Large-scale online learning under partial feedback

Julia Olkhovskaya

DOCTORAL THESIS UPF / 2021

THESIS SUPERVISORS
Dr. Gergely Neu
Prof. Gábor Lugosi

Dept. of Information and Communication Technologies

Universitat
Pompeu Fabra
Barcelona

Creative Commons Attribution-ShareAlike 4.0 International License

You are free to copy and redistribute the material in any medium or format, remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Under the following terms: a) Attribution – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. b) ShareAlike – If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. The complete terms of the license can be found at: `http://creativecommons.org/licenses/by-sa/4.0/legalcode`

Abstract

Sequential decision making under uncertainty covers a broad class of problems. Real-world applications require the algorithms to be computationally efficient and scalable. We study a range of sequential learning problems, where the learner observe only partial information about the rewards we develop the algorithms that are robust and computationally efficient in large-scale settings.

First problem that we consider is an online influence maximization problem in which a decision maker sequentiaonally selects a node in the graph in order to spread the information throughout the graph by placing the information in the chosen node. The available feedback is only some informartion about a small neighbourhood of the selected vertex. Our results show that such partial local observations can be sufficient for maximizing global influence. We propose sequential learning algorithms that aim at maximizing influence, and provide their theoretical analysis in both the subcritical and supercritical regimes of broadly studied graph models. Thus this is the first algorithms in the sequential influence maximization setting, that perform efficiently in the graph with a huge number of nodes.

In another line of work, we study the contextual bandit problem, where the reward function is allowed to change in an adversarial manner and the learner only gets to observe the rewards associated with its actions. We assume that the number of arms is finite and the context space can be infinite. We develop a computationally efficient algorithm under the assumption that the d-dimensional contexts are generated i.i.d. at random from a known distribution. We also propose an algorithm that is shown to be robust to misspecification in the setting where the true reward function is linear up to an additive nonlinear error. To our knowledge, our performance guarantees constitute the very first results on this problem setting. We also provide an extension when the context is an element of a reproducing kernel Hilbert space.

Finally, we consider an extension of the contextual bandit problem described above. We study a setting where the learner interacts with a Markov decision process in a sequence of episodes, where an adversary chooses the reward function and the reward observations are available only for the selected action. We allow the state space to be arbitrarily large, but we assume that all action-value functions can be represented as linear functions in terms of a known low-dimensional feature map, and that the learner at least has access to the simulator of the trajectories in the MDP. Our main contributions are the first algorithms that are shown to be robust and efficient in this problem setting.

Contents

Chapter 1

Introduction

This work is related to the field of sequential learning and, more specifically, to frameworks such as multi-armed bandits and reinforcement learning. Sequential learning covers the family of problems where information is revealed incrementally, and the learner must make decisions before all information is available. This approach is required when the system has to be adaptive to the environment's changes and provides scalable algorithms, which is necessary for problems with a dynamic nature. Therefore, the field of sequential learning completes this deficiency. This field already exists for many decades, but it has attracted a lot of attention during the last years. In this work we study a range of problems, that belong to sequential learning framework. Our results have theoretical nature, with a focus on providing formal guarantees to the proposed methods.

The current chapter consists of two parts. In the first part, we describe the context and all necessary notions, and in the second part, we summarise the contributions of the different chapters.

1.1 Online learning

First, we describe the common online learning framework. It can be stated as a sequential game, in which for time steps $t = 1, \dots, T$:

1. The learner observes the context X_t,

2. the learner selects an action A_t,

3. the environment simultaneously selects an action B_t,
4. the learner suffers loss $\ell(A_t, B_t, X_t)$,
5. the learner observes the feedback $f(A_t, B_t, X_t)$.

This sequential game has been studied under various assumptions on the nature of the environment, choice of loss function, the type of feedback that the learner observes, presence of the context and the set of actions. Among this variations, the most fundamental factor is the assumption on the behavior of the environment. There are two diametrically opposite settings, in one setting the actions of the environment are generated as the outcomes of the stochastic process and in the other the environment may be adaptive to the strategy of the learner by looking on the history of the learner's actions. These different settings are considered for various problems. In such examples of practical problems as forecasting the weather or playing on the stock exchange, we do not consider the opponent dependent on our actions. In such settings, the learner's strategy usually involves estimating the parameters of the underlying stochastic process. In the opposite setting, the environment may be considered as an opponent with the goal of maximizing the learner's loss. Such an opponent is referred to as the adversary, the study of the sequential problems with adversarial environment was first done in the works of Blackwell [27] and Hannan [67]. The adversary the learner to maneuver between minimizing losses based on the observations and avoiding being tricked by an opponent. While the assumption on the presence of the adversary may be seen as too strong, one can use it in settings where the environment can be so complex so the learner could give up on estimating the stochastic model. Then, considering that the situation could turn out in the worst possible way for the learner, the learner may assume facing the adversarial opponent.

In the sequential game the learner's task is to minimize the cumulative loss and perform not much worse than some strategy that achieves the minimal loss in the hindsight. Thus the performance of the learner may be quantified by a measure called *regret*, which is the difference between the loss accumulated by the learner and the loss that the learner would get by following the fixed strategy that achieves the smallest cumulative loss in hindsight. This strategy is referred to as a comparator. In the setting where the loss does not depend on the context, the comparator is simply one of the actions, fixed for all time steps. In the more complex setting, different actions may have smaller losses for different contexts, and then the ideal comparator would choose the action with the smallest loss for the current context. Nevertheless, it is unfeasible to compete with this comparator if the set of contexts is large. In such problems, the comparator is usually restricted to be chosen from some set of policies, which is a map from the context to the set of distributions over actions.

For now we turn to the discussion of the feedback the learner observes. In the simplest case, the learner observes the losses of all actions. This setting with this type of feedback and with the adversarial environment is called the prediction with expert advice, and it was first introduced in the papers of Vovk [144], Littlestone and Warmuth [98]. Prediction with expert advice is simple, but a fundamental model and many sequential learning techniques were first developed for this setting. Each action here corresponds to some expert, and by choosing the actions, we decide to follow the appropriate expert. The practical example of the problem of prediction with expert advice is the aggregation of models for predicting the sequence since each model can be evaluated after the sequence element is revealed. Still, the assumption on the observations of losses of all actions may be too strong for some problems. The other widely studied feedback model assumes that the learner observes only the loss of the chosen action. This framework is called a multi-armed bandits, which comes from the name of a slot machine, a one-armed bandit. The name refers to the gambler places his bet each time on a possibly different slot machine he can observe the reward only of the chosen slot machine. This problem is more challenging than the prediction with expert advice since the learner at each time step has to choose between choosing actions with small cumulative loss and exploring the unknown actions. Thus the framework of multi-armed bandit is the simplest mathematical formulation of the explore-exploit dilemma. In almost the whole thesis, we consider problems with bandit form of feedback.

In many sequential learning problems, the learner has access to the context, which is a piece of additional information that may help to predict the losses of the actions. For example, imagine a system that has to choose which advertisements to show to a website visitor. This system has to consider the information available about the user to select the ad that the user will likely click on. On the one side, the presence of context makes the learning problem easier since the learner gets more information. On the other side, as we already discussed, the comparator for the settings with contexts gets "stronger" since it also may adapt to the context. In the same manner, as we considered the nature of the environment, the context may be assumed to be generated from some distribution in an i.i.d manner or be chosen by an adversary. The widely used assumption on the context is that it is generated according to the dynamics of the Markov decision process (MDP). This assumption is widely used in reinforcement learning, where the context is referred to as a state. This setting has received a particular interest since, in many real-world problems, the transitions between the states in the Markov decision process depend on the learner's past action. At the same time, the learner may estimate the parameters of the MDP, which makes the setting less challenging than when an adversary chooses the context.

All sequential learning problems can be classified either as loss games or as gain games. In the vast majority of settings, it is only a lexical difference. The goal of minimizing the loss can be replaced by maximizing the reward, although sometimes this replacement is non-trivial. Often the loss/reward function is assumed to be nice in some way (convex or smooth on some argument, for example). In many cases, significantly tighter performance bounds may be shown under specific assumptions on the loss/reward function.

We have described above the general sequential learning setting and some possible variations of its components. Still, we have not covered of the possible assumptions for feedback, environment, and other factors. Now we state more formally the specific frameworks that we consider further in this work.

Multi-armed bandits

The multi-armed bandit problem was first considered by Thomson [138] in 1933, as a solution of the sequential treatment allocation problem. Recently multi-armed bandits became an active area of research, since the vast of real-valued problems lie under this framework. For the overview of the bandit algorithms, see the resent book of Lattimore and Szepesvári [92]. This problem has been widely studied both in a stochastic and in a adversarial setting and now we separately describe the mathematical formulation of each setting .

The *stochastic* bandits setting was first considered in the papers of Thomson [138] and [124].In the stochastic bandits each action a from a finite set of actions $[K]$ corresponds to an unknown probability distribution P_a. The learner sequentially chooses one of K actions over T time steps for some fixed finite T. At each time step t, the learner chooses an action $A_t \in [K]$ and then the reward $Z_t(A_t) \sim P_{A_t}$ is revealed to the learner. For an action a, we denote the mean of the distribution P_a as μ_a and the expected reward of optimal action is $\mu_* = \arg\max_a \mu_a$. The goal of the learner is to maximize the cumulative rewards or minimize the pseudo regret, defined as

$$R_T = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} Z_t(A_t)\right],$$

where the expectation is taken with respect to the strategy of the learner and the randomness of the reward distributions for different actions.

While in the stochastic bandit framework, the rewards are drawn i.i.d for each action, so the reward at time t depends only on the chosen action. In the *adversarial* bandit framework, there is almost no assumption on how the rewards are generated, it can change arbitrarily and can depend on the history of actions taken by the learner. This setting was first considered in the works of Banos [21], Megiddo and Avivl [105] and Auer et al. [13]. In contrast to stochastic bandit formulation, where the learner is maximizing her reward, historically, losses have been used more often than rewards in the adversarial setting. So the goal of the learner in the adversarial setting is to minimize the loss, which are assumed to be bounded. So, after choosing action A_t at time t, the learner suffer the loss $Y_t(A_t)$. The pseudo regret is defined as following:

$$R_T = \min_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} Y_t(A_t) - \sum_{t=1}^{T} Y_t(a)\right],$$

Note that the expectation is taken over the randomization of the learner.

There are many domains in which multi-armed bandit problems arise, now we state a few. In the web design, the learner's action would be a page layout, and the cumulative reward is the number of clicks. For the data center design problem, the arm is a server at which the task was sent, and the loss would be the task completion time. The action also can have a complicated structure, as in the problem of the route planning, the set of actions would be the set of all possible routes from sink to the target (see, e.g., Valko [140]).

Contextual bandits

The first paper that studied sequential learning with bandit feedback and available context was Woodroofe [148]. As Thomson did in his work on the classic multi-armed bandit problem, Woodroofe considered the clinical trial of medicines as a motivating example. In this problem, the learner wants to learn how to map features of the patient (e.g. age, gender, symptoms) into one of the available medicines. Although there is no evidence of practical use of bandits for the sequential trial problem, Nowadays the contextual bandits are widely applied in the problem of online advertisement placement and other personalization problems ([96], [3]).

The precise definition of context, action, and loss function depends on the setting. The context X_t are drawn from some context space $\mathcal{S}$. For the finite context space, one simple approach is to use a bandit algorithm for each context. However, if the set of possible contexts is large, this approach won't work very well. Further, we will make use of the concept of *policies* $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$. A policy π prescribes a behavior rule to the learner by assigning probability $\pi(a|x)$ to action a at state x. As we already described for the general sequential learning framework, the learner's goal is to minimize the total loss, which can also be stated as the regret minimization problem. Let Π be the set of policies to which the learner is comparing to. We define the regret for the contextual bandit problem as:

$$R_T = \max_{\pi \in \Pi} \mathbb{E} \left[\sum_{t=1}^{T} \ell_t(\pi(X_t), X_t) - \sum_{t=1}^{T} \ell_t(A_t, X_t) \right],$$

where the expectation is taken over the distribution of the context and the randomness of the learner's policy. The case when the set Π is finite was studied in Auer et al. [14], but their algorithm becomes computationally infeasible if the set of policies is large. This difficulty has been overcomed by assuming that there is some structure on how reward may depend on the context. For example, [1] considers the setting where losses a linear function of the context with a presence of some noise. This is a strong assumption, but it makes the problem more computationally tractable and works well in practice. [61], [60], [135] consider a more complicated dependences of losses on the context.

Episodic Markov decision processes

An episodic Markov Decision Process (MDP), denoted by $M = (\mathcal{S}, \mathcal{A}, H, P, r)$ is defined by a state space $\mathcal{S}$, action space $\mathcal{A}$, episode length $H \in \mathbb{Z}_+$, transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ and a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We will assume that the MDP has a layered structure, satisfying the following conditions:

- The state set $\mathcal{S}$ can be decomposed into H disjoint sets: $\mathcal{S} = \cup_{h=1}^{H} \mathcal{S}_h$,

- $\mathcal{S}_1 = \{x_1\}$ and $\mathcal{S}_H = \{x_H\}$ are singletons,

- transitions are only possible between consecutive layers, that is, for any $x_h \in \mathcal{S}_h$, the distribution $P(\cdot|x, a)$ is supported on $\mathcal{S}_{h+1}$ for all a and $h \in [H-1]$.

These assumptions are common in the related literature (e.g., 111, 152, 126) and are not essential for the analysis; their primary role is simplifying the notation. The learner observes the state $X_{t,h}$ at each step $h \in [H-1]$, picks an action $A_{t,h}$ and observes the reward $r_{t,h}(X_{t,h}, A_{t,h})$. Then, unless $h = H$, the learner moves to the next state $X_{t,h+1}$, which is generated from the distribution $P(\cdot|X_{t,h}, A_{t,h})$. At the end of step H, the episode terminates, and a new one begins. The aim of the learner is to select its actions so that the cumulative sum of rewards is as large as possible.

The learner starts interacting with the MDP in each episode from the initial state $X_{t,1} = x_1$. At each consecutive step $h \in [H-1]$ within the episode, the learner observes the state $X_{t,h}$, picks an action $A_{t,h}$ and observes the reward $r_{t,h}(X_{t,h}, A_{t,h})$. Then, unless $h = H$, the learner moves to the next state $X_{t,h+1}$, which is generated from the distribution $P(\cdot|X_{t,h}, A_{t,h})$. At the end of step H, the episode terminates, and a new one begins. The aim of the learner is to select its actions so that the cumulative sum of rewards is as large as possible.

Let $\tau^\pi = ((X_1, A_1), (X_2, A_2), \ldots, (X_H, A_H))$ be a trajectory generated by following the policy π through the MDP. The expected total reward of a policy π is defined as

$$\rho_t(\pi) = \mathbb{E}\left[\sum_{(X,A)\in\tau^\pi} r_t(X, A)\right].$$

Using this notation, we define the learner's goal as minimizing the total expected regret defined as

$$R_T = \max_\pi \sum_{t=1}^{T} \mathbb{E}\left[\rho_t(\pi) - \rho_t(\pi_t)\right], \tag{1.1}$$

where the maximum is taken over the complete set of stochastic stationary policies.

1.2 Main contributions

1.2.1 Online influence maximization.

The problem of finding the most influential nodes in networks has been broadly studied under various definitions of influence, the information spreading model, and the feedback available to a decision maker [79, 43, 143, 145, 80, 120]. The most studied influence maximization setup is an offline discrete optimization problem of finding the most influential nodes in a network. This setup assumes that the probability of influencing is known, or at least data is available that allows one to estimate these probabilities. However, such information is often not available or is difficult to obtain. Also, the network over which information spreads is rarely fixed. To avoid such assumptions, we introduce a novel model of influence maximization in a sequential setup, where the underlying network changes every time and the learner has only partial information about the set of influenced nodes.

Specifically, we study a family online influence maximization problems where at each time step the decision maker selects a node from a large number of vertices with the goal of maximizing influence. After choosing a node, the decision maker places a piece of information there. The information then spreads to other nodes in the graph, in a way that the information transmits between two nodes with some fixed but unknown probability. The goal of the decision maker is to reach as many nodes as possible.

The decision maker has to learn about the probabilities of spreading information in the graph on the fly, while simultaneously attempting to maximize the total reward. This gives rise to a dilemma of exploration versus exploitation, commonly studied within the framework of multi-armed bandit problems [92]. Indeed, if the decision maker could observe the size of the set of all influenced nodes in every round, the sequential influence maximization problem outlined above could be naturally formulated as a stochastic multi-armed bandit problem. The main drawback of this approach is that in the most of applications tracking down the set of all influenced agents may be difficult or downright impossible due to privacy and computational considerations. This motivates the study of a more restrictive setting where the decision maker has to manage with only partial observations of the set of influenced nodes. Our results show that partial local observations can be sufficient for maximizing global influence under a set of assumption on the information spreading. We propose sequential learning algorithms that aim at maximizing influence under the partial observations on the set of influenced nodes, and provide their theoretical analysis in both the subcritical and supercritical regimes of all considered models.

1.2.2 Adversarial contextual bandits

The contextual bandit problem is one of the most important sequential decision-making problems studied in the machine learning literature. This framework can be used to address a broad range of challenging real-world problems, such as recommendation systems, healthcare and finance. These applications require the algorithm to be robust, which can be guaranteed by providing formal performance guarantees.

The limitation of virtually all known algorithms for linear contextual bandits is that they crucially rely on assuming that the loss function is fixed during the learning procedure. This is in stark contrast with the literature on multi-armed bandits, where there is a rich literature on both stochastic bandit models assuming i.i.d. rewards and adversarial bandit models making no assumptions on the sequence of loss functions. Our main contribution in this work is addressing this gap by designing and analysing algorithms that are guaranteed to work for arbitrary sequences of loss functions (Cesa-Bianchi and Lugosi [39]). While it is tempting to think that the our bandit problem can be directly addressed by a minor adaptation of algorithms developed for adversarial linear bandits, this is unfortunately not the case: all algorithms developed for such problems require a fixed decision set, whereas reducing the linear contextual bandit problem to a linear bandit problem requires the use decision sets that change as a function of the contexts. As a crucial step in our analysis, we will assume that the contexts are generated in an i.i.d. fashion and that the loss function in each round is statistically independent from the context in the same round. This assumption will allow us to relate the contextual bandit problem to a set of auxiliary bandit problems with a fixed action sets, and reduce the scope of the analysis to these auxiliary problems.

1.2.3 Learning in episodic MDP

While for the contextual bandit problem we were assuming that the context is generated i.i.d. from a fixed distribution, the reinforcement learning framework typically allows the actions taking in the past to affect the states that the decision maker will reach in the future. More concretely, the state of the environment changes according to the transition function of the underlying MDP, as a function of the previous state and the action taken by the learner. This is more general and much more challenging setting than the contextual bandit framework.

Online learning in MDP has been first considered in the works of Burnetas and Katehakis [32], Auer and Ortner [16]. We focus on the setting where the reward function can chosen by the adversary in the beginning of each episode. Most of practical applications requires the state space to be very large, so we are interested in computationally efficient algorithm with meaningful guarantees for arbitrarily large state space. Our main contribution is designing algorithms that obtain strong theoretical guarantees in the setting where the reward function is chosen by adversary, the state space is infinite and only feedback on the chosen action is revealed to the learner.

1.3 Notation

We use $\langle \cdot, \cdot \rangle$ to denote inner products in Euclidean space and by $\|\cdot\|$ we denote the Euclidean norm for vectors and the operator norm for matrices. For a symmetric positive semidefinite matrix A, we use $\lambda_{\min}(A)$ to denote its smallest eigenvalue. We use $\|A\|_{\mathrm{op}}$ to denote the operator norm of A and we write $\operatorname{tr}(A)$ for the trace of a matrix A. Finally, we use $A \succcurlyeq 0$ to denote that an operator A is positive semi-definite, and we use $A \succcurlyeq B$ to denote $A - B \succcurlyeq 0$. For a d-dimensional vector v, we denote the corresponding $d \times d$ diagonal matrix by $\mathrm{diag}(v)$. For a positive integer N, we use $[N]$ to denote the set of positive integers $\{1, 2, \ldots, N\}$. Finally, we will denote the set of all probability distributions over any set $\mathcal{X}$ by $\Delta_{\mathcal{X}}$.

Chapter 2

Sequential influence maximization

2.1 Introduction

Finding influential nodes in networks has a long history of study. The problem has been cast in a variety of different ways according to the notion of influence and the information available to a decision maker. We refer the reader to Kempe et al. [79], Chen et al. [43, 42], Vaswani et al. [143], Carpentier and Valko [38], Wen et al. [147], Wang and Chen [145], Khim et al. [80], Perrault et al. [120] and the references therein for recent progress in various directions.

The most studied influence maximization setup is an offline discrete optimization problem of finding the set of the most influential nodes in a network. This setup assumes that the probability of influencing is known, or at least data is available that allows one to estimate these probabilities. However, such information is often not available or is difficult to obtain. Also, the network over which information spreads is rarely fixed. To avoid such assumptions, we introduce a novel model of influence maximization in a sequential setup, where the underlying network changes every time and the learner has only partial information about the set of influenced nodes.

Specifically, we define and explore a sequential decision-making model in which the goal of a decision maker is to find one among a set of n agents with maximal (expected) influence. We parametrize the information spreading mechanism by a symmetric $n \times n$ matrix P, whose entries $p_{i,j} \in [0, 1]$ express “affinity” or “probability of communication” between agents i and j. We assume that $p_{i,i} = 0$ for all $i \in [n]$. The matrix P defines an inhomogeneous random graph G in a natural way: an (undirected) edge is present between nodes $i < j$ with probability $p_{i,j}$ and all edges are independent. When two nodes are connected by an edge, information flows between the corresponding agents. Hence, a piece of information placed at a node i spreads to the nodes of the entire connected component of i in G.

In the sequential decision-making process we study, an independent random graph is formed at each time instance $t = 1, \dots, T$ on the vertex set $[n]$. The random graph formed at time t is denoted by G_t. Hence, $G_1, \dots, G_T$ is an independent, identically distributed sequence of random graphs on the vertex set $[n]$, whose distribution is determined by the matrix P. If the decision maker selects a node $a \in [n]$ at time t, then the information placed at the node spreads to every node of the connected component of a in the graph G_t. The goal of the decision maker is to spread information as much as possible, that is, to reach as many agents as possible. The *reward* of the decision maker at time t is the number of nodes in the connected component containing the selected node in G_t.

In this paper, we study a setting where the decision maker has no prior knowledge of the distribution P, so she has to learn about this distribution on the fly, while simultaneously attempting to maximize the total reward. This gives rise to a dilemma of *exploration versus exploitation*, commonly studied within the framework of *multi-armed bandit* problems (for a survey, see 30 or 88). Indeed, if the decision maker could observe the size of the set of all influenced nodes in every round, the sequential influence maximization problem outlined above could be naturally formulated as a *stochastic multi-armed bandit* problem [86, 12]. However, this direct approach has multiple drawbacks. First of all, in many applications, the number n of nodes is so large that one cannot even hope to maintain individual statistics about each of them, let alone expect any algorithm to identify the most influential node in reasonable time. More importantly, in most cases of interest, tracking down the set of *all* influenced agents may be difficult or downright impossible due to privacy and computational considerations. This motivates the study of a more restrictive setting where the decision maker has to manage with only partial observations of the set of influenced nodes.

We address this latter challenge by considering a more realistic observation model, where after selecting an agent A_t to be influenced, the learner only observes a local neighbourhood of A_t in the realized random graph G_t , or even only the number of immediate neighbours of A_t (i.e., the degree of vertex A_t in G_t). This model raises the following question: is it possible to maximize global influence while only having access to such local measurements? Our key technical result is answering this question in the positive for some broadly studied random graph models.

The rest of the paper is structured as follows. In Section 3.2 we formalize the sequential influence maximization problem. In Section 2.1.2 a general model of inhomogeneous random graphs is described and the crucial notions of sub-, and super-criticality are formally introduced. Section 2.2 is dedicated to the general case when the underlying random graph is an arbitrary inhomogenous random graph and the learner only knows whether it is in the subcritical or supercritical regime. We show that in both cases online influence maximization is possible by only observing a small "local" neighborhood of the selected node. We provide two separate algorithms and regret bounds for the subcritical and supercritical cases, respectively. In Section 2.6, we consider the situation when the learner has even less information about the underlying random graph. In particular, we assume that the learner only observes the degree of the selected node in the realized random graph. We study three well-known special cases of inhomogeneous random graphs that are commonly used to model large social networks, namely stochastic block models, the Chung–Lu model, and Kronecker random graphs. We prove that in these three random graph models, degree observations are sufficient to maximize global influence both in the subcritical and supercritical regimes.

2.1.1 Problem setup

We now describe our problem and model assumptions formally. We consider the problem of sequential influence maximization on the set of nodes $V = [n]$, formalized as a repeated interaction scheme between a learner and its environment. We assume that node i influences node j with (unknown) probability $p_{i,j} (= p_{j,i})$. At each iteration, a new graph G_t is generated on the vertex set V by independent draws of the edges such that edge (i, j) is present with probability $p_{i,j}$ and all edges are independent. The set of nodes influenced by the chosen node A_t is the connected component of G_t that contains A_t. $C_{i,t}$ denotes the connected component containing vertex i:

$$C_{i,t} = \{v \in V : v \text{ is connected to } i \text{ by a path in } G_t\} \ .$$

The feedback that the decision maker receives after choosing a node is some "local" information around the chosen vertex A_t in G_t. We consider several feedback models. In the simplest case, the feedback is the degree of vertex A_t in G_t. In another model, the information might consist of the vertices found after a few steps of depth-first exploration of G_t started from vertex A_t. In a general framework, we may define a "local neighborhood" of A_t, denoted by $\widehat{C}_{A_t,t}$, where $\widehat{C}_{A_t,t} \subset C_{A_t,t}$. For each model considered below, we specify later what exactly $\widehat{C}_{A_t,t}$ is. In the general setup, the following steps are repeated for each round $t = 1, 2, \ldots$:

1. the learner picks a vertex $A_t \in V$,
2. the environment generates a random graph G_t,
3. the learner observes the local neighborhood $\widehat{C}_{A_t,t}$,
4. the learner earns the reward $r_{t,A_t} = |C_{A_t,t}|$.

We stress that the learner does *not* observe the reward, only the local neighborhood $\widehat{C}_{A_t,t}$. Define c_i as the expected size of the connected component associated with the node i: $c_i = \mathbb{E}\left[|C_{i,1}|\right]$. Ideally, one would like to minimize the *expected regret* defined as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{i\in V} c_i - c_{A_t}\right)\right]. \tag{2.1}$$

Since we are interested in settings where the total number of nodes n is very large, even with a fully known random graph model, finding the optimal node maximizing c_i is infeasible both computationally and statistically. Such intractability issues have lead to alternative definitions of the regret such as the *approximation regret* [76, 44, 133] or the *quantile regret* [41, 45, 100, 83].

