s}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s\right\}\right] \\
& \quad \leq \max _{j \in \mathcal{K}_{i}} \frac{(1+\varepsilon) \Delta_{j, i} \log (T)}{\mathrm{kl}\left(\theta_{i}, \theta_{j}\right)}+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}
\end{aligned}
$$

where in the last inequality, we used the fact that

$$
\sum_{j \in \mathcal{K}_{i}} \sum_{n=s}^{T} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, t_{i}(n)=s\right\} \leq 1
$$

since $\tau_{n}$ is a bijection for any $n$. Lemma 4.1, proven at the end of this section, implies that

$$
\max _{j \in \mathcal{K}_{i}} \frac{\Delta_{j, i}}{\operatorname{kl}\left(\theta_{i}, \theta_{j}\right)}=\frac{\theta_{\sigma(i)}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)},
$$

which gives

$$
\mathbb{E}\left[\sum_{n=1}^{T} \sum_{j \in \mathcal{K}_{i}} \Delta_{j, i} \mathbb{1}\left\{M_{i}(n)=1, \tau_{n}(i)=j, \omega_{i}(n) \geq \theta_{j}\right\}\right] \leq(1+\varepsilon) \log (T) \frac{\theta_{\sigma(i)}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{\sigma(i)}\right)}+\frac{C_{2}(\varepsilon)}{T^{\beta(\varepsilon)}}
$$

This completes the proof of inequality (4.9) and hence concludes the proof.

## 4.D Proof of Supporting Lemmas

We prove the following lemma about the KL-divergence of two Bernoulli distributions.

Lemma 4.1. Let $j^{\star}=\operatorname{argmin}_{j: \theta_{j}>\theta_{i}} \theta_{j}$. Then:

$$
\begin{equation*}
\max _{j: \theta_{j}>\theta_{i}} \frac{\theta_{j}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{j}\right)}=\frac{\theta_{j^{\star}}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{j}^{\star}\right)} \tag{4.11}
\end{equation*}
$$

Proof. We prove the lemma by contradiction. Assume that $j^{\star}$ is not the maximizer of (4.11), namely there exists some $k \neq j^{\star}$ such that $\theta_{k}>\theta_{i}$ and

$$
\begin{equation*}
\frac{\theta_{j^{\star}}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{j^{\star}}\right)}<\frac{\theta_{k}-\theta_{i}}{\mathrm{kl}\left(\theta_{i}, \theta_{k}\right)}, \tag{4.12}
\end{equation*}
$$

or equivalently,

$$
\frac{\theta_{j^{\star}}-\theta_{i}}{\theta_{k}-\theta_{i}}<\frac{\mathrm{kl}\left(\theta_{i}, \theta_{j^{\star}}\right)}{\mathrm{kl}\left(\theta_{i}, \theta_{k}\right)} \leq 1
$$

Letting $\alpha=\frac{\theta_{j^{\star}-\theta_{i}}}{\theta_{k}-\theta_{i}}$, we may write $\theta_{j^{\star}}=\alpha \theta_{k}+(1-\alpha) \theta_{i}$. Observe that $\alpha \in(0,1)$. Convexity of $z \mapsto \operatorname{kl}(p, z)$ for any $p$ implies that

$$
\mathrm{kl}\left(\theta_{i}, \theta_{j^{\star}}\right) \leq(1-\alpha) \mathrm{kl}\left(\theta_{i}, \theta_{i}\right)+\alpha \mathrm{kl}\left(\theta_{i}, \theta_{k}\right)=\alpha \mathrm{kl}\left(\theta_{i}, \theta_{k}\right),
$$

and thus

$$
\frac{\theta_{j^{\star}}-\theta_{i}}{\operatorname{kl}\left(\theta_{i}, \theta_{j^{\star}}\right)} \geq \frac{\alpha\left(\theta_{k}-\theta_{i}\right)}{\alpha \operatorname{kl}\left(\theta_{i}, \theta_{k}\right)}
$$

which contradicts 4.12. Thus, $j^{\star}$ is the maximizer of 4.11 and the proof is concluded.

## Chapter 5

## Stochastic Combinatorial MABs: Geometric Rewards

In this chapter we study online shortest-path routing problems as introduced in Chapter 11 which fall in the class of combinatorial MAB problems. Here we consider instances of these problems with geometrically distributed rewards. We consider two types of routing: source routing in which the path is determined at the source and hop-by-hop routing, where routing decisions are taken at intermediate nodes. Using the machinery of Chapter 3, we derive lower bounds on the regret. We consider three algorithms for this class of problems and provide upper bounds on their regret. These upper bounds are the best one proposed so far in the literature for combinatorial MABs with geometric rewards. We also provide numerical experiments, which show that our algorithms outperform existing ones.

This main part of this chapter is based on the work [72] and is organized as follows: Section 5.1 outlines our contributions in this chapter and discusses related works. Section 5.2 describes the network model, feedback models, and objectives. In Section 5.3 we present regret lower bounds for various types of feedbacks. In Section 5.4 we present routing policies for the case of source routing with semibandit feedback along with their regret analysis. Section 5.5 presents numerical experiments. In Section 5.6, we give a brief summary of the materials presented in this chapter. All proofs are provided in the appendix.

### 5.1 Contributions and Related Work

We make the following contributions for stochastic online shortest-path routing problem:
(a) We derive tight asymptotic (when the number of packets $N$ grows large) regret lower bounds. The two first 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 it turns out, the regret lower bounds for source routing policies with semi-bandit feedback and that for hop-by-hop routing policies

| Algorithm |  | Regret | Complexity |
| :---: | :---: | :---: | :---: |
| Implementation |  |  |  |
| CUCB [56] |  | $\mathcal{O}\left(\frac{d m^{2}}{\Delta_{\min } \theta_{\min }^{4}} \log (N)\right)$ | $\mathcal{O}(\|V\| d)$ |
| Distributed |  |  |  |
| GEOCombUCB-1 (Theorem 5.5 | $\mathcal{O}\left(\frac{d \sqrt{m}}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ | $\mathcal{O}(\|\mathcal{M}\|)$ | Centralized |
| GEOCombUCB-2 (Theorem 5.5 | $\mathcal{O}\left(\frac{d \sqrt{m}}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ | $\mathcal{O}(\|\mathcal{M}\|)$ | Centralized |
| KL-SR (Theorem 5.6 | $\mathcal{O}\left(\frac{d m}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ | $\mathcal{O}(\|V\| d)$ | Distributed |

Table 5.1: Comparison of various algorithms for shortest-path routing under semibandit feedback.
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.
(b) In the case of semi-bandit feedback, we propose two online source routing policies, namely GeoCombUCB-1 and GeoCombUCB-2. Geo refers to the fact that the delay on a given link is geometrically distributed, Сомв stands for combinatorial, and UCB (Upper Confidence Bound) indicates that these policies are based on the same "optimism in face of uncertainty" principle as the celebrated UCB1 algorithm designed for classical MAB problems [19]. Moreover, we improve the regret upper bound of KL-SR 30] to $\mathcal{O}\left(\frac{d m}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)^{1}$, where $m$ denotes the length (number of links) of the longest path in the network from the source to the destination, $\theta_{\text {min }}$ 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 the optimal and of a sub-optimal path. We further show that the regret under GeoCombUCB-1 and GeoCombucb-2 scales at most as $\mathcal{O}\left(\frac{d \sqrt{m}}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$. Our routing policies strike an interesting trade-off between computational complexity and performance, and exhibit better regret upper bounds than that of the CUCB algorithm [56, which was, to our knowledge, the best online shortest-path routing algorithm. The regret guarantees of various algorithms and their computational complexity are summarized in Table 5.1 Finally we conduct numerical experiments, showing that our routing policies perform significantly better than CUCB.

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.

[^13]
### 5.1.1 Related Work

We summarize existing results for generic stochastic combinatorial bandits that could be applied to online shortest-path routing. In [56], 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 presented in [56] scales as $\mathcal{O}\left(\frac{d m^{2}}{\Delta_{\min } \theta_{\min }^{4}} \log (N)\right)$ (see Appendix 5.F 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 [55, 57, 50] 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 [73, 31, 74]. Liu and Zhao [73] consider routing with bandit (end-to-end) feedback and propose a forced-exploration algorithm with $\mathcal{O}\left(d^{3} m \log (N)\right)$ regret in which a random barycentric spanner ${ }^{2}$ path is chosen for exploration. He et al. 31 consider 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 present work is the first to consider online routing problems with hop-by-hop decisions. Such a problem can be formulated as a classical Markov Decision Process (MDP), in which the states are the packet locations and the actions are the outgoing links of each node. However, most studies considered MDP problem under stricter assumptions than ours and/or targeted different performance measures. For these problems, Burnetas and Katehakis [35] derive the asymptotic lower bound on the regret and propose an optimal index policy. Their result can be applied only to the so-called ergodic MDP [75], 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. Algorithms for general MDPs with logarithmic regret were also proposed in, e.g., [76, 77]. Nevertheless, these algorithms perform badly when applied to hop-by-hop routing due to loosely constructed confidence intervals, and the asymptotic performance bounds were not studied.

[^14]| Policy Set | Routing Type | Feedback |
| :---: | :---: | :---: |
| $\Pi_{\mathrm{b}}$ | Source routing | Bandit |
| $\Pi_{\mathrm{s}}$ | Source routing | Semi-bandit |
| $\Pi_{\mathrm{h}}$ | Hop-by-hop | Semi-bandit |

Table 5.2: Various policy sets for online shortest path routing.

### 5.2 Model and Objectives

### 5.2.1 Network Model

We consider a network 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. Let $d$ be the cardinality of $E$. 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 $i$ at time $t$ is successful. $\left(X_{i}(t)\right)_{t \geq 1}$ is a sequence of i.i.d. Bernoulli variables with initially unknown mean $\theta_{i}$. 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 $\theta_{\min }=\min _{i \in E} \theta_{i}>0$, and let $\theta=\left(\theta_{i}, i \in E\right)$ 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 $\mathcal{M} \subseteq\{0,1\}^{d}$ the set of loop-free paths from $u$ to $v$ in $G$, where each path $M \in \mathcal{M}$ is a $d$ dimensional binary vector; for any $i \in E, M_{i}=1$ if and only if $i$ belongs to $M$. Hence, for any $M \in \mathcal{M}$, the length of path $M$ is $\|M\|_{1}=\sum_{i \in E} M_{i}$. We let $m$ denote the maximum length of the paths in $\mathcal{M}$, i.e., $m=\max _{M \in \mathcal{M}}\|M\|_{1}$. For brevity, in what follows, for any binary vector $z$, we write $i \in z$ to denote $z_{i}=1$. Moreover, we use the convention that $z^{-1}=\left(z_{i}^{-1}\right)_{i}$.

### 5.2.2 Objectives and Feedback

We assume that the source is fully backlogged (i.e., it always has packets to send), and that the parameter $\theta$ is initially unknown. Packets are sent successively from $u$ to $v$ over various paths to estimate $\theta$, and in turn to learn the path $M^{\star}$ with the minimum average delay: $M^{\star} \in \operatorname{argmin}_{M \in \mathcal{M}} \sum_{i \in M} \frac{1}{\theta_{i}}$. After a packet is sent, we assume that the source gathers some 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). These policy sets are defined below:

- Policy Set $\Pi_{\mathrm{b}}$ : 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 $M^{\pi}(n)$ be the path selected under policy $\pi$, and let $D^{\pi}(n)$ denote the corresponding end-to-end delay. Then $M^{\pi}(n)$ depends on $M^{\pi}(1), \ldots, M^{\pi}(n-1), D^{\pi}(1), \ldots, D^{\pi}(n-1)$.
- Policy Set $\Pi_{s}$ : 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 $\pi, M^{\pi}(n)$ depends on $M^{\pi}(1), \ldots, M^{\pi}(n-1),\left(h_{i}^{\pi}(1), i \in M^{\pi}(1)\right), \ldots,\left(h_{i}^{\pi}(n-\right.$ 1), $\left.i \in M^{\pi}(n-1)\right)$, where $h_{i}^{\pi}(k)$ is the delay experienced on link $i$ for the $k$-th packet (if this packet uses link $i$ at all).
- Policy Set $\Pi_{\mathrm{h}}$ : 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 $t$.

Table 5.2 lists different policy sets for the three types of online routing policies considered. In the case of source routing policies (in $\Pi_{\mathrm{b}} \cup \Pi_{\mathrm{s}}$ ), 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 $\Pi_{h}$ ), 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 $R^{\pi}(N)$ of policy $\pi$ up to the $N$-th packet is the expected difference of delays for the $N$ first packets under $\pi$ and under the policy that always selects the best path $M^{\star}$ for transmission:

$$
R^{\pi}(N)=\mathbb{E}\left[\sum_{n=1}^{N} D^{\pi}(n)\right]-N \mu^{\star}(\theta)
$$

where $D^{\pi}(n)$ denotes the end-to-end delay of the $n$-th packet under policy $\pi$. Moreover, $\mu^{\star}(\theta)=\operatorname{argmin}_{M \in \mathcal{M}} \mu_{M}(\theta)$ where $\mu_{M}(\theta)=\sum_{i \in M} \frac{1}{\theta_{i}}$ is the average packet delay through path $M$ given link success rates $\theta$, and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the random transmission outcomes and possible randomization in the policy $\pi$. For any path $M \in \mathcal{M}$, define $\Delta_{M}=\mu_{M}(\theta)-\mu^{\star}(\theta)=\left(M-M^{\star}\right)^{\top} \theta^{-1}$. Furthermore, let $\Delta_{\text {min }}=\min _{\Delta_{M} \neq 0} \Delta_{M}$.

### 5.3 Regret Lower Bounds

In this section, we provide fundamental performance limits satisfied by any online routing policy in $\Pi_{\mathrm{b}}, \Pi_{\mathrm{s}}$, or $\Pi_{\mathrm{h}}$. By comparing these performance limits, we can quantify the potential performance improvements taking routing decisions at each hop rather than at the source only, and observing per-link delays (semi-bandit feedback) rather than end-to-end delays (bandit feedback).

### 5.3.1 Source Routing with Semi-Bandit (Per-Link) Feedback

Consider routing policies in $\Pi_{\mathrm{s}}$ that make decisions at the source, but have information on the individual link delays. Let $\operatorname{KLG}(u, v)$ denote the KL-divergence between two geometric random variables with parameters $u$ and $v$ :

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

The next theorem provides the regret lower bound for online routing policies in $\Pi_{s}$.

Theorem 5.1. For all $\theta$ and for any uniformly good policy $\pi \in \Pi_{s}$,

$$
\liminf _{N \rightarrow \infty} \frac{R^{\pi}(N)}{\log (N)} \geq c_{\mathrm{s}}(\theta)
$$

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

$$
\begin{array}{cc}
\inf _{x \geq 0} & \sum_{M \neq M^{\star}} \Delta_{M} x_{M} \\
\text { subject to: } & \inf _{\lambda \in B_{\mathrm{s}}(\theta)} \sum_{M \neq M^{\star}} x_{M} \sum_{i \in M} \operatorname{KLG}\left(\theta_{i}, \lambda_{i}\right) \geq 1 . \tag{5.2}
\end{array}
$$

with $B_{\mathrm{s}}(\theta)=\left\{\lambda: \lambda_{i}=\theta_{i}, \forall i \in M^{\star}, \min _{M \in \mathcal{M}} \mu_{M}(\lambda)<\mu^{\star}(\theta)\right\}$.

Proof of the above theorem is quite similar to that of Theorem3.1 and is omitted.
Remark 5.1. The asymptotic lower bound proposed in [31] has a similar expression to ours, but the set $B_{\mathrm{s}}(\theta)$ is replaced by

$$
B_{\mathrm{s}}^{\prime}(\theta)=\bigcup_{i \in E}\left\{\lambda: \lambda_{j}=\theta_{j}, \forall j \neq i, \min _{M \in \mathcal{M}} \mu_{M}(\lambda)<\mu^{\star}(\theta)\right\}
$$

It is noted that $B_{\mathrm{s}}^{\prime}(\theta) \subset B_{\mathrm{s}}(\theta)$, which implies that the lower bound derived in [31] is smaller than ours. In other words, we propose a regret lower bound that improves that in [31], and moreover, our bound is tight (it cannot be improved further).

We note that the lower bound of Theorem 5.1 is unfortunately implicit. Hence, it could be interesting to see how it scales as a function of the problem dimensions $d$ and $m$. Below we consider a particular instance of routing problem in which the underlying topology is a grid. This instance demonstrates that the machinery developed in Chapter 3 to simplify the lower bound of Theorem 3.1 may prove inapplicable for routing problems.

(a)

(b)

(c)

(d)

(e)

Figure 5.1: Routing in a grid: (a) Grid topology with source (red) and destination (blue) nodes, (b) Optimal path $M^{\star}$, (c)-(e) Elements of $\mathcal{H}$.

Routing in a grid. Consider routing in an $N$-by- $N$ directed grid, whose topology is shown in Figure 5.1 (a), where the source (resp. destination) node is shown in red (resp. blue). Here $\mathcal{M}$ is the set of all $\binom{2 N-2}{N-1}$ paths with $m=2(N-1)$ edges. We further have $d=2 N(N-1)$. In this example, elements of any maximal set $\mathcal{H}$ satisfying $P(\theta)$, defined in Chapter 3 do not cover all sub-optimal links. For instance, for the grid shown in Figure 5.1(a), there are 6 links that do not appear in any $\operatorname{arm}$ in $\mathcal{H}$. Moreover, one may easily prove that in this case, $|\mathcal{H}|$ scales as $N$ rather than $N^{2}=d$.

### 5.3.2 Source Routing with Bandit Feedback

We now consider routing policies in $\Pi_{\mathrm{b}}$. Denote by $\psi_{\theta}^{M}(k)$ the probability that the delay of a packet sent on path $M$ is $k$ slots. The end-to-end delay is the sum of several independent random geometric variables. Assuming $\theta_{i} \neq \theta_{j}$ for $i \neq j$, we
have for all $k \geq\|M\|_{1}[78]$ :

$$
\psi_{\theta}^{M}(k)=\sum_{i \in M}\left(\prod_{j \in M, j \neq i} \frac{\theta_{j}}{\theta_{j}-\theta_{i}}\right) \theta_{i}\left(1-\theta_{i}\right)^{k-1}
$$

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 $\Pi_{b}$.

Theorem 5.2. For all $\theta$ and for any uniformly good policy $\pi \in \Pi_{\mathrm{b}}$,

$$
\liminf _{N \rightarrow \infty} \frac{R^{\pi}(N)}{\log (N)} \geq c_{\mathrm{b}}(\theta)
$$

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

$$
\begin{equation*}
\inf _{x \geq 0} \sum_{M \in \mathcal{M}} x_{M} \Delta_{M} \tag{5.3}
\end{equation*}
$$

subject to: $\inf _{\lambda \in B_{\mathrm{b}}(\theta)} \sum_{M \neq M^{\star}} x_{M} \sum_{k=\|M\|_{1}}^{\infty} \psi_{\theta}^{M}(k) \log \frac{\psi_{\theta}^{M}(k)}{\psi_{\lambda}^{M}(k)} \geq 1$,
with

$$
B_{\mathrm{b}}(\theta):=\left\{\lambda:\left\{\lambda_{i}, i \in M^{\star}\right\}=\left\{\theta_{i}, i \in M^{\star}\right\}, \min _{M \in \mathcal{M}} \mu_{M}(\lambda)<\mu^{\star}(\theta)\right\} .
$$

Similarly to the case of Bernoulli rewards in Chapter 3 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 routing policies than those we can devise using bandit feedback (i.e., $\Pi_{\mathrm{b}} \subset \Pi_{\mathrm{s}}$ ).

### 5.3.3 Hop-by-hop Routing

Finally, we consider routing policies in $\Pi_{h}$. 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 $\Pi_{\mathrm{S}}$ (source routing with semi-bandit feedback). In other words, we cannot improve the performance by taking routing decisions at each hop.

Theorem 5.3. For all $\theta$ and for any uniformly good rule $\pi \in \Pi_{h}$,

$$
\liminf _{N \rightarrow \infty} \frac{R^{\pi}(N)}{\log (N)} \geq c_{\mathrm{h}}(\theta)=c_{\mathrm{s}}(\theta)
$$



Figure 5.2: Line Topology

For the proof of this theorem, we refer to [72]. As shown in [20, Theorem 2], the asymptotic regret lower bounds derived in Theorems 5.2 5.1 5.3 are tight in the sense that one can design actual routing policies achieving these regret bounds (although these policies might well be extremely complex to compute and impractical to implement). Hence from the fact that $c_{\mathrm{b}}(\theta) \geq c_{\mathrm{s}}(\theta)=c_{\mathrm{h}}(\theta)$, we conclude that:

- The best source routing policy with semi-bandit feedback asymptotically achieves a lower regret than the best source routing policy with bandit feedback;
- The best hop-by-hop routing policy asymptotically obtains the same regret as the best source routing policy with semi-bandit feedback.


### 5.3.4 Line Networks

We now consider shortest-path routing in a line network whose topology is shown in Figure 5.2 Define

$$
\mathcal{I}=\{F \subset E: \text { every pair of elements in } F \text { share at most one vertex }\} .
$$

It can be easily verified that $G(\mathcal{G})=(E, \mathcal{I})$ is a matroid, and that each basis of $G$ is a path between the left-most and the right-most vertices of $\mathcal{G}$. As a consequence, we have the following lower bound on the regret of any uniformly good policy in $\Pi_{\mathrm{s}}$ in line networks:

Corollary 5.1. For all $\theta \in \Theta$ and for any uniformly good policy $\pi \in \Pi_{\mathrm{s}}$ in line networks,

$$
\liminf _{N \rightarrow \infty} \frac{R^{\pi}(N)}{\log (N)} \geq \sum_{i \in E \backslash M^{\star}} \frac{1}{\operatorname{KLG}\left(\theta_{i}, \theta_{\sigma(i)}\right)}\left(\frac{1}{\theta_{i}}-\frac{1}{\theta_{\sigma(i)}}\right) .
$$

The proof of this result is similar to that of Theorem 4.2; it is thus omitted.