In the present paper, we consider the α-quantile regret as our performance measure, which, instead of measuring the learner's performance against the single best decision, uses a near-optimal action as a baseline. For a more technical definition, let $i_1, i_2, \ldots, i_n$ be an ordering of the nodes satisfying $c_{i_1} \leq c_{i_2} \leq \cdots \leq c_{i_n}$, and denote the α-quantile over the mean rewards as $c^*_\alpha = c_{i_{\lceil(1-\alpha)n\rceil}}$. Then, defining the set $V^*_\alpha = \{i_{\lceil(1-\alpha)n\rceil}, \ldots, i_n\}$ as the set of α-near-optimal nodes, we define the α-quantile regret as

$$R^\alpha_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\min_{i\in V^*_\alpha} c_i - c_{A_t}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(c^*_\alpha - c_{A_t}\right)\right]. \tag{2.2}$$

2.1.2 Inhomogeneous Erdős–Rényi random graphs

Next we discuss the random graph models considered in this paper. All belong to the *inhomogeneous Erdős–Rényi model*, that is, edges are present independently of each other, with possibly different probabilities. Moreover, the graphs we consider are *sparse* graphs, that is, the average degree is bounded. We will formulate our random graph model following the work of [28], whose framework is particularly useful for handling large values of n. To this end, let κ be a bounded symmetric non-negative measurable function on $[0, 1] \times [0, 1]$. Each edge (i, j) for $1 \leq i < j \leq n$ is present with probability $p_{i,j} = \min(\kappa(i/n, j/n)/n, 1)$, independently of all other edges. When n is fixed, we will often use the notation $A_{i,j} = \kappa(i/n, j/n)$ so that $p_{i,j} = \min(A_{i,j}/n, 1)$. We are interested in random graphs where the average degree is $O(1)$ (as $n \to \infty$). This assumption makes the problem both more realistic and challenging: denser graphs are connected with high probability, making the problem essentially vacuous. A random graph drawn from the above distribution is denoted by $G(n, \kappa)$. This model is sometimes called the *binomial random graph* and was first considered by [84].

We consider two fundamentally different regimes of the parameters $G(n, \kappa)$: the *subcritical* case in which the size of the largest connected component is sublinear in n (with high probability), and the *supercritical* case where the largest connected component is at least of size cn for some constant $c > 0$, with high probability. (We say that an event holds *with high probability* if its probability converges to one as $n \to \infty$.) Such a connected component of linear size is called a *giant component*. These regimes can be formally characterized with the help of the integral operator T_κ, defined by

$$(T_\kappa f)(x) = \int_{(0,1]} \kappa(x, y) f(y) d\mu(y) ,$$

for any measurable bounded function f, where μ is the Lebesgue measure. We call κ subcritical if $\|T_\kappa\|_2 < 1$ and supercritical if $\|T_\kappa\|_2 > 1$. We use the same expressions for a random graph $G(n, \kappa)$. It follows from Bollobás et al. [28, Theorem 3.1] that, with high probability, $G(n, \kappa)$ has a giant component if it is supercritical, while the number of vertices in the largest component is $o(n)$ with high probability if it is subcritical.

2.2 Observations of censored component size

First we study a natural feedback model in which the decision maker, unable to explore the entire connected component $C_{i,t}$ of the influenced node i in G_t, resorts to exploring the connected component up to a certain (small) number of nodes. More precisely, we define feedback as the result of counting the number of nodes in $C_{i,t}$ by (say, depth-first search) exploration of the connected component, which stops after revealing K nodes, or before, if $|C_{i,t}| < K$. Here K is a fixed positive integer, independent of the number of nodes n.

The main results of this section show that this type of feedback is sufficient for sequential influence maximization. However, the subcritical and supercritical cases need to be treated separately as they are quite different. In the subcritical case, the expected size of the connected component of any vertex is of constant order while in the supercritical case there exist vertices whose connected component is linear in n. This also means that the rewards – and therefore the per-round regrets – are of different order of magnitude (as a function of n) in the subcritical and supercritical cases. For simplicity, we assume that the decision maker knows in advance whether the function κ defining the inhomogeneous random graph is subcritical or supercritical, as we propose different algorithms for both cases. We believe that this is a mild assumption, since in typical applications it is possible to set the two settings apart based on prior data. We also assume that $\|T_\kappa\|_2 \neq 1$, that is, the random graph is not exactly critical.

2.2.1 Subcritical case

First we study the subcritical case, that is, we assume that $||T_\kappa||_2 < 1$. In this case the proposed influence-maximization algorithm uses the censored size of the connected component of the selected node. That is, for a node $i \in [n]$, we define $u_{i,t}(K)$ as the result of counting the number of nodes in $C_{i,t}$ by exploration of the connected component, which stops after revealing K nodes or before, if $|C_{i,t}| < K$. Hence, the feedback is $u_{i,t}(K) = \min\left(|C_{i,t}|, K\right)$.

A key ingredient in our analysis in the subcritical case is an estimate for the lower tail of the size of the connected component containing a fixed vertex. We state it in the following lemma:

Lemma 1. *For any subcritical κ, there exist positive constants $\lambda(\kappa), g(\kappa)$ and $n_0(\kappa)$, such that for any $n \geq n_0$, for any node i in $G(n, \kappa)$, the size of the* *connected component C_i of a vertex i satisfies*

$$\mathbb{P}\left[|C_i| > u\right] \leq e^{-\lambda(\kappa)u} g(\kappa)\ . \tag{2.3}$$

Unfortunately, there is no closed-form expression for the dependence $\lambda(\kappa)$ and $g(\kappa)$ on κ. The idea of the proof of this lemma relies on the proof of Theorem 12.5 in Bollobás et al. [28]. To obtain this result, we show that the size of the connected component in $G(n, \kappa)$ is stochastically dominated by the total progeny of the multitype Poisson branching process with carefully chosen parameters. We introduce branching processes in Section 2.4 and prove Lemma 1 in Section 2.5.1.

Now we are ready to define an estimate of $c_i = \mathbb{E}|C_i|$ in the sequential decision game. For a fixed a constant K, we define the estimate $\widehat{u}_{i,t}(K) = (1/t)\sum_{s=1}^{t} u_{i,s}(K)\mathbb{I}_{\{A_s=i\}}$. Using the concentration inequality (2.3), with the choice of the threshold parameter $K = \frac{\log(T)}{\lambda}$ with $\lambda > \lambda(\kappa)$, we get that the bias of $\widehat{u}_{i,t}(K)$ is at most $\frac{g(\kappa)}{T}$. We state this result more formally in Lemma 2. The censored observations are bounded, since $u_{i,t}(K) \in [1, K]$. We use those observations as rewards in our bandit problem and we feed them to an instance of the UCB algorithm [15]. We call the resulting algorithm Local UCB(V_0), defined in Algorithm 2.1 below.

A minor challenge is that, since we are interested in very large values of n, it is infeasible to use *all* nodes as separate actions in our bandit algorithm. To address this challenge, we propose to *subsample* a set of representative nodes for UCB to play on. The size of the subsampled nodes depends on the quantile α targeted in the regret definition (2.2) and the time horizon T. Our algorithm uniformly samples a subset V_0 of size

$$|V_0| = \left\lceil \frac{\log T}{\log(1/(1-\alpha))} \right\rceil \tag{2.4}$$

and plays Local UCB(V_0) for the corresponding regime on the resulting set. Note that the size of V_0 is chosen such that the probability that V_0 does not contain any of the αn notes with the largest values of c_i is at most $1/T$.

To simplify the presentation, we introduce some more notation. Analogously to the α-optimal reward c^*_α, we define the α-optimal censored component size $u_{*,\alpha}(K) = \min_{i\in V^*_\alpha} u_i(K)$ and we define the corresponding gap parameters $\Delta_{\alpha,i} = (c^*_\alpha - c_i)_+$, $\delta^{sub}_{\alpha,i}(K) = (u_{*,\alpha}(K) - u_i(K))_+$ and $\Delta_{\alpha,\max} = \max_i \Delta_{\alpha,i}$. $N_{i,t} = \sum_{s=1}^{t} \mathbb{I}_{\{A_s=i\}}$ denotes the number of times node i is selected up to time t.

Algorithm 2.1 Local UCB(V_0) for subcritical $G(n, \kappa)$.

```
**Parameters:** A set of nodes $V_0 \subseteq V$, $K > 0$.
**Initialization:** Select each node in $V_0$ once. For each $i \in V_0$, set $N_{i,|V_0|} = 1$ and $\widehat{u}_{i,|V_0|} = u_{i,i}(K)$.
**For** $t = |V_0|, \ldots T$, **repeat**
    1. Select any node $A_{t+1} \in \arg\max_i \widehat{u}_{i,t}(K) + K\sqrt{\frac{\log t}{N_{i,t}}}$.
    2. Observe $u_{A_{t+1},t+1}(K)$ , update $\widehat{u}_{i,t+1}$ and $N_{i,t+1}$ for all $i \in [n]$.
```

For the subcritical case, Local UCB(V_0) has the following performance guarantee:

Theorem 2.2.1 (Subcritical inhomogeneous random graph)**.** *Assume that κ is subcritical. Let V_0 be a uniform subsample of V with size given in (2.4) and define the event $\mathcal{E} = \{V_0 \cap V_\alpha^* \neq \emptyset\}$. Then for any $G(n, \kappa)$ with $n > n_0(\kappa)$ and any K, the expected α-quantile regret of Local UCB(V_0) satisfies*

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i}\left(\frac{4K^2 \log T}{(\delta_{\alpha,i}^{sub}(K))^2} + 8\right) \middle| \mathcal{E}\right],$$

where the expectation is taken over the random choice of V_0. Furthermore, if κ is such that $\lambda(\kappa) > \lambda$, $g(\kappa) < g$, then, taking $K = \frac{\log T}{\lambda}$, we have

$$R_T^\alpha \leq \frac{4 \log T}{\lambda}\sqrt{\frac{T}{\log(1/(1-\alpha))}} + 8\Delta_{\alpha,\max}\left\lceil\frac{\log T}{\log(1/(1-\alpha))}\right\rceil + 2g.$$

We prove Theorem 2.2.1 in Section 2.5.1. Observe that one may choose the value of K as a constant, regardless of the number n of the nodes. This means that the feedback information is truly “local” in the sense that only a constant number of vertices of the connected component of the selected node need to be explored. How large K needs to be depends on the parameter λ. An undesirable feature of Local UCB(V_0) is that the learner needs to know the parameter λ that depends on the unknown function κ. To resolve this problem we propose a version of a "doubling trick" (see, e.g., Section 2.3 [40]). While in our problem it is not possible to control the range of $\lambda(\kappa)$ explicitly, we still can control the frequency with which $|C_i|$ is censored by choosing the range of K. In order to do this, we propose a variation of Local UCB(V_0), such that we split time T into episodes $q = 1, 2 \ldots$ in the following way. At the beginning of each episode q, the learner starts a new instance of Local UCB(V_0) with a threshold parameter $K_q = 2^q \log T$ and starts a new time counter t_q. Then, at each time step of the current episode, the learner computes the empirical probability $\widehat{p}_q = \frac{1}{t_q} \sum_{\tau=t_{q-1}+1}^{t_q} \mathbb{I}_{\{|C_{A_\tau,\tau}|>K_q\}}$, that is updated each time when the size of connected component of the chosen node exceeds K_q. Once $\widehat{p}_q$ gets larger than $\frac{1}{T} + \sqrt{\frac{\ln T}{2(t_q+1)}}$, the episode q finishes and the next episode begins. In this way, the length of each episode and the total number of episodes $Q_{\max}$ are random. We call this algorithm UCB(V_0)-DOUBLE , and show that it has the following performance guarantee:

Algorithm 2.2 UCB(V_0)-DOUBLE for subcritical $G(n, \kappa)$.

```
Parameters: A set of nodes $V_0 \subseteq V, T > 0$.
Initialization: $K_0 = \log T$, $t = 1$, $q = 0$, $t_q = 0$, $\widehat{p}_q = 0$.
While $t \leq T$, repeat:
    • Select each node in $V_0$ once. For each $i \in V_0$, set $N_{i,t} = 1$, $\widehat{u}_{i,t} = u_{i,t}(K_q)$
      and $\widehat{p}_q = \frac{1}{|V_0|} \sum_{\tau=t_{q-1}+1}^{t_{q-1}+1+|V_0|} \mathbb{I}_{\{|C_{A_\tau,\tau}|>K_q\}}$.
      While $\widehat{p}_q \leq \frac{1}{T} + \sqrt{\frac{\ln T}{2(t_q+1)}}$, repeat:
        1. Select any node $A_{t_q+1} \in \arg\max_i \widehat{u}_{i,t_q}(K_q) + K_q\sqrt{\frac{\log T}{N_{i,t_q}}}$.
        2. Observe $u_{A_{t_q+1},t+1}$,
        3. Update $\widehat{u}_{A_{t_q+1},t_q+1} = \frac{1}{t_q} \sum_{\tau=t_{q-1}+1}^{t_q} u_{i,\tau}(K_q)\mathbb{I}_{\{A_\tau = A_{t_q+1}\}}$,
           $N_{A_{t_q+1},t_q+1} = N_{A_{t_q},t_q} + 1$, $\widehat{p}_q = \frac{1}{t_q} \sum_{\tau=t_{q-1}+1}^{t_q} \mathbb{I}_{\{|C_{A_\tau,\tau}|>K_q\}}$,
        4. Update $t_q = t_q + 1$ and $t = t + 1$.
    • Set $t_{q+1} = 0$, $\widehat{p}_{q+1} = 0$, $K_{q+1} = 2K_q$ and $q = q + 1$.
```

Theorem 2.2.2. *Assume that κ is subcritical and $n > n_0(\kappa)$. Let V_0 be a uniform subsample of V with size given in (2.4) and define the event $\mathcal{E} = \{V_0 \cap V_\alpha^* \neq \emptyset\}$. Then for $G(n, \kappa)$ with $n > n_0(\kappa)$, the expected α-quantile regret of* UCB(V_0)-

DOUBLE *satisfies*

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \frac{64}{3}\mathbb{E}\left[\sum_{i\in V_0} \Delta_{\alpha,i}\left(\frac{\log^3 T}{(\lambda(\kappa)\cdot \min_{q\in[Q_{\max}]}\{\delta_{\alpha,i}^{sub}(2^q \log T)\})^2} + 8\right)\middle| \mathcal{E}\right],$$

where the expectation is taken over the random choice of V_0*, and*

$$R_T^\alpha = \mathcal{O}\left(\frac{\sqrt{T(\log(1/\lambda(\kappa))+1)}}{\lambda(\kappa)\sqrt{\ln(1/(1-\alpha))}}\log^2 T\right).$$

The proof of Theorem 2.2.2 may be found in Section 2.5.1. Note that both Theorems 2.2.1 and 2.2.2 present two types of regret bounds. The first set of these bounds are polylogarithmic[1] in the time horizon T, but show strong dependence on the parameters of the distribution of the graphs G_t. Such bounds are usually called *instance-dependent*, as they are typically interesting in the regime where T grows large and the problem parameters are fixed independently of T. However, these bounds become vacuous for smaller values of T as the gap parameters $\delta_{\alpha,i}^{sup}(\cdot)$ and $\delta_{\alpha,i}^{sub}(\cdot)$ approach zero. This issue is addressed by our second set of guarantees, which offer a bounds of $\widetilde{O}(\sqrt{|U|T})$ for some set $U \subseteq V$ that holds simultaneously for all problem instances without becoming vacuous in any regime. Such bounds are commonly called *worst-case*, and they are often more valuable when optimizing performance over a fixed horizon T.

A notable feature of our bounds is that they show no explicit dependence on the number of nodes n. This is enabled by our notion of α-quantile regret, which allows us to work with a small subset of the total nodes as our action set. Instead of n, our bounds depend on the size of some suitably chosen set of nodes U, which is of the order $\operatorname{polylog} T/\log(1/(1-\alpha))$. Notice that this gives rise to a subtle tradeoff: choosing smaller values of α inflates the regret bounds, but, in exchange, makes the baseline of the regret definition stronger (thus strengthening the regret notion itself).

2.2.2 Supercritical case

Next we address the supercritical case, that is, when $||T_\kappa||_2 > 1$. Here the proposed algorithm uses $v_{i,t}(K)$ defined as the indicator whether $|C_i|$ is larger

[1] Upon first glance, the bound of Theorem 2.2.1 may appear to be logarithmic, however, notice that the sum involved in the bound has $\Theta(\log T)$ elements, thus technically resulting in a bound of order $\log^2 T$.

than K, that is, $v_{i,t}(K) = \mathbb{I}_{\{|C_i(G_t)|>K\}}$. Since the observation is an indicator function, $v_{i,t}(K) \in \{0,1\}$. Similarly to the subcritical case, we propose a variant of UCB algorithm, Local UCB(V_0), played over a random subsample of nodes of size defined in (2.4). We define $v_i(K) = \mathbb{E}\left[v_{i,t}(K)\right]$ and $v_*(K) = \max_i v_i(K)$. Analogously to the notation introduced for the subcritical regime, we denote $v_{*,\alpha}(K) = \min_{i\in V_\alpha^*} v_i(K)$ and $\delta_{\alpha,i}^{sup}(K) = (v_i(K) - v_{*,\alpha}(K))_+$.

In the supercritical case, the learner receives $v_{i,t}(K)$ as a reward and we design a bandit algorithm based on this form of indicator observations. Note, that $v_{i,t}(K)$ is a Bernoulli random variable with parameter $\mathbb{P}\left[|C_i(G_t)| > K\right]$. The following algorithm is a variant of the UCB algorithm of Auer et al. [15]. Just like before, $N_{i,t}$ denotes the number of times node i is selected up to time t by the algorithm.

```
**Algorithm 2.3** Local UCB$(V_0)$ for supercritical $G(n,\kappa)$.
**Parameters:** A set of nodes $V_0 \subseteq V$, $k(n)$.
**Initialization:** Select each node in $V_0$ once. For each $i \in V_0$, set $N_{i,|V_0|} = 1$ and $\widehat{v}_{i,|V_0|}(k(n)) = v_{i,i}(k(n))$.
**For** $t = |V_0|, \dots T$, **repeat**
    1. Select any node $A_{t+1} \in \arg\max_i \widehat{v}_{i,t}(k(n)) + \sqrt{\frac{\log t}{N_{i,t}}}$.
    2. Observe the feedback $v_{i,t}(k(n))$ , update $\widehat{v}_{i,t+1}(k(n))$ and $N_{i,t+1}$ for all $i \in [n]$.
```

Local UCB(V_0) for supercritical $G(n,\kappa)$ satisfies the following regret bound:

Theorem 2.2.3. *Let V_0 be a uniform subsample of V with size given in (2.4) and define the event $\mathcal{E} = \{V_0 \cap V_\alpha^* \neq \emptyset\}$. For any $G(n,\kappa)$ with supercritical κ and $n > n_0(\kappa)$, for any function $k : \mathbb{N} \to \mathbb{N}$ such that $\lim_{n\to\infty} k(n) = \infty$, we get*

$$\frac{R_T^\alpha}{n} \leq \frac{1}{n}\Delta_{\alpha,\max} + \frac{1}{n}\mathbb{E}\left[\sum_{i\in V_0} \Delta_{\alpha,i}\left(\frac{4\log T}{(\delta_{\alpha,i}^{sup}(k(n)))^2} + 8\right)\middle| \mathcal{E}\right],$$

where the expectation is taken over the random choice of V_0, and

$$\frac{R_T^\alpha}{n} \leq 9\left(\frac{\mathbb{E}\left[|C_1|\right]}{n} + 1\right)\left\lceil\frac{\log T}{\log(1/(1-\alpha))}\right\rceil\sqrt{T\log T}.$$

For the proof of Theorem 2.2.3, see Section 2.5.2. Note that for a supercritical κ, $\mathbb{E}\left[C_1\right] = \Theta_n(n)$. Therefore, R_T^α scales linearly with n and hence it is natural to normalize the regret by the number of nodes. Both in the subcritical and supercritical regimes, our bounds scale linearly with the maximal expected reward c^*, which is of $\Theta_n(1)$ in the subcritical case, but is $\Theta_n(n)$ in the supercritical case. The dependence of the obtained bounds on the time horizon T is similar in both regimes. Note that unlike in the subcritical case, the censoring level K is not a constant anymore as we choose it to be $K = k(n)$ for some function k. Hence, strictly speaking, the feedback is not local as the number of vertices that need to be explored is not independent of the number of nodes even if $k(n)$ can grow arbitrarily slowly. Similarly to the subcritical case, a sufficiently large constant value of K would suffice. The value of the constant should be so large that for any vertex i, the conditional probability – conditioned on the event that i is not in the giant component – that the component of i has size larger than K is sufficiently small. Such a constant exists, see (2.10) below. However, this value depends on the unknown distribution of the underlying random graph. In the subcritical case we solved this problem by applying a doubling trick. This is made possible by the fact that in the subcritical case one observes the "bad" event that a component has size larger than K and therefore censoring occurs. By "trying" increasingly large values of K one eventually finds a value such that the probability of censoring is sufficiently small. However, in the supercritical case, the "bad" event is that even though the selected vertex is not in the giant component, the size of the component is larger than K. Unfortunately, one cannot decide whether the bad event occurs or simply the vertex lies in the giant component. For this reason, we have been unable to apply an analogous doubling trick in the supercritical case. To circumvent this difficulty, we choose K to be growing with n. This guarantees that the bad event occurs with small probability. The price to pay is that the observation is not entirely local in the strict sense.

2.3 Degree observations

The results of the previous section show that it is possible to learn to maximize influence under very general conditions if the learner has access to the censored size of the connected component, where the size of censoring may be kept much smaller than the size of the entire network. In this section we consider the case when the learner has access to significantly less information. In particular, we study the case when the learner only observes the degree of the selected vertex A_t (i.e., the number of edges adjacent to A_t) in the graph G_t. Under such a restricted feedback, one cannot hope to learn to maximize influence in the full generality of sparse inhomogeneous random graphs as in Section 2.2. However, we show that in several well-known models of real networks, degree information suffices for influence maximization. In particular, we study three random graph models that have been introduced to replicate properties of large (social) networks appearing in a variety of applications. These are (1) stochastic block models; (2) the Chung–Lu model; and (3) Kronecker random graphs.

2.3.1 Three random graph models

We start by introducing the three models we study. All of them are special cases of inhomogeneous Erdős–Rényi graphs.

Stochastic block model

In the stochastic block model, the probabilities $p_{i,j}$ are defined through the notion of *communities*, defined as elements of a partition $H_1, \dots, H_S$ of the set of vertices V. We refer to the index m of community H_m as the *type* of a vertex belonging to H_m. Each community H_m contains $\alpha_m n$ nodes (assuming without loss of generality that $\alpha_m n$ is an integer). With the help of the community structure, the probabilities $p_{i,j}$ are constructed as follows: if $i \in H_\ell$ and $j \in H_m$, the probability of i and j being connected is given by $p_{i,j} = \frac{K_{\ell,m}}{n}$, where K is a symmetric matrix of size $S \times S$, with positive elements. The random graph from the above distribution is denoted as $G(n, \alpha, K)$.

In the stochastic block model, identifying a node with maximal reward amounts to finding a node from the most influential community. Consequently, it is easy to see that choosing α such that $\alpha > \min_m \alpha_m$, the near-optimal set V^*_α exactly corresponds to the set of optimal nodes, and thus the quantile regret (2.2) coincides with the regret (2.1). We consider the stochastic block models satisfying the following simplifying assumptions:

Assumption 1. $K_{l,m} = k > 0$ *for all* $l \neq m$.

This assumption requires that nodes i, j belonging to different communities are connected with the same probability. Additionally, in our analysis in the supercritical case we make the following natural assumptions:

Assumption 2. *For all l, $K_{l,l} > k$.*

In plain words, this assumption requires that the density of edges within communities is larger than the density of edges between communities.

Chung–Lu model

Another thoroughly studied special case of the inhomogeneous Erdős–Rényi model is the so-called *Chung–Lu model* (sometimes referred to as *rank-1 model*) as first defined by Chung and Lu [49] (see also [50, 28]). In this model the edge probabilities are defined by a vector $w \in \mathbb{R}^n$ with positive components, representing the "weight" of each vertex. Then the matrix defining the edge probabilities has entries $A_{ij} = w_i w_j$. We assume that the vector w is such that $w_i w_j / n < 1$ for all i, j. In other words, the Chung–Lu model considers rank-1 matrices of the form $A = ww^\intercal$. The random graph from the Chung–Lu model is denoted by $G(n, w)$. Chung–Lu random graphs replicate some key properties of certain real networks. For instance, if w is a sequence satisfying a power law, then $G(n, w)$ is a power law model, which allows one to model social networks, see [50].

Kronecker graphs

Kronecker random graphs were introduced by [94, 93, 95] as models of large networks appearing in various applications, including social networks. The matrix P of the edge probabilities of a Kronecker random graph $G_{n,P^{[k]}}$ is defined recursively. The model is parametrized by the constants $\zeta, \beta, \gamma \in [0, 1]$. Here one assumes that the number of vertices n is a power of 2. Starting from a 2×2 seed matrix,

$$P^{[1]} = \begin{bmatrix} \zeta & \beta \\ \beta & \gamma \end{bmatrix} ,$$

we define the matrices $P^{[2]}, \dots, P^{[k]}$ such that for each $i = 2, \dots, k$, $P^{[i]}$ is a $2^i \times 2^i$ matrix obtained from $P^{[i-1]}$ by

$$P^{[i]} = \begin{bmatrix} \zeta P^{[i-1]} & \beta P^{[i-1]} \\ \beta P^{[i-1]} & \gamma P^{[i-1]} \end{bmatrix} .$$

Finally $P = P^{[k]}$. Hence, the Kronecker random graph $G_{n,P^{[k]}}$ has $n = 2^k$ vertices, where each vertex i is characterised by a binary string $s_i \in \{0, 1\}^k$, such that the probability of an edge between nodes i and j is equal to $p_{i,j} = \zeta^{\langle s_i, s_j \rangle} \gamma^{\langle \bar{1} - s_i, \bar{1} - s_j \rangle} \beta^{k - \langle s_i, s_j \rangle - \langle \bar{1} - s_i, \bar{1} - s_j \rangle}$, where $\bar{1} = (1, \ldots, 1) \in \{0, 1\}^k$ denotes the all-one vector and $\langle \cdot, \cdot \rangle$ is the usual inner product. [94] show that a Kronecker graph with properly tuned values of ζ, β, γ replicates properties of real world networks, such as small diameter, clustering, and heavy-tailed degree distribution.

2.3.2 Learning with degree feedback in stochastic block models and Chung–Lu graphs

In this section we introduce an online influence maximization algorithm that only uses the degree of the selected node as feedback information. The algorithm is a variant of the kl-UCB algorithm, that was proposed and analyzed by [65, 102, 37, 87]. The main reason why learning is possible based on degree observations only is that nodes with the largest expected degrees μ^* are exactly the ones with the largest influence c^*. This (nontrivial) fact holds in both the stochastic block model (under Assumptions 1 and 2) and the Chung–Lu model, across both the subcritical and supercritical regimes. These facts are proven in Sections 2.6.1 and 2.6.2. Further, we define $X_{t,i}$ as the degree of node i in the realized graph G_t, and define $\mu_i = \mathbb{E}\left[X_{1,i}\right]$ as the expected degree of node i. We also define $c^* = \max_i c_i$ and $\mu^* = \max_i \mu_i$.

The learner uses the observed degrees as rewards, and feeds them to an instance of kl-UCB originally designed for Poisson-distributed rewards. A key technical challenge arising in the analysis is that the degree distributions do not actually belong to the Poisson family for finite n. We overcome this difficulty by showing that the degree distributions have a moment generating function bounded by those of Poisson distributions, and that this fact is sufficient for most of the kl-UCB analysis to carry through without changes.

As in the case of the inhomogeneous Erdős–Rényi model, we subsample a set of size given in Equation (2.4) of representative nodes for kl-UCB to play on. For clarity of presentation, we first propose a simple algorithm that assumes prior knowledge of T, and then move on to construct a more involved variant that adds new actions on the fly. We present our kl-UCB variant for a fixed set of nodes V_0 as Algorithm 2.4. We refer to this algorithm as d-UCB(V_0) (short for "degree-UCB on V_0"). Our two algorithms mentioned above use d-UCB(V_0) as a subroutine: they are both based on uniformly sampling a large enough set V_0 of nodes so that the subsample includes at least one node from the top α-quantile,

```
Algorithm 2.4 $d$-UCB$(V_0)$
Parameters: A set of nodes $V_0 \subseteq V$.
Initialization: Select each node in $V_0$ once. Observe the degree $X_{i,i}$ of vertex $i$ in the graph $G_i$ for $i = 1, \dots, |V_0|$. For each $i \in V_0$, set $N_i(|V_0|) = 1$ and $\widehat{\mu}_i(|V_0|) = X_{i,i}$.
For $t = |V_0|, \dots T$, repeat
    1. For each node, compute
       $$U_i(t) = \sup\left\{\mu : \mu - \widehat{\mu}_i(t) + \widehat{\mu}_i(t)\log\left(\frac{\widehat{\mu}_i(t)}{\mu}\right) \le \frac{3\log(t)}{N_i(t)}\right\}.$$
    2. Select any node $A_{t+1} \in \arg\max_i U_i(t)$.
    3. Observe degree $X_{t+1,A_{t+1}}$ of node $A_{t+1}$ in $G_{t+1}$ and update
       $$\widehat{\mu}_{A_{t+1}}(t+1) = \frac{N_{A_{t+1}}(t)\widehat{\mu}_{A_{t+1}}(t+1) + X_{t+1,A_{t+1}}}{N_{A_{t+1}}(t) + 1}.$$
       Update $N_{A_{t+1}}(t+1) = N_{A_{t+1}}(t) + 1$.
```

with high probability. We define the α-optimal degree $\mu^*_\alpha = \min_{i \in V^*_\alpha} \mu_i$ and the gap parameter $\delta_{\alpha,i} = (\mu_i - \mu^*_\alpha)_+$. We first present a performance guarantee of our simpler algorithm that assumes knowledge of T, so the learner plays d-UCB(V_0) on the uniformly sampled a subset of size (2.4). This algorithm satisfies the following performance guarantee:

Theorem 2.3.1. *Assume that the underlying random graph is either (a) a subcritical stochastic block model satisfying Assumption 1; (b) a supercritical stochastic block model satisfying Assumptions 1 and 2; (c) a subcritical Chung–Lu random graph; or (d) a supercritical Chung–Lu random graph.*

Let V_0 be a uniform subsample of V with size given in Equation (2.4) *and define the event $\mathcal{E} = \{V_0 \cap V^*_\alpha \neq \emptyset\}$. If the number of vertices n is sufficiently large, then the expected α-quantile regret of d-UCB(V_0) simultaneously satisfies*

Algorithm 2.5 d-UCB-DOUBLE(β)

```
Parameters: $\beta \geq 2$.
Initialization: $V_0 = \emptyset$.
For $k = 1, 2 \ldots,$ repeat
    1. Sample subset of nodes $U_k$ uniformly such that $|U_k| = \left\lceil \frac{\log \beta}{\log(1/(1-\alpha))} \right\rceil$.
    2. Update action set $V_k = V_{k-1} \cup U_k$.
    3. For rounds $t = \beta^{k-1}, \beta^{k-1} + 1, \ldots, \beta^k - 1$, run a new instance of $d$-UCB ($V_k$).
```

$$R_T^\alpha \leq \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i} \left(\frac{\mu_\alpha^* \left(18 + 27 \log T\right)}{\delta_{\alpha,i}^2} + 3\right) \middle| \, \mathcal{E} \right] + \Delta_{\alpha,\max},$$

where the expectation is taken over the random choice of V_0, *and*

$$R_T^\alpha \leq 18c^* \sqrt{\frac{T\mu^* \left(2 + 3 \log T\right)^2}{\log(1/(1-\alpha))}} + \left(\frac{3 \log T}{\log(1/(1-\alpha))} + 4\right) \Delta_{\alpha,\max}.$$

In contrast to the results obtained in the general setting of Section 2.2, where we have to run different algorithms in the subcritical and supercritical cases, for the models considered in this section the learner can run the Algorithm 2.5 without prior knowledge of the regime.

For unknown values of T, we propose the d-UCB-DOUBLE(β) algorithm (presented as Algorithm 2.5) that uses a doubling trick to estimate T. The following theorem gives a performance guarantee for this algorithm:

Theorem 2.3.2. *Assume that the underlying random graph is either (a) a subcritical stochastic block model satisfying Assumption 1; (b) a supercritical stochastic block model satisfying Assumptions 1 and 2; (c) a subcritical Chung–Lu random graph; or (d) a supercritical Chung–Lu random graph.*

Fix T, let $k_{\max}$ *be the value of k on which d-UCB-*DOUBLE(β) *terminates, and define the event* $\mathcal{E} = \{V_{k_{\max}} \cap V_\alpha^* = \emptyset\}$. *If the number of vertices* n *is sufficiently large, then the* α*-quantile regret of d-UCB-*DOUBLE(β) *simultaneously satisfies*

$$R_T^\alpha \le \mathbb{E}\left[\sum_{i\in V_{k_{\max}}} \Delta_i \left(\left(\frac{18\mu^*}{\delta_{\alpha,i}^2} + 3 \right) (\log_\beta T + 1) + \frac{27 \log \beta (\log_\beta T + 1)^2}{2\delta_{\alpha,i}^2} \right) \middle| \, \mathcal{E} \right]$$
$$+ \Delta_{\alpha,\max} \log_\beta T,$$

where the expectation is taken over the random choice of the sets $V_1, V_2, \ldots$, and

$$R_T^\alpha \le 36c^* \sqrt{\frac{T\left(\mu^* + \log\left(\beta T\right)\right)\log^2 T}{\log(1/(1-\alpha))}} + \left(\frac{3\log^2 T}{\log(1/(1-\alpha))} + 4 \right) \Delta_{\alpha,\max}.$$

2.3.3 Learning with degree feedback in Kronecker random graphs

In this section we study influence maximization when the underlying random network is a Kronecker random graph. We set this model apart as the properties of Kronecker random graphs differ significantly from those of the stochastic block model and the Chung–Lu model. At the same time, we show that observing the degree of the selected nodes is enough to maximize the total influence in this graph model as well. In particular, the same algorithm d-UCB(V_0) introduced above achieves a small regret.

Since subcritical Kronecker random graphs contain only $o(n)$ non-isolated vertices with high probability, we consider only supercritical regime with parameters are such that $(\zeta + \beta)(\beta + \gamma) > 1$. Denote by H the subgraph of $G_{n,P^{[k]}}$, induced by the vertices of weight $l \ge k/2$. We exploit the property that for the graph $G_{n,P^{[k]}}$with parameters $(\zeta + \beta)(\beta + \gamma) > 1$, there exists a constant $b(P)$ such that a subgraph of $G_{n,P^{[k]}}$, induced by the vertices of H, is connected with probability at least $1 - n^{-b(P)}$, see Frieze and Karonski [63, Theorem 9.10]. This means that on this event, the connected components C_i are the same for all $i \in H$. This allows us to prove the following:

Theorem 2.3.3. *Let V_0 be a uniform subsample of V of size $\left\lceil \frac{\log(nT)}{\log(2)} \right\rceil$. Let $G_{n,P^{[k]}}$be such that $(\zeta + \beta)(\beta + \gamma) > 1$ and $\zeta > \gamma > \beta$. Then there exists a constant $b(P)$ such that the quantile regret of d-UCB(V_0) satisfies*

$$\frac{R_T^\alpha}{n} \le \left\lceil \frac{\log(nT)}{\log(2)} \right\rceil \left(\frac{\mu^*(2 + 6\log T)}{\left(1 - \frac{\beta+\gamma}{\zeta+\beta}\right)^2} + 3 + n^{-b(P)} \frac{\mu^*(2 + 6\log T)(\zeta + \beta)^2}{(\zeta - \gamma)^2} + 3n^{-b(P)} \right) +$$

2.4 Multi-type branching processes

One of the most important technical tools for analyzing the component structure of random graphs is the theory of *branching processes*, see Bollobás et al. [28], van der Hofstad [141]. Indeed, while the connected components of an inhomogenous random graph $G(n, \kappa)$ have a complicated structure, many of their key properties may be analyzed through the concept of multi-type Galton–Watson processes. Recall the notation introduced in Section 2.1.2. Consider a Galton–Watson process, where an individual $x \in (0, 1]$ is replaced in the next generation by a set of particles distributed as a Poisson process on $(0, 1]$ with intensity $\kappa(x, y)d\mu(y)$ and the number of children has a Poisson distribution with mean $\int_{(0,1]} \kappa(x, y)d\mu(y)$. We denote this branching process, started with a single particle x by $W_\kappa(x)$. [28] establishes a connection between the sizes of connected components of $G(n, \kappa)$, the survival probability of a branching process $W_\kappa(x)$, and the function κ. As shown in [28], the operator Φ_κ can be directly used for characterizing the probability $\rho(x)$ of survival of the process $W_\kappa(x)$ for all $x \in (0, 1]$. By their Theorem 6.2, the function ρ is the maximum fixed point of the non-linear equation $\Phi_\kappa(f) = f$. Furthermore, as was shown in Bollobás et al. [28, Lemma 5.8.], if $\|T_\kappa\|_2 < 1$, then $\rho(x) = 0$ for all x and when $\|T_\kappa\|_2 > 1$, $\rho(x) > 0$ for all x.

To analyze the random graph $G(n, \kappa)$, we use Poisson multi-type Galton–Watson branching processes with n types, parametrized by an $n \times n$ matrix A with positive elements. Therefore, each node corresponds to its own type. The branching process tracks the evolution of a set of *individuals* of various types. Starting in round $n = 0$ from a single individual of type i, each further generation in the Galton–Watson process $W_\kappa(i)$ is generated by each individual of each type i producing $X_{i,j} \sim Poisson(A_{i,j}/n)$ new individuals of each type j. Therefore, the number of offsprings of the individual of type i is $\sum_{j=1}^{n} X_{i,j} \sim Poisson(\sum_{j=1}^{n} A_{i,j}/n)$.

Our analysis below makes use of the following quantities associated with the multi-type branching process:

1. $Z_n(i)$ is the number of individuals in generation n of $W_\kappa(i)$ (where $Z_0(i) = 1$);

2. $B(i)$ is the *total progeny*, that is, the total number of individuals generated by $W_\kappa(i)$ and its expectation is denoted by $x_i = \mathbb{E}[B(i)]$;

3. $\rho(i)$ is the *probability of survival*, that is, the probability that $B(i)$ is infinite.

2.5 Analysis of inhomogeneous random graph model

2.5.1 Proofs of Theorem 2.2.1 and 2.2.2.

The connected components C_i of an individual i have a complicated structure, but many key properties can be analyzed through the concept of multi-type Galton-Watson branching processes with n types. Fix an arbitrary node i and let $Y_{i,1}, Y_{i,2}, \dots, Y_{i,n}$ be independent Bernoulli random variables with respective parameters $A_{i,j}/n$ for $i, j \in [n]$. Consider a multitype binomial branching process where an individual of type i produces an individual j with probability $A_{i,j}/n$, and let $B_{Ber}(i)$ denote its total progeny when started from an individual i. In the same way, consider a multitype Poisson branching process where an individual of type i produces $X_{i,j} \sim Poisson(A_{i,j}/n)$ individuals, and let $B(i)$ denote its total progeny when started from an individual i. We use the concept of *stochastic dominance* between random variables. The random variable X is *stochastically dominated* by the random variable Y when, for every $x \in \mathbb{R}$, $\mathbb{P}[X \leq x] \geq \mathbb{P}[Y \leq x]$. We denote this by $X \preceq Y$.

Proof of Lemma 1. First, we define an upper approximation to κ. We choose an integer m and we partition the interval $(0, 1]$ into m sets $\mathcal{A}_1, \dots, \mathcal{A}_m$, where $\mathcal{A}_k = ((k-1)/m, k/m]$, $k \in [1, m]$. Also we denote by $\mathcal{A}_m(x)$ the set $\mathcal{A}_k$ for which $x \in \mathcal{A}_k$. Then we bound κ from above by

$$\kappa_m^+(x, y) = \sup\{\kappa(x', y') : x' \in \mathcal{A}_m(x), y' \in \mathcal{A}_m(y)\} .$$

As κ is bounded, there exists a sufficiently large m such that $\|T_{\kappa_m^+}\| < 1$:

$$\|T_{\kappa_m^+}\| \leq \|T_\kappa\| + \|T_{\kappa_m^+} - T_\kappa\| \leq \|T_\kappa\| + \left(\int_{(0,1]\times(0,1]} (\kappa_m^+(x, y) - \kappa(x, y))^2 dxdy \right)^{1/2} .$$

Then for any node i in $G(n, \kappa)$, we define a type $k_i = k$ if $(k-1)/m < i/n \leq k/m$ holds. By our definition of κ_m^+, we have

$$\mathbb{P}[|C_i| > u] \leq \mathbb{P}[B_{Ber}(k_i) > u] .$$

For $k, \ell \in [m]$ we define $p_{k,\ell} = \frac{1}{m}\kappa_m^+(k/m, \ell/m)$. Notice, that for random variables $Y \sim Ber(p)$ and $X \sim Poisson(p')$ with $p' = -\log(1-p) > p$, $Y \preceq X$ holds. This follows from the observation that $\mathbb{P}[Y > 0] = p$ and $\mathbb{P}[X > 0] = p$. It follows that $Ber(p_{k,\ell}) \preceq Poisson((1+\varepsilon)p_{k,\ell})$. Then there exists $\varepsilon > 0$ such that the multitype Poisson branching process $\widetilde{B}(k)$ with parameters $(1+\varepsilon)p_{k,\ell}$ is such that $\mathbb{P}[B_{Ber}(k) > u] < \mathbb{P}\left[\widetilde{B}(k) > u\right]$ and it is subcritical. We also define a random variable $\widetilde{X}_{k,\ell} \sim Poisson\left((1+\varepsilon)p_{k,\ell}\right)$. Since the total number of descendants of individuals in the first generation are independent, we can write the following recursive equation on the number of descendants of type k:

$$|\widetilde{B}(k)| = 1 + \sum_{\ell=1}^{m} \widetilde{X}_{k,\ell}|\widetilde{B}(\ell)|.$$

For any type k, for $z_k > 1$, the probability generating function of $|\widetilde{B}(k)|$ is $g(k) = \mathbb{E}\left[z_k^{|\widetilde{B}(k)|}\right]$ and we denote $g = (g(1), \ldots, g(m))^T$. Using that for $X \sim Poisson(\gamma)$ for some $\gamma > 0$, $y > 1$ the probability generating function is $\mathbb{E}\left[y^X\right] = e^{\gamma(y-1)}$, we have

$$\begin{aligned} g(k) = \mathbb{E}\left[z_k^{|\widetilde{B}(k)|}\right] &= z_k\mathbb{E}\left[z_k^{\widetilde{X}_{k,1}|\widetilde{B}(1)|} \ldots z_k^{\widetilde{X}_{k,\mathcal{M}}|\widetilde{B}(m)|}\right] = z_k \prod_\ell \mathbb{E}\left[z_k^{\widetilde{X}_{k,\ell}|\widetilde{B}(\ell)|}\right] \\ &= z_k \prod_\ell \mathbb{E}\left[\left(\mathbb{E}\left[z_k^{|\widetilde{B}(\ell)|}\right]\right)^{\widetilde{X}_{k,\ell}}\right] = z_k \exp\left((1+\varepsilon)\sum_\ell p_{k,\ell}(g(\ell)-1)\right). \end{aligned}$$

Recall that P denotes the $m \times m$ matrix with entries $p_{k,\ell}$. Our next aim is to study the fixed point of the operator G_P, defined as

$$g = G_P g := z \exp\left((1+\varepsilon)P(g - \bar{1})\right). \tag{2.5}$$

Define the function $F(z, g) = z\exp\left((1+\varepsilon)P(g-\bar{1})\right) - g$. This function is smooth and the entries of the Jacobian matrix are

$$J_{k,\ell}(z, g) := \frac{\partial F_k}{\partial g_\ell} = z_k(1+\varepsilon)p_{k,\ell}\exp\left((1+\varepsilon)\sum_\ell p_{k,\ell}(g_\ell - 1)\right) - \mathbb{I}_{\{k=\ell\}}.$$

Let $P'(g,z)$ be the matrix with elements $z_k(1+\varepsilon)p_{k,\ell}\exp\left((1+\varepsilon)\sum_\ell p_{k,\ell}(g_\ell - 1)\right)$. Then, at point $(\bar{1},\bar{1})$, $P'_{k,\ell}(\bar{1},\bar{1}) = (1+\varepsilon)p_{k,\ell}$. Since ε is chosen such that the branching process $\widetilde{B}(k)$ is subcritical, $P'(\bar{1},\bar{1})$ is smaller than one. This means, that we can find $z' = 1+\delta, g' > 0$, such that the largest eigenvalue of $P'(g',z')$ is smaller than one as well, and therefore $J(z',g')$ is invertible. Then, by the implicit function theorem there exists an open set $U_z \subset (1,+\infty)^m$ and a function $q : U_z \to (0,+\infty)^m$ such that $F(z,q(z)) = \bar{0}$.

Finally, the statement of the lemma is obtained by applying the Chernoff bound:

$$\mathbb{P}\left[\widetilde{B}(k) > u\right] = \mathbb{P}\left[z_k^{\widetilde{B}(k)} > z_k^u\right] \leq \frac{\mathbb{E}\left[z_k^{\widetilde{B}(k)}\right]}{z_k^u}.$$

Denote $\lambda_k = \ln(z_k) > 0$. Then,

$$\frac{\mathbb{E}\left[z_k^{\widetilde{B}(k)}\right]}{z_k^u} = \frac{g_k}{z_k^u} = \exp(-\lambda_k u)g_k.$$

Then taking any $\lambda(\kappa) = \min_k \lambda_k$, $g(\kappa) = \max_k g_k$, we get the statement of the lemma. □

Armed with this concentration result, we can see that the typical size $|C_i|$ of the connected component of any vertex i is $O(1)$. Recall that the learning algorithm has only access to a censored value of $|C_i|$, truncated by a constant K. Our main technical result shows that nodes with the largest expected censored observations $u_*(K)$ are exactly the ones with the largest influence c_*. We formally state this result next:

Lemma 2. *For $G(n,\kappa)$ with subcritical κ, and $n > n_0(\kappa)$, for any node i we have $c_* - c_i \leq u_*(K) - u_i(K) + e^{-\lambda(\kappa)K}g$. Then, for $K = \frac{\log T}{\lambda}$, with $\lambda < \lambda(\kappa)$ we have $c_* - c_i \leq u_*(K) - u_i(K) + \frac{g(\kappa)}{T}$.*

Proof. The expected bias of $u_i(K)$ is, using the result of Lemma 1:

$$\begin{aligned} c_i - u_i(K) &= \mathbb{E}\left[|C_i(G_t)| - u_i(K)\right] = \mathbb{E}\left[(|C_i| - K)_+\right] \\ &\leq \int_0^\infty \mathbb{P}\left[|C_i| - K > u\right] du \leq \int_0^n e^{-\lambda(u+K)} du \leq e^{-\lambda K} g(\kappa). \end{aligned}$$

Set $K = \frac{\log T}{\lambda}$. Then,

$$c_* - c_i \leq u_*(\log T/\lambda) - u_i(\log T/\lambda) + \frac{g(\kappa)}{T}.$$

□

Proof of Theorem 2.2.1. In order not to overload notation we write $\delta_{\alpha,i}^{sub}$ for $\delta_{\alpha,i}^{sub}(K)$. We first note that, with high probability, the size of V_0 guarantees that the subset contains at least one node from the set V_α^*: $\mathbb{P}[\mathcal{E}] \geq 1 - 1/T$. Then, the regret can be bounded as

$$R_T^\alpha \leq \mathbb{P}[\mathcal{E}^c]\, T\Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V_0}\mathbb{I}[A_t = i]\Delta_{\alpha,i}\middle|\, \mathcal{E}\right]\mathbb{P}[\mathcal{E}] \tag{2.6}$$

$$\leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i\in V_0}\Delta_{\alpha,i}\mathbb{E}[N_{i,T}]\middle|\, \mathcal{E}\right]. \tag{2.7}$$

By Hoeffding's inequality,

$$\mathbb{P}\left[A_{t+1} = i|\, N_{i,t} \geq \frac{4K^2\log t}{(\delta_{\alpha,i}^{sub})^2}\right] \leq \frac{4}{t^2}.$$

Then,

$$\mathbb{E}[N_{i,T}] \leq \frac{4K^2\log T}{(\delta_{\alpha,i}^{sub})^2} + \sum_{t=|V_0|}^{T}\mathbb{P}\left[A_{t+1} = i|\, N_{i,t} \geq \frac{4K^2\log t}{(\delta_{\alpha,i}^{sub})^2}\right]$$
$$\leq \frac{4K^2\log T}{(\delta_{\alpha,i}^{sub})^2} + \sum_{t=|V_0|}^{T}\frac{4}{t^2} \leq \frac{4K^2\log T}{(\delta_{\alpha,i}^{sub})^2} + 8.$$

Now, observing that $\delta_{\alpha,i}^{sub} \leq \max_{j\in V_0} u_j(K) - u_i(K)$ holds under event $\mathcal{E}$, we obtain

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i\in V_0}\Delta_{\alpha,i}\left(\frac{4K^2\log T}{(\delta_{\alpha,i}^{sub})^2} + 8\right)\middle|\, \mathcal{E}\right], \tag{2.8}$$

thus proving the first statement.

Next, we turn to proving the second statement regarding worst-case guarantees. To do this, we appeal to Proposition 2 and take $K = \frac{\log T}{\lambda}$, where λ is any number, satisfying conditions of Lemma 1. To proceed, let us fix an arbitrary $\varepsilon > 0$ and split the set V_0 into two subsets: $U(\varepsilon) = \left\{a \in V_0 : \delta^{sub}_{\alpha,i} \leq \varepsilon\right\}$ and $W(\varepsilon) = V_0 \setminus U(\varepsilon)$. Then, under event $\mathcal{E}$, we have

$$\begin{aligned}
\sum_{i \in V_0} \Delta_{\alpha,i} \mathbb{E}\left[N_{i,T}\right] &= \sum_{i \in U(\varepsilon)} \Delta_{\alpha,i} \mathbb{E}\left[N_{i,T}\right] + \sum_{i \in W(\varepsilon)} \Delta_{\alpha,i} \mathbb{E}\left[N_{i,T}\right] \\
&\leq \varepsilon \sum_{i \in U(\varepsilon)} \mathbb{E}\left[N_{i,T}\right] + \frac{g}{T} \sum_{i \in U(\varepsilon)} \mathbb{E}\left[N_{i,T}\right] \\
&+ \sum_{i \in W(\varepsilon)} \delta^{sub}_{\alpha,i} \left(\frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{(\delta^{sub}_{\alpha,i})^2} \right) + \frac{g}{T} \sum_{i \in W(\varepsilon)} \frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{(\delta^{sub}_{\alpha,i})^2} \\
&+ 8|W(\varepsilon)|\Delta_{\alpha,\max} \\
&\leq \varepsilon T + g + \sum_{i \in W(\varepsilon)} \left(\frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{\delta^{sub}_{\alpha,i}} \right) + \frac{g}{T} \sum_{i \in W(\varepsilon)} \frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{(\delta^{sub}_{\alpha,i})^2} \\
&+ 8|W(\varepsilon)|\Delta_{\alpha,\max} \\
&\leq \varepsilon T + g + |V_0| \frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{\varepsilon} + \frac{g}{T}|V_0| \frac{4\left(\frac{\log T}{\lambda}\right)^2 \log T}{\varepsilon^2} \\
&+ 8|V_0|\Delta_{\alpha,\max} \\
&\leq 4\left(\frac{\log T}{\lambda}\right) \sqrt{|V_0| T \log T} + 2g + 8|V_0|\Delta_{\alpha,\max}.
\end{aligned}$$

where the last step uses the choice $\varepsilon = 2\left(\frac{\log T}{\lambda}\right)\sqrt{|V_0| \log T / T}$. Plugging in the choice of $|V_0|$ concludes the proof. □

Proof of Theorem 2.2.2. To simplify the notation, we use λ instead of $\lambda(\kappa)$. Let T_q be the length of the q-th iterate. The expected regret over each period q can be bounded as an expected regret of Local UCB(V_0) with parameters $\lambda_q = 2^{-q}$ and T_q time steps. Appealing to Theorem 2.2.1, we can bound the expected regret as

$$
\begin{aligned}
R_T^\alpha &\leq \mathbb{P}\left[\mathcal{E}^c\right] T \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in V_0} \mathbb{I}_{\{A_t = i\}} \Delta_{\alpha,i} \middle| \mathcal{E} \right] \\
&\leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{q=1}^{Q_{\max}} \sum_{i \in V_0} \Delta_{\alpha,i} \mathbb{E}\left[N_{i,T_q}\right] \middle| \mathcal{E} \right]
\end{aligned}
$$

Following the analysis of Theorem 2.2.1, by (2.8), we get

$$
R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{q=1}^{Q_{\max}} \sum_{i \in V_0} \Delta_{\alpha,i} \left(\frac{4K_q^2 \log T}{(\delta_{\alpha,i}^{sub}(K_q))^2} + 8 \right) \middle| \mathcal{E} \right].
$$

We have $Q_{\max} = \lceil \log_2(1/\lambda) \rceil \leq \log_2(1/\lambda) + 1$, and

$$
\sum_{q=0}^{Q_{\max}} K_q^2 = \log^2 T \sum_{q=0}^{Q_{\max}} 4^q = \log^2 T \frac{4^{Q_{\max}+1} - 1}{3} \leq \frac{16}{3} \frac{1}{\lambda^2} \log^2 T.
$$

This gives us

$$
R_T^\alpha \leq \Delta_{\alpha,\max} + \frac{64}{3} \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i} \left(\frac{\log^3 T}{(\lambda \cdot \min_{q \in [Q_{\max}]} \{\delta_{\alpha,i}^{sub}(K_q)\})^2} + 8 \right) \middle| \mathcal{E} \right].
$$

Next, we prove the second statement regarding worst-case guarantees. To proceed, let us take $\varepsilon_q = \frac{\log T}{\lambda} \sqrt{|V_0| \log T / T_q}$ and split the set V_0 into two subsets: $U(\varepsilon_q) = \left\{ a \in V_0 : \delta_{\alpha,i}^{sub}(K_q) \leq \varepsilon_q \right\}$ and $W(\varepsilon_q) = V_0 \setminus U(\varepsilon_q)$. Then, under event $\mathcal{E}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V_0}\mathbb{I}_{\{A_t=i\}}\Delta_{\alpha,i}\right] \\
&\quad\leq \mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in V_0}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|\leq K_q\right\}}\right]+\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in V_0}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|>K_q\right\}}\right] \\
&\quad= \underbrace{\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in U(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|\leq K_q\right\}}\right]}_{\text{Term 1}}+\underbrace{\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in U(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|>K_q\right\}}\right]}_{\text{Term 2}} \\
&\quad+ \underbrace{\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in W(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|\leq K_q\right\}}\right]}_{\text{Term 3}}+\underbrace{\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in W(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|>K_q\right\}}\right]}_{\text{Term 4}}.
\end{aligned}$$

Term 1:

$$\begin{aligned}
\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in U(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|\leq K_q\right\}}\right] &\leq |V_0|\frac{\log T}{\lambda}\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sqrt{|V_0|T_q\log T}\right] \\
&\leq |V_0|^{3/2}\frac{\log T}{\lambda}\sqrt{(\log_2(1/\lambda)+1)T}.
\end{aligned}$$

Term 2: The expected bias of $\mu_{i,t}^{sub}(K_q)$ is, using the result of Lemma 1:

$$\begin{aligned}
\mathbb{E}\left[(|C_i|-K_q)_+\right] \leq \int_0^\infty \mathbb{P}\left[|C_i|-K_q>u\right]du \leq \int_0^n e^{-\lambda(u+K_q)}du \leq e^{-\lambda K_q}g \\
= \left(\frac{1}{T}\right)^{\lambda 2^q} g \leq \left(\frac{1}{T}\right)^{2^{q-Q_{max}}} g.
\end{aligned}$$