### 5.4 Algorithms

In this section, we present online routing policies for semi-bandit feedback, which are simple to implement and yet approach the performance limits identified in the

| Index | Type | Computation | Algorithm |
| :---: | :---: | :---: | :---: |
| $b_{M}$ | Path | Line search | GeoCombUCB-1 |
| $c_{M}$ | Path | Explicit | GEOCoMBUCB-2 |
| $\omega_{i}$ | Edge | Line search | KL-SR |

Table 5.3: Summary of indexes.
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 define $s_{i}(n)$ as the number of packets routed through link $i$ before the $n$-th packet is sent. 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 $(n-1)$ first packets; namely $\hat{\theta}_{i}(n)=s_{i}(n) / t_{i}(n)$ if $t_{i}(n)>0$ and $\hat{\theta}_{i}(n)=0$ otherwise. We define the corresponding vectors $t(n)=\left(t_{i}(n)\right)_{i \in E}, s(n)=\left(s_{i}(n)\right)_{i \in E}$, and $\hat{\theta}(n)=\left(\hat{\theta}_{i}(n)\right)_{i \in E}$.

### 5.4.1 Path Indexes

The proposed policies rely on indexes attached to individual paths. Next we introduce two indexes used in our policies. They depend on the round, i.e., on the number $n$ of packets already sent, and on the estimated link parameters $\hat{\theta}(n)$. The two indexes and their properties (i.e., in which policy they are used, and how one can compute them) are summarized in Table 5.3. Let $n \geq 1$ and assume that $n$-th packet is to be sent. The indexes are defined as follows.

The first index, denoted by $b_{M}(n, \hat{\theta}(n))$ for path $M \in \mathcal{M}$, or for short $b_{M}(n)$, is motivated by the index defined in Chapter 3 and is defined as the optimal value of the following optimization problem:

$$
\begin{aligned}
\inf _{u \in(0,1]^{d}} & M^{\top} u^{-1} \\
\text { subject to: } & \sum_{i \in M} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), u_{i}\right) \leq f_{1}(n),
\end{aligned}
$$

where $f_{1}(n)=\log (n)+4 m \log (\log (n))$.
The second index is denoted by $c_{M}(n, \hat{\theta}(n))$, or for short $c_{M}(n)$, and explicitly defined for path $M \in \mathcal{M}$ as:

$$
c_{M}(n)=M^{\top} \hat{\theta}(n)^{-1}-\sqrt{\frac{f_{1}(n)}{2} \sum_{i \in M} \frac{1}{s_{i}(n) \hat{\theta}_{i}(n)^{3}}} .
$$

The next theorem provides generic properties of the two indexes $b_{M}$ and $c_{M}$.
Theorem 5.4. (i) For all $n \geq 1, M \in \mathcal{M}$, and $\lambda \in(0,1]^{d}$, we have $b_{M}(n, \lambda) \geq$ $c_{M}(n, \lambda)$. (ii) There exists a constant $K_{m}>0$ depending on $m$ only such that, for

## Algorithm 5.1 GeoCombUCB

for $n \geq 1$ do
Select path $M(n) \in \operatorname{argmin}_{M \in \mathcal{M}} \xi_{M}(n)$ (ties are broken arbitrarily), where $\xi_{M}(n)=b_{M}(n)$ for GeoCombUCB-1, and $\xi_{M}(n)=c_{M}(n)$ for GeoCombUCB-2.

Collect feedbacks on links $i \in M(n)$, and update $\hat{\theta}_{i}(n)$ for $i \in M(n)$.
end for
all $M \in \mathcal{M}$ and $n \geq 2$ :

$$
\mathbb{P}\left[b_{M}(n, \hat{\theta}(n)) \geq M^{\top} \theta\right] \leq K_{m} n^{-1}(\log (n))^{-2} .
$$

Corollary 5.2. We have:

$$
\sum_{n \geq 1} \mathbb{P}\left[b_{M^{\star}}(n, \hat{\theta}(n)) \geq M^{\star \top} \theta^{-1}\right] \leq 1+K_{m} \sum_{n \geq 2} n^{-1}(\log (n))^{-2}:=K_{m}^{\prime}<\infty .
$$

### 5.4.2 The GeoCombUCB Algorithm

We present two routing policies, referred to as GeoCombUCB-1 and GeoCombUCB-2, respectively. For the transmission of the $n$-th packet, GeoCombUCB-1 (resp. GeoCombUCB-2) selects the path $M$ with the lowest index $b_{M}(n)$ (resp. $c_{M}(n)$ ). The pseudo-code of GeoCombUCB-1 and GeoCombUCB-2 algorithms is presented in Algorithm 5.1

In the following theorem, we provide a finite-time analysis of the GeoCombUCB algorithm.

Theorem 5.5. There exists a constant $K_{m}^{\prime} \geq 0$ that only depends on $m$ such that for every $\delta \in(0,1)$, the regret under policy $\pi \in\{$ GeoCombUCB-1, GeoCombUCB-2\} satisfies for any $N$ :

$$
R^{\pi}(N) \leq \frac{4(1+\delta)^{2}}{(1-\delta)^{5}} \cdot \frac{d \sqrt{m} f_{1}(N)}{\Delta_{\min } \theta_{\min }^{3}}+m \theta_{\min }^{-1}\left(K_{m}^{\prime}+\frac{2}{\delta^{2} \min \left(\frac{\Delta_{\min }}{M^{*} \theta^{-1}}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2}\right) .
$$

Hence, $R^{\pi}(N)=\mathcal{O}\left(\frac{d \sqrt{m}}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ when $N \rightarrow \infty$.

### 5.4.3 Improved Regret Bound for The KL-SR Algorithm

We now present an improved regret upper bound for the KL-SR algorithm. KL-SR, initially proposed in [30, relies on an index attached to links. More precisely, for each link $i \in E$, KL-SR maintains index $\omega_{i}$ defined for the $n$-th packet as:

$$
\omega_{i}(n)=\min \left\{\frac{1}{q}: q \in\left[\hat{\theta}_{i}(n), 1\right], t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), q\right) \leq f_{2}(n)\right\},
$$

```
Algorithm 5.2 KL-SR
    for \(n \geq 1\) do
        Select path \(M(n) \in \operatorname{argmin}_{\in \mathcal{M}} M^{\top} \omega(n)\) (ties are broken arbitrarily).
        Collect feedbacks on links \(i \in M(n)\), and update \(\hat{\theta}_{i}(n)\) for \(i \in M(n)\).
    end for
```

where $f_{2}(n)=\log (n)+3 \log (\log (n))$. For the transmission of the $n$-th packet, KL-SR selects the path $M(n) \in \operatorname{argmin}_{M \in \mathcal{M}} M^{\top} \omega(n)$, where $\omega(n)=\left(\omega_{i}(n)\right)_{i \in E}$. The pseudo-code of KL-SR is presented in Algorithm 5.2

In the following theorem, we provide an improved finite-time analysis of the KL-SR algorithm.

Theorem 5.6. There exists a constant $K^{\prime \prime} \geq 0$ such that for every $\delta \in(0,1)$, the regret under policy $\pi=K L-S R$ satisfies for any $N$ :

$$
R^{\pi}(N) \leq \frac{45(1+\delta)^{2}}{(1-\delta)^{5}} \cdot \frac{m d f_{2}(N)}{\Delta_{\min } \theta_{\min }^{3}}+m \theta_{\min }^{-1}\left(K^{\prime \prime}+\frac{2}{\delta^{2} \min \left(\frac{\Delta_{\min }}{M^{*} \theta^{-1}}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2}\right) .
$$

Hence, $R^{\pi}(N)=\mathcal{O}\left(\frac{m d}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ when $N \rightarrow \infty$.

Remark 5.2. Theorem 5.6 holds even when the delays on the various links can be arbitrarily correlated as considered in [56, 57].

GeoCombUCB and KL-SR have better performance guarantees than existing routing algorithms. Indeed, as shown in Appendix 5.F the best regret upper bound for the CUCB algorithm derived in [56] is

$$
R^{\mathrm{CUCB}}(N)=\mathcal{O}\left(\frac{d m^{2}}{\Delta_{\min } \theta_{\min }^{4}} \log (N)\right)
$$

We believe that applying the proof techniques presented in 57] (see the proof of Theorem 5 there), one might provide a regret upper bound for CUCB scaling as $\mathcal{O}\left(\frac{d m}{\Delta_{\min } \theta_{\min }^{4}} \log (N)\right)$, which still 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.

### 5.4.4 Implementation

Next we discuss the implementation of our routing policies and KL-SR, and in particular propose simple methods to compute the various indexes involved in these policies. Note first that the path index $c_{M}$ is explicit, and easy to compute. The link index $\omega_{i}$ is also straightforward as it amounts to finding the roots of a strictly
convex and increasing function in one variable (note that $v \mapsto \mathrm{kl}(u, v)$ is strictly convex and increasing for $v \geq u$ ). Hence, the index $\omega_{i}$ can be computed by a simple line search. The path index $b_{M}$ can also be computed using a slightly more complicated line search, as shown in the following proposition.

For $\lambda>0, w \in[0,1]$, and $v \in \mathbb{N}$ define:

$$
g(\lambda, w, v)=\frac{\lambda v w-1+\sqrt{(1-\lambda v w)^{2}+4 \lambda v}}{2 \lambda v} .
$$

Fix $n \geq 1, M \in \mathcal{M}, \hat{\theta}(n)$, and $t(n)$. Define $I=\left\{i \in M: \hat{\theta}_{i}(n) \neq 1\right\}$, and for $\lambda>0$, define:

$$
F(\lambda)=\sum_{i \in I} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), g\left(\lambda, \hat{\theta}_{i}(n), t_{i}(n)\right)\right) .
$$

Proposition 5.1. (i) $\lambda \mapsto F(\lambda)$ is strictly increasing, and $F\left(\mathbb{R}^{+}\right)=\mathbb{R}^{+}$. (ii) If $I=$ $\emptyset, b_{M}(n)=\sum_{i \in E} M_{i}$. Otherwise, $b_{M}(n)=\sum_{i \in E} M_{i}-|I|+\sum_{i \in I} g\left(\lambda^{\star}, \hat{\theta}_{i}(n), t_{i}(n)\right)$, where $\lambda^{\star}$ is the unique solution to $F(\lambda)=f_{1}(n)$.

As stated in Proposition 5.1, $\lambda^{\star}$ can be computed efficiently by a simple line search, and $b_{M}(n)$ 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 BellmanFord 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 $\mathcal{O}(|\mathcal{M}|)$ in each round.

## 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 which is a hop-by-hop version of the KL-SR algorithm and hence belongs to the set of policies $\Pi_{h}$. We first introduce the necessary notations. For any node $v \in V$, we let $\mathcal{M}_{v}$ denote the set of loop-free paths from node $v$ to the destination. For any time slot $\tau$, 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 $\tau$, that is $\tilde{\theta}_{i}(\tau)=s_{i}(n(\tau)) / t_{i}^{\prime}(\tau)$, where $t_{i}^{\prime}(\tau)$ denotes the total number of transmission attempts on link $i$ up to time slot $\tau$. Moreover, with slight abuse of notation, we denote the index of link $i$ at time $\tau$ by $\omega_{i}\left(\tau, \tilde{\theta}_{i}(\tau)\right)$. Note that by definition $t_{i}^{\prime}(\tau) \geq t_{i}(n)$ and $\tilde{\theta}_{i}(\tau)$ is a more accurate estimate of $\theta_{i}$ than $\hat{\theta}_{i}(n(\tau))$.

We define $J_{v}(\tau)$ as the minimum cumulative index from node $v$ to the destination:

$$
J_{v}(\tau)=\min _{M \in \mathcal{M}_{v}} \sum_{i \in M} \omega_{i}\left(\tau, \tilde{\theta}_{i}(\tau)\right) .
$$

We note that $J_{v}(\tau)$ can be computed using the Bellman-Ford algorithm. KL-HHR works based on the following idea: at time $\tau$ if the current packet is at node $v$, it will be sent to node $v^{\prime}$ with $\left(v, v^{\prime}\right) \in E$ such that $\omega_{\left(v, v^{\prime}\right)}\left(\tau, \tilde{\theta}_{v}(\tau)\right)+J_{v^{\prime}}(\tau)$ is minimal over all outgoing links of node $v$. The pseudo-code of KL-HHR is given in Algorithm 5.3

```
Algorithm 5.3 KL-HHR for node \(v\)
    for \(\tau \geq 1\) do
        Select link \(\left(v, v^{\prime}\right) \in E\), where
```

            \(v^{\prime} \in \arg \min _{w \in V:(v, w) \in E}\left(\omega_{(v, w)}\left(\tau, \tilde{\theta}_{v}(\tau)\right)+J_{w}(\tau)\right)\).
        Update index of the link \(\left(v, v^{\prime}\right)\).
    end for
    The regret analysis of $\mathrm{KL}-\mathrm{HHR}$ is more complicated than that of $\mathrm{KL}-\mathrm{SR}$ and is let for future work.

## Line Networks

We now revisit the case of line networks. Recall that in a line network, each path between the left-most and the right-most vertices of $\mathcal{G}$ may be seen as a basis of a matroid. The following corollary shows that KL-SR is an asymptotically optimal routing policy in line networks:

Corollary 5.3. For any $\varepsilon>0$, there exist positive constants $C_{1}, C_{2}(\varepsilon)$, and $\beta(\varepsilon)$ such that the regret under algorithm $\pi=\mathrm{KL}-\mathrm{SR}$ in line networks satisfies:
$R^{\pi}(N) \leq \sum_{i \in E \backslash M^{\star}} \frac{(1+\varepsilon) \log (N)}{\operatorname{KLG}\left(\theta_{i}, \theta_{\sigma(i)}\right)}\left(\frac{1}{\theta_{i}}-\frac{1}{\theta_{\sigma(i)}}\right)+(d-m)\left(\frac{m C_{1}}{\theta_{\min }} \log (\log (N))+\frac{C_{2}(\varepsilon)}{N^{\beta(\varepsilon)}}\right)$.
For line networks we have that $\Delta_{\min }=\min _{i \notin M^{*}}\left(\theta_{i}^{-1}-\theta_{\sigma(i)}^{-1}\right)$. Applying Pinsker's inequality and Lemma B. 3 together give

$$
\begin{aligned}
\sum_{i \in E \backslash M^{\star}} \frac{\frac{1}{\theta_{i}}-\frac{1}{\theta_{\sigma(i)}}}{\operatorname{KLG}\left(\theta_{i}, \theta_{\sigma(i)}\right)} & \leq \sum_{i \in E \backslash M^{\star}} \frac{1}{2 \theta_{\sigma(i)}\left(\theta_{\sigma(i)}-\theta_{i}\right)} \\
& \leq \sum_{i \in E \backslash M^{\star}} \frac{1}{2 \theta_{\sigma(i)}^{2} \theta_{i}}\left(\frac{1}{\theta_{i}}-\frac{1}{\theta_{\sigma(i)}}\right)^{-1} \\
& \leq \frac{1}{2 \min _{i \notin M^{\star}} \theta_{\sigma(i)}^{2}} \cdot \frac{d-m}{\Delta_{\min } \theta_{\min }}
\end{aligned}
$$

Let us consider a problem instance for routing in line networks in which $\theta_{i}=$ $1, \forall i \in M^{\star}$. Then,

$$
\limsup _{N \rightarrow \infty} \frac{R^{\mathrm{KL}-\mathrm{SR}}(N)}{\log (N)} \leq \frac{d-m}{2 \Delta_{\min } \theta_{\min }}
$$

which implies a regret scaling of $\mathcal{O}\left((d-m) \Delta_{\min }^{-1} \theta_{\min }^{-1} \log (N)\right)$.
The proof of Corollary 5.3 is similar to that of Theorem 4.4, it is thus omitted.

### 5.5 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 [56] applied to our online routing problem. The CUCB algorithm is an index policy in $\Pi_{\mathrm{s}}$ (set of source routing policies with semi-bandit feedback), and selects path $M(n)$ for the transmission of the $n$-th packet:

$$
M(n) \in \arg \min _{M \in \mathcal{M}} \sum_{i \in M} \frac{1}{\hat{\theta}_{i}(n)+\sqrt{1.5 \log (n) / t_{i}(n)}} .
$$

We consider a grid network whose topology is depicted in Figure 5.1 a), where the node in red (resp. blue) is the source (resp. destination). In this network, there are $\binom{6}{3}=20$ possible paths from the source to the destination.

In Figure 5.3 (a)-(c), we plot the regret against the number of the packets $N$ under the various routing policies, and for three sets of link parameters $\theta$. For each set, we choose a value of $\theta_{\min }$ and generate the values of $\theta_{i}$ independently, uniformly at random in $\left[\theta_{\min }, 1\right]$. 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 yields the smallest regret. The comparison between GeoCombUCB-2 and KL-SR is more subtle and depends on the links parameters. KL-SR seems to perform better than GeoCombUCB-2 in scenarios where $\Delta_{\text {min }}$ is large.

### 5.6 Summary

In this chapter we investigated stochastic combinatorial MABs with geometrically distributed rewards. 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 [30]. 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,


Figure 5.3: Regret versus number of received packets.
through numerical experiments we demonstrated that these policies outperform state-of-the-art algorithms in practice.

## 5.A Proof of Theorem 5.2

The proof of this theorem is quite similar to that of Theorem 3.3 The following lemma however is required.

Lemma 5.1. Consider $\left(X_{i}\right)_{i}$ independent with $X_{i} \sim \operatorname{Geo}\left(\theta_{i}\right)$ and $0<\theta_{i} \leq 1$. Consider $\left(Y_{i}\right)_{i}$ independent with $Y_{i} \sim \operatorname{Geo}\left(\lambda_{i}\right)$ and $0<\lambda_{i} \leq 1$. Define $\bar{X}=\sum_{i} X_{i}$ and $\bar{Y}=\sum_{i} Y_{i}$. Then $X \stackrel{d}{=} Y$ if and only if $\left(\theta_{i}\right)_{i}=\left(\lambda_{i}\right)_{i}$ up to a permutation ${ }^{3}$.

Proof. If $\left(\theta_{i}\right)_{i}=\left(\lambda_{i}\right)_{i}$, up to a permutation then $X \stackrel{d}{=} Y$ by inspection. Assume that $X \stackrel{d}{=} Y$. Define $z_{m}=\min _{i} \min \left(1 /\left(1-\theta_{i}\right), 1 /\left(1-\lambda_{i}\right)\right)$. For all $z$ such that $|z|<z_{m}$ we have $\mathbb{E}\left[z^{\bar{X}}\right]=\mathbb{E}\left[z^{\bar{Y}}\right]$ so that

$$
\prod_{i} \frac{\theta_{i}}{1-\left(1-\theta_{i}\right) z}=\prod_{i} \frac{\lambda_{i}}{1-\left(1-\lambda_{i}\right) z} .
$$

Hence:

$$
P_{X}(z):=\prod_{i} \theta_{i}\left(1-\left(1-\lambda_{i}\right) z\right)=\prod_{i} \lambda_{i}\left(1-\left(1-\theta_{i}\right) z\right):=P_{Y}(z) .
$$

Both $P_{X}(z)$ and $P_{X}(z)$ are polynomials and are equal on an open set. So they are equal everywhere, and the sets of their roots are equal $\left\{1 /\left(1-\theta_{i}\right), i\right\}=\{1 /(1-$ $\left.\left.\lambda_{i}\right), i\right\}$. Hence, $\left(\theta_{i}\right)_{i}=\left(\lambda_{i}\right)_{i}$ up to a permutation as announced.

## 5.B Proof of Theorem 5.4

Proof. Statement (i): Let $M \in \mathcal{M}, u, \lambda \in(0,1]^{d}$, and $s(n) \in \mathbb{R}^{d}$. Define $t_{i}(n)=$ $s_{i}(n) / \lambda_{i}$ for any $i$. By convexity of $u \mapsto M^{\top} u^{-1}$, we have:

$$
M^{\top} \lambda^{-1}-M^{\top} u^{-1} \leq \sum_{i \in M} \frac{u_{i}-\lambda_{i}}{\lambda_{i}^{2}}
$$

By Cauchy-Schwarz and Pinsker's inequalities, we have that

$$
\sum_{i \in M} \frac{u_{i}-\lambda_{i}}{\lambda_{i}^{2}} \sqrt{\frac{s_{i}(n)}{s_{i}(n)}} \leq \sqrt{\sum_{i \in M} t_{i}(n)\left(\lambda_{i}-u_{i}\right)^{2}} \sqrt{\sum_{i \in M} \frac{1}{s_{i}(n) \lambda_{i}^{3}}}
$$

[^15]$$
\leq \sqrt{\sum_{i \in M} \frac{t_{i}(n) \mathrm{kl}\left(\lambda_{i}, u_{i}\right)}{2}} \sqrt{\sum_{i \in M} \frac{1}{s_{i}(n) \lambda_{i}^{3}}} .
$$

Hence,

$$
M^{\top} \lambda^{-1}-M^{\top} u^{-1} \leq \sqrt{\sum_{i \in M} \frac{t_{i}(n) \mathrm{kl}\left(\lambda_{i}, u_{i}\right)}{2}} \sqrt{\sum_{i \in M} \frac{1}{s_{i}(n) \lambda_{i}^{3}}}
$$

Thus, $\sum_{i \in M} t_{i}(n) \operatorname{kl}\left(\lambda_{i}, u_{i}\right) \leq f_{1}(n)$ implies:

$$
M^{\top} u^{-1} \geq M^{\top} \lambda^{-1}-\sqrt{\frac{f_{1}(n)}{2} \sum_{i \in M} \frac{1}{s_{i}(n) \lambda_{i}^{3}}}=c_{M}(n, \lambda)
$$

so that, by definition of $b_{M}(n, \lambda)$, we have $b_{M}(n, \lambda) \geq c_{M}(n, \lambda)$.
Statement (ii): If $\sum_{i \in M} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right) \leq f_{1}(n)$, then we have $b_{M}(n) \leq M^{\top} \theta^{-1}$ by definition of ${ }_{M}(n)$. Therefore, using Lemma A.4 there exists $K_{m}$ such that for all $n \geq 2$ we have:

$$
\mathbb{P}\left[b_{M}(n)>M^{\top} \theta^{-1}\right] \leq \mathbb{P}\left[\sum_{i \in M} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f_{1}(n)\right] \leq K_{m} n^{-1}(\log (n))^{-2}
$$

which concludes the proof.