Then,

$$c_* - c_i \leq \mu_*^{sub}(K_q) - \mu_i^{sub}(K_q) + \left(\frac{1}{T}\right)^{2^{q-Q_{max}}} g. \tag{2.9}$$

According to the stopping rule, we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in U(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{|C_{A_{t_q}}|>K_q\right\}}\right]\\
&\quad\leq |V_0|\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\left(\varepsilon_q+g\left(\frac{1}{T}\right)^{2^{q-Q_{max}}}\right)\left(\frac{1}{T}+\sqrt{\frac{\log T}{2T_q}}\right)T_q\right]\\
&\quad\leq |V_0|\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\left(\frac{\log T}{\lambda}\sqrt{|V_0|\log T/T_q}+g\right)\left(\frac{1}{T}+\sqrt{\frac{\log T}{2T_q}}\right)T_q\right]\\
&\quad\leq |V_0|\left(\frac{\log T}{\lambda}\sqrt{\frac{(\log_2(1/\lambda+1))|V_0|\log T}{T}}+\frac{g}{T}\right)+|V_0|^{3/2}(\log_2(1/\lambda)+1)\frac{\log^2 T}{\lambda}\\
&\quad+|V_0|g\left(\sqrt{(\log_2(1/\lambda)+1)T\log T}\frac{\log T}{\lambda}\right).
\end{aligned}
$$

Term 3: Following the analysis of Theorem 2.2.1 and by (2.9), we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in W(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{A_t=i,|C_{A_t}|\leq K_q|\right\}}\right]\\
&\quad\leq \mathbb{E}\left[\sum_{i\in W(\varepsilon_q)}\sum_{q=0}^{Q_{\max}}\delta_{\alpha,i}^{sub}(K_q)\left(\frac{4\left(\frac{\log T_q}{\lambda}\right)^2\log T}{(\delta_{\alpha,i}^{sub}(K_q))^2}\right)\right]+8|V_0|\Delta_{\alpha,\max}\\
&\quad\leq 4\sqrt{|V_0|}\mathbb{E}\left[\sum_{q=0}^{Q_{\max}}\left(\sqrt{T_q}\frac{\log^{3/2}T}{\lambda}\right)\right]+8|V_0|\Delta_{\alpha,\max}\\
&\quad\leq 4\sqrt{|V_0|}\left(\sqrt{(\log(1/\lambda)+1)T}\frac{\log^{3/2}T}{\lambda}\right)+8|V_0|\Delta_{\alpha,\max}.
\end{aligned}
$$

Term 4:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{q=1}^{Q_{\max}}\sum_{t_q=1}^{T_q}\sum_{i\in W(\varepsilon_q)}\Delta_{\alpha,i}\mathbb{I}_{\left\{A_t=i,|C_{A_t}|>K_q|\right\}}\right]\\
&\le \mathbb{E}\left[\sum_{i\in W(\varepsilon_q)}\sum_{q=0}^{Q_{\max}}\left(\delta_{\alpha,i}^{sub}(K_q)+g\left(\frac{1}{T}\right)^{2^{q-Q_{\max}}}\right)\left(\frac{4\left(\frac{\log T_q}{\lambda}\right)^2\log T}{(\delta_{\alpha,i}^{sub}(K_q))^2}\right)\left(\frac{1}{T}+\sqrt{\frac{\log T}{2T_q}}\right)\right]\\
&+8|W(\varepsilon_q)|\Delta_{\alpha,\max}\\
&\le 4\sqrt{|V_0|}(\log_2(1/\lambda)+1)\frac{\log^{3/2}T}{\sqrt{T}\lambda}+4\sqrt{|V_0|}(\log_2(1/\lambda)+1)\frac{\log^2 T}{\lambda}\\
&+4g(\log_2(1/\lambda)+1)\frac{\log T}{\lambda}+4g\frac{\log^{3/2}T}{\lambda}\sqrt{(\log_2(1/\lambda)+1)T}+8|V_0|\Delta_{\alpha,\max}.
\end{aligned}
$$

Putting everything together, we conclude that

$$
\begin{aligned}
R_t^{\alpha} \le{}& 4\frac{1}{\sqrt{\ln(1/(1-\alpha))}}\left(\sqrt{(\log(1/\lambda)+1)T}\frac{\log^2 T}{\lambda}\right)+16\Delta_{\alpha,\max}\frac{\log T}{\sqrt{\ln(1/(1-\alpha))}}\\
&+4g(\log_2(1/\lambda)+1)\frac{\log T}{\lambda}+4g\frac{\log^{3/2}T}{\lambda}\sqrt{(\log_2(1/\lambda)+1)T}\\
&+\log T\cdot\sqrt{\frac{\log_2(1/\lambda+1)}{T\ln(1/(1-\alpha))}}+\frac{g\sqrt{\log T}}{T\sqrt{\ln(1/(1-\alpha))}}+\frac{2(\log_2(1/\lambda)+1)\log^{5/2}T}{(\ln(1/(1-\alpha)))^{3/2}}\\
&+g\sqrt{\frac{T(\log_2(1/\lambda)+1)}{\ln(1/(1-\alpha))}}\log T.
\end{aligned}
$$

2.5.2 Proof of Theorem 2.2.3.

The proof relies on some known properties of the largest connected component in $G(n,\kappa)$ for supercritical κ. We denote the largest and second-largest connected components of G_t by $C_1(G_t)$ and $C_2(G_t)$, respectively. The survival probability of the branching process $W_\kappa(x)$ is denoted as $\rho(x)$. The expected size of the connected component containing vertex i can be estimated in terms of $\rho(i/n)$ and $\mathbb{E}\left[|C_1|\right]$ as

$$c_i=\rho(i/n)\mathbb{E}\left[|C_1|\right]+o_n(n)\ ,$$

see Bollobás et al. [28, Chapter 9]. The following properties are proved by Bollobás et al. [28]:

- If $G(n, \kappa)$ is supercritical, then, with high probability, $C_1 = \Theta_n(n)$;
- $C_1(G_n) \to \sum_{i \in V} \rho(i/n)$ in probability;
- $C_2(G_n) = o_n(n)$ with high probability.

Recall from Section 2.2 that in the supercritical case the feedback $v_{i,t}(K)$ is the indicator whether $|C_i|$ is larger than K. In the following lemma we show that taking $K = k(n)$ for an arbitrary function of n that diverges to infinity, it is enough to control the bias of the estimate of c_i:

Lemma 3. *For any supercritical κ, for any node i satisfying $c_i < c_*$ and for any $K = k(n)$, where $k : \mathbb{N} \to \mathbb{N}$ is an arbitrary positive function satisfying $\lim_{n\to\infty} k(n) = \infty$, there exist a positive function $f_\kappa : \mathbb{N} \to \mathbb{R}$, such that $\lim_{n\to\infty} f_\kappa(n) = 0$ and*

$$\frac{c_* - c_i}{n} \leq (v_*(k(n)) - v_i(k(n)))\frac{\mathbb{E}\left[C_1\right]}{n} + f_\kappa(n).$$

Proof. Define a kernel $\bar{\kappa}(x, y) = (1 - \rho(y))\kappa(x, y)$, where ρ is defined in Section 2.4. By Theorem 6.7 in [28], the branching process W_κ conditional on extinction is subcritical and has the same distribution as the branching process with parameters $W_{\bar{\kappa}}$. Then, by Lemma 1,

$$\mathbb{P}\left[B(i) > K | B(i) < \infty\right] \leq e^{-\lambda(\bar{\kappa})k(n)} g(\bar{\kappa}) \,. \tag{2.10}$$

We relate the size of the connected component to the total progeny of branching process. Following the stochastic dominance $C_i \preceq B(i)$,

$$v_i(k(n)) = \mathbb{P}\left[|C_i| > k(n)\right] \leq \rho_i + \mathbb{P}\left[B(i) > k(n) | B(i) < \infty\right].$$

This implies, for $n > n_0(\bar{\kappa})$,

$$\begin{aligned} v_i(k(n))\mathbb{E}\left[|C_1|\right] - c_i &< \mathbb{P}\left[B(i) > k(n) | B(i) < \infty\right] \mathbb{E}\left[|C_1|\right] + \rho_i \mathbb{E}\left[|C_1|\right] - c_i \\ &\leq e^{-\lambda(\bar{\kappa})k(n)} g(\bar{\kappa}) \mathbb{E}\left[|C_1|\right] + o_n(n). \end{aligned}$$

Finally, using that $\rho_* \leq v_*(k(n))$, we get

$$\begin{aligned}\frac{c_* - c_i}{n} &\leq (v_*(k(n)) - v_i(k(n)))\frac{\mathbb{E}\left[|C_1|\right]}{n} + e^{-\lambda k(n)} g(\bar{\kappa})\frac{\mathbb{E}\left[|C_1|\right]}{n} + o_n(1) \\ &= \delta_i^{sup}(k(n))\mathbb{E}\left[|C_1|\right] + f_\kappa(n).\end{aligned}$$

□

Proof of Theorem 2.2.3. First, by (2.6),

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i}\mathbb{E}\left[N_{i,T}\right]\middle| \mathcal{E}\right].$$

As we mentioned before, with high probability, $C_2(G_n) = o_n(n)$, which means that if $A_t \notin C_1(G_t)$, then $|C_{A_t}(G_t)| = o_n(n)$. Since $G(n,\kappa)$ is supercritical, $\arg\max_a \mu_a = \arg\max_a \rho_a$. Then, we can approximate distribution of rewards of arm a by a Bernoulli distribution with parameter ρ_a. Using the result of Proposition 3, we reduce the initial problem to the analysis of a multi-armed problem with arms $Z_1, \ldots, Z_{|V_0|}$, where $Z_i \sim Ber(u_i)$, for p_i defined in Proposition 3.

By Hoeffding's inequality,

$$\mathbb{P}\left[A_{t+1} = i|\, N_{i,t} \geq \frac{4\log t}{(\delta_{\alpha,i}^{sup}(K)^2}\right] \leq \frac{4}{t^2}.$$

Then

$$\begin{aligned}\mathbb{E}\left[N_{i,T}\right] &\leq \frac{4\log T}{(\delta_{\alpha,i}^{sup}(k(n)))^2} + \sum_{t=|V_0|}^{T}\mathbb{P}\left[A_{t+1} = i|\, N_{i,t} \geq \frac{4\log t}{(\delta_{\alpha,i}^{sup}(k(n)))^2}\right] \\ &\leq \frac{4\log T}{(\delta_{\alpha,i}^{sup}(k(n)))^2} + 8.\end{aligned}$$

Now, observing that $\delta_{\alpha,i}^{sup}(k(n)) \leq \max_{j\in V_0} v_j - v_i$ holds under the event $\mathcal{E}$, we obtain

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i\in V_0}\Delta_{\alpha,i}\left(\frac{4\log T}{(\delta_{\alpha,i}^{sup}(k(n)))^2} + 8\right)\middle| \mathcal{E}\right], \tag{2.11}$$

thus proving the first statement.

Now we fix an arbitrary $\varepsilon > 0$, we split the set V_0 into two subsets: $U(\varepsilon) = \left\{a \in V_0 : \delta^{sub}_{\alpha,i}(k(n)) \leq \varepsilon\right\}$ and $W(\varepsilon) = V_0 \setminus U(\varepsilon)$, where we use the choice $\varepsilon = 2\sqrt{|V_0|\mathbb{E}\left[C_1\right]\log T/T}$. Lemma 3 shows that $\frac{c_* - c_i}{n} \leq \frac{v_*(k(n)) - v_i(k(n))}{n}\mathbb{E}\left[|C_1|\right] + f_\kappa(n)$. Then there exists $n_0(\kappa)$, such that for any $G(n,\kappa)$ with $n > n_0(\kappa)$, $f_\kappa(n) \leq \varepsilon$ holds. Then, under the event $\mathcal{E}$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i\in V_0}\Delta_{\alpha,i}\mathbb{E}\left[N_{i,T}\right] &= \sum_{i\in U(\varepsilon)}\frac{\Delta_{\alpha,i}}{n}\mathbb{E}\left[N_{i,T}\right] + \sum_{i\in W(\varepsilon)}\frac{\Delta_{\alpha,i}}{n}\mathbb{E}\left[N_{i,T}\right] \\
&\leq \left(\frac{\varepsilon\mathbb{E}\left[|C_1|\right]}{n} + \varepsilon\right)\sum_{i\in U(\varepsilon)}\mathbb{E}\left[N_{i,T}\right] + \sum_{i\in W(\varepsilon)}\frac{\Delta_{\alpha,i}}{n}\mathbb{E}\left[N_{i,T}\right] \\
&\leq \left(\frac{\varepsilon\mathbb{E}\left[|C_1|\right]}{n} + \varepsilon\right)|V_0|T + \sum_{i\in W(\varepsilon)}\delta^{sub}_{\alpha,i}(k(n))\frac{\mathbb{E}\left[C_1\right]}{n}\left(\frac{4\log T}{(\delta^{sub}_{\alpha,i}(k(n)))^2}\right) \\
&\quad + \varepsilon\sum_{i\in W(\varepsilon)}\left(\frac{4\log T}{(\delta^{sub}_{\alpha,i}(k(n)))^2}\right) \\
&\leq \left(\frac{\varepsilon\mathbb{E}\left[|C_1|\right]}{n} + \varepsilon\right)|V_0|T + |V_0|\left(\frac{\mathbb{E}\left[|C_1|\right]}{n} + 1\right)\frac{8\log T}{\varepsilon n} \\
&\leq 9\left(\frac{\mathbb{E}\left[|C_1|\right]}{n} + 1\right)|V_0|\sqrt{T\log T}\,,
\end{aligned}$$

where the last step uses the choice $\varepsilon = \sqrt{\log T/T}$. Plugging in the choice of $|V_0|$ concludes the proof. □

2.6 Degree observations.

2.6.1 Subcritical case

Our main technical result is proving that nodes with the largest expected degrees μ^* are exactly the ones with the largest influence c^*, in both the stochastic block model and the Chung–Lu model, across both the subcritical and supercritical regimes. The following lemma states this result for the subcritical case.

Lemma 4. *Suppose that*

1. *G is generated from a subcritical $G(n, \alpha, K)$ satisfying Assumption 1, or*
2. *G is generated from a subcritical $G(n, w)$.*

Then, for any i satisfying $\mu_i < \mu^*$, we have $c^* - c_i \leq 2c^* (\mu^* - \mu_i) + O(1/n)$.

Before stating and proving the lemma, we introduce some useful technical tools. Since we suppose that $G(n, \kappa)$ is subcritical, we have $\mathbb{P}[B(i) = \infty] = 0$ and $x_i = \mathbb{E}[B(i)]$ is finite. First observe that the vector x of expected total progenies satisfies the system of linear equations

$$x = e + \frac{1}{n} Ax \ ,$$

where e is the vector with $e_i = 1$ for all i.

For the analysis of the stochastic block model we define the vector $b \in \mathbb{R}^S$ with coordinates $b_l = \mu_l, l = 1, \ldots, S$, where by μ_l we define the expected degree of the node from community H_l. Also we define vector $x' \in \mathbb{R}^S$ with coordinates $x'_l = \mathbb{E}[B(l)]$, $l = 1, \ldots, S$, where by $B(l)$ we define the total progeny of the individual of type l. We define $x^* = \max_{i \in [n]} x_i$. Armed with this notation, we begin the proof Lemma 4, which consists of the following steps:

- proving that for any $i, j \in V$, $x_i - x_j \leq 2x^* (\mu_i - \mu_j)$, (Lemma 5, 6),
- proving that for any $i, j \in V$, $c_i - c_j = x_i - x_j + O(1/n)$ (Lemmas 7, 8).

These facts together lead to Lemma 4, given that n is large enough to suppress the effects of the residual terms. We begin with analysing the relation between b_l and x'_l in a straightforward way:

Lemma 5 (Coordinate order for mean of the total progeny in the SBM)**.** *Assume that $G(n, \alpha, K)$ is subcritical and that $K_{m\ell} = k > 0$ holds for all $m \neq \ell$. If two coordinates of b are such that $b_l > b_m$, then we have $x'_l > x'_m$, and $x'_l - x'_m \leq 2x^* (b_l - b_m)$.*

Proof. For the stochastic block model with S blocks, the system of equations $x = e + Ax$ can be equivalently written as $x' = e + Mx'$, for $M = K\text{diag}(\alpha) \in \mathbb{R}^{S \times S}$, and $x' \in \mathbb{R}^S$, with x'_m now standing for the expected total progeny associated with any node of type m. Similarly, we define b'_m as the expected degree of any node of type m. Notice that the system of equations $x' = e + Mx'$ satisfied by x' can be rewritten as $(I - M)x' = e$, where I is the $S \times S$ identity matrix. By exploiting our assumption on the matrix K and defining $\gamma_m = K_{m,m} - k$, this can be further rewritten as

$$\left(\begin{pmatrix} 1 - \alpha_1\gamma_1 & & \\ & \ddots & \\ & & 1 - \alpha_S\gamma_S \end{pmatrix} - k \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_S \\ \alpha_1 & \alpha_2 & \cdots & \alpha_S \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_S \end{pmatrix} \right) x' = e,$$

which means that for any m, x'_m satisfies

$$x'_m = \frac{1 + k(\alpha^\intercal x')}{1 - \alpha_m\gamma_m}.$$

Also observe that

$$b'_m = k(\alpha^T \bar{1}) + \alpha_m\gamma_m,$$

so, for any pair of types m and ℓ, we have

$$x'_m - x'_\ell = \frac{(1 + k(\alpha^\intercal x'))(\alpha_m\gamma_m - \alpha_\ell\gamma_\ell)}{(1 - \alpha_m\gamma_m)(1 - \alpha_\ell\gamma_\ell)},$$

which proves the first statement.

To prove the second statement, observe that for any pair ℓ and m of communities, we have either $\alpha_m \leq \frac{1}{2}$ or $\alpha_\ell \leq \frac{1}{2}$ (otherwise we would have $\alpha_m + \alpha_\ell > 1$). To proceed, let ℓ and m be such that $x'_m \geq x'_\ell$, and let us study the case $\alpha_\ell \leq \frac{1}{2}$ first. Here, we get

$$\begin{aligned} x'_m - x'_\ell = \frac{(1 + k(\alpha^\intercal x'))(\alpha_m\gamma_m - \alpha_\ell\gamma_\ell)}{(1 - \alpha_m\gamma_m)(1 - \alpha_\ell\gamma_\ell)} &= \frac{(\alpha_m\gamma_m - \alpha_\ell\gamma_\ell)}{(1 - \alpha_\ell\gamma_\ell)} x'_m \\ \leq \frac{(\alpha_m\gamma_m - \alpha_\ell\gamma_\ell)}{(1 - \gamma_\ell/2)} x'_m &\leq 2x'_m(b'_m - b'_\ell). \end{aligned}$$

In the other case where $\alpha_m \leq \frac{1}{2}$, we can similarly obtain

$$x'_m - x'_\ell \leq 2x'_\ell(b'_m - b'_\ell) \leq 2x'_m(b'_m - b'_\ell).$$

This concludes the proof. □

For the analysis of the Chung–Lu model, we define $\mu \in \mathbb{R}^n$ as the vector of mean degrees. Then we may prove the following.

Lemma 6 (Coordinate order for mean of the total progeny in the Chung–Lu model)**.** *Assume that $G(n, w)$ is subcritical. If two nodes are such that $\mu_i > \mu_j$, then we have $x_i > x_j$ and $x_i - x_j \leq x^*(\mu_i - \mu_j)$.*

Proof. From the system of equations $x = e + \frac{1}{n}Ax$, the coordinates x_i have the form

$$x_i = 1 + \frac{1}{n} \cdot w_i \left(\sum_{j=1}^{n} w_j x_j \right),$$

which implies that $w_i \geq w_j$ holds if and only if $x_i \geq x_j$. This observation implies for $x^* = \max_i x_i$

$$x_i - x_j \leq \frac{1}{n} \cdot (w_i - w_j) \left(\sum_{j=1}^{n} w_j \right) x^* = (\mu_i - \mu_j)\, x^*,$$

thus concluding the proof. □

The next two lemmas establish the relationship between the expected component size c_i of vertex i and the expected total progeny x_i of the multi-type branching process seeded at vertex i.

Lemma 7. *For any i, the mean of the connected component associated with type i is bounded by the mean of the total progeny: $c_i \leq x_i$.*

Proof. Now fix an arbitrary $i \in [n]$ and let $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n}$ be independent Bernoulli random variables with respective parameters $(A_{i,1}/n, A_{i,2}/n,$
$\ldots, A_{i,i}/n, \ldots, A_{i,n}/n)$. Consider a multitype binomial branching process where the individual of type i produces $Y_{i,j}$ individuals of type j, and let $B_{\text{Ber}}(i)$ denote its total progeny when started from an individual of type i. Recalling the Poisson branching process defined in Section 2.4 with offspring-distributions $X_{i,j}$, we can show $B_{\text{Ber}}(i) \preceq B(i)$ using the relation $Y_{i,j} \preceq X_{i,j}$.

Considering a node a of type i, we can use Theorem 4.2 of van der Hofstad [141] to bound the size of the the connected component C_a as $|C_a| \preceq B_{\text{Ber}}(i)$, which implies by transitivity of $\preceq$ that $|C_{a_i}| \preceq B(i)$. The proof is concluded by appealing to Theorem 2.15 of [141] that shows that stochastic domination implies an ordering of the means. □

Next we upper bound the excess that appears in the domination by the branching process:

Lemma 8. $x_i - c_i = O(\frac{1}{n})$.

Proof. As in Lemma 7, $B_{\text{Ber}}(i)$ denotes the total progeny of a Bernoulli branching process whose set of parameters corresponds to $G(n, \kappa)$. Then we may decompose the difference as

$$x_i - c_i = x_i - \mathbb{E}\left[B_{\text{Ber}}(i)\right] + \mathbb{E}\left[B_{\text{Ber}}(i)\right] - c_i.$$

Denote the set of edges in the connected component C_a as $E(C_a)$ and the set of edges containing a vertex v as $E(v)$. We call $|\mathcal{S}|$ the *surplus*, which is the number of edges to be deleted from $E(C_a)$ such that the graph C_a becomes a tree. Then, we have $\mathbb{E}\left[B_{Ber}(i)\right] - c_i \leq \mathbb{E}\left[|\mathcal{S}|\right]$. The expectation of the surplus may be written as

$$\begin{aligned}\mathbb{E}\left[|\mathcal{S}|\right] = \mathbb{E}\left[\sum_{e \in E(C_a)} \mathbb{I}\{e \in \mathcal{S}\}\right] &= \sum_{k=1}^{\infty} \mathbb{P}\left[|C_a| = k\right] \sum_{e \in E(C_a)} \mathbb{E}\left[\mathbb{I}\{e \in \mathcal{S}\} \mid |C_a| = k\right] \\ &= \frac{1}{2} \sum_{v \in C_a} \sum_{e \in E(v)} \mathbb{E}\left[\mathbb{I}\{e \in \mathcal{S}\} \mid |C_a| = k\right].\end{aligned}$$

Define $A_{\max} = \max_{i,j} A_{i,j}$ as the maximal element of the matrix A. Then for an arbitrary vertex, the probability of an edge $e \in E(v)$ being in the surplus can be upper bounded as

$$\sum_{e \in E(v)} \mathbb{E}\left[\mathbb{I}\{e \in \mathcal{S}\} \mid |C_a| = k\right] \leq \frac{A_{\max} k}{n}.$$

Then we may upper bound the sum as

$$\frac{1}{2} \sum_{v \in C_a} \sum_{e \in E(v)} \mathbb{E}\left[\mathbb{I}\{e \in \mathcal{S}\} \mid |C_a| = k\right] \leq \frac{A_{\max} k^2}{n}.$$

Using our expression for $\mathbb{E}\left[|\mathcal{S}|\right]$, we get

$$\mathbb{E}\left[|\mathcal{S}|\right] \leq \sum_{k=1}^{\infty} \mathbb{P}\left[|C_a| = k\right] \frac{A_{\max} k^2}{n} = \frac{A_{\max} \mathbb{E}|C_a|^2}{n}.$$

Now we notice that, by Le Cam's theorem, the total variation distance between the sum of independent Bernoulli random variables with parameters $(A_{i,1}/n, \dots, A_{i,n/n})$ and the Poisson distribution Poi$(\sum_{j=1}^{n} A_{i,j}/n)$ is at most $2(\sum_{j=1}^{n} A_{i,j}^2)/n$. Using this fact and that the moments of the total progeny of a subcritical branching process do not scale with n (cf. Theorem 1 of 68), we have $x_i - \mathbb{E}\left[B_{Ber}(i)\right] = O\left(\frac{1}{n}\right)$, thus proving the lemma. □

2.6.2 Supercritical case

Lemma 9. *Suppose that*

1. *G is generated from a supercritical $G(n, \alpha, K)$ satisfying Assumptions 1 and 2, or*

2. *G is generated from a supercritical $G(n, w)$.*

Then, for any node i satisfying $\mu_i < \mu^*$, we have $c^* - c_i \le c^*\left(\mu^* - \mu_i\right) + o_n(n)$.

The proof of Lemma 9 follows from the following lemmas for the stochastic block model and the Chung–Lu model and from the following relation between c_i and ρ_i:

$$c_i = \rho_i \mathbb{E}\left[|C_1|\right] + o_n(n) \,,$$

see Bollobás et al. [28, Chapter 9].