## 5.C Proof of Theorem 5.5

Proof. For any $n \in \mathbb{N}, w \in \mathbb{N}^{d}, M \in \mathcal{M}$, and $\lambda \in(0,1]^{d}$ define

$$
h_{n, w, M, \lambda}=\sqrt{f_{1}(n) \sum_{i \in p} \frac{1}{2 w_{i} \lambda_{i}^{3}}} .
$$

Fix $\delta \in(0,1)$ and define $\varepsilon=\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)$ with $\mu^{\star}=M^{\star} \theta^{\top} \theta^{-1}$. For any $n$, introduce the following events:

$$
\begin{aligned}
A_{n} & =\left\{\sum_{i \in M^{\star}} t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f_{1}(n)\right\}, \\
B_{n, i} & =\left\{M_{i}(n)=1,\left|\hat{\theta}_{i}(n)-\theta_{i}\right| \geq \varepsilon \theta_{i}\right\}, \quad B_{n}=\bigcup_{i \in E} B_{n, i}, \\
F_{n} & =\left\{\Delta_{M(n)} \leq(1+\varepsilon) h_{N, s(n), M(n),(1-\varepsilon) \theta_{\min } 1}+\varepsilon \mu^{\star}\right\},
\end{aligned}
$$

where 1 is a vector all of whose elements are one.
Let $N>0$. Then the regret can be upper bounded as:

$$
R^{\pi}(N)=\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)}\right]
$$

$$
\begin{aligned}
& \leq \mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)}\left(\mathbb{1}\left\{A_{n}\right\}+\mathbb{1}\left\{B_{n}\right\}\right)\right]+\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}\right\}\right] \\
& \left.\leq m \theta_{\min }^{-1} \sum_{n=1}^{N}\left(\mathbb{P}\left[A_{n}\right]+\mathbb{P}\left[B_{n}\right]\right)\right]+\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}\right\}\right]
\end{aligned}
$$

where the last inequality follows from $\Delta_{M(n)} \leq m \theta_{\min }^{-1}$. Consider $n$ such that $M(n) \neq M^{\star}$. Next we show that $\overline{A_{n} \cup B_{n}} \subset F_{n}$. Recall that $c_{M}(n) \leq b_{M}(n)$ for any $n$ and $p$ (Theorem 5.4). Moreover, $\sum_{i \in M^{\star}} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right) \leq f_{1}(n)$ implies: $b_{M^{\star}}(n) \leq \mu^{\star}$. For brevity, let us define $h(n)=h_{n, s(n), M(n), \hat{\theta}(n)}$ and $h^{\prime}(n)=$ $h_{n, s(n), M(n),(1-\varepsilon) \theta_{\min } \mathbf{1}}$. Hence we have:

$$
\begin{align*}
\mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}, M(n) \neq M^{\star}\right\} & =\mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}, \xi_{M(n)}(n) \leq \xi_{M^{\star}}(n)\right\} \\
& \leq \mathbb{1}\left\{\overline{B_{n}}, c_{M(n)}(n) \leq \mu^{\star}\right\} \\
& =\mathbb{1}\left\{\overline{B_{n}}, M(n)^{\top} \hat{\theta}(n)^{-1}-h(n) \leq \mu^{\star}\right\} \\
& \leq \mathbb{1}\left\{(1+\varepsilon)^{-1} M(n)^{\top} \theta^{-1}-h^{\prime}(n) \leq \mu^{\star}\right\}  \tag{5.4}\\
& \leq \mathbb{1}\left\{(1+\varepsilon)^{-1} \Delta_{M(n)} \leq h^{\prime}(n)+\varepsilon \mu^{\star} /(1+\varepsilon)\right\} \\
& \leq \mathbb{1}\left\{\Delta_{M(n)} \leq h_{N, s(n), M(n),(1-\varepsilon) \theta_{\min } 1}+\varepsilon \mu^{\star}\right\} \\
& =\mathbb{1}\left\{F_{n}\right\},
\end{align*}
$$

where in 5.4 we use the fact that $\overline{B_{n}}$ implies for any $i \in M(n): \hat{\theta}_{i}(n)^{-1}<$ $(1-\varepsilon)^{-1} \theta_{i}^{-1} \leq(1-\varepsilon)^{-1} \theta_{\min }^{-1}$. Hence, the regret is upper bounded by:

$$
R^{\pi}(N) \leq m \theta_{\min }^{-1} \sum_{n=1}^{N}\left(\mathbb{P}\left[A_{n}\right]+\mathbb{P}\left[B_{n}\right]\right)+\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right] .
$$

We will prove the following inequalities:
(i) $\sum_{n=1}^{N} \mathbb{P}\left[B_{n}\right] \leq \frac{2}{\delta^{2} \min \left(\Delta_{\min } / \mu^{\star}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2}$,
(ii) $\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right] \leq \frac{4(1+\delta)^{2}}{(1-\delta)^{5}} d \sqrt{m} \Delta_{\text {min }}^{-1} \theta_{\min }^{-3} f_{1}(N)$.

Moreover, Corollary 5.2 implies that: $\sum_{n=1}^{N} \mathbb{P}\left[A_{n}\right] \leq K_{m}^{\prime}$, with $K_{m}^{\prime}$ depending only on $m$. Putting these together, we obtain the announced result

$$
\begin{aligned}
R^{\pi}(N) & \leq \frac{4(1+\delta)^{2}}{(1-\delta)^{5}} d \sqrt{m} \Delta_{\min }^{-1} \theta_{\min }^{-3} f_{1}(N) \\
& +m \theta_{\min }^{-1}\left(K_{m}^{\prime}+\frac{2}{\delta^{2} \min \left(\Delta_{\min } / \mu^{\star}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2}\right) .
\end{aligned}
$$

Inequality (i): Fix $i$ and $n$. Define $\tau=\sum_{n^{\prime}=1}^{n} \mathbb{1}\left\{B_{n^{\prime}, i}\right\}$. Observe that $B_{n^{\prime}, i}$ implies $M_{i}\left(n^{\prime}\right)=1$, hence $s_{i}(n) \geq \tau$. Therefore, applying [61, Lemma B.1], we
have that $\sum_{n=1}^{N} \mathbb{P}\left[B_{n, i}\right] \leq 2 \varepsilon^{-2} \theta_{i}^{-2}$. Applying union bound, we get

$$
\sum_{n=1}^{N} \mathbb{P}\left[B_{n}\right] \leq 2 \varepsilon^{-2} \sum_{i \in E} \theta_{i}^{-2}=\frac{2}{\delta^{2} \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2} .
$$

Inequality (ii): Let $\ell>0$. For any $n$, define $U=\frac{(1+\varepsilon)^{2} f_{1}(N)}{(1-\varepsilon)^{3} \theta_{\text {min }}^{3}}$ and introduce the following events:

$$
\begin{aligned}
S_{n} & =\left\{i \in M(n): s_{i}(n) \leq m U\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{-2}\right\} \\
G_{n} & =\left\{\left|S_{n}\right| \geq \ell\right\} \\
L_{n} & =\left\{\left|S_{n}\right|<\ell,\left[\exists i \in M(n): s_{i}(n) \leq \ell U\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{-2}\right]\right\}
\end{aligned}
$$

We claim that for any $n$ such that $M(n) \neq M^{\star}$, we have $F_{n} \subset\left(G_{n} \cup L_{n}\right)$. To prove this, we show that when $F_{n}$ holds and $M(n) \neq M^{\star}$, the event $\overline{G_{n} \cup L_{n}}$ cannot happen. Let $n$ be such that $M(n) \neq M^{\star}$ and $F_{n}$ holds, and assume that

$$
\overline{G_{n} \cup L_{n}}=\left\{\left|S_{n}\right|<\ell,\left[\forall i \in M(n): s_{i}(n)>\ell U\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{-2}\right]\right\}
$$

happens. Then $F_{n}$ implies:

$$
\begin{align*}
\Delta_{M(n)} & \leq(1+\varepsilon) h_{N, s(n), M(n),(1-\varepsilon) \theta_{\min } 1}+\varepsilon \mu^{\star} \\
& =\varepsilon \mu^{\star}+(1+\varepsilon) \sqrt{\frac{f_{1}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}}} \sqrt{\sum_{i \in M(n) \backslash S_{n}} \frac{1}{s_{i}(n)}+\sum_{i \in S_{n}} \frac{1}{s_{i}(n)}} \\
& <\varepsilon \mu^{\star}+(1+\varepsilon) \sqrt{\frac{f_{1}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}}} \sqrt{\frac{m\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{2}}{m U}+\frac{\left|S_{n}\right|\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{2}}{\ell U}} \\
& <\varepsilon \mu^{\star}+(1+\varepsilon) \sqrt{\frac{f_{1}(N)\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{2}}{(1-\varepsilon)^{3} \theta_{\min }^{3} U}} \\
& =\Delta_{M(n)}, \tag{5.5}
\end{align*}
$$

where the last inequality follows from the observation that $\overline{G_{n} \cup L_{n}}$ implies $\left|S_{n}\right|<\ell$. Clearly, 5.5 is a contradiction. Thus $F_{n} \subset\left(G_{n} \cup L_{n}\right)$ and consequently:

$$
\begin{equation*}
\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\} \leq \sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{G_{n}\right\}+\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{L_{n}\right\} \tag{5.6}
\end{equation*}
$$

To further bound the r.h.s. of the above, we introduce the following events for any $i$ :

$$
\begin{aligned}
G_{i, n} & =G_{n} \cap\left\{i \in M(n), s_{i}(n)\right. \\
L_{i, n} & =L_{n} \cap\left\{i \in M(n), s_{i}(n) \leq \ell U\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{-2}\right\}, \\
& \left.\left.\leq \varepsilon \mu^{\star}\right)^{-2}\right\} .
\end{aligned}
$$

Note that $\sum_{i \in E} \mathbb{1}\left\{G_{i, n}\right\}=\mathbb{1}\left\{G_{n}\right\} \sum_{i \in E} \mathbb{1}\left\{i \in S_{n}\right\}=\left|S_{n}\right| \mathbb{1}\left\{G_{n}\right\} \geq \ell \mathbb{1}\left\{G_{n}\right\}$, and hence we have $\mathbb{1}\left\{G_{n}\right\} \leq \frac{1}{\ell} \sum_{i \in E} \mathbb{1}\left\{G_{i, n}\right\}$. Moreover $\mathbb{1}\left\{L_{n}\right\} \leq \sum_{i \in E} \mathbb{1}\left\{L_{i, n}\right\}$. Suppose each link $i$ belongs to $K_{i}$ sub-optimal paths, and these sub-optimal paths are ordered according to the gap between their average delays and $\mu^{\star}$ as: $\Delta^{i, 1} \geq \cdots \geq$ $\Delta^{i, K_{i}}>0$. Also define $\Delta^{i, 0}=\infty$.

Plugging the above inequalities into (5.6), we obtain

$$
\begin{align*}
\sum_{n=1}^{N} \Delta_{M(n)} & \mathbb{1}\left\{F_{n}\right\} \leq \sum_{n=1}^{N} \sum_{i \in E} \Delta_{M(n)}\left(\frac{1}{\ell} \mathbb{1}\left\{G_{i, n}\right\}+\mathbb{1}\left\{L_{i, n}\right\}\right) \\
& =\sum_{n=1}^{N} \sum_{i \in E} \Delta_{M(n)}\left(\frac{1}{\ell} \mathbb{1}\left\{G_{i, n}, M(n) \neq M^{\star}\right\}+\mathbb{1}\left\{L_{i, n}, M(n) \neq M^{\star}\right\}\right) \\
& \leq \sum_{n=1}^{N} \sum_{i \in E} \sum_{k \in\left[K_{i}\right]} \Delta^{i, k}\left(\frac{1}{\ell} \mathbb{1}\left\{G_{i, n}, M(n)=k\right\}+\mathbb{1}\left\{L_{i, n}, M(n)=k\right\}\right) \\
& \leq \sum_{i \in E} \sum_{n=1}^{N} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{\ell} \mathbb{1}\left\{i \in M(n), s_{i}(n) \leq \frac{m U}{\left(\Delta^{i, k}-\varepsilon \mu^{\star}\right)^{2}}, M(n)=k\right\} \\
& +\sum_{i \in E} \sum_{n=1}^{N} \sum_{k \in\left[K_{i}\right]} \Delta^{i, k} \mathbb{1}\left\{i \in M(n), s_{i}(n) \leq \frac{\ell U}{\left(\Delta^{i, k}-\varepsilon \mu^{\star}\right)^{2}}, M(n)=k\right\} \\
& \leq \sum_{i \in E} \sum_{n=1}^{N} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{\ell} \mathbb{1}\left\{s_{i}(n) \leq \frac{1}{\left(1-\varepsilon \mu^{\star} / \Delta_{\text {min }}\right)^{2}} \frac{m U}{\left(\Delta^{i, k}\right)^{2}}, M(n)=k\right\} \\
& +\sum_{i \in E} \sum_{n=1}^{N} \sum_{k \in\left[K_{i}\right]} \Delta^{i, k_{1}} \mathbb{1}\left\{s_{i}(n) \leq \frac{1}{\left(1-\varepsilon \mu^{\star} / \Delta_{\text {min }}\right)^{2}} \frac{\ell U}{\left(\Delta^{i, k}\right)^{2}}, M(n)=k\right\} \\
& \leq \frac{2 U d}{\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2} \Delta_{\min }}\left(\frac{m}{\ell}+\ell\right), \tag{5.7}
\end{align*}
$$

where the second last inequality uses the observation that we have for any $i$ and $k$ :

$$
\Delta^{i, k}-\varepsilon \mu^{\star}=\Delta^{i, k}\left(1-\varepsilon \mu^{\star} / \Delta^{i, k}\right) \geq \Delta^{i, k}\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)
$$

and the last inequality follows from the fact that for any $i \in E$ with $K_{i} \geq 1$ and $C>0$ that does not depend on $n$, we have from Lemma 3.1

$$
\sum_{n=1}^{N} \sum_{k=1}^{K_{i}} \mathbb{1}\left\{s_{i}(n) \leq C\left(\Delta^{i, k}\right)^{-2}, M(n)=k\right\} \Delta^{i, k} \leq \frac{2 C}{\Delta_{\min }}
$$

Setting $\ell=\sqrt{m}$ in 5.7) yields

$$
\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\} \leq \frac{4(1+\varepsilon)^{2}}{(1-\varepsilon)^{3}\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2}} \cdot \frac{d \sqrt{m} f_{1}(N)}{\Delta_{\min } \theta_{\min }^{3}}
$$

The proof of inequality (iii) is completed by observing that

$$
\begin{aligned}
\frac{(1+\varepsilon)^{2}}{(1-\varepsilon)^{3}\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2}} & =\frac{\left(1+\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)\right)^{2}}{\left(1-\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)\right)^{3}\left(1-\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right) \frac{\mu^{\star}}{\Delta_{\min }}\right)^{2}} \\
& \leq \frac{(1+\delta)^{2}}{(1-\delta)^{3}\left(1-\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right) \frac{\mu^{\star}}{\Delta_{\min }}\right)^{2}} \leq \frac{(1+\delta)^{2}}{(1-\delta)^{5}},
\end{aligned}
$$

where the last inequality follows from $\min (z, 1) \leq z$.

## 5.D Proof of Theorem 5.6

Proof. The proof leverages some of the ideas in the proof of [57, Theorem 5].
For any $n, w \in \mathbb{N}$ and $\lambda \in(0,1]$ define $g_{n, w, \lambda}=\sqrt{\frac{f_{2}(n)}{2 w \lambda^{3}}}$. Fix $\delta \in(0,1)$ and define $\varepsilon=\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)$ with $\mu^{\star}=M^{\star} \theta^{-1}$. For any $n$, introduce the following events:

$$
\begin{aligned}
A_{n, i} & =\left\{t_{i}(n) \mathrm{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right)>f_{2}(n)\right\}, \quad A_{n}=\bigcup_{i \in M^{\star}} A_{n, i} \\
B_{n, i} & =\left\{M_{i}(n)=1,\left|\hat{\theta}_{i}(n)-\theta_{i}\right| \geq \varepsilon \theta_{i}\right\}, \quad B_{n}=\bigcup_{i \in E} B_{n, i}, \\
F_{n} & =\left\{\Delta_{M(n)} \leq(1+\varepsilon) \sum_{i \in M(n)} g_{N, s_{i}(n),(1-\varepsilon) \theta_{\min }}+\varepsilon \mu^{\star}\right\} .
\end{aligned}
$$

Let $N>0$. Then the regret satisfies:

$$
R^{\pi}(N)=\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)}\right] \leq \mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)}\left(\mathbb{1}\left\{A_{n}\right\}+\mathbb{1}\left\{B_{n}\right\}\right)\right]+\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}\right\}\right] .
$$

Consider $n$ such that $M(n) \neq M^{\star}$. Next we show that $\overline{A_{n} \cup B_{n}} \subset F_{n}$. From the definition of $\omega_{i}(n)$, observe that $t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), \theta_{i}\right) \leq f_{2}(n)$ for any $i \in M^{\star}$ implies $M^{\star \top} \omega(n) \leq \mu^{\star}$. Hence we have:

$$
\begin{align*}
\mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}},\right. & \left.M(n) \neq M^{\star}\right\}=\mathbb{1}\left\{\overline{A_{n}}, \overline{B_{n}}, M(n)^{\top} \omega(n) \leq M^{\star \top} \omega(n)\right\} \\
& \leq \mathbb{1}\left\{\overline{B_{n}}, M(n)^{\top} \omega(n) \leq \mu^{\star}\right\} \\
& \leq \mathbb{1}\left\{\overline{B_{n}}, M(n)^{\top} \hat{\theta}(n)^{-1}-\sum_{i \in M(n)} g_{n, s_{i}(n), \hat{\theta}_{i}(n)} \leq \mu^{\star}\right\}  \tag{5.8}\\
& \leq \mathbb{1}\left\{(1+\varepsilon)^{-1} M(n)^{\top} \theta^{-1}-\sum_{i \in M(n)} g_{n, s_{i}(n),(1-\varepsilon) \theta_{\min }} \leq \mu^{\star}\right\}
\end{align*}
$$

$$
\begin{aligned}
& \leq \mathbb{1}\left\{\Delta_{M(n)} \leq(1+\varepsilon) \sum_{i \in M(n)} g_{N, s_{i}(n),(1-\varepsilon) \theta_{\min }}+\varepsilon \mu^{\star}\right\} \\
& =\mathbb{1}\left\{F_{n}\right\}
\end{aligned}
$$

where (5.8) follows from Lemma 5.2, proven at the end of this section, and the rest follows in the similar line as in the proof of Theorem5.5 Hence, the regret is upper bounded by:

$$
\begin{equation*}
\left.R^{\pi}(N) \leq m \theta_{\min }^{-1} \sum_{n=1}^{N}\left(\mathbb{P}\left[A_{n}\right]+\mathbb{P}\left[B_{n}\right]\right)\right]+\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right] \tag{5.9}
\end{equation*}
$$

We will prove the following inequalities:
(i) $\sum_{n=1}^{N} \mathbb{P}\left[A_{n}\right] \leq m K^{\prime \prime}$ for some constant $K^{\prime \prime}$,
(ii) $\mathbb{E}\left[\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\}\right] \leq \frac{45(1+\delta)^{2}}{(1-\delta)^{5}} m d \Delta_{\min }^{-1} \theta_{\min }^{-3} f_{2}(N)$.

Furthermore, from the proof of Theorem 5.5 we have:

$$
\sum_{n=1}^{N} \mathbb{P}\left[B_{n}\right] \leq \frac{2}{\delta^{2} \min \left(\Delta_{\min } / \mu^{\star}, 1\right)^{2}} \sum_{i \in E} \theta_{i}^{-2}
$$

Putting these together with 5.9, we obtain the announced result.
Inequality (i): By Lemma A.4 there exists a number $K^{\prime}$ such that for all $i$ and all $n \geq 2: \mathbb{P}\left[A_{n, i}\right] \leq K^{\prime} n^{-1}(\log (n))^{-2}$. Therefore:

$$
\sum_{n=1}^{N} \mathbb{P}\left[A_{n, i}\right] \leq 1+K^{\prime} \sum_{n \geq 2} n^{-1}(\log (n))^{-2} \equiv K^{\prime \prime}<\infty
$$

Applying the union bound, we get: $\sum_{n=1}^{N} \mathbb{P}\left[A_{n}\right] \leq m K^{\prime \prime}$.
Inequality (ii): Let $\left(\alpha_{\ell}\right)_{\ell \geq 1}$ and $\left(\beta_{\ell}\right)_{\ell \geq 0}$ be decreasing sequences such that (i) $\beta_{0}=1$, (ii) $\alpha_{\ell}, \beta_{\ell} \rightarrow_{\ell \rightarrow \infty} 0$, and (iii) $\sum_{\ell=1}^{\infty} \frac{\beta_{\ell-1}-\beta_{\ell}}{\sqrt{\alpha_{\ell}}} \leq 1$.