Lemma 10 (Coordinate order preserving in the stochastic block model.)**.** *Assume the conditions of Lemma 9 and let $l_* =$ arg max$_l$ b_l. Let $a \in \mathbb{R}^S$ be any vector such that $a_l \in [0, a_{l_*}]$ for all l. Then $\left(\Phi_M(a)\right)_{l_*} \ge \left(\Phi_M(a)\right)_l$.*

Proof. Let us fix two arbitrary indices l and l'. By the definition of Φ_M, we have

$$\left(\Phi_M(a)\right)_l = 1 - e^{-((\sum_{m \neq l} \alpha_m a_m)k + \alpha_l K_{l,l} a_l)} \,,$$
$$\left(\Phi_M(a)\right)_{l'} = 1 - e^{-((\sum_{m \neq l'} \alpha_m a_m)k + \alpha_{l'} K_{l',l'} a_{l'})} \,.$$

Notice that if l and l' satisfy

$$\left(\sum_{m \neq l} \alpha_m a_m\right) k + \alpha_l K_{l,l} a_l \ge \left(\sum_{m \neq l'} \alpha_m a_m\right) k + \alpha_{l'} K_{l',l'} a_{l'},$$

we have $\left(\Phi_M(a)\right)_l \ge \left(\Phi_M(a)\right)_{l'}$. Now, using the facts that

- $\sum_{m\neq l}\alpha_m a_m - \sum_{m\neq l'}\alpha_m a_m = \alpha_{l'}a_{l'} - \alpha_l a_l$,
- $\alpha_l K_{l,l} \geq \alpha_l k$,
- $\alpha_l K_{l,l} + \alpha_{l'}k \geq \alpha_{l'}k_{l',l'} + \alpha_l k$ and
- $a_l - a_{l'} \geq 0$,

we can verify that

$$\begin{aligned}&\alpha_l K_{l,l}a_l + \alpha_{l'}ka_{l'} - \alpha_l ka_l - \alpha_{l'}K_{l',l'}a_{l'}\\ &\quad = \quad (\alpha_l K_{l,l} + \alpha_{l'}k)a_{l'} + (a_l - a_{l'})\alpha_l K_{l,l} - (\alpha_{l'}K_{l',l'} + \alpha_i k)a_{l'} - (a_l - a_{l'})\alpha_l k \geq 0,\end{aligned}$$

thus proving the lemma. □

Lemma 11 (Order of coordinates of eigenvector in the SBM)**.** *Let a be the eigenvector corresponding to the largest eigenvalue λ of the matrix $M = K diag(\alpha)$. Then if $l_* =$ arg max$_l$ b_l, we have $a_{l_*} \geq a_l$ for $l \neq l_*$.*

Proof. If a is an eigenvector of M, then for coordinates l, l':

$$\begin{cases}\left(\sum_{m\neq l}\alpha_m a_m\right)k + \alpha_l k_{l,l}a_l = \lambda a_l,\\ \left(\sum_{m\neq m'}\alpha_m a_m\right)k + \alpha_l K_{l',l'}a_{l'} = \lambda a_{l'}\end{cases}$$

By the Perron–Frobenius theorem and our conditions on matrix M, λ is a real number larger than one. Denote $C = k\sum_{m\neq l, m\neq l'}\alpha_m a_m$, $x = a_l$, $y = a_{l'}$, $a = \alpha_l K_{l,l}$, $b = \alpha_{l'}k$, $c = \alpha_l k$, $d = \alpha_{l'}k_{l',l'}$. Then,

$$\begin{cases}C + ax + by = \lambda x,\\ C + cx + dy = \lambda y\end{cases} \tag{2.12}$$

Let $r = 1 + \epsilon$ be such that $y = rx = (1+\epsilon)x$. Then

$$\begin{cases}\frac{C}{x} + a + b + b\epsilon = \lambda,\\ \frac{C}{x} + c + d + d\epsilon = \lambda + \lambda\epsilon\end{cases}$$

and therefore
$$\frac{C}{x} + c + d + d\epsilon = \frac{C}{x} + a + b + b\epsilon + \lambda\epsilon\ .$$

Rearranging the terms and using the fact that $a + b \geq c + d$, we have

$$0 \leq (a + b) - (c + d) = (d - b - \lambda)\epsilon\ .$$

Since $K_{l,l} \geq k$, we have $\alpha_l k_{l,l} \geq \alpha_l k$ and $a \geq c$.

We consider two cases separately: First, if $b \geq d$, we have $d - b - \lambda < 0$, which implies $\epsilon < 0$ and $y < x$, therefore proving $a_l > a_{l'}$ for this case. In the case when $b < d$, we have $a + b \geq c + d$ and $\frac{d-b}{a-c} \leq 1$. Subtracting the two equalities of the linear system 2.12, we get

$$\lambda(1 - r) = (a - c)\left(1 - \frac{d - b}{a - c} r\right)\ .$$

Now, since $\frac{d-b}{a-c} \leq 1$, we have $\lambda \geq a - c$, which implies $\lambda \geq d - b$ and $d - b - \lambda \leq 0$, thus leading to $\epsilon \leq 0$ and $y \leq x$, therefore proving $a_l \geq a_{l'}$ for this case. □

Lemma 12 (Order of coordinates of eigenvector in the Chung–Lu model)**.** *Let a be the eigenvector corresponding to the largest eigenvalue λ of the matrix A. Then if $i_* = \arg\max_m b_m$, we have $a_{i_*} \geq a_j$ for $j \neq i_*$.*

Proof. It is easy to see that the only eigenvector of A corresponding to a non-zero eigenvalue is $a = w$ with $\lambda_{max} = w^\intercal w/n$:

$$\frac{1}{n}Aw = \frac{1}{n} \cdot (ww^\intercal)w = \frac{w^\intercal w}{n} \cdot w.$$

The proof is concluded by observing that the maximum coordinate of the vector b corresponds to the maximum coordinate of w, due to the equality

$$b_i = \frac{1}{n} \cdot w_i \sum_{j=1}^{n} w_j.$$

□

Lemma 13 (Coordinate order preserving in the Chung–Lu model)**.** *Assume the conditions of Lemma 9 and let $i_* = \arg\max_i b_i$. Let $a = (a_1, \ldots, a_n)$ be such that $a_j \in [0, a_{i_*}]$ for all j. Then $(\Phi_A(a))_{i_*} \geq (\Phi_A(a))_j$.*

Proof. Let us fix two arbitrary indices i and i'. By the definition of Φ_A, we have

$$(\Phi_A(a))_i = 1 - e^{-w_i(\sum_{j=1}^n w_j a_j)} .$$

Then, using the fact that $w = a$, we have $(\Phi_A(a))_{i_*} \geq (\Phi_A(a))_j$, thus proving the lemma. □

We finally study the maximal fixed point of the operator Φ_A, keeping in mind this fixed point is exactly the survival-probability vector ρ of the multi-type Galton–Watson branching process Bollobás et al. [28]. By Lemma 5.9 of Bollobás et al. [28], this is the unique fixed point satisfying $\rho_i > 0$ for all i. The following lemma shows that ρ_i takes its maximum at $i_* = \arg\max_i b_i$, concluding the proof of Lemma 9.

Lemma 14 (Fixed point coordinate domination). *Let ρ be the unique non-zero fixed point of Φ_A, and let $i_* =$ arg max$_i$ b_i. Then, $\rho_{i_*} \geq \rho_j$ and $\rho_{i_*} - \rho_j \leq \rho^* \left(b_{i_*} - b_j\right)$ holds for all $j \neq i_*$.*

Proof. Letting a be the eigenvector of A that corresponds to the largest eigenvalue λ, Lemma 12 and 11 guarantee $a_{i_*} \geq a_j$ for $j \neq i^*$. Let $\epsilon > 0$ be such that $\epsilon \leq \frac{1-1/\lambda}{a^*}$, where $a^* = \max_{i=1,\dots,S} a_i$. Then by Lemma 5.13 of Bollobás et al. [28], $\Phi_M(\epsilon a) \geq \epsilon a$ holds elementwise for the two vectors.

Since the coordinates of the vector ϵa are positive, we can appeal to Lemma 5.12 of Bollobás et al. [28] to show that iterative application of Φ_A converges to the fixed point ρ: letting Φ_A^m be the operator obtained by iterative application of Φ_A for m times, we have $\lim_{m\to\infty} \Phi_A^m(\epsilon a) = \rho$, where ρ satisfies $\rho \geq \epsilon a \geq 0$ and $\Phi_A(\rho) = \rho > 0$. By Lemmas 12 and 11 we have $\rho_{i_*} \geq \rho_j$, for $i_* \neq j$ for both the SBM and the Chung–Lu models, proving the first statement.

The second statement can now be proven directly as

$$\begin{aligned}
\rho_{i_*} - \rho_i = e^{-(\frac{1}{n}A\rho)_j} - e^{-(\frac{1}{n}A\rho)_{i*}} &= e^{-\frac{1}{n}\sum_j^n A_{i*j}\rho_j} - e^{-\frac{1}{n}\sum_j^n A_{ij}\rho_j} \\
&= \quad e^{-\frac{1}{n}\sum_j^n A_{i*j}\rho_j}\big(1 - e^{-\frac{1}{n}\sum_j^n A_{ij}\rho_j - A_{i*j}\rho_j}\big) \leq e^{-\frac{1}{n}\sum_j^n A_{i*j}\rho_j} \left(\frac{1}{n}\sum_j^n (A_{i_*j} - A_{ij})\rho_{i_*}\right) \\
&\leq \quad \rho^*(b_{i_*} - b_i),
\end{aligned}$$

where the first inequality uses the relation $1 - e^{-z} \leq z$ that holds for all $z \in \mathbb{R}$, and the last step uses the fact that $A\rho$ has positive elements. □

2.6.3 Proofs of Theorems 2.3.1, 2.3.2 and 2.3.3 .

Having established that, in order to minimize regret in our setting, it is sufficient to design an algorithm that quickly identifies the nodes with the highest degree. It remains to show that our algorithms indeed achieve this goal. We do this below by providing a bound on the expected number of times $\mathbb{E}\left[N_{T,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_{\{A_t=i\}}\right]$ that the algorithm picks a suboptimal node i such that $c_i < c^*$, and then using this guarantee to bound the regret.

Without loss of generality, we assume that $V_0 = \{1, 2, \ldots, |V_0|\}$. The key to our regret bounds is the following guarantee on the number of suboptimal actions taken by d-UCB(V_0).

Theorem 2.6.1 (Number of suboptimal node plays in d-UCB). *Define* $\eta_i = \left(\max_{j\in V_0} \mu_j - \mu_i\right)/3$*. The number of times that any node* $i \in \{i : \mu_i < \max_{j\in V_0} \mu_j\}$ *is chosen by d-UCB*(V_0) *satisfies*

$$\mathbb{E}N_{T,i} \leq \frac{\mu^* \left(2 + 6 \log T\right)}{\eta_i^2} + 3\ . \tag{2.13}$$

The proof is largely based on the analysis of the kl-UCB algorithm due to Cappé et al. [37], with some additional tools borrowed from Ménard and Garivier [106], crucially using that the degree distribution of each node is stochastically dominated by an appropriately chosen Poisson distribution. Specifically, letting Z_i be a Poisson random variable with mean $\mathbb{E}\left[X_{t,i}\right]$, we have $\mathbb{E}\left[e^{sX_{t,i}}\right] \leq \mathbb{E}\left[e^{sZ_i}\right]$ for all s. It turns out that this property is sufficient for the kl-UCB analysis to go through in our case, which is an observation that may be of independent interest.

Before delving into the proof, we introduce some useful notation. We start by defining $Y_{i,1}, \ldots, Y_{i,n}$ as independent Bernoulli random variables with respective parameters $\mathbb{B} = (A_{i,1}/n,\ A_{i,2}/n, \ldots, A_{i,n}/n)$, and noticing that the degree $X_{t,i}$ can be written as a sum $X_i = \sum_{j\neq i} Y_{i,j}$. The following lemma, used several times in our proofs, relates this quantity to a Poisson distribution with the same mean.

Lemma 15. *Let* $i \in [S]$ *and let* $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n}$ *be independent Bernoulli random variables with respective parameters* $p_{i,1}, p_{i,2}, \ldots, p_{i,n}$*, and let* Z_i *be a Poisson random variable with parameter* $\mu_i = \sum_{j\neq i} p_{i,j}$*. Defining* $X_i = \sum_{j\neq i} Y_{i,j}$*, we have* $\mathbb{E}\left[e^{sX_i}\right] \leq \mathbb{E}\left[e^{sZ_i}\right]$ *for all* $s \in \mathbb{R}$*.*

Proof. Fix an arbitrary $s \in \mathbb{R}$ and $i \in [n]$. By direct calculations, we obtain

$$\mathbb{E}e^{sX_i} = \prod_{j=1}^{n} \left(\mathbb{E}e^{sY_{i,j}}\right) \leq \prod_{j=1}^{n} (1 + p_{i,j}(e^s - 1)) \leq \prod_{j=1}^{n} \exp\left(p_{i,j} \cdot (e^s - 1)\right),$$

where the last step follows from the elementary inequality $1 + x \leq e^x$ that holds for all $x \in \mathbb{R}$. The proof is concluded by observing that $\mathbb{E}e^{sZ_i} = \exp\left(\mu\left(e^s - 1\right)\right)$ and using the definition of μ. □

For simplicity, we also introduce the notation $\psi_{\mathbb{B}}(s) = \log \mathbb{E}\left[e^{sX}\right]$ and $\phi_\lambda(s) = \log \mathbb{E}e^{sZ_i} = \lambda(e^s - 1)$. The proof below repeatedly refers to the Fenchel conjugate of ϕ_λ defined as

$$\phi^*_\lambda(z) = \sup_{s \in \mathbb{R}} \{sz - \phi(s)\} = z \log\left(\frac{z}{\lambda}\right) + \lambda - z$$

for all $z \in \mathbb{R}$. Finally, we define $d(\mu, \mu') = \mu' - \mu + \mu \log\left(\frac{\mu}{\mu'}\right)$ for all $\mu, \mu' > 0$, noting that $\phi^*_\lambda(z) = d(z, \lambda)$.

Proof of Theorem 2.6.1. The statement is proven in four steps. Within this proof, we refer to nodes as *arms* and use K to denote the size of V_0. We use the notation $f(t) = 3 \log t$.

Step 1. We begin by rewriting the expected number of draws $\mathbb{E}[N_i]$ for any suboptimal arm i as

$$\mathbb{E}N_i = \mathbb{E}\left[\sum_{t=K}^{T-1} \mathbb{I}\{A_{t+1} = i\}\right] = \sum_{t=K}^{T-1} \mathbb{P}\{A_{t+1} = i\}.$$

By definition of our algorithm, at rounds $t > K$, we have $A_{t+1} = i$ only if $U_i > U_{i^*i}$. This leads to the decomposition:

$$\begin{aligned}\{A_{t+1} = a\} &\subseteq \{\mu^* \geq U_{i^*}(t)\} \cup \{\mu^* < U_{i^*}(t) \text{ and } A_{t+1} = a\} \\ &\subseteq \{\mu^* \geq U_{i^*}(t)\} \cup \{\mu^* < U_i(t) \text{ and } A_{t+1} = a\}\end{aligned}$$

Steps 2 and 3 are devoted to bounding the probability of the two events above.

Step 2. Here we aim to upper bound

$$\sum_{t=K}^{T-1} \mathbb{P}\left[\mu^* \geq U_{i^*}(t)\right]. \tag{2.14}$$

Note, that $\{U_{i^*}(t) \leq \mu^*\} = \{\hat{\mu}_{i^*}(t) \leq U_{i^*}(t) \leq \mu^*\}$. Since $d(\mu,\mu') = \mu' - \mu + \mu \log(\frac{\mu}{\mu'})$ is non-decreasing in its second argument on $[\mu, +\infty)$, and by definition of $U_{i^*} = \sup\{\mu : d(\hat{\mu}_{i^*}(t), \mu) \leq \frac{f(t)}{N_{i^*}(t)}\}$ we have

$$\{\mu^* \geq U_{i^*}(t)\} \subseteq \left\{\hat{\mu}_{i^*}(t) \leq U_{i^*}(t) \leq \mu^* \text{ and } d(\hat{\mu}_{i^*}(t), \mu^*) \geq \frac{f(t)}{N_{i^*}(t)}\right\},$$

Taking a union bound over the possible values of $N_{i^*}(t)$ yields

$$\{\mu^* \geq U_{i^*}(t)\} \subseteq \bigcup_{n=1}^{t-K+1} \left\{\mu^* \geq \hat{\mu}_{i^*,n} \text{ and } d(\hat{\mu}_{i^*,n}, \mu^*) \geq \frac{f(t)}{n}\right\} = \bigcup_{n=1}^{t-K+1} D_n(t),$$

where the event $D_n(t)$ is defined through the last step. Since $d(\mu, \mu^*)$ is decreasing and continuous in its first argument on $[0, \mu^*)$, either $d(\hat{\mu}_{i^*,n}, \mu^*) < \frac{f(t)}{n}$ on this interval and $D_n(t)$ is the empty set, or there exists a unique $z_n \in [0, \mu^*)$ such that $d(z_n, \mu^*) = \frac{f(t)}{n}$. Thus, we have

$$\bigcup_{n=1}^{t-K+1} D_n(t) \subseteq \bigcup_{n=1}^{t-K+1} \{\hat{\mu}_{i^*,n} \leq z_n\}.$$

For $\lambda < 0$, let us define $\psi(\lambda)$ as the cumulant-generating function of the sum of binomials with parameters $\mathbb{B}$, and let $\phi(\lambda)$ be the cumulant-generating function of a Poisson random variable with parameter μ^*. With this notation, we have for *any* $\lambda < 0$ that

$$\begin{aligned}\mathbb{P}\left[\hat{\mu}_{i^*,n} \leq z_n\right] &= \mathbb{P}\left[\exp(\lambda \hat{\mu}_{i^*,n}) \geq \exp(\lambda z_n)\right] \\ &= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^{n} X_{i^*,i} - n\psi(\lambda)\right) \geq \exp(n\lambda z_n - n\psi(\lambda))\right] \\ &\leq \left(\frac{\mathbb{E}e^{\lambda X_{i^*,1}}}{e^{\psi(\lambda)}}\right)^n e^{-n(\lambda z_n - \psi(\lambda))} \leq e^{-n(\lambda z_n - \psi(\lambda))},\end{aligned}$$

where the last step uses the definition of $\psi(\lambda)$. Now fixing $\lambda^* = \arg\max_\lambda\{\lambda z_n - \phi(\lambda)\} = \log(z_n/\mu^*) < 0$, we get by Lemma 15 that

$$e^{-n(\lambda^* z_n - \psi(\lambda^*))} \leq e^{-n(\lambda^* z_n - \phi(\lambda^*))} = e^{-n\phi^*_{\mu^*}(z_n)} = e^{-nd(z_n,\mu^*)} .$$

In view of the definition of z_n and $f(t)$, this gives the bound

$$e^{-nd(z_n,\mu^*)} = e^{-f(t)} = \frac{1}{t^3},$$

which leads to

$$\sum_{t=K}^{T-1} \mathbb{P}\left[\mu^* \geq U_{i^*}(t)\right] \leq \sum_{t=K}^{T-1} \sum_{n=1}^{t-K+1} \frac{1}{t^3} < 2,$$

thus concluding this step.

Step 3. In this step, we borrow some ideas by Ménard and Garivier [106, Proof of Theorem 2, step 2] to upper bound the sum

$$B = \sum_{t=K}^{T-1} \mathbb{P}\left[\mu^* < U_i(t) \text{ and } A_{t+1} = i\right]. \tag{2.15}$$

Writing $\eta = \eta_i = \{\mu^* - \mu_i\}/3$ for ease of notation, we have

$$\begin{aligned}\{\mu^* < U_i(t) \text{ and } A_{t+1} = i\} &\subseteq \{\mu^* - \eta < U_i(t) \text{ and } A_{t+1} = i\} \\ &\subseteq \{d(\hat{\mu}_i(t), \mu^* - \eta) \leq f(t)/N_i(t) \text{ and } A_{t+1} = i\}.\end{aligned}$$

Thus, we have

$$\begin{aligned} B &\leq \sum_{t=K}^{T-1} \mathbb{P}\left[d(\hat{\mu}_i(t), \mu^* - \eta) \leq f(t)/N_i(t) \text{ and } A_{t+1} = i\right] \\ &\leq \sum_{n=1}^{T} \mathbb{P}\left[d(\hat{\mu}_{i,n}, \mu^* - \eta) \leq f(T)/n\right] \end{aligned}$$

Defining the integer $n(\eta)$ as

$$n(\eta) = \left\lceil \frac{f(T)}{d(\mu_i + \eta, \mu^* - \eta)} \right\rceil,$$

we have $f(T)/n \leq d(\mu_i + \eta, \mu^* - \eta)$ for all $n \geq n(\eta)$. Thus, we may further upper bound B as

$$\begin{aligned} B &\leq n(\eta) - 1 + \sum_{n=n(\eta)}^{T} \mathbb{P}\left[d(\hat{\mu}_{i,n}, \mu^* - \eta) \leq f(T)/n\right] \\ &\leq \frac{f(T)}{d(\mu_i + \eta, \mu^* - \eta)} + \sum_{n=n(\eta)}^{T} \mathbb{P}\left[d(\hat{\mu}_{i,n}, \mu^* - \eta) \leq d(\mu_i + \eta, \mu^* - \eta)\right]. \end{aligned}$$

By definition of η, we have

$$\{\hat{\mu}_{i,n}, \mu^* - \eta) \leq d(\mu_i + \eta, \mu^* - \eta)\} \subseteq \{\hat{\mu}_{i,n} \geq \mu_i + \eta\},$$

which implies

$$\sum_{n=n(\eta)}^{T} \mathbb{P}\left[d(\hat{\mu}_{i,n}, \mu^* - \eta) \leq d(\mu_i + \eta, \mu^* - \eta)\right] \leq \sum_{n=n(\eta)}^{T} \mathbb{P}\left[\hat{\mu}_{i,n} \geq \mu_i + \eta\right].$$

By an argument analogous to the one used in the previous step, we get for a well-chosen λ that

$$\begin{aligned} \sum_{n=n(\eta)}^{T} \mathbb{P}\left[\hat{\mu}_{i,n} \geq \mu_i + \eta\right] &\leq \mathbb{P}\left[\exp(\lambda\hat{\mu}_{i,n}) \geq \exp(\lambda(\mu_i + \eta))\right] \\ &= \sum_{n=n(\eta)}^{T} \mathbb{P}\left[\exp(\lambda\sum_{j=1}^{n} X_{i,j} - n\psi(\lambda)) \geq \exp(n\lambda(\mu_i + \eta) - n\psi(\lambda))\right] \\ &\leq \sum_{n=n(\eta)}^{T} \left(\frac{\mathbb{E}\left[e^{\lambda X_{i,j}}\right]}{e^{\psi(\lambda)}}\right)^{n} e^{-n(\lambda(\mu_i+\eta)-\psi(\lambda))} \\ &\leq \sum_{n=n(\eta)}^{T} e^{-n(\lambda(\mu_i+\eta)-\phi(\lambda))} = \sum_{n=n(\eta)}^{T} e^{-nd(\mu_i+\eta,\mu_i)} \\ &\leq \sum_{n=n(\eta)}^{\infty} e^{-nd(\mu_i+\eta,\mu_i)} \leq \frac{1}{e^{d(\mu_i+\eta,\mu_i)} - 1} \leq \frac{1}{d(\mu_i + \eta, \mu_i)}, \end{aligned}$$

where the last step uses the elementary inequality $1 + x \leq e^x$ that holds for all $x \in \mathbb{R}$.

Step 4. Putting together the results from the first three steps, we get

$$\mathbb{E}N_i \leq 3 + \frac{1}{d(\mu_i + \eta, \mu_i)} + \frac{3 \log T}{d(\mu_i + \eta, \mu^* - \eta)}.$$

We conclude by taking a second-order Taylor-expansion of $d(\mu_i + \eta, \mu_i)$ in η to obtain for some $\eta' \in [0, \eta]$ that

$$d(\mu_i + \eta, \mu_i) = \frac{\eta^2}{2(\mu_i + \eta')} \geq \frac{\eta^2}{2(\mu_i + \eta)}.$$

Taking into account the definition of η, we get

$$\frac{1}{d(\mu_i + \eta, \mu_i)} \leq \frac{2\mu^*}{\eta^2}.$$

An identical argument can be used to bound $(d(\mu_i + \eta, \mu^* - \eta))^{-1} \leq 2\mu^*/\eta^2$. □

The remainder of the section uses Theorem 2.6.1 to prove Theorem 2.3.1. The proof of Theorem 2.3.2 follows from similar ideas and some additional technical arguments.

Proof of Theorem 2.3.1. First, by (2.6),

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i} \mathbb{E}\left[N_{T,i}\right] \middle| \mathcal{E}\right].$$

Now, observing that $\delta_{\alpha,i} \leq 3\eta_i$ holds under event $\mathcal{E}$, we appeal to Theorem 2.6.1 to obtain

$$R_T^\alpha \leq \Delta_{\alpha,\max} + \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i} \left(\frac{\mu^* \left(18 + 27 \log T\right)}{\delta_{i,\alpha}^2} + 3\right) \middle| \mathcal{E}\right], \qquad (2.16)$$

thus proving the first statement.

Next, we turn to proving the second statement regarding worst-case guarantees. To do this, we appeal to Propositions 4 and 9 that respectively show $\Delta_i \leq 2c^*\delta_i + O(1/n)$ and $\Delta_i \leq c^*\delta_i + o(n)$ for the sub- and supercritical settings, and we use our assumption that n is large enough so that we have $\Delta_i \leq 3c^*\delta_i$ in both settings. Specifically, we observe that $\delta_i = \Theta_n(1)$ by our sparsity assumption and c^* is $\Theta_n(1)$ in the subcritical and $\Theta_n(n)$ supercritical settings, so, for large enough n, the superfluous $O(1/n)$ and $o(n)$ terms can be respectively bounded by $c^*\delta_i$. To proceed, let us fix an arbitrary $\varepsilon > 0$ and split the set V_0 into two subsets: $U(\varepsilon) = \{i \in V_0 : \delta_{\alpha,i} \leq \varepsilon\}$ and $W(\varepsilon) = V_0 \setminus U(\varepsilon)$. Then, under event $\mathcal{E}$, we have

$$\begin{aligned}
\sum_{i\in V_0} \Delta_{\alpha,i}\mathbb{E}\left[N_{T,i}\right] &= \sum_{i\in U(\varepsilon)} \Delta_{\alpha,i}\mathbb{E}\left[N_{T,i}\right] + \sum_{i\in W(\varepsilon)} \Delta_{\alpha,i}\mathbb{E}\left[N_{T,i}\right] \\
&\leq 3c^*\varepsilon \sum_{i\in U(\varepsilon)} \mathbb{E}\left[N_{T,i}\right] + 3c^* \sum_{i\in W(\varepsilon)} \delta_{\alpha,i}\left(\frac{\mu^*\left(18+27\log T\right)}{\delta_{\alpha,i}^2}\right) \\
&\quad + 3|W(\varepsilon)|\Delta_{\alpha,\max} \qquad \text{(by Theorem 2.6.1)} \\
&\leq 3c^*\varepsilon T + 3c^* \sum_{i\in W(\varepsilon)} \frac{\mu^*\left(18+27\log T\right)}{\delta_{\alpha,i}} + 3|V_0|\Delta_{\alpha,\max} \\
&\leq 3c^*\left(\varepsilon T + |V_0|\frac{\mu^*\left(18+27\log T\right)}{\varepsilon}\right) + 3|V_0|\Delta_{\alpha,\max} \\
&\leq 6c^*\sqrt{T|V_0|\mu^*\left(18+27\log T\right)} + 3|V_0|\Delta_{\alpha,\max},
\end{aligned}$$

where the last step uses the choice $\varepsilon = \sqrt{|V_0|\mu^*\left(18+27\log T\right)/T}$. Plugging in the choice of $|V_0|$ concludes the proof. □

Proof of Theorem 2.3.2. We start by assuming that $\alpha < 1/2$. Also notice that for a uniformly sampled set of nodes U, the probability of U not containing a vertex from V_α^* is bounded as

$$\mathbb{P}\left[U \cap V_\alpha^* = \emptyset\right] \leq (1-\alpha)^{|U|}.$$

By the definition of V_k, this gives that the probability of not having sampled a node from V_α^* in period k of the algorithm is bounded as

$$\mathbb{P}\left[V_k \cap V_\alpha^* = \emptyset\right] \leq (1-\alpha)^{|V_k|} \leq \beta^{-k}.$$

For each period k, the expected regret can bounded as the weighted sum of two terms: the expected regret of d-UCB (V_k) in period k whenever $V_k \cap V_\alpha^*$ is not empty, and the trivial bound $\Delta_{\alpha,\max}\beta^k$ in the complementary case. Using the above bound on the probability of this event and appealing to Theorem 2.6.1 to bound the regret of d-UCB (V_k), we can bound the expected regret as

$$\begin{aligned}
\mathbb{E}\left[R_T^\alpha\right] &\leq \sum_{k=1}^{k_{\max}} \left(\beta^k \frac{1}{\beta^k}\Delta_{\alpha,\max} + \sum_{i\in V_k} \Delta_{\alpha,i}\left(\frac{\mu^*\left(2+3\log\beta^k\right)}{\delta_{\alpha,i}^2}+3\right)\right) \\
&\leq k_{\max}\Delta_{\alpha,\max} + \sum_{k=1}^{k_{\max}}\left(\sum_{i\in V_k}\Delta_{\alpha,i}\left(\frac{\mu^*\left(2+3k\log\beta\right)}{\delta_{\alpha,i}^2}+3\right)\right) \\
&\leq k_{\max}\Delta_{\alpha,\max} + \sum_{i\in\overline{V}}\Delta_{\alpha,i}\left(\left(3+\frac{2\mu^*}{\delta_{\alpha,i}^2}\right)(k_{\max}+1)+\frac{3\log\beta(k_{\max}+1)^2}{2\delta_{\alpha,i}^2}\right) .
\end{aligned}$$

The proof of the first statement is concluded by upper bounding the number of restarts up to time T as $k_{\max} \leq \frac{\log T}{\log\beta}$.