For any $\ell \in \mathbb{N}$ and $n$, define

$$
m_{\ell, n}=\frac{\alpha_{\ell}(1+\varepsilon)^{2} m^{2} f_{2}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right)^{2}}
$$

and introduce the following events:

$$
\begin{aligned}
S_{\ell, n} & =\left\{i \in M(n): s_{i}(n) \leq m_{\ell, n}\right\}, \\
G_{\ell, n} & =\bigcap_{j=1}^{\ell-1}\left\{\left|S_{j, n}\right|<\beta_{j} m\right\} \cap\left\{\left|S_{\ell, n}\right| \geq \beta_{\ell} m\right\}
\end{aligned}
$$

For any $n$ such that $M(n) \neq M^{\star}$, we claim that $F_{n} \subset \cup_{\ell \geq 1} G_{\ell, n}$. We prove this claim by showing that if $F_{n}$ occurs, then the event $\overline{U_{\ell \geq 1} G_{\ell, n}}=\cap_{\ell \geq 1}\left\{\left|S_{\ell, n}\right|<m \beta_{\ell}\right\}$ does not happen. Observe that occurrence of $F_{n}$ gives:

$$
\begin{aligned}
\Delta_{M(n)} & \leq(1+\varepsilon) \sum_{i \in M(n)} g_{N, s_{i}(n),(1-\varepsilon) \theta_{\min }}+\varepsilon \mu^{\star} \\
& =(1+\varepsilon) \sqrt{\frac{f_{2}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}}} \sum_{i \in M(n)} \frac{1}{\sqrt{s_{i}(n)}}+\varepsilon \mu^{\star} \\
& <(1+\varepsilon) \sqrt{\frac{f_{2}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}}} \sum_{\ell=1}^{\infty} \frac{m\left(\beta_{\ell-1}-\beta_{\ell}\right)}{\sqrt{m_{\ell, n}}}+\varepsilon \mu^{\star} \\
& =\left(\Delta_{M(n)}-\varepsilon \mu^{\star}\right) \sum_{\ell=1}^{\infty} \frac{\beta_{\ell-1}-\beta_{\ell}}{\sqrt{\alpha_{\ell}}}+\varepsilon \mu^{\star} \leq \Delta_{M(n)},
\end{aligned}
$$

where the second inequality follows from the proof of [57, Lemma 3] that relies on the condition that the event $\overline{U_{\ell \geq 1} G_{\ell, n}}$ happens. The above result is a contradiction. Hence, $F_{n} \subseteq \cup_{\ell \geq 1} G_{\ell, n}$ and consequently: $\mathbb{1}\left\{F_{n}\right\} \leq \sum_{\ell=1}^{\infty} \mathbb{1}\left\{G_{\ell, n}\right\}$.

To provide an upper bound for the r.h.s. of the latter inequality, we introduce for any $i$ :

$$
G_{i, \ell, n}=G_{\ell, n} \cap\left\{i \in M(n), s_{i}(n) \leq m_{\ell, n}\right\} .
$$

Observe that:

$$
\sum_{i \in E} \mathbb{1}\left\{G_{i, \ell, n}\right\}=\mathbb{1}\left\{G_{\ell, n}\right\} \sum_{i \in E} \mathbb{1}\left\{i \in S_{\ell, n}\right\}=\left|S_{\ell, n}\right| \mathbb{1}\left\{G_{\ell, n}\right\} \geq m \beta_{\ell} \mathbb{1}\left\{G_{\ell, n}\right\},
$$

and hence:

$$
\mathbb{1}\left\{G_{\ell, n}\right\} \leq \frac{1}{m \beta_{\ell}} \sum_{i \in E} \mathbb{1}\left\{G_{i, \ell, n}\right\} .
$$

Putting these inequalities together, we obtain

$$
\begin{aligned}
& \sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\} \leq \sum_{n=1}^{N} \sum_{\ell=1}^{\infty} \Delta_{M(n)} \mathbb{1}\left\{G_{\ell, n}\right\} \\
& \quad \leq \sum_{n=1}^{N} \sum_{i \in E} \sum_{\ell=1}^{\infty} \frac{\Delta_{M(n)}}{m \beta_{\ell}} \mathbb{1}\left\{G_{i, \ell, n}, M(n) \neq M^{\star}\right\} \\
& \quad \leq \sum_{n=1}^{N} \sum_{i \in E} \sum_{\ell=1}^{\infty} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{m \beta_{\ell}} \mathbb{1}\left\{G_{i, \ell, n}, M(n)=k\right\} \\
& \quad \leq \sum_{n=1}^{N} \sum_{i \in E} \sum_{\ell=1}^{\infty} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{m \beta_{\ell}} \mathbb{1}\left\{s_{i}(n) \leq \frac{\alpha_{\ell}(1+\varepsilon)^{2} m^{2} f_{2}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}\left(\Delta^{i, k}-\varepsilon \mu^{\star}\right)^{2}}, M(n)=k\right\}
\end{aligned}
$$

$$
\begin{align*}
& \leq \sum_{n=1}^{N} \sum_{i \in E} \sum_{\ell=1}^{\infty} \sum_{k \in\left[K_{i}\right]} \frac{\Delta^{i, k}}{m \beta_{\ell}} \\
& \quad \times \mathbb{1}\left\{s_{i}(n) \leq \frac{\alpha_{\ell}(1+\varepsilon)^{2} m^{2} f_{2}(N)}{2(1-\varepsilon)^{3} \theta_{\min }^{3}\left(\Delta^{i, k}\right)^{2}} \frac{1}{\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2}}, M(n)=k\right\} \\
& \leq \frac{(1+\varepsilon)^{2}}{(1-\varepsilon)^{3}\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2}} \frac{m d f_{2}(N)}{\Delta_{\min } \theta_{\min }^{3}} \sum_{\ell=1}^{\infty} \frac{\alpha_{\ell}}{\beta_{\ell}} \\
& \leq \frac{(1+\delta)^{2}}{(1-\delta)^{5}} \cdot \frac{m d f_{2}(N)}{\Delta_{\min } \theta_{\min }^{3}} \sum_{\ell=1}^{\infty} \frac{\alpha_{\ell}}{\beta_{\ell}}, \tag{5.10}
\end{align*}
$$

where the last inequality follows by plugging $\varepsilon=\delta \min \left(\frac{\Delta_{\min }}{\mu^{\star}}, 1\right)$ and using the inequality $\frac{(1+\varepsilon)^{2}}{(1-\varepsilon)^{3}\left(1-\varepsilon \mu^{\star} / \Delta_{\min }\right)^{2}} \leq \frac{(1+\delta)^{2}}{(1-\delta)^{5}}$, as derived in the proof of Theorem 5.5 Similarly to the proof of [57, Theorem 5], we select $\alpha_{\ell}=\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^{2} \alpha^{\ell}$ and $\beta_{\ell}=\beta^{\ell}$ for all $\ell$, with $0<\alpha<\beta<\sqrt{\alpha}<1$. With this choice of $\alpha_{\ell}$ and $\beta_{\ell}$, it follows that:

$$
\sum_{\ell=1}^{\infty} \frac{\beta_{\ell-1}-\beta_{\ell}}{\sqrt{\alpha_{\ell}}}=\left(\beta^{-1}-1\right) \frac{\sqrt{\alpha}-\beta}{1-\beta} \sum_{\ell=1}^{\infty}\left(\frac{\beta}{\sqrt{\alpha}}\right)^{\ell}=\left(\beta^{-1}-1\right) \frac{\sqrt{\alpha}-\beta}{1-\beta} \frac{\beta}{\sqrt{\alpha}-\beta}=1
$$

Moreover, $\beta_{0}=1$ and $\alpha_{\ell}, \beta_{\ell} \rightarrow_{\ell \rightarrow \infty} 0$, so that conditions (i)-(iii) are satisfied. Furthermore, we have that

$$
\sum_{\ell=1}^{\infty} \frac{\alpha_{\ell}}{\beta_{\ell}}=\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^{2} \sum_{\ell=1}^{\infty} \frac{\alpha^{\ell}}{\beta^{\ell}}=\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^{2} \frac{\alpha}{\beta-\alpha}
$$

which gives

$$
\sum_{n=1}^{N} \Delta_{M(n)} \mathbb{1}\left\{F_{n}\right\} \leq \frac{(1+\delta)^{2}}{(1-\delta)^{5}} \cdot \frac{m d f_{2}(N)}{\Delta_{\min } \theta_{\min }^{3}} \cdot\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^{2} \frac{\alpha}{\beta-\alpha}
$$

Given the constraint $0<\alpha<\beta<\sqrt{\alpha}<1$, the r.h.s. of the above inequality is minimized at $\left(\alpha^{\star}, \beta^{\star}\right)=(0.1459,0.2360)$. The proof is concluded by observing that $\left(\frac{1-\beta^{\star}}{\sqrt{\alpha^{\star}}-\beta^{\star}}\right)^{2} \frac{\alpha^{\star}}{\beta^{\star}-\alpha^{\star}}<45$.

Lemma 5.2. For all $n \geq 1$, $i$, and $\lambda \in(0,1]$, we have:

$$
\omega_{i}(n, \lambda) \geq \frac{1}{\lambda_{i}}-\sqrt{\frac{f_{2}(n)}{2 s_{i}(n) \lambda_{i}^{3}}} .
$$

Proof. Let $i \in E, q \in(0,1], \lambda \in(0,1]^{d}$, and $s(n) \in \mathbb{R}^{d}$. Define $t_{i}(n)=s_{i}(n) / \lambda_{i}$ for any $i \in E$. We have:

$$
\begin{aligned}
\frac{1}{q} \geq \frac{1}{\lambda_{i}}-\frac{1}{\lambda_{i}^{2}}\left(q-\lambda_{i}\right) & \geq \frac{1}{\lambda_{i}}-\sqrt{\frac{\left(q-\lambda_{i}\right)^{2}}{\lambda_{i}^{4}}} \sqrt{\frac{s_{i}(n)}{s_{i}(n)}} \\
& \geq \frac{1}{\lambda_{i}}-\sqrt{\frac{t_{i}(n) \mathrm{kl}\left(\lambda_{i}, q\right)}{2}} \sqrt{\frac{1}{s_{i}(n) \lambda_{i}^{3}}},
\end{aligned}
$$

where the first inequality follows from convexity of $q \mapsto \frac{1}{q}$ and the last one is due to Pinsker's inequality. Hence, $t_{i}(n) \mathrm{kl}\left(\lambda_{i}, q\right) \leq f_{2}(n)$ implies:

$$
\frac{1}{q} \geq \frac{1}{\lambda_{i}}-\sqrt{\frac{f_{2}(n)}{2 s_{i}(n) \lambda_{i}^{3}}}
$$

so that by definition of $\omega_{i}$,

$$
\omega_{i}(n, \lambda) \geq \frac{1}{\lambda_{i}}-\sqrt{\frac{f_{2}(n)}{2 s_{i}(n) \lambda_{i}^{3}}} .
$$

## 5.E Proof of Proposition 5.1

Proof. It is easy to verify that the function $F(\lambda)$ is strictly increasing. The rest of the proof follows the similar lines as in the proof of Theorem 3.5. In this case, the KKT conditions for index $b_{M}(n)$ read:

$$
\frac{1}{u_{i}^{2}}-\lambda t_{i}(n) \frac{d}{d u_{i}} \operatorname{kl}\left(\hat{\theta}_{i}(n), u_{i}\right)=0, \quad \sum_{i \in I} t_{i}(n) \operatorname{kl}\left(\hat{\theta}_{i}(n), u_{i}\right)-f_{1}(n)=0
$$

Hence, replacing the derivative of kl by its expression in the first equation, we obtain the following quadratic equation

$$
u_{i}^{2}+u_{i}\left(\frac{1}{\lambda t_{i}(n)}-\hat{\theta}_{i}(n)\right)-\frac{1}{\lambda t_{i}(n)}=0
$$

Solving for $u_{i}$, we obtain $u_{i}(\lambda)=g\left(\lambda, \hat{\theta}_{i}(n), t_{i}(n)\right)$. Plugging $u_{i}(\lambda)$ into the second KKT condition yields $F(\lambda)=f_{1}(n)$. The results then follow directly.

## 5.F Regret Upper Bound of The CUCB Algorithm

The regret of CUCB satisfies [56, Theorem 1]:

$$
\limsup _{N \rightarrow \infty} \frac{R^{\mathrm{CUCB}}(N)}{\log (N)} \leq \min _{f \in \mathcal{F}} \frac{12 d \Delta_{\min }}{\left(f^{-1}\left(\Delta_{\min }\right)\right)^{2}}
$$

where

$$
\begin{aligned}
\mathcal{F}=\{f: \forall \lambda, & \lambda^{\prime} \in \Theta, \forall M \in \mathcal{M}, \forall \Lambda>0 \\
& \left.\left(\max _{i \in M}\left|\lambda_{i}-\lambda_{i}^{\prime}\right| \leq \Lambda\right) \Rightarrow\left(\left|\mu_{M}(\lambda)-\mu_{M}\left(\lambda^{\prime}\right)\right| \leq f(\Lambda)\right)\right\}
\end{aligned}
$$

Next we show that the regret upper bound of the GeoCombUCB algorithm is smaller than that of CUCB. To this end, we claim that

$$
\min _{f \in \mathcal{F}} \frac{\Delta_{\min }}{\left(f^{-1}\left(\Delta_{\min }\right)\right)^{2}} \geq \frac{\sqrt{m}}{\theta_{\min }^{3} \Delta_{\min }} .
$$

The claim is proved as follows. Let $f \in \mathcal{F}$, and let $\Lambda>0$. We have that:

$$
f(\Lambda) \geq \frac{m \Lambda}{\theta_{\min }\left(\Lambda+\theta_{\min }\right)} .
$$

Indeed, let us take $\lambda=\theta_{\min } \mathbf{1}, \lambda^{\prime}=\lambda+\Lambda \mathbf{1}$, and $M$ of length $m$. Then, we have $\max _{i \in M}\left|\lambda_{i}-\lambda_{i}^{\prime}\right| \leq \Lambda$. Thus,

$$
f(\Lambda) \geq\left|\mu_{M}(\lambda)-\mu_{M}\left(\lambda^{\prime}\right)\right|=\frac{m \Lambda}{\theta_{\min }\left(\Lambda+\theta_{\min }\right)}:=f_{0}(\Lambda)
$$

Now we deduce that:

$$
\min _{f \in \mathcal{F}} \frac{\Delta_{\min }}{\left(f^{-1}\left(\Delta_{\min }\right)\right)^{2}} \geq \frac{\Delta_{\min }}{\left(f_{0}^{-1}\left(\Delta_{\min }\right)\right)^{2}}=\frac{\Delta_{\min }}{\left(\frac{\Delta_{\min } \theta_{\min }^{2}}{m-\Delta_{\min } \theta_{\min }}\right)^{2}}=\frac{\left(m-\Delta_{\min } \theta_{\min }\right)^{2}}{\Delta_{\min } \theta_{\min }^{4}}
$$

which concludes the proof of the claim. This latter result also verifies that $R^{\mathrm{CUCB}}(N)=$ $\mathcal{O}\left(\frac{d m^{2}}{\Delta_{\min } \theta_{\min }^{4}} \log (N)\right)$ as $N$ grows large.

## Chapter 6

## Adversarial Combinatorial MABs

In this chapter we study adversarial combinatorial MAB problems, namely online combinatorial problems in which rewards are bounded yet no statistical assumption on their generation is made. We consider bandit feedback and concentrate on the case where all arms have the same number of basic actions but are otherwise arbitrary. Our main contribution is CombEXP, an OSMD type algorithm, whose regret is as good as state-of-the-art algorithms, while it has lower computational complexity.

This chapter is based on the work [50], and is organized as follows. Section 6.1 outlines the contributions of this chapter and provides an overview of related literature on online combinatorial optimization in the adversarial setting. Section 6.2 describes our model and problem formulation. In Section 6.3 we present the Combexp algorithm and provide its regret and computational complexity analyses. Finally, Section 6.4 summarizes the chapter. Proof are presented in the appendix.

### 6.1 Contributions and Related Work

Consider a set of $d$ basic actions. Let $\mathcal{M}$ be an arbitrary subset of $\{0,1\}^{d}$, such that each of its element $M$ has $m$ basic actions. The main contribution of this chapter is an algorithm, which we may call CombEXP, for adversarial combinatorial MABs with $\mathcal{M}$ being the set of arms. Under bandit feedback, CombEXP achieves a regret scaling as

$$
\mathcal{O}\left(\sqrt{m^{3} T\left(d+\frac{m^{1 / 2}}{\underline{\lambda}}\right) \log \mu_{\min }^{-1}}\right),
$$

where $\mu_{\text {min }}=\min _{i \in[d]} \frac{1}{m|\mathcal{M}|} \sum_{M \in \mathcal{M}} M_{i}$ and $\underline{\lambda}$ is the smallest nonzero eigenvalue of the matrix $\mathbb{E}\left[M M^{\top}\right]$ when $M$ is uniformly distributed over $\mathcal{M}$ (Theorem 6.1). For most problems of interest, $m(d \underline{\lambda})^{-1}=\mathcal{O}(1)$ [79] and $\mu_{\text {min }}^{-1}=\mathcal{O}(\operatorname{poly}(d / m))$, so that CombEXP has $\mathcal{O}\left(\sqrt{m^{3} d T \log (d / m)}\right)$ regret. A known regret lower bound is $\Omega(m \sqrt{d T})$ [21], so the regret gap between CombEXP and this lower bound scales
at most as $m^{1 / 2}$ up to a logarithmic factor. This is the same regret scaling of state-of-the-art algorithms for these problems.

CombEXP relies on projections using the KL-divergence, which we may not be able to compute exactly. Nonetheless, we prove that under a carefully chosen precision for the projection, CombEXP attains the same regret scaling as if the projection would have been computed exactly (Proposition 6.1). Furthermore, we present an analysis of the computational complexity of CombEXP (Theorem 6.2 and show that Combexp has lower computational complexity than state-of-the-art algorithms for some problems of interest. The presented computational complexity analysis extends in an straightforward manner to other OSMD type algorithms and hence could be of independent interest.

### 6.1.1 Related Work

Online combinatorial optimization problems in the adversarial setting have been extensively investigated recently; see [21] and references therein. Here we give an overview of the most important contributions.

## Full Information

Full information setting constitutes the more tractable case for online combinatorial problems, which is very well understood by now. A known lower bound on the regret under this feedback model is $\Omega(m \sqrt{T \log (d / m)})$ 80. As one of the early instances of adversarial combinatorial problems, Takimoto and Warmuth 81] studied online shortest path problem where they propose a computationally efficient algorithm based on the Hedge algorithm of Freund and Schapire [82]. Kalai and Vempala [83] consider online linear optimization over a discrete set of arms and propose a computationally efficient algorithm based on FPL (Follow-the-Perturbed-Leader) for a more generic problem. Helmbold and Warmuth 84 introduce a computationally efficient algorithm for learning perfect matchings in the full information setting. Koolen et al. [80] propose Component Hedge that extends the latter work to generic combinatorial structures. The work by Koolen et al. has a two-fold importance: Firstly, it is one of the earliest works that targeted generic combinatorial problems. Secondly, the proposed algorithm Component Hedge achieves the optimal regret bound of $\mathcal{O}(m \sqrt{T \log (d / m)})$.

## Semi-Bandit Feedback

The most notable contributions in the semi-bandit setting include [28, 29, 85, 86, 21, 87, 88]. A known regret lower bound in this case scales as $\Omega(\sqrt{m d T})$ 21. Some contributions consider specific combinatorial problems. For example, the case of fixed-size subsets is investigated by Uchiya et al. [86] and Kale et al. [85], where both propose algorithms with optimal regret bounds up to logarithmic factors. The latter considers ordered subsets which prove useful in ranking problems. György et al. [28, 29] study online shortest path problem and propose an algorithm with
$\mathcal{O}(\sqrt{d T \log (|\mathcal{M}|)})$ regret. Generic problems are investigated in Audibert et al. 21] and Neu and Bartok [87]. The former proposes OSMD which achieves the optimal regret bound $\mathcal{O}(\sqrt{m d T})$. The latter presents an FPL-based algorithm, FPL with GR, which attains a regret $\mathcal{O}(m \sqrt{d T \log (d)})$. The performance of FPL WITH GR is worse than the lower bound by a factor of $m^{1 / 2}$ up to a logarithmic factor. Note however, that when the offline combinatorial optimization is efficiently implementable, FPL WITH GR can be run in polynomial time.

## Bandit Feedback

The work of Awerbuch and Kleinberg [27] is one of the earliest works that considers shortest-path routing as a particular instance of online linear optimization under bandit feedback. Their algorithm obtains $O^{*}\left(T^{2 / 3}\right)$ against an oblivious adversary. This work is followed by MacMahan and Blum [89, where they provide an algorithm achieving $O^{*}\left(T^{3 / 4}\right)$ but against an adaptive adversary. This result was further improved by György et al. [28, 29] to a high probability regret guarantee of $O^{*}\left(T^{2 / 3}\right)$ against an adaptive adversary. The Geometrichedge algorithm of Dani et al. [16], which attains a regret of $O^{*}(\sqrt{T})$, is the first algorithm with the optimal scaling in terms of $T$, thus signifying the sub-optimality of the aforementioned algorithms. Dani et al. [16] indeed consider online linear optimization over a $d$-dimensional compact convex set in the bandit setting and against an oblivious adversary. Bartlett et al. 90 further show that a modified version of Geometrichedge has a regret $O^{*}(\sqrt{T})$ with high probability and also the results holds against an adaptive adversary.