The second statement is proven by an argument analogous to the one used in the proof of Theorem 2.3.1, and straightforward calculations. □

Proof of Theorem 2.3.3. For a node i, such that s_i contains l_i ones, the expected degree is

$$\mu_i = (\zeta+\beta)^{l_i}(\beta+\gamma)^{k-l_i}.$$

Since $\zeta > \gamma > \beta$, we get that $\mu_i > \mu_j$ if $l_i > l_j$. By symmetry of the nodes in the Kronecker graph, if two nodes i and j are such that $l_i = l_j$, then $c_i = c_j$. This implies that for any node i, c_i is a function of l_i. Then we may choose nodes i and j such that $s_i \geq s_j$ coordinate-wise. Then, using the condition $\zeta > \gamma > \beta$, it is straightforward to see that for any vertex k, the probability of the edge (i,k) is greater than that of edge (j,k). This implies that the connected component is a monotone function of the degree.

Theorem 9.10 in [63] shows that for a graph $G_{n,P^{[k]}}$there exists $b(P)$ such that a subgraph of $G_{n,P^{[k]}}$induced by the vertices $i \in H$ of weight $l_i \geq k/2$ is connected with probability at least $1-n^{-b(P)}$. We denote by $\mathcal{H}$ the event that the subgraph of $G_{n,P^{[k]}}$induced by the vertices of H of weight $l \geq k/2$ is connected. This implies that under event $\mathcal{H}$, $|C_i| = |\max_j C_j|$ for all $i \in H$ and

$|C_i| \leq |\max_j C_j|$ for all $i \notin H$. Then we get

$$\begin{aligned}
c_\alpha^* - c_i &= \mathbb{E}\left[\max_j |C_j| \middle| \mathcal{H}\right] \mathbb{P}\left[\mathcal{H}\right] + \mathbb{E}\left[|C_\alpha^*| \middle| \mathcal{H}^c\right] \mathbb{P}\left[\mathcal{H}^c\right] \\
&\quad - \mathbb{E}\left[\max_j |C_j| \middle| \mathcal{H}\right] \mathbb{P}\left[\mathcal{H}\right] - \mathbb{E}\left[|C_i| \middle| \mathcal{H}^c\right] \mathbb{P}\left[\mathcal{H}^c\right] \\
&\leq \mathbb{E}\left[|C_*| \middle| \mathcal{H}^c\right] \mathbb{P}\left[\mathcal{H}^c\right] \leq n^{1-b(P)}.
\end{aligned}$$

For all $i \in V_0 \backslash H$, $\delta_{\alpha,i} \geq ((\zeta+\beta)(\beta+\gamma))^{k/2} - (\zeta+\beta)^{k/2-1}(\beta+\gamma)^{k/2+1} = ((\zeta+\beta)(\beta+\gamma))^{k/2}\left(1 - \frac{\beta+\gamma}{\zeta+\beta}\right)$. Since we consider the regime, where $(\zeta+\beta)(\beta+\gamma) > 1$, we get that $\delta_{\alpha,i} > \left(1 - \frac{\beta+\gamma}{\zeta+\beta}\right)$. For all $i, j \in H$, $\delta_{\alpha,i} = (\zeta+\beta)^{l_i}(\beta+\gamma)^{k-l_i} \geq (\zeta+\beta)^{k/2-1}(\beta+\gamma)^{k/2}(\zeta-\gamma) \geq (\zeta-\gamma)/(\zeta+\beta)$. In the same way as we analysed the regret of the stochastic block model and Chung–Lu model, we can write

$$R_T^\alpha \leq nT\mathbb{P}\left[\mathcal{E}^c\right] + \mathbb{E}\left[\sum_{i \in V_0} \Delta_{\alpha,i} \mathbb{E}\left[N_{i,T}\right] \middle| \mathcal{E}\right].$$

Applying Theorem 2.6.1, we get

$$\begin{aligned}
\frac{R_T^\alpha}{n} &\leq \mathbb{E}\left[\sum_{i \in V_0 \backslash H} \frac{\Delta_{\alpha,i}}{n}\left(\frac{\mu^*(2 + 6\log T)}{\left(1 - \frac{\beta+\gamma}{\zeta+\beta}\right)^2} + 3\right) \middle| \mathcal{E}\right] \\
&\quad + n^{-b(P)}\left\lceil \frac{\log(nT)}{\log(2)} \right\rceil \left(\frac{\mu^*(2 + 6\log T)(\zeta+\beta)^2}{(\zeta-\gamma)^2} + 3\right) + 1\,.
\end{aligned}$$

Applying $|V_0 \setminus H| \leq |V_0|$ and $\Delta_{\alpha,i} \leq n$, we get the final bound on the regret. □

Chapter 3

Adversarial contextual bandits

3.1 Introduction

The contextual bandit problem is one of the most important sequential decision-making problems studied in the machine learning literature. Due to its ability to account for contextual information, the applicability of contextual bandit algorithms is far superior to that of standard multi-armed bandit methods: the framework of contextual bandits can be used to address a broad range of important and challenging real-world decision-making problems such as sequential treatment allocation [137] and online advertising [97]. On the other hand, the framework is far less complex than that of general reinforcement learning, which allows for proving formal performance guarantees under relatively mild assumptions. As a result, there has been significant interest in this problem within the learning-theory community, resulting in a wide variety of algorithms with performance guarantees proven under a number of different assumptions. In this work, we fill a gap in this literature and design computationally efficient algorithms with strong performance guarantees for an adversarial version of the *linear contextual bandit* problem.

Perhaps the most well-studied variant of the contextual bandit problem is that of *stochastic linear contextual bandits* [11, 127, 48, 1, 89]. First proposed by Abe and Long [2], this version supposes that the loss of each action is a fixed linear function of the vector-valued context, up to some zero-mean noise. Most algorithms designed for this setting are based on some variation of the "optimism in the face of uncertainy" principle championed by Auer [11], Auer et al. [12], or more generally by an appropriate exploitation of the concentration-of-measure phenomenon [29]. By now, this problem setting is very well-understood in many respects: there exist several computationally efficient, easy-to-implement algorithms achieving near-optimal worst-case performance guarantees [1, 8]. These algorithms can be even adapted to more involved loss models like generalized linear models, Gaussian processes, or very large structured model classes while retaining their performance guarantees [59, 131, 34, 61]. That said, most algorithms for stochastic linear contextual bandits suffer from the limitation that they are sensitive to *model misspecification*: their performance guarantees become void as soon as the true loss functions deviate from the postulated linear model to the slightest degree. This issue has very recently attracted quite some attention due to the work of Du et al. [56], seemingly implying that learning an ε-optimal policy in a contextual bandit problem has an extremely large sample complexity when assuming that the linear model is ε-inaccurate (defined formally later). This claim was quickly countered by Van Roy and Dong [142] and Lattimore et al. [91], who both showed that learning a (somewhat worse) $\varepsilon\sqrt{d}$-optimal policy is feasible with the very same sample complexity as learning a near-optimal policy in a well-specified linear model. Yet, since algorithms that are currently known to enjoy these favorable guarantees are quite complex, there is much work left to be done in designing practical algorithms with strong guarantees under model misspecification. This is one of the main issues we address in this work.

Another limitation of virtually all known algorithms for linear contextual bandits is that they crucially rely on assuming that the loss function is *fixed during the learning procedure* [1]. This is in stark contrast with the literature on multi-armed (non-contextual) bandits, where there is a rich literature on both stochastic bandit models assuming i.i.d. rewards and adversarial bandit models making no assumptions on the sequence of loss functions—see Bubeck and Cesa-Bianchi [30] and Lattimore and Szepesvári [90] for an excellent overview of both lines of work. Our main contribution is addressing this gap by designing and analyzing algorithms that are guaranteed to work for arbitrary sequences of loss functions. While it is tempting to think that the our bandit problem can be directly addressed by a minor adaptation of algorithms developed for adversarial linear bandits, this is unfortunately not the case: all algorithms developed for such problems require a *fixed decision set*, whereas reducing the linear contextual bandit problem to a

[1] Or make other stringent assumptions about the losses, such as supposing that their total variation is bounded—see, e.g., Cheung et al. [46], Russac et al. [128], Kim and Tewari [81].

linear bandit problem requires the use *decision sets that change as a function of the contexts* [90, Section 18]. As a crucial step in our analysis, we will assume that the contexts are generated in an i.i.d. fashion and that the loss function in each round is statistically independent from the context in the same round. This assumption will allow us to relate the contextual bandit problem to a set of auxiliary bandit problems with a *fixed action sets*, and reduce the scope of the analysis to these auxiliary problems.

Our main results are the following. We consider a K-armed linear contextual bandit problem with d-dimensional contexts where in each round, a loss function mapping contexts and actions to real numbers is chosen by an adversary in a sequence of T rounds. The aim of the learner is to minimize its regret, defined as the gap between the total incurred by the learner and that of the best decision-making policy π^* fixed in full knowledge of the loss sequence. We consider two different assumptions on the loss function. Assuming that the loss functions selected by the adversary are all linear, we propose an algorithm achieving a regret bound of order $\sqrt{KdT}$, which is known to be minimax optimal even in the simpler case of i.i.d. losses (cf. 48). Second, we consider loss functions that are "nearly linear" up to an additive nonlinear function uniformly bounded by ε. For this case, we design an algorithm that guarantees regret bounded by $(Kd)^{1/3}T^{2/3} + \varepsilon\sqrt{d}T$. Notably, these latter bounds hold against *any* class of policies and the $\varepsilon\sqrt{d}T$ overhead paid for nonlinearity is optimal when K is large [91]. Both algorithms are computationally efficient, but require some prior knowledge to the distribution of the contexts.

There exist numerous other approaches for contextual bandit problems that do not rely on modeling the loss functions, but rather make use of a class of *policies* that map contexts to actions. Instead of trying to fit the loss functions, these approaches aim to identify the best policy in the class. A typical assumption in this line of work is that one has access to a computational oracle that can perform various optimization problems over the policy class (such as returning an optimal policy given a joint distribution of context-loss pairs for each action). Given access to such an oracle, there exist algorithms achieving near-optimal performance guarantees when the loss function is fixed [57, 5]. More relevant to our present work are the works of Rakhlin and Sridharan [123] and Syrgkanis et al. [134, 135] who propose efficient algorithms with guaranteed performance for adversarial loss sequences and i.i.d. contexts. Unlike the algorithms we present, these methods fail to guarantee optimal performance guarantees of order $\sqrt{T}$.

Yet another line of work considers optimizing surrogate losses, where achieving regret of order $\sqrt{T}$ is indeed possible, with the caveat that the bounds only hold for the surrogate loss [77, 26, 60].

This chapter is organized as follows. After defining some basic notation, Section 3.2 presents our problem formulation and states our assumptions. We present our algorithms and main results in Section 3.3 and provide the proofs in Section 3.4.

3.2 Preliminaries

We consider a sequential interaction scheme between a *learner* and its *environment*, where the following steps are repeated in a sequence of rounds $t = 1, 2, \dots, T$:

1. For each action $a = 1, 2, \dots, K$, the environment chooses a loss vector $\theta_{t,a} \in \mathbb{R}^d$,
2. independently of the choice of loss vectors, the environment draws the context vector $X_t \in \mathbb{R}^d$ from the context distribution $\mathcal{D}$, and reveals it to the learner,
3. based on X_t and possibly some randomness, the learner chooses action $A_t \in [K]$,
4. the learner incurs and observes loss $\ell_t(X_t, A_t) = \langle X_t, \theta_{t,A_t} \rangle$.

The goal of the learner is to pick its actions in a way that its total loss is as small as possible. Since we make no statistical assumptions about the sequence of losses (and in fact we allow them to depend on all the past interaction history), the learner cannot actually hope to incur as little loss as the best sequence of actions. A more reasonable aim is to match the performance of the *best fixed policy* that maps contexts to actions in a static way. Formally, the learner will consider the set Π of all policies $\pi : \mathbb{R}^d \to [K]$, and aim to minimize its *total expected regret* (or, simply, *regret*) defined as

$$R_T = \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T}\bigl(\ell_t(X_t, A_t) - \ell_t(X_t, \pi(X_t))\bigr)\right]$$
$$= \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_t, \theta_{t,A_t} - \theta_{t,\pi(X_t)}\right\rangle\right]$$

where the expectation is taken over the randomness injected by the learner, as well as the sequence of random contexts. For stating many of our technical results, it will be useful to define the filtration $\mathcal{F}_t = \sigma(X_s, A_s, \forall s \le t)$ and the notations $\mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot|\mathcal{F}_{t-1}\right]$ and $\mathbb{P}_t\left[\cdot\right] = \mathbb{P}\left[\cdot|\mathcal{F}_{t-1}\right]$. We will also often make use of a *ghost sample* $X_0 \sim \mathcal{D}$ drawn independently from the entire interaction history $\mathcal{F}_T$ for the sake of analysis. For instance, we can immediately show using this technique that for any policy π, we have

$$\mathbb{E}\left[\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right]\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\langle X_0, \theta_{t,\pi(X_0)}\right\rangle\right]\right]$$
$$= \mathbb{E}\left[\left\langle X_0, \mathbb{E}\left[\theta_{t,\pi(X_0)}\right]\right\rangle\right],$$

where the last expectation emphasizes that the loss vector $\theta_{t,a}$ may depend on the past random contexts and actions. This in turn can be used to show

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_t, \theta_{t,\pi(X_t)}\right\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left\langle X_0, \mathbb{E}\left[\theta_{t,\pi(X_0)}\right]\right\rangle\right] \ge \mathbb{E}\left[\min_a \sum_{t=1}^{T}\left\langle X_0, \mathbb{E}\left[\theta_{t,a}\right]\right\rangle\right],$$

so the optimal policy π_T^* that the learner compares itself to is the one defined through the rule

$$\pi_T^*(x) = \operatorname*{arg\,min}_a \sum_{t=1}^{T}\left\langle x, \mathbb{E}\left[\theta_{t,a}\right]\right\rangle \qquad (\forall x \in \mathbb{R}^d). \tag{3.1}$$

We will refer to policies of the above form as *linear-classifier policies* and are defined through the rule $\pi_\theta(x) = \operatorname{arg\,min}_a \left\langle x, \theta_a\right\rangle$ for any collection of parameter vectors $\theta \in \mathbb{R}^{K\times d}$. We will also rely on the notion of *stochastic policies* that assign probability distributions over the action set to each state, and use $\pi(a|x)$ to denote the probability that the stochastic policy π takes action a in state x.

Our analysis will rely on the following assumptions. We will suppose the context distribution is supported on the bounded set $\mathcal{X}$ with each $x \in \mathcal{X}$ satisfying

$\|x\|_2 \leq \sigma$ for some $\sigma > 0$, and also that $\|\theta_{t,a}\|_2 \leq R$ for some positive R for all t, a. Additionally, we suppose that the loss function is bounded by one in absolute value: $\left|\ell_t(x, a)\right| \leq 1$ for all t, a and all $x \in \mathcal{X}$. We will finally assume that the covariance matrix of the contexts $\Sigma = \mathbb{E}\left[X_t X_t^\mathsf{T}\right]$ is positive definite with its smallest eigenvalue being $\lambda_{\min} > 0$.

3.3 Algorithms and main results

Our main algorithmic contribution is a natural adaptation of the classic EXP3 algorithm of Auer et al. [12] to the linear contextual bandit setting. The key idea underlying our method is to design a suitable estimator of the loss vectors and use these estimators to define a policy for the learner as follows: letting $\widehat{\theta}_{t,a}$ be an estimator of the true loss vector $\theta_{t,a}$ and their cumulative sum $\widehat{\Theta}_{t,a} = \sum_{k=1}^{t} \widehat{\theta}_{k,a}$, our algorithm will base its decisions on the values $\langle X_t, \widehat{\Theta}_{t-1,a}\rangle$ serving as estimators of the cumulative losses $\langle X_t, \Theta_{t-1,a}\rangle = \sum_{k=1}^{t-1} \langle X_t, \theta_{k,a}\rangle$. The algorithm then uses these values in an exponential-weights-style algorithm and plays action a with probability proportional to $\exp\bigl(-\eta\langle X_t, \widehat{\Theta}_{t-1,a}\rangle\bigr)$, where $\eta > 0$ is a *learning-rate* parameter. We present a general version of this method as Algorithm 4.3. As a tribute to the LINUCB algorithm, a natural extension of the classic UCB algorithm to linear contextual bandits, we refer to our algorithm as LINEXP3.

As presented above, LINEXP3 is more of a template than an actual algorithm since it does not specify the loss estimators $\widehat{\theta}_{t,a}$. Ideally, one may want to use *unbiased* estimators that satisfy $\mathbb{E}\bigl[\widehat{\theta}_{t,a}\bigr] = \theta_{t,a}$ for all t, a. Our key contribution is designing two different (nearly) unbiased estimators that will allow us to prove performance guarantees of two distinct flavors. Both estimators are efficiently computable, but require some prior knowledge the context distribution $\mathcal{D}$. In what follows, we describe the two variants of LINEXP3 based on the two estimators and state the corresponding performance guarantees, and relegate the proof sketches to Section 3.4. We also present two simple variants of our algorithms that work with various degrees of full-information feedback in Section 3.5.

3.3.1 Algorithm for nearly-linear losses: ROBUSTLINEXP3

We begin by describing the simpler one of our two algorithms, which will be seen to be robust to misspecification of the linear loss model. We will accordingly refer to this algorithm as ROBUSTLINEXP3. Specifically, we suppose in this section

Algorithm 3.1 LINEXP3

Parameters: Learning rate $\eta > 0$, exploration parameter $\gamma \in (0, 1)$, Σ
Initialization: Set $\theta_{0,i} = 0$ for all $i \in [K]$.
For $t = 1, \dots, T$, **repeat:**

```
1. Observe $X_t$ and, for all $a$, set
       $$w_t(X_t, a) = \exp\left(-\eta \sum_{s=0}^{t-1} \langle X_t, \widehat{\theta}_{s,a} \rangle\right),$$
2. draw $A_t$ from the policy defined as
       $$\pi_t\,(a|X_t) = (1-\gamma)\frac{w_t(X_t, a)}{\sum_{a'} w_t(X_t, a')} + \frac{\gamma}{K},$$
3. observe the loss $\ell_t(X_t, A_t)$ and compute $\widehat{\theta}_{t,a}$ for all $a$.
```

that $\ell_t(x, a) = \langle x, \theta_{t,a} \rangle + \varepsilon_t(x, a)$, where $\varepsilon_t(x, a) : \mathbb{R}^d \times K \to \mathbb{R}$ is an arbitrary nonlinear function satisfying $|\varepsilon_t(x, a)| \le \varepsilon$ for all t, x and a and some $\varepsilon > 0$. Also supposing that we have perfect knowledge of the covariance matrix Σ, we define the loss estimator used by ROBUSTLINEXP3 for all actions a as

$$\widehat{\theta}_{t,a} = \frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)} \Sigma^{-1} X_t \ell_t(X_t, A_t). \tag{3.2}$$

In case the loss is truly linear, it is easy to see that the above is an unbiased estimate since

$$\begin{aligned}\mathbb{E}_t\left[\widehat{\theta}_{t,a}\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[\left.\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)}\Sigma^{-1}X_t\left\langle X_t, \theta_{t,a}\right\rangle\right| X_t\right]\right] \\ &= \mathbb{E}_t\left[\mathbb{E}_t\left[\left.\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)}\right| X_t\right]\Sigma^{-1}X_tX_t^{\mathsf{T}}\theta_{t,a}\right] = \mathbb{E}_t\left[\Sigma^{-1}X_tX_t^{\mathsf{T}}\theta_{t,a}\right] = \theta_{t,a},\end{aligned}$$

where we used the definition of Σ and the independence of $\theta_{t,a}$ from X_t in the last step. A key result in our analysis will be that, for nonlinear losses, the estimate above satisfies

$$\left|\mathbb{E}_t\left[\left\langle X_t, \widehat{\theta}_{t,a}\right\rangle - \ell_t(X_t, a)\right]\right| \le \varepsilon\sqrt{d}.$$

Our main result regarding the performance of RobustLinExp3 is the following:

Theorem 3.3.1. *For any positive* $\eta \leq \frac{\gamma \lambda_{\min}}{K\sigma^2}$ *and for any* $\gamma \in (0,1)$ *the expected regret of* RobustLinExp3 *satisfies*

$$R_T \leq 2\sqrt{d}\varepsilon T + 2\gamma T + \frac{2\eta K d T}{\gamma} + \frac{\log K}{\eta}.$$

Furthermore, letting $\eta = T^{-2/3}\left(Kd\right)^{-1/3}\left(\log K\right)^{2/3}$, $\gamma = T^{-1/3}\left(Kd\log K\right)^{1/3}$ *and supposing that* T *is large enough so that* $\eta \leq \frac{\gamma\lambda_{\min}}{K\sigma^2}$ *holds, the expected regret of* RobustLinExp3 *satisfies*

$$R_T \leq 5T^{2/3}\left(Kd\log K\right)^{1/3} + 2\varepsilon\sqrt{d}T.$$

3.3.2 Algorithm for linear losses: RealLinExp3

Our second algorithm uses a more sophisticated estimator based on the covariance matrix

$$\Sigma_{t,a} = \mathbb{E}_t\left[\mathbb{I}_{\{A_t=a\}} X_t X_t^\mathsf{T}\right],$$

which is used to define the estimate

$$\widetilde{\theta}^*_{t,a} = \mathbb{I}_{\{A_t=a\}}\Sigma_{t,a}^{-1} X_t \left\langle X_t, \theta_{t,a}\right\rangle.$$

This can be easily shown to be unbiased as

$$\mathbb{E}_t\left[\widetilde{\theta}^*_{t,a}\right] = \mathbb{E}_t\left[\mathbb{I}_{\{A_t=a\}}\Sigma_{t,a}^{-1}X_t\left\langle X_t,\theta_{t,a}\right\rangle\right] = \mathbb{E}_t\left[\Sigma_{t,a}^{-1}\mathbb{I}_{\{A_t=a\}}X_tX_t^\mathsf{T}\theta_{t,a}\right] = \theta_{t,a},$$

where we used the conditional independence of $\theta_{t,a}$ and X_t once again. Unfortunately, unlike the estimator used by RobustLinExp3, the bias of this estimator cannot be bounded when the losses are misspecified. However, its variance turns out to be much smaller for well-specified linear losses, which will enable us to prove tighter regret bounds for this case.

One downside of the estimator defined above is that it is very difficult to compute: the matrix $\Sigma_{t,a}$ depends on the joint distribution of the context X_t and the action A_t, which has a very complicated structure. While it is trivially easy to design an unbiased estimator of $\Sigma_{t,a}$, it is very difficult to compute a reliable-enough estimator of its inverse. To address this issue, we design an alternative estimator based on a matrix generalization of the Geometric Resampling method of Neu and Bartók [109, 110]. The method that we hereby dub *Matrix Geometric Resampling* (MGR) has two parameters $\beta > 0$ and $M \in \mathbb{Z}_+$, and constructs an estimator of $\Sigma_{t,a}^{-1}$ through the following procedure:

```
Matrix Geometric Resampling
Input: data distribution $\mathcal{D}$, policy $\pi_t$, action $a$.
For $k = 1, \dots, M$, repeat:
    1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
    2. compute $B_{k,a} = \mathbb{I}_{\{A(k)=a\}} X(k) X(k)^\mathsf{T}$,
    3. compute $A_{k,a} = \prod_{j=1}^{k} (I - \beta B_{k,a})$.
Return $\widehat{\Sigma}_{t,a}^{+} = \beta I + \beta \sum_{k=1}^{M} A_{k,a}$.
```

Clearly, implementing the MGR procedure requires sampling access to the distribution $\mathcal{D}$. The rationale behind the estimator $\widehat{\Sigma}_{t,a}^{+}$ is the following. Assuming that $M = \infty$ and $\beta \leq \frac{1}{\sigma^2}$, we can observe that $\mathbb{E}_t \left[B_{k,a}\right] = \Sigma_{t,a}$ and, due to independence of the contexts $X(k)$ from each other,

$$\mathbb{E}_t \left[A_{k,a}\right] = \mathbb{E}_t \left[\prod_{j=1}^{k} (I - \beta B_{k,a})\right] = \left(I - \beta \Sigma_{t,a}\right)^k,$$

we can see that $\widehat{\Sigma}_{t,a}^{+}$ is a good estimator of $\Sigma_{t,a}^{-1}$ on expectation:

$$\mathbb{E}_t \left[\widehat{\Sigma}_{t,a}^{+}\right] = \beta I + \beta \sum_{k=1}^{\infty} \left(I - \beta \Sigma_{t,a}\right)^k = \beta \sum_{k=0}^{\infty} \left(I - \beta \Sigma_{t,a}\right)^k = \beta \left(\beta \Sigma_{t,a}\right)^{-1} = \Sigma_{t,a}^{-1}. \tag{3.3}$$

As we will see later in the analysis, the bias introduced by setting a finite M can be controlled relatively easily.

Based on the above procedure, we finally define our loss estimator used in this section as

$$\widetilde{\theta}_{t,a} = \widehat{\Sigma}_{t,a}^{+} X_t \ell(X_t, A_t) \mathbb{I}_{\{A_t = a\}}. \tag{3.4}$$

Via a careful incremental implementation, the estimator can be computed in $O(MKd)$ time and M calls to the oracle generating samples from the context distribution $\mathcal{D}$. We present the details of this efficient computation procedure in

Section 4.3.4. We will refer to the version of LINEXP3 using the estimates above as REALLINEXP3, alluding to its favorable guarantees obtained for realizable linear losses. Our main result in this section is the following guarantee regarding the performance of REALLINEXP3:

Theorem 3.3.2. *For $\gamma \in (0,1)$, $M \geq 0$, any positive $\eta \leq \frac{1}{\sigma^2\beta(M+1)}$ and any positive $\beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, the expected regret of* REALLINEXP3 *satisfies*

$$R_T \leq 2T\sigma R \cdot \exp\left(-\frac{\gamma\beta\lambda_{\min}M}{K}\right) + 2\gamma T + (3+5d)\eta KT + \frac{\log K}{\eta}.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, $M = \left\lceil \frac{K^2\sigma^4 d\log^2(T\sigma^2R^2)}{\gamma^2\lambda_{\min}^2}\right\rceil$, $\gamma = \sqrt{\frac{\log(T\sigma^2R^2)}{T}}$, and $\eta = \sqrt{\frac{\log K}{dKT\log(T\sigma^2R^2)}}$ and supposing that T is large enough so that the above constraints are satisfied, we also have

$$R_T \leq 4\sqrt{T} + \sqrt{dKT\log K}\left(8 + \sqrt{\log(T\sigma^2R^2)}\right).$$

3.4 Analysis

This section is dedicated to proving our main results, Theorems 3.3.1 and 3.3.2. We present the analysis in a modular fashion, first proving some general facts about the algorithm template LINEXP3, and then treat the two variants separately in Sections 3.4.1 and 3.4.2 that differ in their choice of loss estimator.