For generic combinatorial problems, Cesa-Bianchi and Lugosi [79] propose ComBand and derive a regret upper bound which depends on the structure of the of set of arms $\mathcal{M}$ and the choice of exploration distribution. For most problems of interest ${ }^{1}$, under uniform exploration, the regret under ComBand is upper-bounded by $\mathcal{O}\left(\sqrt{m^{3} d T \log (d / m)}\right)$. Another algorithm for generic problems is EXP2 with John's Exploration (henceforth, EXP2-John), which is proposed by Bubeck et al. [15]. EXP2-Jонn is an algorithm based on exponential weights with a novel way of exploration. It achieves a regret of $O\left(m^{3 / 2} \sqrt{d T \log (|\mathcal{M}|)}\right)$ for a discrete set $\mathcal{M} \subset\{0,1\}^{d}$.

As shown by Audibert et al. [21], for the setup we considered, the problem admits a minimax lower bound of $\Omega(m \sqrt{d T})$ if $d \geq 2 m$. This lower bound signifies that ComBand and EXP2-John are off the optimal regret bound by a factor of $m^{1 / 2} \sqrt{\log (d / m)}$. We remark however that if the assumption $\|M\|_{1}=m, \forall M \in \mathcal{M}$ is relaxed to $\|M\|_{1} \leq m$, i.e., when arms can have different number of basic actions, the regret upper bound of Comband is tight as it matches a lower bound proposed in [16] ${ }^{2}$. As we show next, for many combinatorial structures of interest (e.g., fixed-

[^16]| Algorithm | Regret |
| :---: | :---: |
| Lower Bound [21] | $\Omega(m \sqrt{d T})$, if $d \geq 2 m$ |
| ComBand [79] | $\mathcal{O}\left(\sqrt{m^{3} d T \log \frac{d}{m}\left(1+\frac{2 m}{d \underline{\lambda}}\right)}\right)$ |
| EXP2-John [15] | $\mathcal{O}\left(\sqrt{m^{3} d T \log \frac{d}{m}}\right)$ |
| CombEXP (Theorem 6.1 | $\mathcal{O}\left(\sqrt{m^{3} d T\left(1+\frac{m^{1 / 2}}{d \underline{\lambda}}\right) \log \mu_{\min }^{-1}}\right)$ |

Table 6.1: Regret of various algorithms for adversarial combinatorial bandits with bandit feedback. Note that for most combinatorial classes of interests, $m(d \underline{\lambda})^{-1}=$ $\mathcal{O}(1)$ and $\mu_{\min }^{-1}=\mathcal{O}(\operatorname{poly}(d / m))$.
size subsets, matchings, spanning trees, cut sets), CombEXP yields the same regret scaling as ComBand and EXP2-Jонn, but with lower computational complexity for a large class of problems. Table 6.1 summarises known regret bounds.

### 6.2 Model and Objectives

Given a set of basic actions $[d]=\{1, \ldots, d\}$, consider a set of arms $\mathcal{M}$ such that every element in $\mathcal{M}$ has exactly $m$ basic actions. For $i \in[d]$, let $X_{i}(n)$ denote the reward of basic action $i$ in round $n$. We consider an oblivious adversary. Namely, we assume that the reward vector $X(n)=\left(X_{1}(n), \ldots, X_{d}(n)\right)^{\top} \in[0,1]^{d}$ is arbitrary, and the sequence $(X(n), n \geq 1)$ is decided (but unknown) at the beginning of the experiment. We identify each arm $M$ with a binary column vector $\left(M_{1}, \ldots, M_{d}\right)^{\top}$, and we have $\|M\|_{1}=m, \forall M \in \mathcal{M}$. At the beginning of each round $n$, an algorithm or policy $\pi$, selects an $\operatorname{arm} M^{\pi}(n) \in \mathcal{M}$ based on the arms chosen in previous rounds and their observed rewards. The reward of $\operatorname{arm} M^{\pi}(n)$ selected in round $n$ is $X^{M^{\pi}(n)}(n)=M^{\pi}(n)^{\top} X(n)$.

We consider bandit feedback, where at the end of round $n$ and under policy $\pi$, the decision maker only observes $M^{\pi}(n)^{\top} X(n)$. Her objective is to identify a policy $\pi$ maximizing the cumulative expected reward over a finite time horizon $T$, or equivalently minimizing the regret defined by:

$$
R^{\pi}(T)=\max _{M \in \mathcal{M}} \mathbb{E}\left[\sum_{n=1}^{T} X^{M}(n)\right]-\mathbb{E}\left[\sum_{n=1}^{T} X^{M^{\pi}(n)}(n)\right]
$$

where the expectation is here taken with respect to the possible randomization in the policy.

### 6.2.1 The OSMD Algorithm

In order to have better understanding of the main contribution of this chapter, namely the CombEXP algorithm, here we describe the Online Stochastic Mirror De-
scent (OSMD) algorithm [21, 15].
OSMD is based on the mirror descent algorithm of Nemirovski and Yudin 91 ] for solving convex optimization problems. The underlying idea of mirror descent is that instead of performing gradient descent in the primal space, one can perform it in another space referred to as the dual space (see below for a precise definition). From an optimization perspective, mirror descent proves very efficient for largescale optimization problems as its convergence rate grows logarithmically with the dimension of the problem [92]. Mirror descent in online learning is of special interest since popular learning algorithms such as online gradient descent [93] and the weighted majority algorithm [47] can be seen as special cases of mirror descent.

To describe OSMD, we collect some definitions from convex analysis which can be found in, e.g., [94]. Consider an open convex set $\mathcal{D} \subset \mathbb{R}^{d}$ and its closure $\operatorname{cl}(\mathcal{D})$. Mirror descent performs gradient descent in a dual space of $\mathcal{D}$. To define a dual space for $\mathcal{D}$, we next introduce the notion of Legendre functions. A continuous function $F: \operatorname{cl}(\mathcal{D}) \rightarrow \mathbb{R}$ is called Legendre if (i) it is strictly convex and continuously differentiable on $\mathcal{D}$, and (ii) $\|\nabla F(x)\| \rightarrow_{x \rightarrow \operatorname{cl}(\mathcal{D}) \backslash \mathcal{D}} \infty$. Now, $\mathcal{D}^{\star}=\nabla F(\mathcal{D})$ is called the dual space of $\mathcal{D}$ under $F$.

To any Legendre function $F$, we may associate the Bregman divergence $D_{F}$ : $\operatorname{cl}(\mathcal{D}) \times \mathcal{D} \rightarrow \mathbb{R}$ defined by

$$
D_{F}(x, y)=F(x)-F(y)-(x-y)^{\top} \nabla F(y) .
$$

It is noted that $D_{F}(x, y) \geq 0$ for all $x, y$ since $F$ is convex. Moreover, it is convex in the first argument. Finally, the Legendre-Fenchel transform of $F$, denoted by $F^{\star}$, is defined by

$$
F^{\star}(y)=\sup _{x \in \operatorname{cl}(\mathcal{D})}\left(x^{\top} y-F(x)\right)
$$

A result in convex analysis states that Bregman divergence in the primal space is equivalent to Bregman divergence of the Legendre-Fenchel transform in the dual space; namely $D_{F}(x, y)=D_{F^{\star}}(\nabla F(y), \nabla F(x))$ for all $x, y \in \mathcal{D}$ (see, e.g., [21, Lemma 2.1] and [95, Chapter 11]).

As an example consider the function $F(x)=\sum_{i=1}^{d} x_{i} \log \left(x_{i}\right)-\sum_{i=1}^{d} x_{i}$ and $\mathcal{D}=\mathbb{R}_{++}^{d}$. The dual space of $\mathcal{D}$ under this choice of $F$ is $\mathcal{D}^{\star}=\mathbb{R}^{d}$. Moreover, the Bregman divergence in this case becomes the generalized KL-divergence:

$$
\begin{aligned}
D_{F}(x, y) & =\sum_{i=1}^{d} x_{i} \log \left(x_{i}\right)-\sum_{i=1}^{d} x_{i}-\sum_{i=1}^{d} y_{i} \log \left(y_{i}\right)+\sum_{i=1}^{d} y_{i}-\sum_{i=1}^{d} \log \left(y_{i}\right)\left(x_{i}-y_{i}\right) \\
& =\sum_{i=1}^{d} x_{i} \log \left(x_{i} / y_{i}\right)-\sum_{i=1}^{d}\left(x_{i}-y_{i}\right) .
\end{aligned}
$$

We are now ready to describe OSMD. The corresponding pseudo-code for loss minimization problem is presented in Algorithm6.1. It performs a gradient update

```
Algorithm 6.1 OSMD [21]
    Initialization: Set \(x_{0} \in \operatorname{argmin}_{x \in \operatorname{conv}(\mathcal{M})} F(x)\).
    for \(n \geq 1\) do
        Select a distribution \(p_{n-1}\) over \(\mathcal{M}\) such that \(\sum_{M} p_{n-1}(M) M=x_{n-1}\).
        Select a random arm \(M(n)\) with distribution \(p_{n-1}\) and observe the feedback.
        Compute the estimate reward vector \(\tilde{X}(n)\).
        Find \(w_{n}\) such that \(\nabla F\left(w_{n}\right)=\nabla F\left(x_{n-1}\right)-\eta \tilde{X}(n)\).
        Project \(w_{n}\) onto \(\operatorname{conv}(\mathcal{M})\) :
            \(x_{n}=\arg \min _{x \in \operatorname{conv}(\mathcal{M})} D_{F}\left(x, w_{n}\right)\).
    end for
```

with estimated loss vector in the dual space. Then it projects back the updated vector onto the primal space.

Audibert et al. [21] originally proposed OSMD for loss minimization under semibandit feedback. They showed that under a suitable choice of Legendre function $F$, OSMD achieves a regret $\mathcal{O}(\sqrt{m d T})$ which is minimax optimal.

### 6.3 The CombEXP Algorithm

In this section we propose the CombEXP algorithm, which is inspired by OSMD described in the previous section (see Algorithm 6.1). In order to have better understanding of the various steps of the algorithm, we start with the following observation:

$$
\begin{aligned}
\max _{M \in \mathcal{M}} X^{M} & =\max _{M \in \mathcal{M}} M^{\top} X \\
& =\max _{p(M) \geq 0, \sum_{M \in \mathcal{M}} p(M)=1} \sum_{M \in \mathcal{M}} p(M) M^{\top} X \\
& =\max _{\mu \in \operatorname{conv}(\mathcal{M})} \mu^{\top} X .
\end{aligned}
$$

We can embed $\mathcal{M}$ in the simplex of distributions in $\mathbb{R}^{d}$ by multiplying all the entries by $1 / m$. Let $\mathcal{P}$ be this scaled version of $\operatorname{conv}(\mathcal{M})$.

We propose the CombEXP algorithm as an OSMD type algorithm where the corresponding Legendre function is negative entropy, i.e. $F(x)=\sum_{i=1}^{d} x_{i} \log \left(x_{i}\right)$, and thus the corresponding Bregman divergence is the KL-divergence. As a result, the projection step in OSMD reduces to projection using the KL-divergence: By definition, the projection of a distribution $q$ onto a closed convex set $\Xi$ of distributions is

$$
p^{\star}=\arg \min _{p \in \Xi} \operatorname{KL}(p, q),
$$

## Algorithm 6.2 CombEXP

Initialization: Set $q_{0}=\mu^{0}, \gamma=\frac{\sqrt{m \log \mu_{\min }^{-1}}}{\sqrt{m \log \mu_{\min }^{-1}}+\sqrt{C\left(C m^{2} d+m\right) T}}$ and $\eta=\gamma C$, with $C=\frac{\lambda}{m^{3 / 2}}$.
for $n \geq 1$ do
Mixing: Let $q_{n-1}^{\prime}=(1-\gamma) q_{n-1}+\gamma \mu^{0}$.
Decomposition: Select a distribution $p_{n-1}$ over $\mathcal{M}$ such that $\sum_{M} p_{n-1}(M) M=$ $m q_{n-1}^{\prime}$.

Sampling: Select a random arm $M(n)$ with distribution $p_{n-1}$ and incur a reward $Y_{n}=\sum_{i} X_{i}(n) M_{i}(n)$.

Estimation: Let $\Sigma_{n-1}=\mathbb{E}\left[M M^{\top}\right]$, where $M$ has law $p_{n-1}$. Set $\tilde{X}(n)=$ $Y_{n} \Sigma_{n-1}^{+} M(n)$, where $\Sigma_{n-1}^{+}$is the pseudo-inverse of $\Sigma_{n-1}$.

Update: Set $\tilde{q}_{n}(i) \propto q_{n-1}(i) \exp \left(\eta \tilde{X}_{i}(n)\right), \forall i \in[d]$.
Projection: Set $q_{n}$ to be the projection of $\tilde{q}_{n}$ onto the set $\mathcal{P}$ using the KL-divergence:

$$
q_{n}=\underset{p \in \mathcal{P}}{\operatorname{argmin}} \operatorname{KL}\left(p, \tilde{q}_{n}\right) .
$$

end for
where the uniqueness of the minimizer follows from strict convexity of $z \mapsto \operatorname{KL}(z, q)$. For a thorough description of projection using the KL-divergence we refer to 96 , Chapter 3].

The pseudo-code of CombEXP is shown in Algorithm 6.2 At each round $n$, the projected vector $q_{n-1}$ of the previous round is mixed with the exploration-inducing distribution $\mu^{0} \in \mathcal{P}$ defined as:

$$
\mu_{i}^{0}=\frac{1}{m|\mathcal{M}|} \sum_{M \in \mathcal{M}} M_{i}, \quad \forall i \in[d] .
$$

We remark that $q_{n-1} \in \mathcal{P}$ since it is the projection of some vector onto $\mathcal{P}$ using the KL-divergence. As a result, $q_{n-1}^{\prime} \in \mathcal{P}$. The mixed distribution $q_{n-1}^{\prime}$ is then decomposed to some probability distribution $p_{n-1}$ over the set of arms $\mathcal{M}$. Of course, the linear equation system in the decomposition step is always consistent since $q_{n-1}^{\prime} \in \mathcal{P}$ implies $m q_{n-1}^{\prime} \in \operatorname{conv}(\mathcal{M})$ and hence, there must exist a probability vector $\kappa$ such that $m q_{n-1}^{\prime}=\sum_{M} \kappa(M) M$. Clearly $\mu^{0}$ defines a distribution over basic actions $[d]$ that induces uniform distribution over $\mathcal{M}$. Therefore, CombEXP uses uniform sampling for exploration. After playing a randomly generated arm $M(n)$ with law $p_{n-1}$, Combexp constructs the reward estimate vector $\tilde{X}(n)$ and updates the probability vector $q_{n-1}$. The resulting vector $\tilde{q}_{n}$ does not necessarily belong to $\mathcal{P}$, and hence it is projected onto $\mathcal{P}$ using the KL-divergence. For more details on the implementation of the various steps, we refer to Section 6.3.2

Let $\underline{\lambda}$ be the smallest nonzero eigenvalue of $\mathbb{E}\left[M M^{\top}\right]$, where $M$ is uniformly distributed over $\mathcal{M}$. Moreover, let $\mu_{\text {min }}=\min _{i} m \mu_{i}^{0}$. The following theorem provides a regret upper bound for CombEXP.

Theorem 6.1. For all $T \geq 1$, we have:

$$
R^{\mathrm{CombEXP}}(T) \leq 2 \sqrt{m^{3} T\left(d+\frac{m^{1 / 2}}{\underline{\lambda}}\right) \log \mu_{\min }^{-1}}+\frac{m^{5 / 2}}{\underline{\lambda}} \log \mu_{\min }^{-1}
$$

As will be discussed in the next subsection, for most classes of $\mathcal{M}$, we have $\mu_{\min }^{-1}=\mathcal{O}(\operatorname{poly}(d / m))$. Furthermore, $m(d \underline{\lambda})^{-1}=\mathcal{O}(1)$ holds for most classes of $\mathcal{M}$ [79]. For these classes, CombEXP has a regret of $\mathcal{O}\left(\sqrt{m^{3} d T \log (d / m)}\right)$, which is a factor $\sqrt{m \log (d / m)}$ off the minimax lower bound (see Table 6.1).

### 6.3.1 Examples

In this subsection, we compare the performance of CombEXP against state-of-the-art algorithms (refer to Table 6.1 for the summary of regret of various algorithms).

Fixed-size subsets. In this case, $\mathcal{M}$ is the set of all $d$-dimensional binary vectors with $m$ ones. We have

$$
\mu_{\min }=\min _{i} \frac{1}{\binom{d}{m}} \sum_{M} M_{i}=\frac{\binom{d-1}{m-1}}{\binom{d}{m}}=\frac{m}{d}
$$

Moreover, according to [79, Proposition 12], we have $\underline{\lambda}=\frac{m(d-m)}{d(d-1)}$. When $d \geq$ $2 m$, the regret of Combexp becomes $O\left(\sqrt{m^{3} d T \log (d / m)}\right)$, namely it has the same performance as ComBand and EXP2-John.

Matching. Let $\mathcal{M}$ be the set of perfect matchings in $\mathcal{K}_{m, m}$, where we have $d=m^{2}$ and $|\mathcal{M}|=m$ !. We have

$$
\mu_{\min }=\min _{i} \frac{1}{m!} \sum_{M} M_{i}=\frac{(m-1)!}{m!}=\frac{1}{m},
$$

Furthermore, from [79, Proposition 4] we have that $\underline{\lambda}=\frac{1}{m-1}$, thus giving $R^{\operatorname{CombEXP}}(T)=$ $O\left(\sqrt{m^{5} T \log (m)}\right)$, which is the same as the regret of ComBand and EXP2-John in this case.

Spanning trees. In our next example, we assume that $\mathcal{M}$ is the set of spanning trees in the complete graph $\mathcal{K}_{N}$. In this case, we have $d=\binom{N}{2}, m=N-1$, and by Cayley's formula $\mathcal{M}$ has $N^{N-2}$ elements. Observe that

$$
\mu_{\min }=\min _{i} \frac{1}{N^{N-2}} \sum_{M} M_{i}=\frac{(N-1)^{N-3}}{N^{N-2}}
$$

which gives for $N \geq 2$

$$
\begin{aligned}
\log \mu_{\min }^{-1} & =\log \left(\frac{N^{N-2}}{(N-1)^{N-3}}\right)=(N-3) \log \left(\frac{N}{N-1}\right)+\log N \\
& \leq(N-3) \log 2+\log (N) \leq 2 N
\end{aligned}
$$

From [79, Corollary 7], we also get $\underline{\lambda} \geq \frac{1}{N}-\frac{17}{4 N^{2}}$. For $N \geq 6$, the regret of Comband takes the form $O\left(\sqrt{N^{5} T \log (N)}\right)$ since $\frac{m}{d \lambda}<7$ when $N \geq 6$. Further, EXP2-Jонn attains the same regret. On the other hand, we get

$$
R^{\mathrm{ComвEXP}}(T)=O\left(\sqrt{N^{5} T \log (N)}\right), \quad N \geq 6,
$$

namely, CombEXP gives the same regret as ComBand and EXP2-John.

Cut sets. Consider the case where $\mathcal{M}$ is the set of balanced cuts of the complete graph $\mathcal{K}_{2 N}$, where a balanced cut is defined as the set of edges between a set of $N$ vertices and its complement. It is easy to verify that $d=\binom{2 N}{2}$ and $m=N^{2}$. Moreover, $\mathcal{M}$ has $\binom{2 N}{N}$ balanced cuts and hence

$$
\mu_{\min }=\min _{i} \frac{1}{\binom{2 N}{N}} \sum_{M} M_{i}=\frac{\binom{2 N-2}{N-1}}{\binom{2 N}{N}}=\frac{N}{4 N-2},
$$

Moreover, by [79, Proposition 9], we have

$$
\underline{\lambda}=\frac{1}{4}+\frac{8 N-7}{4(2 N-1)(2 N-3)}, \quad N \geq 2,
$$

and consequently, the regret of CombEXP becomes $O\left(N^{4} \sqrt{T}\right)$ for $N \geq 2$, which is the same as that of ComBand and EXP2-John.

### 6.3.2 Implementation

We propose to perform the projection step, namely to find the KL projection of $\tilde{q}$ onto $\mathcal{P}$, using interior-point methods $[97]^{3}$. We also remark that the Decomposition step can be efficiently implemented using the algorithm of [98].

It might not be possible to exactly compute the projection step of CombEXP in a finite number of operations. Thus, in round $n$, this step can be solved up to some accuracy $\varepsilon_{n}$; namely we find $q_{n}$ such that $\operatorname{KL}\left(q_{n}, \tilde{q}_{n}\right) \leq \operatorname{KL}\left(u_{n}, \tilde{q}_{n}\right)+\varepsilon_{n}$, where $u_{n}=\arg \min _{p \in \Xi} \operatorname{KL}\left(p, \tilde{q}_{n}\right)$. Proposition 6.1 shows that for $\varepsilon_{n}=\mathcal{O}\left(n^{-2} \log ^{-3}(n)\right)$, the upper bound of the regret will have the same order as if projection was computed exactly (i.e., $\varepsilon_{n}=0$ ).

[^17]Proposition 6.1. Assume that the projection step of CombEXP is solved up to accuracy $\varepsilon_{n}=\mathcal{O}\left(n^{-2} \log ^{-3}(n)\right)$ for all $n \geq 1$. Then

$$
R^{\mathrm{ComвEXP}}(T) \leq 2 \sqrt{2 m^{3} T\left(d+\frac{m^{1 / 2}}{\underline{\lambda}}\right) \log \mu_{\min }^{-1}}+\frac{2 m^{5 / 2}}{\underline{\lambda}} \log \mu_{\min }^{-1}
$$

Theorem 6.2 provides the computational complexity of the CombEXP algorithm under this choice of accuracy. In particular, it asserts that when $\operatorname{conv}(\mathcal{M})$ is described by polynomially (in $d$ ) many linear equalities/inequalities, CombEXP is efficiently implementable and its running time scales (almost) linearly in $T$.