The main challenge in the contextual bandit setting is that the comparator term in the regret definition features actions that depend on the observed contexts, which is to be contrasted with the classical multi-armed bandit setting where the comparator strategy always plays a fixed action. The most distinctive element of our analysis is the following lemma that tackles this difficulty by essentially reducing the contextual bandit problem to a set of auxiliary online learning problems defined separately for each context x:

Lemma 16. *Let π^* be any fixed stochastic policy and let X_0 be sample from the context distribution $\mathcal{D}$ independent from $\mathcal{F}_T$. Suppose that $\pi_t \in \mathcal{F}_{t-1}$ and that*

$\mathbb{E}_t\left[\widehat{\theta}_{t,a}\right] = \theta_{t,a}$ *for all* t, a. *Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\big(\pi_t(a|X_t) - \pi^*(a|X_t)\big)\left\langle X_t, \theta_{t,a}\right\rangle\right] \tag{3.5}$$
$$= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\big\langle X_0, \widehat{\theta}_{t,a}\big\rangle\right].$$

Proof. Fix any t and a. Then, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\big\langle X_0, \widehat{\theta}_{t,a}\big\rangle\right] \\
&= \mathbb{E}_t\left[\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\big\langle X_0, \widehat{\theta}_{t,a}\big\rangle\Big| X_0\right]\right] \\
&= \mathbb{E}_t\left[\mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \theta_{t,a}\right\rangle\big| X_0\right]\right] \\
&= \mathbb{E}_t\left[\big(\pi_t(a|X_0) - \pi^*(a|X_0)\big)\left\langle X_0, \theta_{t,a}\right\rangle\right] \\
&= \mathbb{E}_t\left[\big(\pi_t(a|X_t) - \pi^*(a|X_t)\big)\left\langle X_t, \theta_{t,a}\right\rangle\right],
\end{aligned}$$

where the first step uses the tower rule of expectation, the second that $\mathbb{E}_t\left[\widehat{\theta}_{t,a}\middle|X_0\right] = \theta_{t,a}$ that holds due to the independence of $\widehat{\theta}_t$ and θ_t on X_0, the third step is the tower rule again, and the last step uses that X_0 and X_t have the same distribution and both are conditionally independent on θ_t. Summing up for all actions concludes the proof. □

Notably, the lemma above is not specific to our algorithm LINEXP3 and only uses the properties of the loss estimator. Applying the lemma to the policies π_t produced by LINEXP3 and using *any* comparator π^*, we can notice that the term on the right hand side is the regret R_T of the algorithm. We stress here that the above result is in fact very powerful since it does not assume *anything* (except measurability) about π^*, even allowing it to be non-smooth—we provide a more detailed discussion of this issue in Section 5.2. In order to interpret the term on the right-hand side of Equation (3.5), let us consider an auxiliary online learning problem for a fixed x with K actions and losses $\widehat{\ell}_t(x,a) = \big\langle x, \widehat{\theta}_{t,a}\big\rangle$ for each t, a, and consider running a copy of the classic exponential-weights algorithm[2] of Littlestone and Warmuth [99] fed with these

[2] For the sake of clarity, we omit the step of mixing in the uniform distribution in this expository discussion.

losses. The probability distribution played by this algorithm over the actions a is given as $\pi_t(a|x) \propto \exp\left(-\eta \sum_{s=1}^{t-1} \widehat{\ell}_s(x,a)\right)$, which implies that the regret in the auxiliary game against comparator π^* at x can be written as

$$\widehat{R}_T(x) = \sum_{t=1}^{T} \sum_{a} \big(\pi_t(a|x) - \pi^*(a|x)\big)\big\langle x, \widehat{\theta}_{t,a}\big\rangle.$$

This brings us to the key observation that the term on the right-hand side of the equality in Lemma 25 is exactly $\mathbb{E}\left[R_T(X_0)\right]$. Thus, our proof strategy will be to prove an almost-sure regret bound for the auxiliary games defined at each x and take expectation of the resulting bounds with respect to the law of X_0, thus achieving a bound on the regret R_T. The following lemma provides the desired bounds for the auxiliary games:

Lemma 17. *Fix any $x \in \mathcal{X}$ and suppose that $\widehat{\theta}_{t,a}$ is such that $\left|\eta\big\langle x, \widehat{\theta}_{t,a}\big\rangle\right| < 1$. Then, the regret of* LinExp3 *in the auxiliary game at x satisfies*

$$\widehat{R}_T(x) \leq \frac{\log K}{\eta} + 2\gamma U_T(x) + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} \pi_t(a|x)\big\langle x, \widehat{\theta}_{t,a}\big\rangle^2,$$

where $U_T(x) = \sum_{t=1}^{T} \left(\frac{1}{K} \sum_a \big\langle x, \widehat{\theta}_{t,a}\big\rangle - \big\langle x, \widehat{\theta}_{t,\pi^*(x)}\big\rangle\right)$.

In the above bound, $U_T(x)$ is the regret of the uniform policy, which can be bounded by T for all algorithms on expectation. The proof is a straightforward application of standard ideas from the classical Exp3 analysis due to Auer et al. [14], and we include it in Section 4.4.4 for completeness.

The lemmas above suggest that all we need to do is to bound the expectation of the second-order terms on the right-hand side, $\mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0)\big\langle X_0, \widehat{\theta}_{t,a}\big\rangle^2\right]$. This, however, is not the only challenge due to the fact that the estimators our algorithms use are not necessarily all unbiased. Specifically, supposing that our estimator can be written as $\widehat{\theta}_{t,a} = \widehat{\theta}^*_{t,a} + b_{t,a}$, where $\widehat{\theta}^*_{t,a}$ is such that $\mathbb{E}_t\big[\widehat{\theta}^*_{t,a}\big] = \theta_{t,a}$ and $b_{t,a}$ is a bias term, we can directly deduce the following bound from Lemma 25:

$$R_T \leq \mathbb{E}\big[\widehat{R}_T(X_0)\big] + 2\sum_{t=1}^{T} \max_a \left|\mathbb{E}\left[\langle X_t, b_{t,a}\rangle\right]\right|. \tag{3.6}$$

The rest of the section is dedicated to finding the upper bounds on the bias term above and on the expectation of the second-order term discussed right before for both estimators (3.2) and (3.4), therefore completing the proofs of our main results, Theorems 3.3.1 and 3.3.2.

3.4.1 Proof of Theorem 3.3.1

We first consider RobustLinExp3 which uses the estimator $\widehat{\theta}_{t,a}$ defined in Equation (3.2). While we have already shown in Section 3.3.1 that the estimator is unbiased, we now consider the case where the true loss function may be nonlinear and can be written as $\ell_t(x,a) = \langle x, \theta_{t,a}\rangle + \varepsilon_t(x,a)$ for some nonlinear function ε_t uniformly bounded on $\mathcal{X}$ by ε. Then, we can see that our estimator satisfies

$$\begin{aligned}\mathbb{E}_t\left[\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right] &= \mathbb{E}_t\left[\frac{\mathbb{I}_{\{A_t=a\}}}{\pi(a|X_t)} X_0^\mathsf{T}\Sigma^{-1}X_t\big(\langle X_t, \theta_{t,a}\rangle + \varepsilon_t(X_t,a)\big)\right] \\ &= \mathbb{E}_t\left[\langle X_0, \theta_{t,a}\rangle\right] + \mathbb{E}_t\left[X_0^\mathsf{T}\Sigma^{-1}X_t\varepsilon_t(X_t,a)\right],\end{aligned}$$

and thus the bias can be bounded using the Cauchy–Schwarz inequality as

$$\begin{aligned}&\left|\mathbb{E}_t\left[X_0^\mathsf{T}\Sigma^{-1}X_t\varepsilon_t(X_t,a)\right]\right| \leq \sqrt{\mathbb{E}_t\left[\mathrm{tr}\left(X_0X_0^\mathsf{T}\Sigma^{-1}X_tX_t^\mathsf{T}\Sigma^{-1}\right)\right]} \cdot \sqrt{\mathbb{E}_t\left[\left(\varepsilon_t(X_t,a)\right)^2\right]} \\ &\leq \sqrt{d}\varepsilon. \qquad (3.7)\end{aligned}$$

Here, we used $\mathbb{E}_t\left[X_0X_0^\mathsf{T}X_tX_t^\mathsf{T}\right] = \Sigma^2$, which follows from the conditional independence of X_0 and X_t and the definition of Σ, and the boundedness of ε_t in the last step. The other key component of the proof is the following bound:

$$\begin{aligned}&\mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_0)\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle^2\right] \\ &= \mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_0)\frac{\mathbb{I}_{\{A_t=a\}}\ell_t(X_t,a)^2}{\pi_t^2(a|X_t)} X_0^\mathsf{T}\Sigma^{-1}X_tX_t^\mathsf{T}\Sigma^{-1}X_0\right] \qquad (3.8) \\ &\leq \mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_0)\cdot\frac{K}{\gamma}\cdot\frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)}\cdot \mathrm{tr}\left(\Sigma^{-1}X_tX_t^T\Sigma^{-1}X_0X_0^\mathsf{T}\right)\right] \leq \frac{Kd}{\gamma},\end{aligned}$$

where we used $\pi_t(a|X_t) \geq \frac{\gamma}{K}$ in the first inequality and the conditional independence of X_t and X_0 in the last step. The problem we are left with is to prove that $\eta\left|\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right| \leq 1$:

$$\left|\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right| = \frac{\mathbb{I}_{\{A_t=a\}}}{\pi_t(a|X_t)} \left|X_0^\mathsf{T}\Sigma^{-1}X_t\right| \ell_t(X_t, A_t) \leq \frac{K\sigma^2}{\gamma\lambda_{\min}},$$

where we used the conditions $\pi_t(a|X_t) \geq \frac{\gamma}{K}$ and $|\ell_t(x,a)| \leq 1$ and the Cauchy–Schwarz inequality to show $\left|X_0^\mathsf{T}\Sigma^{-1}X_t\right| \leq \frac{\sigma^2}{\lambda_{\min}}$. Having satisfied its condition, we may now appeal to Lemma 30, and the proof is concluded by combining and Equations (3.6), (3.7), and (3.8).

3.4.2 Proof of Theorem 3.3.2

We now turn to analyzing REALLINEXP3 which uses the slightly more complicated loss estimator $\widetilde{\theta}_{t,a}$ defined to the MGR procedure. Although we have already seen in Section 3.3.2 that MGR could result in an unbiased estimate if we could set $M = \infty$. However, in order to keep computation at bay, we need to set M to be a finite (and hopefully relatively small) value. Following the same steps as in Equation (3.3), we can show

$$\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,a}\right] = \beta\sum_{k=0}^{M}(I-\beta\Sigma_{t,a})^k = \Sigma^{-1}_{t,a} - (I-\beta\Sigma_{t,a})^M\Sigma^{-1}_{t,a}.$$

Combining this insight with the definition of $\widetilde{\theta}_{t,a}$ and using some properties of our algorithm, we can prove the following useful bound on the bias of the estimator:

Lemma 18. *Suppose that* $M \geq \frac{K\sigma^2\log T}{\gamma\lambda_{\min}}$, $\beta = \frac{1}{2\sigma^2}$. *Then,* $\left|\mathbb{E}_t\left[\left\langle X_t, \theta_{t,a} - \widetilde{\theta}_{t,a}\right\rangle\right]\right| \leq \frac{\sigma R}{\sqrt{T}}$.

Proof. We first observe that the bias of $\widetilde{\theta}_{t,a}$ can be easily expressed as

$$\begin{aligned}\mathbb{E}_t\left[\widetilde{\theta}_{t,a}\right] &= \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,a}X_tX_t^\mathsf{T}\theta_{t,a}\mathbb{I}_{\{A_t=a\}}\right] = \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,a}\right]\mathbb{E}_t\left[X_tX_t^\mathsf{T}\mathbb{I}_{\{A_t=a\}}\right]\theta_{t,a}\\ &= \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,a}\right]\Sigma_{t,a}\theta_{t,a} = \theta_{t,a} - (I-\beta\Sigma_{t,a})^M\theta_{t,a},\end{aligned}$$

where we have used our expression for $\mathbb{E}_t\left[\widehat{\Sigma}_{t,a}^{+}\right]$ derived above. Thus, the bias is bounded as

$$\left|\mathbb{E}_t\left[X_t^{\mathsf{T}}(I-\beta\Sigma_{t,a})^M\theta_{t,a}\right]\right| \le \|X_t\|_2 \cdot \|\theta_{t,a}\|_2 \left\|(I-\beta\Sigma_{t,a})^M\right\|_{\text{op}}.$$

In order to bound the last factor above, observe that $\Sigma_{t,a} \succcurlyeq \frac{\gamma}{K}\Sigma$ due to the uniform exploration used by LINEXP3, which implies that

$$\left\|(I-\beta\Sigma_{t,a})^M\right\|_{\text{op}} \le \left(1-\frac{\gamma\beta\lambda_{\min}}{K}\right)^M \le \exp\left(-\frac{\gamma\beta}{K}\lambda_{\min}M\right) \le \frac{1}{\sqrt{T}},$$

where the second inequality uses $1-z \le e^{-z}$ that holds for all z, and the last step uses our condition on M. This concludes the proof. □

The other key term in the regret bound is bounded in the following lemma:

Lemma 19. *Suppose that X_t is satisfying $\|X_t\|_2 \le \sigma$, $0 < \beta \le \frac{1}{2\sigma^2 d\sqrt{M+1}}$ and $M > 0$. Then for each t,* REALLINEXP3 *guarantees*

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_0)\big\langle X_0,\widetilde{\theta}_{t,a}\big\rangle^2\right] \le K(3+5d).$$

Unfortunately, the proof of this statement is rather tedious, so we have to relegate it to Section 3.4.4. As a final step, we need to verify that the condition of Lemma 30 is satisfied, that is, that $\eta\left|\big\langle X_0,\widetilde{\theta}_{t,a}\big\rangle\right| < 1$ holds. To this end, notice that

$$\begin{aligned}\eta\cdot\left|\big\langle X_0,\widetilde{\theta}_{t,a}\big\rangle\right| &= \eta\cdot\left|X_0^{\mathsf{T}}\widehat{\Sigma}_{t,a}^{+}X_t\left\langle X_t,\theta_{t,a}\right\rangle\mathbb{I}_{\{A_t=a\}}\right| \le \eta\cdot\left|X_0^{\mathsf{T}}\widehat{\Sigma}_{t,a}^{+}X_t\right| \\ &\le \eta\sigma^2\left\|\widehat{\Sigma}_{t,a}^{+}\right\|_{\text{op}} \le \eta\sigma^2\beta\left(1+\sum_{k=1}^{M}\|A_{k,a}\|_{\text{op}}\right) \le \eta(M+1)/2,\end{aligned}$$

where we used the fact that our choice of β ensures that $\|A_{k,a}\|_{\text{op}} = \left\|\prod_{j=0}^{k}(I-\beta B_{j,a})\right\|_{\text{op}} \le 1$. Thus, the condition $\eta \le 2/(M+1)$ allows us to use Lemma 30, so we can conclude the proof of Theorem 3.3.2 by applying Lemma 21, Lemma 31 and the bound of Equation (3.6).

3.4.3 Proof of Lemma 30

The proof follows the standard analysis of EXP3 originally due to Auer et al. [14]. We begin by recalling the notation $w_t(x,a) = \exp\bigl(-\eta\sum_{s=1}^{t-1}\langle x, \widehat{\theta}_{t,a}\rangle\bigr)$ and introducing $W_t(x) = \sum_{a=1}^{K} w_t(x,a)$. The proof is based on analyzing $\log W_{T+1}(x)$, which can be thought of as a potential function in terms of the cumulative losses. We first observe that $\log W_{T+1}(x)$ can be lower-bounded in terms of the cumulative loss:

$$\log\left(\frac{W_{T+1}(x)}{W_1(x)}\right) \geq \log\left(\frac{w_{T+1}(x,\pi^*(x))}{W_1(x)}\right) = -\eta\sum_{t=1}^{T} x^{\mathsf{T}}\widehat{\theta}_{t,\pi^*(x)} - \log K.$$

On the other hand, for any t, we can prove the upper bound

$$\begin{aligned}
\log\frac{W_{t+1}(x)}{W_t(x)} &= \log\left(\sum_{a=1}^{K}\frac{w_{t+1}(x,a)}{W_t(x)}\right) = \log\left(\sum_{a=1}^{K}\frac{w_t(x,a)e^{-\eta\langle x,\widehat{\theta}_{t,a}\rangle}}{W_t(x)}\right)\\
&= \log\left(\sum_{i=1}^{K}\frac{\pi_t(a|x)-\gamma/K}{1-\gamma}\cdot e^{-\eta\langle x,\widehat{\theta}_{t,a}\rangle}\right)\\
&\overset{(a)}{\leq} \log\left(\sum_{i=1}^{K}\frac{\pi_t(a|x)-\gamma/K}{1-\gamma}\Bigl(1-\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle + \bigl(\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle\bigr)^2\Bigr)\right)\\
&\overset{(b)}{\leq} \sum_{a=1}^{K}\frac{\pi_t(a|x)}{1-\gamma}\Bigl(-\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle + \bigl(\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle\bigr)^2\Bigr) + \frac{\eta\gamma}{K(1-\gamma)}\sum_{a}\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle,
\end{aligned}$$

where in step (a) we used the inequality $e^{-z} \leq 1 - z + z^2$, which holds for $z \geq -1$, and in step (b) we used the inequality $\log(1+z) \leq z$ that holds for any z.

Noticing that $\sum_{t=1}^{T}\log\frac{W_{t+1}}{W_t} = \log\frac{W_{T+1}}{W_1}$, we can sum both sides of the above inequality for all $t = 1,\dots,T$ and compare with the lower bound to get

$$\begin{aligned}
-\eta\sum_{t=1}^{T} x^{\mathsf{T}}\widehat{\theta}_{t,\pi^*(x)} - \ln K &\leq \sum_{t=1}^{T}\sum_{a=1}^{K}\frac{\pi_t(a|x)}{1-\gamma}\Bigl(-\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle + \bigl(\eta\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle\bigr)^2\Bigr)\\
&\quad + \frac{\eta\gamma\sum_a\bigl\langle x,\widehat{\theta}_{t,a}\bigr\rangle}{K(1-\gamma)}.
\end{aligned}$$

Reordering and multiplying both sides by $\frac{1-\gamma}{\eta}$ gives

$$\sum_{t=1}^{T}\left(\sum_{a=1}^{K}\pi_t(a|x)\big\langle x,\widehat{\theta}_{t,a}\big\rangle-\big\langle x,\widehat{\theta}_{t,\pi^*(x)}\big\rangle\right)$$
$$\leq\frac{(1-\gamma)\log K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\left(\langle x,\widehat{\theta}_{t,a}\rangle\right)^2$$
$$+\gamma\sum_{t=1}^{T}\left(\frac{1}{K}\sum_{a}\big\langle x,\widehat{\theta}_{t,a}\big\rangle-\big\langle x,\widehat{\theta}_{t,\pi^*(x)}\big\rangle\right).$$

This concludes the proof. □

3.4.4 Proof of Lemma 19

First, we prove the following statement, that would be helpful further in the proof.

Lemma 20. *Let $\widetilde{H}$ be a symmetric positive definite matrix, commuting with $\Sigma_{t,a}$, let V be a matrix where the columns are from the orthonormal system of eigenvectors of $\Sigma_{t,a}$, then*

$$tr\left(\widetilde{H}A_k\Sigma_{t,a}A_k\right)\leq\frac{\beta\sigma^2 tr\left(\Sigma_{t,a}\right)}{2}tr\left(\widetilde{H}\Sigma_{t,a}\left(\left(I-\beta\Sigma_{t,a}\right)^k\right)\right)$$
$$+\frac{\beta\sigma^2 tr\left(\Sigma_{t,a}\right)}{2}tr\left(\widetilde{H}\Sigma_{t,a}\left(Vdiag\left(\frac{1}{\lambda_j(\Sigma_{t,a})}\exp\left(\beta^2\sigma^2 tr\left(\Sigma_{t,a}\right)k+2\beta\lambda_j(\Sigma_{t,a})\right)\right)V^T\right)\right),$$

for $j=1,\ldots,d$.

Proof. To simplify the notation, we omit indices a,t in this proof. Let H be a symmetric positive definite matrix, commuting with $\Sigma_{t,a}$, X be a random vector with $\|X\|\leq\sigma$, then we will show the following inequality that holds almost surely:

$$XX^\intercal HXX^\intercal\preccurlyeq\sigma^2\mathrm{tr}\left(H\right)XX^\intercal.$$

To prove this, we first notice that, since H is symmetric positive definite, we can write $H=\sum_{i=1}^{d}\lambda_i v_i v_i^\intercal$, and thus

$$XX^\intercal HXX^\intercal=\sum_{i=1}^{d}\lambda_i XX^\intercal v_i v_i^\intercal XX^\intercal.$$

To proceed, we will fix i and study the corresponding term in the above sum. Fixing an arbitrary vector $a \in \mathbb{R}^d$ and letting $b_i = v_i X^\intercal a$, we write

$$\begin{aligned} a^\intercal X X^\intercal v_i v_i^\intercal X X^\intercal a = a^\intercal X v_i^\intercal X X^\intercal v_i X^\intercal a = b_i^\intercal X X^\intercal b_i \leq \sigma^2 \left\| b_i \right\|_2^2 = \sigma^2 a^\intercal X v_i^\intercal v_i X^\intercal a \\ = \sigma^2 \left(a^\intercal X \right)^2 \end{aligned}$$

where the inequality is Cauchy–Schwartz and we used that $\left\| v_i \right\|_2 = 1$. Multiplying by λ_i and summing up on both sides, we get

$$\begin{aligned} a^\intercal X X^\intercal H X X^\intercal a &= \sum_{i=1}^{d} \lambda_i a^\intercal X X^\intercal v_i v_i^\intercal X X^\intercal a \\ &\leq \sigma^2 \sum_{i=1}^{d} \lambda_i a^\intercal X v_i^\intercal v_i X^\intercal a \\ &= \sigma^2 \left(a^\intercal X \right)^2 \mathrm{tr}\left(H \right). \end{aligned}$$

Since the inequality holds for arbitrary a, this implies that $X X^\intercal H X X^\intercal \preccurlyeq \sigma^2 \mathrm{tr}\left(H \right) X X^\intercal$. Using the above result and the definition of $A_k = A_{k-1}(I - \beta B_k)$, we get

$$\mathrm{tr}\left(\mathbb{E}\left[\Sigma A_k H A_k \right] \right) \leq \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_{k-1} H \left(I - 2\beta\Sigma \right) A_{k-1} \right] \right) + \beta^2 \sigma^2 \mathrm{tr}\left(H \right) \mathrm{tr}\left(\mathbb{E}\left[\Sigma A_{k-1} \Sigma A_{k-1} \right] \right). \tag{3.9}$$

To proceed, let us introduce some shorthand notations: $\alpha(H) = \beta^2 \sigma^2 \mathrm{tr}\left(H \right)$, $\alpha = \alpha(\Sigma)$, and $U = I - \beta\Sigma$. Thus, we can rewrite (3.9) as

$$\mathrm{tr}\left(\mathbb{E}\left[\widetilde{H} A_k H A_k \right] \right) \leq \mathrm{tr}\left(\mathbb{E}\left[\widetilde{H} A_{k-1} \left(\alpha(H)\Sigma + HU \right) A_{k-1} \right] \right). \tag{3.10}$$

We show that the following holds:

$$\mathrm{tr}\left(\mathbb{E}\left[\widetilde{H} A_k \Sigma A_k \right] \right) \leq \mathrm{tr}\left(\widetilde{H} \Sigma \left(U^k + \alpha \cdot \left(\sum_{i=0}^{k-1} (\alpha+1)^i U^{k-1-i} \right) \right) \right). \tag{3.11}$$

To prove the inequality above, we show by induction the following:

$$\mathrm{tr}\left(\mathbb{E}\left[\widetilde{H} A_k \Sigma A_k \right] \right) \leq \mathrm{tr}\left(\widetilde{H} A_{k-j} \Sigma \left(U^j + \alpha \cdot \left(\sum_{i=0}^{(j-1)\wedge 0} (\alpha+1)^i U^{j-1-i} \right) A_{k-j} \right) \right).$$

First, for $j = 0$, the inequality clearly holds as an equality. Now we will show that if the inequality above holds for j, then it also holds for $j + 1$:

$$
\begin{aligned}
&\operatorname{tr}\left(\widetilde{H}A_{k-j}\Sigma\left(U^j + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i U^{j-1-i}\right)A_{k-j}\right)\right)\\
&\quad\le \operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma\left(U^j + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i U^{j-1-i}\right)\right)UA_{k-j-1}\right)\\
&\quad+\alpha\left(U^j + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i U^{j-1-i}\right)\right)\operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma A_{k-j-1}\right)\\
&\quad\le \operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma\left(U^j + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i U^{j-1-i}\right)\right)UA_{k-j-1}\right)\\
&\quad+\left(\alpha + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i\right)\right)\operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma A_{k-j-1}\right)\\
&\quad\le \operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma\left(U^j + \alpha\cdot\left(\sum_{i=0}^{(j-1)\wedge 0}(\alpha+1)^i U^{j-1-i}\right)\right)UA_{k-j-1}\right)\\
&\quad+(\alpha+1)^j\operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma A_{k-j-1}\right)\\
&\quad\le \operatorname{tr}\left(\widetilde{H}A_{k-j-1}\Sigma\left(U^{j+1} + \alpha\cdot\left(\sum_{i=0}^{j}(\alpha+1)^i U^{j-1-i}\right)\right)A_{k-j-1}\right),
\end{aligned}
$$

which is exactly what we wanted to show. Thus, we proved (3.11).