Theorem 6.2. Let c (resp. s) be the number of linear equalities (resp. inequalities) that defines the convex hull of $\mathcal{M}$. Then, if the projection step of CombEXP is solved up to accuracy $\varepsilon_{n}=\mathcal{O}\left(n^{-2} \log ^{-3}(n)\right)$, for all $n \geq 1$, CombEXP with time horizon $T$ has time complexity $\mathcal{O}\left(T\left[\sqrt{s}(c+d)^{3} \log (T)+d^{4}\right]\right)$.

Proposition 6.1 and Theorem 6.2 easily extend to other OSMD-type algorithms and thus might be of independent interest.

Remark 6.1. We remark that the time complexity of CombEXP can be reduced by exploiting the structure of $\mathcal{M}$ (see [97, page 545]). In particular, if inequality constraints describing $\operatorname{conv}(\mathcal{M})$ are box constraints, the time complexity of CombEXP can be improved to $\mathcal{O}\left(T\left[c^{2} \sqrt{s}(c+d) \log (T)+d^{4}\right]\right)$.

Theorem 6.2 signifies that the computational complexity of CombEXP is determined by the representation of convex hull of $\mathcal{M}$. In contrast, that of Comband depends on the complexity of procedures to sample from $\mathcal{M}$. In turn, Comband might have a time complexity that is super-linear in $T$ (we refer to [99, page 217] for a related discussion). On the other hand, our algorithm is guaranteed to have $\mathcal{O}(T \log (T))$ time complexity (thanks to the efficiency of the interior-point method). The time complexity of CombEXP is examined next through two examples.

Matching. When $\mathcal{M}$ is the set of matchings in $\mathcal{K}_{m, m}, \operatorname{conv}(\mathcal{M})$ is the convex hull of all $m \times m$ permutation matrices:

$$
\operatorname{conv}(\mathcal{M})=\left\{Z \in \mathbb{R}_{+}^{m \times m}: \sum_{k=1}^{m} z_{i k}=1, \forall i \in[m], \sum_{k=1}^{m} z_{k j}=1, \forall j \in[m]\right\}
$$

This set is referred to as the Birkhoff polytope and indeed is the set of all doubly stochastic $m \times m$ matrices ${ }^{4}$ [100], since every doubly stochastic matrix living in $\mathbb{R}^{m \times m}$ can be expressed as the convex combination of at most $m^{2}-2 m+2$ permutation matrices. This result is known as Birkhoff-von Neumann Theorem.

[^18]In this case, there are $c=2 m$ linear equalities and $s=m^{2}$ linear inequalities in the form of box constraints. Using Algorithm 1 in [84], the cost of decomposition in this case is $\mathcal{O}\left(m^{4}\right)$. Hence the total time complexity of CombEXP becomes: $\mathcal{O}\left(m^{5} T \log (T)\right)$ on the account of Remark 6.1. On the other hand, ComBand has a time complexity of $\mathcal{O}\left(m^{10} F(T)\right)$ after $T$ rounds for some super-linear function $F$, as it requires to approximate a permanent, requiring $\mathcal{O}\left(m^{10}\right)$ operations per round. Thus, CombEXP exhibits much lower computational complexity than ComBand while achieving the same regret.

Spanning trees. Consider a connected graph $G=(V, E)$ and let $\mathcal{M}$ be the set of all spanning trees in $G$. The convex hull of $\mathcal{M}$ becomes the spanning tree polytope [23, Corollary 50.7c], that is

$$
\operatorname{conv}(\mathcal{M})=\left\{z \in \mathbb{R}_{+}^{|E|}: \sum_{i \in \mathcal{E}(U)} z_{i} \leq|U|-1, \forall U \subset V, \sum_{i \in E} z_{i}=|V|-1\right\}
$$

where for any $U \subset V, \mathcal{E}(U)$ denotes the set of edges whose both endpoints are in $U$. This description is indeed a consequence of base polytope description of matroids due to Edmonds [62], since spanning trees of graph $G$ are bases of the graphic matroid associated to $G$. There is only one linear equality $(c=1)$, but the number of linear inequalities $s$ grows exponentially with $|E|$. In particular, when $G=\mathcal{K}_{N}$, we have $N^{N-2}$ spanning trees on the account of Cayley's formula, and hence $s=N^{N-2}+N(N-1) / 2$. Hence, CombEXP does not have polynomial time complexity in this case. On the other hand, ComBand needs to sample a spanning tree from $\mathcal{M}$ in each round. Hence, it is not efficient for this case either since there is no polynomial time sampling scheme for this task.

### 6.4 Summary

In this chapter we investigated adversarial combinatorial MAB problems under bandit feedback. We considered oblivious adversary and assumed that all arms have the same number of basic actions. We presented CombEXP, an OSMD-type algorithm, and provided its regret analysis and computational complexity. We have shown that CombEXP achieves a regret that has the same scaling as state-of-the-art algorithms. Yet it admits lower computational complexity for some problems of interests.

## 6.A Proof of Theorem 6.1

Proof. We first prove a simple result:
Lemma 6.1. For all $x \in \mathbb{R}^{d}$, we have $\Sigma_{n-1}^{+} \Sigma_{n-1} x=\bar{x}$, where $\bar{x}$ is the orthogonal projection of $x$ onto $\operatorname{span}(\mathcal{M})$, the linear space spanned by $\mathcal{M}$.

Proof. Note that for all $y \in \mathbb{R}^{d}$, if $\Sigma_{n-1} y=0$, then we have

$$
\begin{equation*}
y^{\top} \Sigma_{n-1} y=\mathbb{E}\left[y^{\top} M M^{\top} y\right]=\mathbb{E}\left[\left(y^{\top} M\right)^{2}\right]=0 \tag{6.1}
\end{equation*}
$$

where $M$ has law $p_{n-1}$ such that $\sum_{M} M_{i} p_{n-1}(M)=q_{n-1}^{\prime}(i), \forall i \in[d]$ and $q_{n-1}^{\prime}=$ $(1-\gamma) q_{n-1}+\gamma \mu^{0}$. By definition of $\mu^{0}$, each $M \in \mathcal{M}$ has a positive probability. Hence, by 6.1), $y^{\top} M=0$ for all $M \in \mathcal{M}$. In particular, we see that the linear application $\Sigma_{n-1}$ restricted to $\operatorname{span}(\mathcal{M})$ is invertible and is zero on $\operatorname{span}(\mathcal{M})^{\perp}$, hence we have $\Sigma_{n-1}^{+} \Sigma_{n-1} x=\bar{x}$.

Lemma 6.2. We have for any $\eta \leq \frac{\gamma \lambda}{m^{3 / 2}}$ and any $q \in \mathcal{P}$,

$$
\sum_{n=1}^{T} q^{\top} \tilde{X}(n)-\sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}(n) \leq \frac{\eta}{2} \sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{\operatorname{KL}\left(q, q_{0}\right)}{\eta}
$$

where $\tilde{X}^{2}(n)$ is the vector that is the coordinate-wise square of $\tilde{X}(n)$.
Proof. We have

$$
\mathrm{KL}\left(q, \tilde{q}_{n}\right)-\mathrm{KL}\left(q, q_{n-1}\right)=\sum_{i \in[d]} q(i) \log \frac{q_{n-1}(i)}{\tilde{q}_{n}(i)}=-\eta \sum_{i \in[d]} q(i) \tilde{X}_{i}(n)+\log Z_{n},
$$

with

$$
\begin{align*}
\log Z_{n} & =\log \sum_{i \in[d]} q_{n-1}(i) \exp \left(\eta \tilde{X}_{i}(n)\right) \\
& \leq \log \sum_{i \in[d]} q_{n-1}(i)\left(1+\eta \tilde{X}_{i}(n)+\eta^{2} \tilde{X}_{i}^{2}(n)\right)  \tag{6.2}\\
& \leq \eta q_{n-1}^{\top} \tilde{X}(n)+\eta^{2} q_{n-1}^{\top} \tilde{X}^{2}(n), \tag{6.3}
\end{align*}
$$

where we used $\exp (z) \leq 1+z+z^{2}$ for all $|z| \leq 1$ in 6.2 and $\log (1+z) \leq z$ for all $z>-1$ in 6.3. Later we verify the condition for the former inequality.

Hence we have

$$
\mathrm{KL}\left(q, \tilde{q}_{n}\right)-\mathrm{KL}\left(q, q_{n-1}\right) \leq \eta q_{n-1}^{\top} \tilde{X}(n)-\eta q^{\top} \tilde{X}(n)+\eta^{2} q_{n-1}^{\top} \tilde{X}^{2}(n)
$$

Generalized Pythagorean inequality (see Theorem 3.1 in 96) gives

$$
\operatorname{KL}\left(q, q_{n}\right)+\operatorname{KL}\left(q_{n}, \tilde{q}_{n}\right) \leq \operatorname{KL}\left(q, \tilde{q}_{n}\right) .
$$

Since $\operatorname{KL}\left(q_{n}, \tilde{q}_{n}\right) \geq 0$, we get

$$
\mathrm{KL}\left(q, q_{n}\right)-\mathrm{KL}\left(q, q_{n-1}\right) \leq \eta q_{n-1}^{\top} \tilde{X}(n)-\eta q^{\top} \tilde{X}(n)+\eta^{2} q_{n-1}^{\top} \tilde{X}^{2}(n)
$$

Finally, summing over $n$ gives

$$
\sum_{n=1}^{T}\left(q^{\top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right) \leq \eta \sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{\mathrm{KL}\left(q, q_{0}\right)}{\eta}
$$

To satisfy the condition for the inequality (6.2), i.e., $\eta\left|\tilde{X}_{i}(n)\right| \leq 1, \forall i \in[d]$, we find the upper bound for $\max _{i \in[d]}\left|\tilde{X}_{i}(n)\right|$ as follows:

$$
\begin{aligned}
\max _{i \in[d]}\left|\tilde{X}_{i}(n)\right| & \leq\|\tilde{X}(n)\|_{2} \\
& =\left\|\Sigma_{n-1}^{+} M(n) Y_{n}\right\|_{2} \\
& \leq m\left\|\Sigma_{n-1}^{+} M(n)\right\|_{2} \\
& \leq m \sqrt{M(n)^{\top} \Sigma_{n-1}^{+} \Sigma_{n-1}^{+} M(n)} \\
& \leq m\|M(n)\|_{2} \sqrt{\lambda_{\max }\left(\Sigma_{n-1}^{+} \Sigma_{n-1}^{+}\right)} \\
& =m^{3 / 2} \sqrt{\lambda_{\max }\left(\Sigma_{n-1}^{+} \Sigma_{n-1}^{+}\right)} \\
& =m^{3 / 2} \lambda_{\max }\left(\Sigma_{n-1}^{+}\right) \\
& =\frac{m^{3 / 2}}{\lambda_{\min }\left(\Sigma_{n-1}\right)},
\end{aligned}
$$

where $\lambda_{\max }(A)$ and $\lambda_{\min }(A)$ respectively denote the maximum and the minimum nonzero eigenvalue of matrix $A$. Note that $\mu^{0}$ induces uniform distribution over $\mathcal{M}$. Thus by $q_{n-1}^{\prime}=(1-\gamma) q_{n-1}+\gamma \mu^{0}$ we see that $p_{n-1}$ is a mixture of uniform distribution and the distribution induced by $q_{n-1}$. Note that, we have:

$$
\lambda_{\min }\left(\Sigma_{n-1}\right)=\min _{\|x\|_{2}=1, x \in \operatorname{span}(\mathcal{M})} x^{\top} \Sigma_{n-1} x .
$$

Moreover, we have

$$
x^{\top} \Sigma_{n-1} x=\mathbb{E}\left[x^{\top} M(n) M(n)^{\top} x\right]=\mathbb{E}\left[\left(M(n)^{\top} x\right)^{2}\right] \geq \gamma \mathbb{E}\left[\left(M^{\top} x\right)^{2}\right],
$$

where in the last inequality $M$ has law $\mu^{0}$. By definition, we have for any $x \in$ $\operatorname{span}(\mathcal{M})$ with $\|x\|_{2}=1$,

$$
\mathbb{E}\left[\left(M^{\top} x\right)^{2}\right] \geq \underline{\lambda},
$$

so that in the end, we get $\lambda_{\min }\left(\Sigma_{n-1}\right) \geq \gamma \underline{\lambda}$, and hence $\eta\left|\tilde{X}_{i}(n)\right| \leq \frac{\eta m^{3 / 2}}{\gamma \underline{\lambda}}, \forall i \in[d]$. Finally, we choose $\eta \leq \frac{\gamma \lambda}{m^{3 / 2}}$ to satisfy the condition for the inequality we used in 6.2).

We have

$$
\mathbb{E}_{n}[\tilde{X}(n)]=\mathbb{E}_{n}\left[Y_{n} \Sigma_{n-1}^{+} M(n)\right]=\mathbb{E}_{n}\left[\Sigma_{n-1}^{+} M(n) M(n)^{\top} X(n)\right]
$$

$$
=\Sigma_{n-1}^{+} \Sigma_{n-1} X(n)=\overline{X(n)}
$$

where the last equality follows from Lemma 6.1 and $\overline{X(n)}$ is the orthogonal projection of $X(n)$ onto $\operatorname{span}(\mathcal{M})$. In particular, for any $m q^{\prime} \in \operatorname{conv}(\mathcal{M})$, we have

$$
\mathbb{E}_{n}\left[m q^{\prime \top} \tilde{X}(n)\right]=m q^{\prime \top} \overline{X(n)}=m q^{\prime \top} X(n)
$$

Moreover, we have:

$$
\begin{aligned}
\mathbb{E}_{n}\left[q_{n-1}^{\top} \tilde{X}^{2}(n)\right] & =\sum_{i \in[d]} q_{n-1}(i) \mathbb{E}_{n}\left[\tilde{X}_{i}^{2}(n)\right] \\
& =\sum_{i \in[d]} \frac{q_{n-1}^{\prime}(i)-\gamma \mu^{0}(i)}{1-\gamma} \mathbb{E}_{n}\left[\tilde{X}_{i}^{2}(n)\right] \\
& \leq \frac{1}{m(1-\gamma)} \sum_{i \in[d]} m q_{n-1}^{\prime}(i) \mathbb{E}_{n}\left[\tilde{X}_{i}^{2}(n)\right] \\
& =\frac{1}{m(1-\gamma)} \mathbb{E}_{n}\left[\sum_{i \in[d]} \tilde{M}_{i}(n) \tilde{X}_{i}^{2}(n)\right]
\end{aligned}
$$

where $\tilde{M}(n)$ is a random arm with the same law as $M(n)$ and independent of $M(n)$. Note that $\tilde{M}_{i}^{2}(n)=\tilde{M}_{i}(n)$, so that we have

$$
\begin{aligned}
\mathbb{E}_{n}[ & \left.\sum_{i \in[d]} \tilde{M}_{i}(n) \tilde{X}_{i}^{2}(n)\right] \\
& =\mathbb{E}_{n}\left[X(n)^{\top} M(n) M(n)^{\top} \Sigma_{n-1}^{+} \tilde{M}(n) \tilde{M}(n)^{\top} \Sigma_{n-1}^{+} M(n) M(n)^{\top} X(n)\right] \\
& \leq m^{2} \mathbb{E}_{n}\left[M(n)^{\top} \Sigma_{n-1}^{+} M(n)\right],
\end{aligned}
$$

where we used the bound $M(n)^{\top} X(n) \leq m$. By [79, Lemma 15], $\mathbb{E}_{n}\left[M(n)^{\top} \Sigma_{n-1}^{+} M(n)\right] \leq$ $d$, so that we have:

$$
\mathbb{E}_{n}\left[q_{n-1}^{\top} \tilde{X}^{2}(n)\right] \leq \frac{m d}{1-\gamma}
$$

Observe that

$$
\begin{aligned}
\mathbb{E}_{n}\left[q^{\star \top} \tilde{X}(n)-q_{n-1}^{\prime \top} \tilde{X}(n)\right] & =\mathbb{E}_{n}\left[q^{\star \top} \tilde{X}(n)-(1-\gamma) q_{n-1}^{\top} \tilde{X}(n)-\gamma \mu^{0 \top} \tilde{X}(n)\right] \\
& =\mathbb{E}_{n}\left[q^{\star \top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right]+\gamma q_{n-1}^{\top} X(n)-\gamma \mu^{0 \top} X(n) \\
& \leq \mathbb{E}_{n}\left[q^{\star \top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right]+\gamma q_{n-1}^{\top} X(n) \\
& \leq \mathbb{E}_{n}\left[q^{\star \top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right]+\gamma .
\end{aligned}
$$

Using Lemma 6.2 and the above bounds, we get with $m q^{\star}$ the optimal arm, i.e. $q^{\star}(i)=\frac{1}{m}$ iff $M_{i}^{\star}=1$,

$$
R^{\mathrm{CombEXP}}(T)=\mathbb{E}\left[\sum_{n=1}^{T} m q^{\star \top} \tilde{X}(n)-\sum_{n=1}^{T} m q_{n-1}^{\prime \top} \tilde{X}(n)\right]
$$

$$
\begin{aligned}
& \leq \mathbb{E}\left[\sum_{n=1}^{T} m q^{\star \top} \tilde{X}(n)-\sum_{n=1}^{T} m q_{n-1}^{\top} \tilde{X}(n)\right]+m \gamma T \\
& \leq \frac{\eta m^{2} d T}{1-\gamma}+\frac{m \log \mu_{\min }^{-1}}{\eta}+m \gamma T
\end{aligned}
$$

since

$$
\mathrm{KL}\left(q^{\star}, q_{0}\right)=-\frac{1}{m} \sum_{i \in M^{\star}} \log m \mu_{i}^{0} \leq \log \mu_{\min }^{-1} .
$$

Choosing $\eta=\gamma C$ with $C=\frac{\lambda}{m^{3 / 2}}$ gives

$$
\begin{aligned}
R^{\mathrm{CombEXP}}(T) & \leq \frac{\gamma C m^{2} d T}{1-\gamma}+\frac{m \log \mu_{\min }^{-1}}{\gamma C}+m \gamma T \\
& =\frac{C m^{2} d+m-m \gamma}{1-\gamma} \gamma T+\frac{m \log \mu_{\min }^{-1}}{\gamma C} \\
& \leq \frac{\left(C m^{2} d+m\right) \gamma T}{1-\gamma}+\frac{m \log \mu_{\min }^{-1}}{\gamma C} .
\end{aligned}
$$

The proof is completed by setting

$$
\gamma=\frac{\sqrt{m \log \mu_{\min }^{-1}}}{\sqrt{m \log \mu_{\min }^{-1}}+\sqrt{C\left(C m^{2} d+m\right) T}} .
$$

## 6.B Proof of Proposition 6.1

Proof. Recall that $u_{n}=\arg \min _{p \in \mathcal{P}} \operatorname{KL}\left(p, \tilde{q}_{n}\right)$ and that $q_{n}$ is an $\varepsilon_{n}$-optimal solution for the projection step, that is

$$
\operatorname{KL}\left(u_{n}, \tilde{q}_{n}\right) \geq \operatorname{KL}\left(q_{n}, \tilde{q}_{n}\right)-\varepsilon_{n} .
$$

By Lemma B.1 and [96, Theorem 3.1], we have

$$
\begin{aligned}
\mathrm{KL}\left(q_{n}, \tilde{q}_{n}\right)-\mathrm{KL}\left(u_{n}, \tilde{q}_{n}\right) & \geq\left(q_{n}-u_{n}\right)^{\top} \nabla \mathrm{KL}\left(u_{n}, \tilde{q}_{n}\right)+\frac{1}{2}\left\|q_{n}-u_{n}\right\|_{1}^{2} \\
& \geq \frac{1}{2}\left\|q_{n}-u_{n}\right\|_{1}^{2}
\end{aligned}
$$

where we used $\left(q_{n}-u_{n}\right)^{\top} \nabla \operatorname{KL}\left(u_{n}, \tilde{q}_{n}\right) \geq 0$, which is due to the first-order optimality condition for $u_{n}$. Hence $\operatorname{KL}\left(q_{n}, \tilde{q}_{n}\right)-\operatorname{KL}\left(u_{n}, \tilde{q}_{n}\right) \leq \varepsilon_{n}$ implies that

$$
\left\|q_{n}-u_{n}\right\|_{\infty} \leq\left\|q_{n}-u_{n}\right\|_{1} \leq \sqrt{2 \varepsilon_{n}} .
$$

Consider $q^{\star}$, the distribution over $\mathcal{P}$ for the optimal arm, i.e. $q^{\star}(i)=\frac{1}{m}$ iff $M_{i}^{\star}=1$. Recall that from the proof of Lemma 6.2 for $q=q^{\star}$ we have

$$
\begin{equation*}
\operatorname{KL}\left(q^{\star}, \tilde{q}_{n}\right)-\operatorname{KL}\left(q^{\star}, q_{n-1}\right) \leq \eta q_{n-1}^{\top} \tilde{X}(n)-\eta{q^{\star}}^{\top} \tilde{X}(n)+\eta^{2} q_{n-1}^{\top} \tilde{X}^{2}(n) . \tag{6.4}
\end{equation*}
$$

Generalized Pythagorean Inequality (see [96, Theorem 3.1]) gives

$$
\begin{equation*}
\operatorname{KL}\left(q^{\star}, \tilde{q}_{n}\right) \geq \operatorname{KL}\left(q^{\star}, u_{n}\right)+\operatorname{KL}\left(u_{n}, \tilde{q}_{n}\right) \tag{6.5}
\end{equation*}
$$

Let $\underline{q}_{n}=\min _{i \in M^{\star}} q_{n}(i)$. Observe that

$$
\begin{aligned}
\mathrm{KL}\left(q^{\star}, u_{n}\right) & =\sum_{i \in[d]} q^{\star}(i) \log \frac{q^{\star}(i)}{u_{n}(i)} \\
& =-\frac{1}{m} \sum_{i \in M^{\star}} \log m u_{n}(i) \\
& \geq-\frac{1}{m} \sum_{i \in M^{\star}} \log m\left(q_{n}(i)+\sqrt{2 \varepsilon_{n}}\right) \\
& \geq-\frac{1}{m} \sum_{i \in M^{\star}} \log m q_{n}(i)-\frac{1}{m} \sum_{i \in M^{\star}} \log \left(1+\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}}\right) \\
& \geq-\frac{1}{m} \sum_{i \in M^{\star}} \log m q_{n}(i)-\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}} \\
& =\operatorname{KL}\left(q^{\star}, q_{n}\right)-\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}},
\end{aligned}
$$

where we used $\log (1+z) \leq z$ for all $z>-1$ in the last inequality. Plugging this into (6.5, we get

$$
\mathrm{KL}\left(q^{\star}, \tilde{q}_{n}\right) \geq \mathrm{KL}\left(q^{\star}, q_{n}\right)-\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}}+\mathrm{KL}\left(u_{n}, \tilde{q}_{n}\right) \geq \mathrm{KL}\left(q^{\star}, q_{n}\right)-\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}} .
$$

Putting this together with 6.4 yields

$$
\mathrm{KL}\left(q^{\star}, q_{n}\right)-\mathrm{KL}\left(q^{\star}, q_{n-1}\right) \leq \eta q_{n-1}^{\top} \tilde{X}(n)-\eta q^{\star}{ }^{\top} \tilde{X}(n)+\eta^{2} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{\sqrt{2 \varepsilon_{n}}}{\underline{q}_{n}} .
$$

Finally, summing over $n$ gives

$$
\sum_{n=1}^{T}\left(q^{\star}{ }^{\top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right) \leq \eta \sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{\operatorname{KL}\left(q^{\star}, q_{0}\right)}{\eta}+\frac{1}{\eta} \sum_{n=1}^{T} \frac{\sqrt{2 \varepsilon_{n}}}{q_{n}}
$$

Defining

$$
\varepsilon_{n}=\frac{\left(\underline{q}_{n} \log \mu_{\min }^{-1}\right)^{2}}{32 n^{2} \log ^{3}(n+1)}, \quad \forall n \geq 1
$$

and recalling that $\operatorname{KL}\left(q^{\star}, q_{0}\right) \leq \log \mu_{\min }^{-1}$, we get

$$
\begin{aligned}
\sum_{n=1}^{T}\left(q^{\star}{ }^{\top} \tilde{X}(n)-q_{n-1}^{\top} \tilde{X}(n)\right) & \leq \eta \sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{\log \mu_{\min }^{-1}}{\eta} \\
& +\frac{\log \mu_{\min }^{-1}}{\eta} \sum_{n=1}^{T} \sqrt{\frac{2}{32 n^{2} \log ^{3}(n+1)}} \\
& \leq \eta \sum_{n=1}^{T} q_{n-1}^{\top} \tilde{X}^{2}(n)+\frac{2 \log \mu_{\min }^{-1}}{\eta}
\end{aligned}
$$

where we used the fact $\sum_{n \geq 1} n^{-1}(\log (n+1))^{-3 / 2} \leq 4$. We remark that by the properties of the KL-divergence and since $q_{n-1}^{\prime} \geq \gamma \mu^{0}>0$, we have $\underline{q}_{n}>0$ at every round $n$, so that $\varepsilon_{n}>0$ at every round $n$.

Using the above result and following the same lines as in the proof of Theorem 6.1. we have

$$
R^{\mathrm{ComвEXP}}(T) \leq \frac{\eta m^{2} d T}{1-\gamma}+\frac{2 m \log \mu_{\min }^{-1}}{\eta}+m \gamma T
$$

Choosing $\eta=\gamma C$ with $C=\frac{\lambda}{m^{3 / 2}}$ gives

$$
R^{\mathrm{ComBEXP}}(T) \leq \frac{\left(C m^{2} d+m\right) \gamma T}{1-\gamma}+\frac{2 m \log \mu_{\min }^{-1}}{\gamma C}
$$

The proof is completed by setting $\gamma=\frac{\sqrt{2 m \log \mu_{\min }^{-1}}}{\sqrt{2 m \log \mu_{\min }^{-1}}+\sqrt{C\left(C m^{2} d+m\right) T}}$.

## 6.C Proof of Theorem 6.2

Proof. We calculate the time complexity of the various steps of CombEXP at round $n \geq 1$.
(i) Mixing: This step requires $\mathcal{O}(d)$ time.
(ii) Decomposition: Using the algorithm of [98], the vector $m q_{n-1}^{\prime}$ may be represented as a convex combination of at most $d+1$ arms in $\mathcal{O}\left(d^{4}\right)$ time, so that $p_{n-1}$ may have at most $d+1$ non-zero elements (observe that the existence of such a representation follows from Carathéodory Theorem).
(iii) Sampling: This step takes $\mathcal{O}(d)$ time since $p_{n-1}$ has at most $d+1$ non-zero elements.
(iv) Estimation: The construction of matrix $\Sigma_{n-1}$ is done in time $\mathcal{O}\left(d^{2}\right)$ since $p_{n}$ has at most $d+1$ non-zero elements and $M M^{\top}$ is formed in $\mathcal{O}(d)$ time. Computing the pseudo-inverse of $\Sigma_{n-1} \operatorname{costs} \mathcal{O}\left(d^{3}\right)$.
(v) Update: This step requires $\mathcal{O}(d)$ time.
(vi) Projection: The projection step is equivalent to solving a convex program up to accuracy $\varepsilon_{n}=\mathcal{O}\left(n^{-2} \log ^{-3}(n)\right)$. We use the Interior-Point Method (Barrier method). The total number of Newton iterations to achieve accuracy $\varepsilon_{n}$ is $\mathcal{O}\left(\sqrt{s} \log \left(s / \varepsilon_{n}\right)\right)$ [97, Chapter 11]. Moreover, the cost of each iteration is $\mathcal{O}\left((d+c)^{3}\right)$ [97] Chapter 10], so that the total cost of this step becomes $\mathcal{O}\left(\sqrt{s}(c+d)^{3} \log \left(s / \varepsilon_{n}\right)\right)$. Plugging $\varepsilon_{n}=\mathcal{O}\left(n^{-2} \log ^{-3}(n)\right)$ and noting that $\mathcal{O}\left(\sum_{n=1}^{T} \log \left(s / \varepsilon_{n}\right)\right)=\mathcal{O}(T \log (T))$, the cost of this step is $\mathcal{O}(\sqrt{s}(c+$ d) $\left.{ }^{3} T \log (T)\right)$.

Hence the total time complexity after $T$ rounds is $\mathcal{O}\left(T\left[\sqrt{s}(c+d)^{3} \log (T)+d^{4}\right]\right)$, which completes the proof.

## Chapter 7

## Conclusions and Future Work

In this chapter, we conclude this thesis by summarizing the main results and proposing some directions for future research.

### 7.1 Conclusions

In Chapter 33 we investigated stochastic combinatorial MABs with Bernoulli rewards. Leveraging the theory of optimal 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). In the case of semi-bandit feedback, we then investigated how this lower bound scales with the dimension of $\mathcal{M}$ for some problems of interests. We proposed the ESCB algorithm for the case of semi-bandit feedback and showed that its regret is growing at most as $\mathcal{O}\left(\sqrt{m} d \Delta_{\min }^{-1} \log (T)\right)$. 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 Eросн-ESCB, without providing any performance guarantee.

In Chapter 4, we focused on stochastic combinatorial MAB problems in which the underlying combinatorial structure is a matroid. Specializing the lower bounds of Chapter 3 to the case of matroids, we provided explicit regret lower bounds. In particular, the lower bound for the case of semi-bandit feedback extends the one proposed by Anantharam et al. [33] to the case of matroids. Moreover, for semi-bandit feedback we proposed KL-OSM, an algorithm based on the KL-UCB index and the Greedy algorithm. Thus it has polynomial time complexity in the independence oracle model. Through a finite-time regret analysis, we proved that KL-OSM achieves a regret (asymptotically) growing as the proposed lower bound and therefore, it is asymptotically optimal. To our best knowledge, this is the first optimal algorithm for this class of combinatorial MABs. Numerical experiments for some specific matroid problems validated that KL-OSM significantly outperforms existing algorithms.

In Chapter 5 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 which 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 $\mathcal{O}\left(\sqrt{m} d \Delta_{\min }^{-1} \theta_{\min }^{-3} \log (N)\right)$. Furthermore, we provided an improved regret bound for KL-SR [30] which grows as $\mathcal{O}\left(m d \Delta_{\min }^{-1} \theta_{\min }^{-3} \log (N)\right)$. 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. Numerical experiments also validated that these three policies outperform state-of-the-art algorithms.

Chapter 6 concerns adversarial combinatorial MAB problems in which all arms consist of the same number $m$ of basic actions. The core contribution of that chapter is Combexp, which is an OSMD-based algorithm for the case of bandit feedback. For most problems of interest, ComвEXP has a regret of $\mathcal{O}\left(\sqrt{m^{3} d \log (d / m) T}\right)$. The regret gap between CombeXP and the minimax lower bound $\Omega(m \sqrt{d T})$ [21] scales at most as $m^{1 / 2}$ up to a logarithmic factor. This is the same regret scaling of state-of-theart algorithms for these problems. We presented an analysis of the computational complexity of CombEXP, which can be extended to other OSMD-based algorithms, and hence might be of independent interests. In particular, we established that CombEXP admits lower computational complexity than state-of-the-art algorithms for some problems of interest.

### 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 Thompson Sampling for combinatorial MAB problems. One intriguing direction for future research is to analyze the performance of Thompson Sampling for the stochastic combinatorial MAB problems considered. Despite its popularity in the MAB literature, Thompson Sampling is seldom studied for combinatorial problems except for the recent work of Komiyama et al. [51, which concerns the very simple setting of fixed-size subsets. Regret analysis of Thompson Sampling for generic combinatorial structures proves quite challenging. Nonetheless, it is a promising direction since (i) if the offline problem is polynomial time
solvable, efficient implementation for Thompson Sampling might exist (because arm selection can be cast as the same linear combinatorial problem as the offline problem), (ii) in empirical evaluations Thompson Sampling exhibits superior performance than existing algorithms.

Reducing the gap in the regret analysis of ESCB-1. The current regret analysis of ESCB-1 in Chapter 3 does not seem to be tight. Our simulation experiments in Chapter 3 suggested that the asymptotic performance of ESCB-1 is close to the lower bound. Hence, we conjecture that through a more elegant regret analysis, one may be able to establish an optimal regret order of $\mathcal{O}\left((d-m) \Delta_{\min }^{-1} \log (T)\right)$ for ESCB-1.

Nonlinear reward functions. In this thesis we only concentrated on combinatorial problems with linear objective functions. Nonetheless, a lot of interesting applications may be cast as combinatorial MABs whose average reward function is nonlinear. 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. Numerous applications of combinatorial problems indeed fall within this framework, e.g., bidding in ad exchange [66], product search 67], leader selection in leader-follower multi-agent systems [70], coverage problem, influence maximization 101. We remark that there is an scarcity of results for stochastic MAB problems with submodular reward functions in the stochastic setting, though these problems have received more attention in the adversarial setting [102, 66].

Stochastic combinatorial MABs under full-bandit feedback. Stochastic combinatorial MABs under bandit feedback have seldom been studied, though the problem is very well investigated in the adversarial setting. Nonetheless, the necessity of corresponding policies is evident as bandit feedback makes more sense in many applications of interest. A notable instance is shortest-path routing, where in most scenarios, the decision maker has access to the end-to-end (bandit) feedback rather than per-link (semi-bandit) feedback. In this work, our results for bandit feedback mainly centered on derivation of regret lower bounds. An interesting direction is to devise algorithms that will work with bandit feedback and analyze their performance. This task could be much more complicated than the case of semi-bandit feedback. However, we conjecture that devising such an algorithm for the case of matroids might be relatively straightforward due to the unimodality of these structures.

Projection-free algorithms for adversarial combinatorial MABs. The optimal algorithm for adversarial combinatorial MABs under semi-bandit feedback is OSMD [21], which relies on projection with Bregman divergences (see Chapter 6 for details). Such projections suffer from two drawbacks: Firstly, as our analysis in Chapter 6 shows, the computational complexity of projection is determined by the representation of $\operatorname{conv}(\mathcal{M})$, i.e., the convex hull of the set of arms. There exist
concrete examples of $\mathcal{M}$ where the number of inequalities representing $\operatorname{conv}(\mathcal{M})$ (and hence, the per round time complexity) is not polynomial in $d$, yet the corresponding offline problem is polynomial time solvable. One such examples of $\mathcal{M}$ is spanning trees (see Chapter 6 for a related discussion). Secondly, projections often introduce precision issues which in turn make the cumulative computational complexity (namely, after $T$ rounds) increase super-linearly with $T$. An intriguing research direction is to devise an algorithm which is not relying on projection while it achieves the optimal regret of $\sqrt{m d T}$ [21]. We remark that under semi-bandit feedback and when the offline problem is efficiently solvable, an efficient algorithm called FPL-GR already appears in [87] whose expected time complexity grows linearly with $T$. However, its worst case time complexity might grows super-linearly with $T$ and its regret is worse than the lower bound of $\sqrt{m d T}$ by a factor $\sqrt{m \log (d)}$. We remark that only for the very simple case of fixed-size subsets, an algorithm with optimal regret and linear (in $T$ ) cumulative time complexity has been provided [86].

For the case of bandit feedback, state-of-the-algorithms suffer from both aforementioned drawbacks. For example, when $\mathcal{M}$ is the set of spanning trees, neither Combexp nor Comband admit polynomial time complexity. We note that for learning permutations, Ailon et al. 99 present an algorithm, whose regret is worse than state-of-the-art by factor of $\sqrt{m}$, yet has a time complexity growing linearly with $T$.

Shortest-path routing policies with smaller regret. In Chapter 5, it was shown that the dependence of the proposed algorithms on $\theta_{\min }$ and $m$ may not be tight. In particular, we conjecture that the optimal regret upper bound should grow at most as $\mathcal{O}\left((d-m) \Delta_{\min }^{-1} \theta_{\min }^{-1} \log (N)\right)$. As a future direction, we wish to propose an index policy whose regret upper bound is growing proportional to $\theta_{\min }^{-1} \log (N)$ (rather than $\theta_{\min }^{-3} \log (N)$ ). A possible approach to this end could be to employ the following index, defined similarly to the index $c_{M}$ in Chapter 5 .

$$
c_{M}^{\prime}(n)=M^{\top} \hat{\theta}(n)^{-1}-\sqrt{\frac{f_{1}(n)}{2} \sum_{i \in M} \frac{1}{\hat{\theta}_{i}(n) s_{i}(n)}} .
$$

Numerical experiments, however, have shown that the resulting policy would not exhibit elegant behavior for short packet horizons in all scenarios. Indeed, for some link parameters, this new policy beats our source routing policies only when $N$ grows very large. As a future research direction, we would like to remedy this issue by designing smarter routing policies, which may use warm start phases to circumvent the aforementioned problem.

## Appendix A

## Concentration Inequalities

This appendix is devoted to the overview of some important concentration inequalities used in various chapters of this thesis.

Theorem A. 1 (Chernoff-Hoeffding Bound). Let $X_{1}, \ldots, X_{n}$ be 0-1 independent random variables with $\mathbb{E}\left[X_{i}\right]=p_{i}$. 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 $0<\lambda<1-\mu$,

$$
\mathbb{P}[X \geq \mu+\lambda] \leq e^{-n \mathrm{kl}(\mu+\lambda, \mu)}
$$

and for all $0<\lambda<\mu$,

$$
\mathbb{P}[X \leq \mu-\lambda] \leq e^{-n \mathrm{kl}(\mu-\lambda, \mu)}
$$

Theorem A. 2 (Chernoff-Hoeffding Bound). Let $X_{1}, \ldots, X_{n}$ be random variables with common ranges $[0,1]$ and such that $\mathbb{E}\left[X_{t} \mid X_{1} \ldots, X_{t-1}\right]=\mu$. Let $S_{n}=\sum_{t=1}^{n} X_{t}$. Then for all $a \geq 0$ :

$$
\begin{aligned}
& \mathbb{P}\left[S_{n} \geq n \mu+a\right] \leq e^{-2 a^{2} / n} \\
& \mathbb{P}\left[S_{n} \leq n \mu-a\right] \leq e^{-2 a^{2} / n}
\end{aligned}
$$

The following result from [39] gives the concentration for self-normalized form of bounded random variables.

Theorem A. 3 ([39, Theorem 10]). Let $\left(X_{t}\right)_{t \geq 1}$ be a sequence of independent random variables bounded in $[0,1]$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with common expectation $\mu=\mathbb{E}\left[X_{t}\right]$. Let $\mathcal{F}_{t}$ be an increasing sequence of $\sigma$-fields of $\mathcal{F}$ such that for each $t, \sigma\left(X_{1}, \ldots, X_{t}\right) \subset \mathcal{F}_{t}$ and for $s>t$, $X_{s}$ is independent from $\mathcal{F}_{t}$. Consider a previsible sequence $\left(\varepsilon_{t}\right)_{t \geq 1}$ of Bernoulli variables (for all $t>0, \varepsilon_{t}$ is $\mathcal{F}_{t-1}$-measurable). Let $\delta>0$ and for every $t \in[n]$ let

$$
S(t)=\sum_{s=1}^{t} \varepsilon_{s} X_{s}, \quad N(t)=\sum_{s=1}^{t} \varepsilon_{s}, \quad \hat{\mu}(t)=\frac{S(t)}{N(t)},
$$

$$
u(n)=\max \{q>\hat{\mu}(n): N(n) \operatorname{kl}(\hat{\mu}(n), q) \leq \delta\}
$$

Then

$$
\mathbb{P}[u(n)<\mu] \leq\lceil\delta \log (n)\rceil e^{-(\delta+1)}
$$

The following theorem is a generalization of Theorem A.3 and gives a concentration inequality on sums of empirical KL-divergences.

Theorem A. 4 ([60, Theorem 2]). For all $\delta \geq K+1$ and $n \in \mathbb{N}$ we have:

$$
\mathbb{P}\left[\sum_{i=1}^{K} N_{i}(n) \mathrm{kl}\left(\hat{\mu}_{i}(n), \mu_{i}\right) \geq \delta\right] \leq\left(\frac{\lceil\delta \log (n)\rceil \delta}{K}\right)^{K} e^{-\delta(K+1)}
$$

In particular, the following corollary proves instrumental in the regret analysis of various algorithms for stochastic combinatorial MABs.

Corollary A.1. There exists a constant $C_{K}$ that only depends on $K$, such that for all $n \geq 2$ we have:

$$
\mathbb{P}\left[\sum_{i=1}^{K} N_{i}(n) \mathrm{kl}\left(\hat{\mu}_{i}(n), \mu_{i}\right) \geq \log (n)+4 K \log (\log (n))\right] \leq C_{K} n^{-1}(\log (n))^{-2}
$$

The following lemma proves useful in the proof of various regret bounds throughout the thesis. It asserts that if a set of instants $\Lambda$ can be decomposed into a family of singletons such that the arm $i$ is drawn sufficiently many times, then the number of times in $\Lambda$ (in expectations) at which the empirical average reward of $i$ is badly estimated is finite.

Lemma A. 1 ([103, Theorem B.1]). Let $i \in\{1, \ldots, K\}$ and $\delta>0$. Define $\mathcal{F}_{n}$ the $\sigma$-algebra generated by $\left(X_{i}(t)\right)_{1 \leq t \leq n, 1 \leq i \leq K}$. Let $\Lambda \subset \mathbb{N}$ be a (random) set of instants. Assume that there exists a sequence of (random) sets $(\Lambda(s))_{s \geq 1}$ such that (i) $\Lambda \subset \cup_{s \geq 1} \Lambda(s)$, (ii) for all $s \geq 1$ and all $n \in \Lambda(s), N_{i}(n) \geq \varepsilon s$, (iii) $|\Lambda(s)| \leq 1$, and (iv) the event $n \in \Lambda(s)$ is $\mathcal{F}_{n}$-measurable. Then, for all $\delta>0$ :

$$
\mathbb{E}\left[\sum_{n \geq 1} \mathbb{1}\left\{n \in \Lambda,\left|\hat{\mu}_{i}(n)-\mu_{i}\right| \geq \delta\right\}\right] \leq \frac{1}{\varepsilon \delta^{2}}
$$

The proof of the above lemma leverages a concentration inequality proposed in [103]. 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 $A_{n}$ be the event of sampling basic action $i \in\{1, \ldots, K\}$ at round $n$.

Corollary A.2. For all $i \in\{1, \ldots, K\}$ and all $\delta>0$ :

$$
\mathbb{E}\left[\sum_{n \geq 1} \mathbb{1}\left\{A_{n},\left|\hat{\mu}_{i}(n)-\mu_{i}\right| \geq \delta\right\}\right] \leq \frac{1}{\delta^{2}}
$$

Proof. Let $\Lambda=\left\{n: \mathbb{1}\left\{A_{n}\right\}=1\right\}$. Observe that for each $s \in \mathbb{N}$, there exists at most one time index $\phi_{s} \in \mathbb{N}$ such that $N_{i}\left(\phi_{s}\right)=s$ and $\phi_{s} \in \Lambda$, since $N_{i}(n)=N_{i}(n-1)+1$ for all $n \in \Lambda$. The set $\Lambda$ is included in $\cup_{s \geq 1}\left\{\phi_{s}\right\}$. The announced result is then a direct consequence of Lemma A.1 with $\varepsilon=1$.

We note that a slightly worse bound can be obtained from [43, Lemma 3].

## Appendix B

## Properties of the KL-Divergence

In this appendix we briefly overview some of the properties of the Kullback-Leibler divergence (henceforth, the KL-divergence), which prove instrumental throughout this thesis. The KL-divergence, originally introduced by Kullback and Leibler in [104], 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 results presented here can be found in, e.g., [105] and (96].

## B. 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

$$
\mathrm{KL}(F, G)=\mathbb{E}_{F}\left[\log \frac{F(\mathrm{~d} x)}{G(\mathrm{~d} x)}\right]=\int_{\mathcal{X}} \log \frac{F(\mathrm{~d} x)}{G(\mathrm{~d} x)} F(\mathrm{~d} x),
$$

where $F(\mathrm{~d} x) / G(\mathrm{~d} x)$ denotes the Radon-Nikodym derivative of $F$ with respect to $G$. $\mathrm{KL}(F, G)$ may be derived using densities as well: Let $m(\mathrm{~d} x)$ be an appropriate measure. Then,

$$
\mathrm{KL}(F, G)=\int_{\mathcal{X}} \log \frac{f(x)}{g(x)} f(x) m(\mathrm{~d} x) .
$$

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 $\mathrm{KL}(F, G)=\infty$.

In the discrete case, namely when $F$ and $G$ are probability vectors, the above definition reads

$$
\mathrm{KL}(F, G)=\sum_{i} F_{i} \log \frac{F_{i}}{G_{i}},
$$

with the usual convention where $p \log \frac{p}{q}$ is defined to be 0 if $p=0$ and $+\infty$ if $p>q=0$. Here we mainly concern the KL-divergence between two discrete distributions.

## B. 2 Properties

## B.2.1 Non-negativity

The KL-divergence is always non-negative: $\operatorname{KL}(F, G) \geq 0$ with equality if $F(x)=$ $G(x), \forall x \in \mathcal{X}$. Furthermore, the KL-divergence between two probability vectors $F$ and $G$ is lower bounded as follows:

$$
\mathrm{KL}(F, G) \geq \frac{1}{2}\|F-G\|_{1}^{2}
$$

This inequality is known as Pinsker's inequality.

## B.2.2 Convexity

The KL-divergence $\operatorname{KL}(F, G)$ is convex in both arguments. The following result states strong convexity of the KL-divergence in the first argument.

Lemma B.1. Let $q \in \mathbb{R}_{++}^{d}$ be a probability vector. Then, the KL-divergence $z \mapsto$ $\mathrm{KL}(z, q)$ is 1-strongly convex with respect to the $\|\cdot\|_{1}$ norm.

Proof. To prove the lemma, it suffices to show that for all $d$-dimensional probability vectors $x, y$ :

$$
(\nabla \mathrm{KL}(x, q)-\nabla \mathrm{KL}(y, q))^{\top}(x-y) \geq\|x-y\|_{1}^{2}
$$

We have

$$
\begin{aligned}
(\nabla \mathrm{KL}(x, q)-\nabla \mathrm{KL}(y, q))^{\top}(x-y) & =\sum_{i \in[d]}\left(1+\log \frac{x_{i}}{q_{i}}-1-\log \frac{y_{i}}{q_{i}}\right)\left(x_{i}-y_{i}\right) \\
& =\sum_{i \in[d]}\left(1+\log x_{i}-1-\log y_{i}\right)\left(x_{i}-y_{i}\right) \\
& =\left(\nabla \sum_{i \in[d]} x_{i} \log x_{i}-\nabla \sum_{i \in[d]} y_{i} \log y_{i}\right)^{\top}(x-y) \\
& \geq\|x-y\|_{1}^{2},
\end{aligned}
$$

where the last inequality follows from strong convexity of the negative entropy function $z \mapsto \sum_{i \in[d]} z_{i} \log z_{i}$ with respect to the $\|\cdot\|_{1}$ norm [106, Proposition 5.1].

We recall from Chapter 6 that the KL-divergence is a specific case of Bregman divergence defined for the negative entropy function. More precisely, for two probability vectors $F, G \in \mathbb{R}_{++}^{d}$ :

$$
\mathrm{KL}(F, G)=h(F)-h(G)-(F-G)^{\top} \nabla h(G)
$$

with $h(x)=\sum_{i} x_{i} \log x_{i}$.
We now state a useful inequality, known as Generalized Pythagorean Inequality, which is valid for any Bregman divergence. Consider $y \in \mathbb{R}^{d}$ and let $F^{\star}$ be the projection of $y$ onto a convex set $\Xi \subset \mathbb{R}^{d}$ using the KL-divergence, i.e.

$$
F^{\star}=\underset{F \in \Xi}{\operatorname{argmin}} \operatorname{KL}(F, y) .
$$

Then, Generalized Pythagorean Inequality [107] (see also [96, Theorem 3.1]) states that for all $x \in \Xi$ :

$$
\mathrm{KL}\left(x, F^{\star}\right)+\mathrm{KL}\left(F^{\star}, y\right) \leq \mathrm{KL}(x, y) .
$$

In particular, when $\Xi$ is an affine set, the above inequality holds with equality.

## B.2.3 Chain Rule

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.

Theorem B. 1 (Chain Rule). For two random variables $x, y \in \mathcal{X}$ we have:

$$
\mathrm{KL}(F(x, y), G(x, y))=\mathrm{KL}(F(x), G(x))+\mathrm{KL}(F(y \mid x), G(y \mid x))
$$

where $\operatorname{KL}(F(y \mid x), G(y \mid x))=\mathbb{E}_{x}[\log (F(y \mid x) / G(y \mid x))]$.
A consequence of this result is that the KL-divergence is additive for independent random variables.

## B. 3 The KL-Divergence between Two Bernoulli Distributions

The KL-divergence between two Bernoulli distributions with respective parameters $p$ and $q$, denoted by $\operatorname{kl}(p, q)$, is:

$$
\operatorname{kl}(p, q)=p \log \frac{p}{q}+(1-p) \log \frac{1-p}{1-q}
$$

The function kl is sometimes referred to as the binary relative entropy. The following lemma provides bounds for the KL-divergence of Bernoulli distributions.

Lemma B.2. For any $p, q \in[0,1]$, it holds that

$$
2(p-q)^{2} \leq \mathrm{kl}(p, q) \leq \frac{(p-q)^{2}}{q(1-q)}
$$

We note that the lower bound in this lemma is a consequence of Pinsker's inequality.

Next, we present some properties of the function $\operatorname{kl}(p, q)$. For all $p \in[0,1]$ :
(i) $q \mapsto \operatorname{kl}(p, q)$ is strictly convex on $[0,1]$ and attains its minimum at $p$, with $\operatorname{kl}(p, p)=0$.
(ii) Its derivative with respect to the second parameter $q \mapsto \mathrm{kl}^{\prime}(p, q)=\frac{q-p}{q(1-q)}$ is strictly increasing on $(p, 1)$.
(iii) For $p<1$, we have $\mathrm{kl}(p, q) \underset{q \rightarrow 1^{-}}{\rightarrow} \infty$ and $\mathrm{kl}^{\prime}(p, q) \underset{q \rightarrow 1^{-}}{\rightarrow} \infty$.

The following lemma relates the KL-divergence between two geometric distributions to that of corresponding Bernoulli distributions.

Lemma B.3. For any $u, v \in(0,1]$, we have:

$$
\operatorname{KLG}(u, v)=\frac{\operatorname{kl}(u, v)}{u}
$$

Proof. We have:

$$
\begin{aligned}
\operatorname{KLG}(u, v) & =\sum_{i=1}^{\infty} u(1-u)^{i-1} \log \frac{u(1-u)^{i-1}}{v(1-v)^{i-1}} \\
& =\sum_{i=1}^{\infty} u(1-u)^{i-1} \log \frac{u}{v}+\sum_{i=1}^{\infty}(i-1) u(1-u)^{i-1} \log \frac{1-u}{1-v} \\
& =\log \frac{u}{v}+\frac{1-u}{u} \log \frac{1-u}{1-v}=\frac{\operatorname{kl}(u, v)}{u} .
\end{aligned}
$$

## Bibliography

[1] H. Robbins, "Some aspects of the sequential design of experiments," in Herbert Robbins Selected Papers. Springer, 1985, pp. 169-177.
[2] 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.
[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," Advances in Applied Mathematics, vol. 6, no. 1, pp. 4-22, 1985.
[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS), 1995, pp. 322-331.
[5] ——, "The nonstochastic multiarmed bandit problem," SIAM Journal on Computing, vol. 32, no. 1, pp. 48-77, 2002.
[6] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," The Annals of Statistics, pp. 1091-1114, 1987.
[7] 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. 731745, 2011.
[8] 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.
[9] R. Kleinberg and T. Leighton, "The value of knowing a demand curve: Bounds on regret for online posted-price auctions," in Proceedings of the 44 th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2003, pp. 594-605.
[10] 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.
[11] 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.
[12] 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, 2010, pp. 661-670.
[13] R. Combes, S. Magureanu, A. Proutiere, and C. Laroche, "Learning to rank: Regret lower bounds and efficient algorithms," in Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2015, pp. 231-244.
[14] J. Abernethy, E. Hazan, and A. Rakhlin, "Competing in the dark: An efficient algorithm for bandit linear optimization." in Proceedings of the 21st Annual Conference on Learning Theory (COLT), 2008, pp. 263-274.
[15] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade, "Towards minimax policies for online linear optimization with bandit feedback," in Proceedings of the 25th Annual Conference on Learning Theory (COLT), 2012, pp. 41.1-41.14.
[16] V. Dani, S. M. Kakade, and T. P. Hayes, "The price of bandit information for online optimization," in Advances in Neural Information Processing Systems 20 (NIPS), 2007, pp. 345-352.
[17] 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.
[18] 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.
[19] 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.
[20] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled markov chains," SIAM J. Control and Optimization, vol. 35, no. 3, pp. 715-743, 1997.
[21] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," Mathematics of Operations Research, vol. 39, no. 1, pp. 31-45, 2013.
[22] M. Lelarge, A. Proutiere, and M. S. Talebi, "Spectrum bandit optimization," in Proceedings of Information Theory Workshop (ITW), 2013, pp. 34-38.
[23] A. Schrijver, Combinatorial Optimization: Polyhedra and Efficiency. Springer, 2003.
$[24]$ 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.
[25] 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.
[26] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN), 2010, pp. 1-9.
[27] 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.
[28] A. György, T. Linder, and G. Ottucsak, "The shortest path problem under partial monitoring," in Learning Theory, ser. Lecture Notes in Computer Science, G. Lugosi and H. U. Simon, Eds. Springer Berlin Heidelberg, 2006, vol. 4005, pp. 468-482.
[29] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The on-line shortest path problem under partial monitoring." Journal of Machine Learning Research, vol. 8, no. 10, 2007.
[30] 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.
[31] 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.
[32] E. Kaufmann, "Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources," Doctorat, ParisTech., Jul, vol. 31, 2015.
[33] 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.
[34] ——, "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.
[35] A. Burnetas and M. Katehakis, "Optimal adaptive policies for sequential allocation problems," Advances in Applied Mathematics, vol. 17, no. 2, pp. 122-142, 1996.
[36] A. Shapiro, "Semi-infinite programming, duality, discretization and optimality conditions," Optimization, vol. 58, no. 2, pp. 133-161, 2009.
[37] J. C. Gittins, "Bandit processes and dynamic allocation indices," Journal of the Royal Statistical Society. Series B (Methodological), pp. 148-177, 1979.
[38] R. Agrawal, "Sample mean based index policies with $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem," Advances in Applied Probability, pp. 10541078, 1995.
[39] 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.
[40] 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.
[41] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz et al., "Kullbackleibler upper confidence bounds for optimal sequential allocation," The Annals of Statistics, vol. 41, no. 3, pp. 1516-1541, 2013.
[42] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multiarmed bandit problem," in Proceedings of the 25th Annual Conference on Learning Theory (COLT), 2012, pp. 39.1-39.26.
[43] ——, "Further optimal regret bounds for Thompson sampling," in Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), 2012, pp. 99-107.
[44] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in Algorithmic Learning Theory. Springer, 2012, pp. 199-213.
[45] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of computer and system sciences, vol. 55, no. 1, pp. 119-139, 1997.
[46] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: a meta-algorithm and applications." Theory of Computing, vol. 8, no. 1, pp. 121-164, 2012.
[47] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," Information and computation, vol. 108, no. 2, pp. 212-261, 1994.
[48] V. G. Vovk, "Aggregating strategies," in Proceedings of the 3rd Workshop on Computational Learning Theory. Morgan Kaufmann, 1990, pp. 371-383.
[49] J.-Y. Audibert and S. Bubeck, "Regret bounds and minimax policies under partial monitoring," The Journal of Machine Learning Research, vol. 11, pp. 2785-2836, 2010.
[50] 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.
[51] 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.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] 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.
[57] 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.
[58] 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.
[59] -_, "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.
[60] 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.
[61] 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
[62] J. Edmonds, "Matroids and the greedy algorithm," Mathematical programming, vol. 1, no. 1, pp. 127-136, 1971.
[63] S. Onn, "Convex matroid optimization," SIAM Journal on Discrete Mathematics, vol. 17, no. 2, pp. 249-253, 2003.
[64] Y. Berstein, J. Lee, H. Maruri-Aguilar, S. Onn, E. Riccomagno, R. Weismantel, and H. Wynn, "Nonlinear matroid optimization and experimental design," SIAM Journal on Discrete Mathematics, vol. 22, no. 3, pp. 901-919, 2008.
[65] M. S. Talebi and A. Proutiere, "An optimal algorithm for stochastic matroid bandit optimization," submitted to International Conference on Autonomous Agents and Multiagent systems (AAMAS), 2016.
[66] M. Streeter, D. Golovin, and A. Krause, "Online learning of assignments," in Advances in Neural Information Processing Systems 22 (NIPS), 2009, pp. 1794-1802.
[67] 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, 2013, pp. 32-40.
[68] J. Bragg, A. Kolobov, M. Mausam, and D. S. Weld, "Parallel task routing for crowdsourcing," in Second AAAI Conference on Human Computation and Crowdsourcing, 2014.
[69] F. Lin, M. Fardad, and M. R. Jovanović, "Algorithms for leader selection in large dynamical networks: Noise-corrupted leaders," in Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), 2011, pp. 2932-2937.
[70] 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.
[71] J. G. Oxley, Matroid theory. Oxford university press, 2006, vol. 3.
[72] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, "Stochastic online shortest path routing: The value of feedback," arXiv preprint arXiv:1309.7367, 2015.
[73] 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), 2012, pp. 232-237.
[74] 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.
[75] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley-Interscience, 2005.
[76] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," The Journal of Machine Learning Research, vol. 99, pp. 1563-1600, 2010.
[77] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and kullback-leibler divergence," in Proceedings of the 48 th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2010, pp. 115-122.
[78] A. Sen and N. Balakrishnan, "Convolution of geometrics and a reliability problem," Statistics \& Probability Letters, vol. 43, no. 4, pp. 421-426, Jul. 1999.
[79] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," Journal of Computer and System Sciences, vol. 78, no. 5, pp. 1404-1422, 2012.
[80] W. M. Koolen, M. K. Warmuth, and J. Kivinen, "Hedging structured concepts." in Proceedings of the 23rd Annual Conference on Learning Theory (COLT), 2010, pp. 93-105.
[81] E. Takimoto and M. K. Warmuth, "Path kernels and multiplicative updates," The Journal of Machine Learning Research, vol. 4, pp. 773-818, 2003.
[82] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in Computational learning theory. Springer, 1995, pp. 23-37.
[83] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," Journal of Computer and System Sciences, vol. 71, no. 3, pp. 291-307, 2005.
[84] D. P. Helmbold and M. K. Warmuth, "Learning permutations with exponential weights," The Journal of Machine Learning Research, vol. 10, pp. 1705-1736, Dec. 2009.
[85] S. Kale, L. Reyzin, and R. Schapire, "Non-stochastic bandit slate problems," in Advances in Neural Information Processing Systems 23 (NIPS), 2010, pp. 1054-1062.
[86] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in Algorithmic Learning Theory. Springer, 2010, pp. 375-389.
[87] G. Neu and G. Bartók, "An efficient algorithm for learning with semi-bandit feedback," in Algorithmic Learning Theory. Springer, 2013, pp. 234-248.
[88] G. Neu, "First-order regret bounds for combinatorial semi-bandits," in Proceedings of the 28th Annual Conference on Learning Theory (COLT), pp. 1360-1375.
[89] H. B. McMahan and A. Blum, "Online geometric optimization in the bandit setting against an adaptive adversary," in Proceedings of the 17th Annual Conference on Learning Theory (COLT), 2004, pp. 109-123.
[90] P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari, "High-probability regret bounds for bandit online linear optimization," in Proceedings of the 21st Annual Conference on Learning Theory (COLT), 2008, pp. 335-342.
[91] A. S. Nemirovsky and D. B. Yudin, Problem complexity and method efficiency in optimization. John Wiley \& Sons, 1983.
[92] A. Ben-Tal, T. Margalit, and A. Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," SIAM Journal on Optimization, vol. 12, no. 1, pp. 79-108, 2001.
[93] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in Proceedings of the 20th International Conference on Machine Learning (ICML), 2003, pp. 928-936.
[94] J.-B. Hiriart-Urruty and C. Lemaréchal, Convex analysis and minimization algorithms I: fundamentals. Springer Science \& Business Media, 2013, vol. 305.
[95] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games. Cambridge University Press, 2006.
[96] I. Csiszár and P. Shields, Information theory and statistics: A tutorial. Now Publishers Inc, 2004.
[97] S. Boyd and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.
[98] H. D. Sherali, "A constructive proof of the representation theorem for polyhedral sets based on fundamental definitions," American Journal of Mathematical and Management Sciences, vol. 7, no. 3-4, pp. 253-270, 1987.
[99] N. Ailon, K. Hatano, and E. Takimoto, "Bandit online optimization over the permutahedron," in Algorithmic Learning Theory. Springer, 2014, pp. 215-229.
[100] G. M. Ziegler, Lectures on polytopes. Springer Science \& Business Media, 1995, vol. 152.
[101] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137-146.
[102] 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.
[103] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in Proceedings of the 31st International Conference on Machine Learning (ICML), 2014, pp. 521-529.
[104] S. Kullback, Information theory and statistics. Courier Corporation, 1968.
[105] T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley \& Sons, 2012.
[106] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," Operations Research Letters, vol. 31, no. 3, pp. 167-175, 2003.
[107] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," USSR computational mathematics and mathematical physics, vol. 7, no. 3, pp. 200-217, 1967.


[^0]:    ${ }^{1}$ The first algorithm for MAB problems, however is due to Thompson [2] which dates back to 1933.

[^1]:    ${ }^{2}$ In this thesis, we will use the terms 'online combinatorial optimization', 'combinatorial MAB', and 'combinatorial bandit', interchangeably,
    ${ }^{3}$ The term 'semi-bandit feedback' was introduced by Audibert et al. 18. Note that this type of feedback is only relevant for combinatorial problems.

[^2]:    ${ }^{4}$ This type of adversary is referred to as oblivious adversary. The case of adaptive adversary, which conditions the reward at each round on the history of plays, will not be addressed here.

[^3]:    ${ }^{5}$ In some works, interference graph is referred to as conflict graph.

[^4]:    ${ }^{6}$ This model assumes that the interference graph is the same over the various channels. This assumption, however, can be relaxed.

[^5]:    ${ }^{1}$ With some abuse of notation, hereafter we write $\operatorname{KL}\left(\theta, \theta^{\prime}\right)$ to indicate $\operatorname{KL}\left(\nu(\theta), \nu\left(\theta^{\prime}\right)\right)$.
    ${ }^{2}$ A simplified proof of this result can be found in [32 Chapter 1].

[^6]:    ${ }^{3}$ Of course, for loss minimization we are interested in lower confidence bounds.

[^7]:    ${ }^{4}$ Auer et al. originally provided the following regret upper bound for UCB1:

    $$
    R(T) \leq 8 \sum_{i>1} \frac{\log (T)}{\Delta_{i}}+\left(1+\frac{\pi^{2}}{3}\right) \sum_{i>1} \Delta_{i}
    $$

[^8]:    ${ }^{5}$ Multiplicative weight algorithm has a plethora applications in various domains of optimization and algorithm design beyond online learning. We refer to the survey paper by Arora et al. [46].

[^9]:    ${ }^{1}$ In 57, 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 54.

[^10]:    ${ }^{3}$ See Chapter 2 for a brief summary.

[^11]:    ${ }^{1}$ For any set $X$ and element $\ell$, by a slight abuse of notation, we write $X \backslash \ell$ to imply $X \backslash\{\ell\}$.

[^12]:    ${ }^{2}$ In some papers, the notion of partition matriod is defined with $k_{i}=1$ for every $i \in[l]$.

[^13]:    ${ }^{1}$ This improves over the regret upper bound scaling as $\mathcal{O}\left(\frac{\Delta_{\max } d m^{3}}{\Delta_{\min } \theta_{\min }^{3}} \log (N)\right)$ derived in 30.

[^14]:    ${ }^{2} \mathrm{~A}$ barycentric spanner is a set of paths from which the delay of all other paths can be computed as its linear combination with coefficients in $[-1,1]$ [27].

[^15]:    ${ }^{3}$ The symbol $\stackrel{d}{=}$ denotes equality in distribution.

[^16]:    ${ }^{1}$ One notable exception is shortest path routing.
    ${ }^{2}$ Suppose that $\mathcal{M}=\{0,1\}^{d}$, i.e., $\mathcal{M}$ is the entire $d$-dimensional hypercube. Dani et al. 16 show that the regret for this choice of $\mathcal{M}$ is lower bounded by $\kappa d^{2} \sqrt{T}$ for some constant $\kappa$ which is independent of $d$. In this case, ComBand has a regret of $\mathcal{O}\left(d^{2} \sqrt{T}\right)$ since $m=d$ and $\underline{\lambda}=1 / 4$ [79].

[^17]:    ${ }^{3}$ For a simpler algorithm, we refer to 50

[^18]:    ${ }^{4}$ This polytope has also been given other names such as the assignment polytope, the perfect matching polytope of $\mathcal{K}_{m, m}$, and the polytope of doubly-stochastic matrices 100 page 20].