Now, we use the inequality $1 + x \le e^x$, to bound $\alpha + 1 \le e^\alpha$ and $\lambda_i(I - 2\beta\Sigma) \le e^{-2\beta\lambda_j(\Sigma)}, \forall j \in [d]$, where $\lambda_j(\cdot)$ is the jth eigenvalue. It gives us

$$
\begin{aligned}
\lambda_j\left(\alpha\left(\sum_{i=0}^{k-1}(\alpha+1)^i U^{k-1-i}\right)\right) &\le \alpha\sum_{i=0}^{k-1} e^{\alpha i}\exp(-2\beta\lambda_j(\Sigma)(k-1-i))\\
&= \alpha\exp(-2\beta\lambda_j(\Sigma)(k-1))\sum_{i=0}^{k-1}\exp(\alpha i + 2\beta\lambda_j(\Sigma)i)
\end{aligned}
$$

$$
\begin{aligned}
&= \alpha \exp(-2\beta\lambda_j(\Sigma)(k-1)) \frac{\exp\left(\left(\alpha + 2\beta\lambda_j(\Sigma)\right)k\right) - 1}{\exp(\alpha + 2\beta\lambda_j(\Sigma)) - 1} \\
&\le \alpha \exp(-2\beta\lambda_j(\Sigma)(k-1)) \frac{\exp\left(\left(\alpha + 2\beta\lambda_j(\Sigma)\right)k\right)}{\exp(\alpha + 2\beta\lambda_j(\Sigma)) - 1} \\
&\le \alpha \exp(-2\beta\lambda_j(\Sigma)(k-1)) \exp\left(\left(\alpha + 2\beta\lambda_j(\Sigma)\right)(k-1)\right) \frac{\exp\left(\left(\alpha + 2\beta\lambda_j(\Sigma)\right)\right)}{\exp(\alpha + 2\beta\lambda_j(\Sigma)) - 1} \\
&\le \alpha \exp\left(\alpha(k-1)\right) \frac{\exp\left(\alpha + 2\beta\lambda_j(\Sigma)\right)}{\alpha + 2\beta\lambda_j(\Sigma)} \\
&\le \frac{\alpha}{2\beta\lambda_j(\Sigma)} \exp\left(\alpha(k-1)\right) \exp\left(\alpha + 2\beta\lambda_j(\Sigma)\right) \\
&\qquad\qquad \text{(using } 1 + x \le e^x \text{ again)} \\
&= \frac{\beta\sigma^2 \mathrm{tr}\left(\Sigma\right)}{2\lambda_j(\Sigma)} \exp\left(\alpha k + 2\beta\lambda_j(\Sigma)\right) = \frac{\beta\sigma^2 \mathrm{tr}\left(\Sigma\right)}{2\lambda_j(\Sigma)} \exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma\right) k + 2\beta\lambda_j(\Sigma)\right).
\end{aligned}
$$

The statement of the lemma is obtained by joining the result of last equation with the equation (3.11). □

Equipped with this result, we can prove the upper bound on the quadratic term. The proof relies on a series of matrix operations, and makes repeated use of the following identity that holds for any symmetric positive definite matrix S:

$$
\sum_{k=0}^{M} (I - S)^k = S^{-1} - (I - S)^M S^{-1}.
$$

We start by plugging in the definition of $\widetilde{\theta}_{t,a}$ and writing

$$
\begin{aligned}
\mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0) \big\langle X_0, \widetilde{\theta}_{t,a} \big\rangle^2\right] &= \mathbb{E}_t\left[\sum_{a=1}^{K} \pi_t(a|X_0) \left(X_0^\mathsf{T} \Sigma_{t,a}^{+} X_t X_t^\mathsf{T} \theta_{t,a} \mathbb{I}_{\{A_t = a\}}\right)^2\right] \\
&\le \mathbb{E}_t\left[\mathbb{E}\left[\sum_{a=1}^{K} \mathrm{tr}\left(\pi_t(a|X_0) X_0 X_0^\mathsf{T} \Sigma_{t,a}^{+} X_t X_t^\mathsf{T} \Sigma_{t,a}^{+} \mathbb{I}_{\{A_t = a\}}\right) \middle| X_0\right]\right] \\
&= \sum_{a=1}^{K} \mathbb{E}_t\left[\mathrm{tr}\left(\Sigma_{t,a} \Sigma_{t,a}^{+} \Sigma_{t,a} \Sigma_{t,a}^{+}\right)\right],
\end{aligned}
$$

where we used $\langle X_0, \theta_{t,a} \rangle \leq 1$ in the inequality and observed that $\Sigma_{t,a} = \mathbb{E}_t \left[\pi_t(a|X_0)X_0X_0^\intercal\right]$. In what follows, we suppress the t, a indexes to enhance readability. Using the definition of Σ^+ and elementary manipulations, we can get

$$\begin{aligned}
\mathbb{E}\left[\text{tr}\left(\Sigma\Sigma^+\Sigma\Sigma^+\right)\right] &= \mathbb{E}\left[\beta^2 \cdot \text{tr}\left(\Sigma\left(\sum_{k=0}^{M} A_k\right)\Sigma\left(\sum_{j=0}^{M} A_j\right)\right)\right] \\
&= \beta^2 \sum_{k=0}^{M}\sum_{j=0}^{M} \text{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_j\right]\right) = \beta^2 \sum_{k=0}^{M} \text{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_k\right]\right) \\
&+ 2\beta^2 \sum_{k=0}^{M}\sum_{j=k+1}^{M} \text{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_j\right]\right).
\end{aligned}$$

Now, recalling the definition $A_k = \prod_{j=1}^{k} B_j$ and using the result of Lemma 20 with $\widetilde{H} = \Sigma_{t,a}$, we can obtain

$$\begin{aligned}
\beta^2 \sum_{k=0}^{M} \text{tr}\left(\mathbb{E}\left[\Sigma_t A_k \Sigma_t A_k\right]\right) &\leq \beta^2 \sum_{k=0}^{M} \text{tr}\left(\Sigma^2 \left(I - 2\beta\Sigma\right)^k\right) \\
&+ \beta^2 \sum_{j=1}^{d}\sum_{k=0}^{M} \lambda_j^2(\Sigma) \frac{\beta\sigma^2 \text{tr}\left(\Sigma\right)}{2\lambda_j(\Sigma)} \exp\left(\beta^2\sigma^2 \text{tr}\left(\Sigma\right) k + 2\beta\lambda_j(\Sigma)\right) \\
&= \beta \text{tr}\left(\Sigma(I - (I - \beta\Sigma)^M)\right) \\
&+ \beta^3 \frac{\sigma^2 \text{tr}\left(\Sigma\right)}{2} \sum_{j=1}^{d} \lambda_j(\Sigma) \exp(2\beta\lambda_j(\Sigma)) \frac{\exp\left(\beta^2\sigma^2 \text{tr}\left(\Sigma\right)(M+1)\right) - 1}{\exp\left(\beta^2\sigma^2 \text{tr}\left(\Sigma\right)\right) - 1} \\
&\leq \beta \text{tr}\left(\Sigma\right) + \beta^3 \frac{\sigma^2 \text{tr}\left(\Sigma\right)}{2\beta^2\sigma^2 \text{tr}\left(\Sigma\right)} \sum_{j=1}^{d} \lambda_j(\Sigma) \exp(2\beta\lambda_j(\Sigma)) \exp\left(\beta^2\sigma^2 \text{tr}\left(\Sigma\right)(M+1)\right) \\
&\leq \beta \text{tr}\left(\Sigma\right) + \frac{\beta}{2} \sum_{j=1}^{d} \lambda_j(\Sigma) \exp(2\beta\lambda_j(\Sigma)) \exp\left(\beta^2\sigma^2 \text{tr}\left(\Sigma\right)(M+1)\right) \\
&\leq \beta\sigma^2 d + \frac{\beta}{2} \sum_{j=1}^{d} \sigma^2 \exp(2\beta\sigma^2) \exp\left(\beta^2\sigma^4 d(M+1)\right) \leq 3,
\end{aligned} \tag{3.12}$$

where we used the condition $\beta \le \frac{1}{2\sigma^2 d\sqrt{(M+1)}} \le \frac{1}{2\sigma^2}$ and the fact that $(I - \beta(2-\beta\sigma^2)\Sigma)^M \succcurlyeq 0$ by the same condition. We also used an observation that our assumption on the contexts implies $\mathrm{tr}\,(\Sigma) \le \mathrm{tr}\left(\sigma^2 I\right) = \sigma^2 d$, so again by our condition on β it implies the final bound.

Moving on to the second term, we first note that for any $j > k$, the conditional expectation of B_j given $B_{\le k} = (B_1, B_2, \dots B_k)$ satisfies $\mathbb{E}\left[A_j \middle| B_{\le k}\right] = A_k(I - \beta\Sigma)^{j-k}$ due to conditional independence of all B_i given B_k, for $i > k$. We make use of this equality by writing

$$
\begin{aligned}
\beta^2 \sum_{k=0}^{M} \sum_{j=k+1}^{M} \mathbb{E}\left[\mathrm{tr}\,(\Sigma_t A_k \Sigma_t A_j)\right] &= \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \mathrm{tr}\,(\Sigma_t A_k \Sigma_t A_j) \middle| B_{\le k}\right]\right] \\
&= \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \mathrm{tr}\left(\Sigma_t A_k \Sigma_t A_j (I - \beta\Sigma_t)^{j-k}\right) \middle| B_{\le k}\right]\right] \\
&= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma_t A_k \Sigma_t A_k \Sigma_t^{-1}\left(I - (I - \beta\Sigma_t)^{M-k}\right)\right) \middle| B_{\le k}\right]\right] \\
&\le \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma_t A_k \Sigma_t A_k \Sigma_t^{-1}\right) \middle| B_{\le k}\right]\right] \\
&\qquad \text{(due to } (I - \beta\Sigma_t)^{M-k} \succcurlyeq 0\text{)} \\
&= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathrm{tr}\,(A_k \Sigma A_k)\right] \\
&\le \beta \sum_{k=0}^{M} \mathrm{tr}\left(\Sigma\,(I - 2\beta\Sigma)^k\right) \\
&\quad + \beta \sum_{j=1}^{d} \sum_{k=0}^{M} \lambda_j(\Sigma) \frac{\beta\sigma^2 \mathrm{tr}\,(\Sigma)}{2\lambda_j(\Sigma)} \exp\left(\beta^2\sigma^2 \mathrm{tr}\,(\Sigma)\, k + 2\beta\lambda_j(\Sigma)\right) \\
&\quad \text{(applying Lemma 20 with } \widetilde{H} = I\text{)} \\
&\le d + \frac{\beta^2\sigma^2 \mathrm{tr}\,(\Sigma)}{2} \sum_{j=1}^{d} \frac{\exp\left(\beta^2\sigma^2 \mathrm{tr}\,(\Sigma)\,(M+1) + 2\beta\lambda_j(\Sigma)\right) - 1}{\exp\left(\beta^2\sigma^2 \mathrm{tr}\,(\Sigma) + 2\beta\lambda_j(\Sigma)\right) - 1}
\end{aligned}
$$

$$\leq d + \frac{1}{2}\sum_{j=1}^{d} \exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma\right)\left(M+1\right) + 2\beta\lambda_j(\Sigma)\right)$$

$$\leq d + \frac{1}{2}\sum_{j=1}^{d} \exp\left(\beta^2\sigma^4 d(M+1) + 2\beta\lambda_j(\Sigma)\right) \leq 5d.$$

where the last line again used the condition $\beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}} \leq \frac{1}{2\sigma^2}$ and $(I - \beta(2-\beta\sigma^2)\Sigma)^M \succcurlyeq 0$. The proof of the theorem is concluded by putting everything together. □

3.5 Algorithms for contextual learning with full information

Clearly, our algorithm LINEXP3 can be simply adapted to simpler settings where the learner gets more feedback about the loss functions ℓ_t chosen by the adversary. In this section, we show results for two such natural settings: one where the learner observes the *entire* loss function ℓ_t, and one where the learner observes the losses $\ell_t(X_t, a)$ for each action a. We refer to the first of these observation models as *counterfactual feedback* and call the second one *full-information feedback*. We describe two variants of our algorithm for these settings and give their performance guarantees below. Both results will hold for general nonlinear losses taking values in $[0, 1]$.

In case of counterfactual feedback, we can modify our algorithm so that, in each round t, it computes the weights $w_{t,a}(X_t) = \exp\left(-\eta\sum_{k=1}^{t-1}\ell_k(X_t, a)\right)$ for each action, and then plays action $A_t = a$ with probability proportional to the obtained weight. Using our general analytic tools, this algorithm can be easily shown to achieve the following guarantee:

Proposition 1. *For any $\eta > 0$, the regret of the algorithm described above for counterfactual feedback satisfies*

$$R_T \leq \frac{\log K}{\eta} + \frac{\eta T}{8}.$$

Setting $\eta = \sqrt{\frac{8\log K}{T}}$, the regret also satisfies $R_T \leq \sqrt{(T/2)\log K}$.

Notably, this bound does not depend at all on the dimension of the context space, the complexity of the policy class, or any property of the loss function, and only shows dependence on the number of actions K. The caveat is of course that the counterfactual model provides the learner with a level of feedback that is entirely unrealistic in any practical setting: it requires the ability to evaluate all past loss functions at *any* context-action pair.

The full-information setting is arguably much more realistic in that it only requires evaluating the losses corresponding to the observed context X_t, which which is typically the case in online classification problems. For this setting, we use our LinExp3 algorithm with the loss estimator defined for each action a as

$$\widehat{\ell}_{t,a} = \Sigma^{-1} X_t \ell_t(X_t, a).$$

Using our analysis, we can show that the bias of this estimator is uniformly bounded by $\varepsilon\sqrt{d}$ (cf. Equation 3.7). The following bound is then easy to prove by following the same steps as in Section 3.4.1:

Proposition 2. *For any positive $\eta \leq \frac{\lambda_{\min}}{\sigma^2}$, the regret of the algorithm described above for full-information feedback*

$$R_T \leq \frac{\log K}{\eta} + \eta d T + \varepsilon\sqrt{d}T.$$

Setting $\eta = \sqrt{\frac{d \log K}{T}}$, the regret also satisfies $R_T \leq 2\sqrt{dT \log K} + \varepsilon\sqrt{d}T$ for large enough T.

As expected, this bound scales with the dimension as $\sqrt{d}$ due to the fact that the algorithm has to "estimate" d, parameters, as opposed to the Kd parameters that need to be learned in the contextual bandit problem we consider in the main text. We also note that this online learning setting is closely related to that of prediction with expert advice, with the set of experts being the class of linear-classifier policies [39]. As a result, it is possible to make use of this framework by running any online prediction algorithm on a finely discretized set of policies, resulting in a regret bound of order $\sqrt{dT \log(KT)}$. Our result above improves on this by a logarithmic factor of T, while being efficient to implement.

3.6 Efficient implementation of MGR

The naïve implementation of the MGR procedure presented in the main text requires $O(MKd + Kd^2)$ time due to the matrix-matrix multiplications involved.

In this section we explain how to compute $\widehat{\ell}_t(x,a) = \left\langle x, \widetilde{\theta}_{t,a} \right\rangle$ in $O(MKd)$ time, exploiting the fact that the matrices $\widehat{\Sigma}_{t,a}$ never actually need to be computed, since the algorithm only works with products of the form $\widehat{\Sigma}_{t,a} X_t$ for a fixed vector X_t. This motivates the following procedure:

```
Fast Matrix Geometric Resampling
Input: context vector $x$, data distribution $\mathcal{D}$, policy $\pi_t$.
Initialization: Compute $Y_{0,a} = Ix$.
For $k = 1, \dots, M$, repeat:
    1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
    2. if $a = A(k)$, set
       $Y_{k,a} = Y_{k-1,a} - \beta \langle Y_{k-1,a}, X(k) \rangle X(k)$,
    3. otherwise, set $Y_{k,a} = Y_{k-1,a}$.
Return $q_{t,a} = \beta Y_{0,a} + \beta \sum_{k=1}^{M} Y_{k,a}$.
```

It is easy to see from the above procedure that each iteration k can be computed using $(K+1)d$ vector-vector multiplications: sampling each action $A(k)$ takes Kd time due to having to compute the products $\left\langle X(k), \widehat{\Theta}_{t,a} \right\rangle$ for each action a, and updating $Y_{k,a}$ can be done by computing the product $\langle Y_{k-1,a}, X(k) \rangle$. Overall, this results in a total runtime of order MKd as promised above.

3.7 Kernel methods

3.7.1 Preliminaries

Kernel method is a natural technique to extend linear models to non-linear, which allows to cover quite broad class of loss functions. There is a huge interest in applying kernel methods into sequential learning setting, with a numerous works on analysis of kernel-based bandit problems [132, 85, 129, 47, 130, 139]. In the previous section, we considered a contextual bandit problem, where the loss function is assumed to be an unknown linear function, parameterized by a finite dimensional vector. Now we generalize this setting, we consider a contextual bandit problem, where loss function is assumed to lie in a reproducing kernel Hilbert space (RKHS) with bounded RKHS norm. Such functions can be represented in a linear space, but for many kernels the representation has large dimension or even infinite. Therefore methods of the previous chapter are not applicable in this setting, which calls for the new algorithms that can obtain a regret that scales with a quantity that depends on the kernel function, rather then on the dimension of the loss function in the linear representation. Our regret bounds come in terms of a quantity of effective dimension, which is often used to generalize the results from parametric problems to non-parametric methods.

The problem we consider in this chapter sounds as following. Let context vector X be a random variable, sampled from a probability distribution $\mathcal{D}$ and taking values in a compact set $\mathcal{S}$. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with an inner product denoted as $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and with a kernel $\kappa : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. Then, for any $f \in \mathcal{H}$, $f(x) = \langle \kappa(x, \cdot), f(\cdot) \rangle_{\mathcal{H}}$, $\forall x \in \mathcal{S}, \forall f \in \mathcal{H}$ and we define a feature map $\varphi : \mathcal{S} \to \mathcal{H}$, such that $\varphi(x) = \kappa(\cdot, x)$, $\forall x \in \mathcal{S}$. At the beginning of round t, the learner observes the context vector $X_t \sim \mathcal{D}$ and simultaneously the adversary chooses the loss function $f_{t,a}(\cdot) \in \mathcal{H}$, so the loss corresponding to the action a is $\ell(X_t, a) = \langle \kappa(\cdot, X_t), f_{t,a} \rangle_{\mathcal{H}} = f_{t,a}(X_t)$. We assume that contexts and loss functions are bounded in the norm of $\mathcal{H}$, so $\kappa(X, X) \leq \sigma^2, \forall X \in \mathcal{S}$ and $||\theta_{t,a}||_{\mathcal{H}} \leq R, \forall t \in [T], a \in \mathcal{A}$. We define the covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ as $\Sigma = \mathbb{E}\left[\varphi(X)\varphi(X)^{\mathsf{T}}\right]$, where $\varphi(X)\varphi(X)^{\mathsf{T}}$ is a Hilbert space outer product.We assume that the eigenvalues of Σ, $\lambda_1 \geq \lambda_2 \geq \ldots$ are decaying in a way that there exists a positive number C such that $\sum_i \lambda_i \leq C$.

3.7.2 Algorithm and main result

We begin by describing our algorithm, which is an adaptation of LINEXP3. It uses an estimator, that comes from the modifying the idea of Matrix Geometric Resampling from previous section and constructs an estimation of $f_{t,a}$ through the following procedure:

```
**Kernel Geometric Resampling**
**Input:** Context $X_t$, data distribution $\mathcal{D}$.
**Initialization:** Set $Y_0 = \Phi(X_t)$.
**For** $k = 1, \dots, M$, **repeat**:
    1. Draw $X(k) \sim \mathcal{D}$ and $A(k) \sim \pi_t(\cdot|X(k))$,
    2. set $Y_k = Y_{k-1} - \beta \langle Y_{k-1}, \varphi(X(k)) \rangle_{\mathcal{H}} \varphi(X(k))$,
**Return** $q_t = \beta Y_0 + \beta \sum_{k=1}^{M} Y_k$.
```

Then, estimator of $f_{t,a}$ can be written as

$$\widetilde{\theta}_{t,a} = \frac{1}{\pi(a|X_t)} q_t \ell_t(X_t, A_t) \mathbb{I}_{\{A_t=a\}}.$$

Notice, that due to the form of Kernel Geometric Resampling, q_t is a linear combination of $\varphi(X_t(k)), k = 1, \dots, M$, so q_t can be written as $q_t = \sum_{k=1}^{M} c_{t,k} \varphi(X_t(k))$. Then for a given context X, it is possible to compute the value of $\widetilde{\theta}_{t,a}$ on X, which is $\left\langle \varphi(X), \widetilde{\theta}_{t,a} \right\rangle_{\mathcal{H}} = \ell_t(X_t, A_t) \mathbb{I}_{\{A_t=a\}} \sum_{k=1}^{M} c_{t,k} \kappa(X, X_t(k))$, without having to hold in memory Y_k, which can be an infinite-dimension objects. Let $B_k = X(k)X(k)^\mathsf{T}$, $A_k = \prod_{j=1}^{k}(I - \beta B_k)$ and $\widehat{\Sigma}^+ = \beta I + \beta \sum_{k=1}^{M} A_k$, then $\widetilde{\theta}_{t,a}$ can be expressed as

$$\widetilde{\theta}_{t,a} = \frac{1}{\pi(a|X_t)} \widehat{\Sigma}^+ X_t \ell_t(X_t, A_t) \mathbb{I}_{\{A_t=a\}}. \tag{3.13}$$

Now we are ready to present the algorithm KERLINEXP3, which is presented as Algorithm 3.2. An important quantity in the performance guarantee that we present shortly is an effective dimension

$$d_{\text{eff}}(\Sigma, \lambda) = \text{tr}\left((\Sigma + \lambda I)^{-1} \Sigma \right).$$

Algorithm 3.2 KERLINEXP3

```
Parameters: Learning rate $\eta > 0$, exploration parameter $\gamma \in (0,1)$, $\Sigma$
Initialization: Set $\theta_{0,i} = 0$ for all $i \in [K]$.
For $t = 1, \dots, T$, repeat:
    1. Observe $X_t$ and, for all $a$, set
           $w_t(X_t, a) = \exp\left(-\eta \sum_{s=0}^{t-1} \langle X_t, \widehat{\theta}_{s,a} \rangle\right)$,
    2. draw $A_t$ from the policy defined as
           $\pi_t(a|X_t) = (1-\gamma)\frac{w_t(X_t,a)}{\sum_{a'} w_t(X_t,a')} + \frac{\gamma}{K}$,
    3. observe the loss $\ell_t(X_t, A_t)$ and compute $\widehat{\theta}_{t,a}$ for all $a$.
```

An effective dimension was introduced by Zhang [151], and for the finite dimension space it can be bounded by dimensionality of the space. Our main result regarding the performance of KERLINEXP3 is the following:

Theorem 3.7.1. *For any positive $\eta \leq \frac{2}{M+1}\beta\sigma^2$ and for any $\gamma \in (0,1)$ the expected regret of* KERLINEXP3 *satisfies*

$$R_T \leq 2T\frac{R}{\beta M} + 2\gamma T + \eta\frac{\gamma}{K}\left(3 + 4C + d_{eff}\left(\Sigma, \frac{1}{\beta M}\right)\right)T + \frac{\log K}{\eta}.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2\sqrt{C(M+1)}}$, $M = \left\lceil \frac{K^2\sigma^4 C \log^2(T\sigma^2R^2)}{\gamma^2\lambda_{\min}^2} \right\rceil$, $\eta = T^{-2/3}(KC)^{-1/3} \cdot (\log K)^{2/3}$, $\gamma = T^{-1/3}(KC\log K)^{1/3}$ and supposing that T is large enough so that $\eta \leq \frac{2}{M+1}\beta\sigma^2$ holds, the expected regret of KERLINEXP3 *satisfies*

$$R_T = \mathcal{O}\left(T^{2/3}(KC\log K)^{1/3}\right).$$

3.7.3 Analysis

For the ease of readability, later we will write X for $\varphi(X)$, $\langle\cdot,\cdot\rangle$ for $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|$ for $\|\cdot\|_{\mathcal{H}}$. For the analysis of LINEXP3 in the kernel setting, we adapt the techniques for the finite-dimensional contextual bandits. We can prove the following useful bound on the bias of the estimator:

Lemma 21. $\left|\mathbb{E}_t\left[\langle X_t, \theta_{t,a} - \widetilde{\theta}_{t,a}\rangle\right]\right| \leq \frac{R}{\beta M}$.

Proof. We start our analysis from the bound on the estimation of the loss. We can see that our estimator satisfies

$$\begin{aligned}\mathbb{E}_t\left[X_0^\mathsf{T}\widetilde{\theta}_{t,a}\mathbb{I}_{\{A_t=a\}}\right] &= \mathbb{E}\left[X_0^\mathsf{T}\theta_{t,a}\right] + \mathbb{E}_t\left[X_0^\mathsf{T}\widehat{\Sigma}^+ X_t X_t^\mathsf{T}\theta_{t,a}\mathbb{I}_{\{A_t=a\}}\right] \\ &= \mathbb{E}\left[X_0^\mathsf{T}\theta_{t,a}\right] + \mathbb{E}_t\left[X_0^\mathsf{T}\widehat{\Sigma}^+ X_t X_t^\mathsf{T}\theta_{t,a}(\Sigma + \frac{1}{\beta M})(\Sigma + \frac{1}{\beta M})^{-1}\mathbb{I}_{\{A_t=a\}}\right] \\ &\leq \mathbb{E}\left[X_0^\mathsf{T}\theta_{t,a}\right] + \left\|(1-\beta\Sigma)\left(\Sigma + \frac{1}{\beta M}\right)\right\|\left\|X_0^\mathsf{T}\left(\Sigma + \frac{1}{\beta M}I\right)^{-1}\theta_{t,a}\right\|.\end{aligned}$$

The first term in the bias can be bounded as

$$\left\|(1-\beta\Sigma)\left(\Sigma + \frac{1}{\beta M}I\right)\right\| \leq \frac{1}{\beta M}.$$

Second term can be bounded as following:

$$\begin{aligned}\mathbb{E}\left[\left\|X_0^\mathsf{T}\left(\Sigma + \frac{1}{\beta M}I\right)^{-1}\theta_{t,a}\right\|\right] &= \mathbb{E}\left[\left\|X_0X_0^\mathsf{T}e_1\left(\Sigma + \frac{1}{\beta M}I\right)^{-1}\theta_{t,a}\right\|\right] \\ &\leq \left\|\Sigma\left(\Sigma + \frac{1}{\beta M}I\right)^{-1}\right\|\left\|\theta_{t,a}\right\|\left\|e_1\right\| \leq R.\end{aligned}$$

Summing all bias terms, we get

$$\mathbb{E}_t\left[X_0^\mathsf{T}\widetilde{\theta}_{t,a}\mathbb{I}_{\{A_t=a\}}\right] \leq \mathbb{E}_t\left[X_0^\mathsf{T}\theta_{t,a}\right] + \frac{R}{\beta M}.$$

□

The other key term in the regret bound is bounded in the following lemma:

Lemma 22. *Suppose that X_t is satisfying $\|X_t\| \leq \sigma$, $0 < \beta \leq \frac{1}{2(\sigma^2+\lambda)}$, tr $(\Sigma) \leq C$ and $M > 0$. Then for each t,* RealLinExp3 *guarantees*

$$\mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_0)\langle X_0, \widetilde{\theta}_{t,a}\rangle^2\right] \leq \frac{\gamma}{K}\left(3 + 4C + d_{eff}\left(\Sigma, \frac{1}{\beta M}\right)\right).$$

Proof. We start by plugging in the definition of $\widetilde{\theta}_{t,a}$ and writing

$$\begin{aligned}\mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_0)\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle^2\right] &= \mathbb{E}_t\left[\sum_{a=1}^K \frac{\pi_t(a|X_0)}{\pi_t(a|X_t)^2}\left(X_0^\mathsf{T}\Sigma^+ X_t X_t^\mathsf{T}\theta_{t,a}\mathbb{I}_{\{A_t=a\}}\right)^2\right] \\ &\leq \mathbb{E}_t\left[\sum_{a=1}^K \text{tr}\left(\pi_t(a|X_0)\frac{K}{\gamma}\Sigma\Sigma^+\Sigma\Sigma^+\right)\right],\end{aligned}$$

where we used $\left\langle X_0, \theta_{t,a}\right\rangle \leq 1$ in the inequality. Using the definition of Σ^+ and elementary manipulations, we can get

$$\begin{aligned}\mathbb{E}\left[\text{tr}\left(\Sigma\Sigma_t^+\Sigma\Sigma_t^+\right)\right] &= \mathbb{E}\left[\beta^2\cdot\text{tr}\left(\Sigma\left(\sum_{k=0}^M A_k\right)\Sigma\left(\sum_{j=0}^M A_j\right)\right)\right] \\ &= \beta^2\sum_{k=0}^M\sum_{j=0}^M \text{tr}\left(\mathbb{E}\left[\Sigma_t A_k \Sigma_t A_j\right]\right) = \beta^2\sum_{k=0}^M \text{tr}\left(\mathbb{E}\left[\Sigma A_k \Sigma A_k\right]\right) \\ &+ 2\beta^2\sum_{k=0}^M\sum_{j=k+1}^M \text{tr}\left(\mathbb{E}\left[\Sigma A_k\Sigma A_j\right]\right).\end{aligned}$$

Let us first address the first term on the right hand side. Recalling the definition $A_k = \prod_{j=1}^k B_j$ and adapting the result of Lemma 20 to the infinite-dimentional case, with $\widetilde{H} = \Sigma$, we can obtain

$$\begin{aligned}\beta^2\sum_{k=0}^M \text{tr}\left(\mathbb{E}\left[\Sigma_t A_k\Sigma_t A_k\right]\right) &\leq \beta^2\sum_{k=0}^M \text{tr}\left(\Sigma^2\left(I-2\beta\Sigma\right)^k\right) \\ &+ \beta^2\sum_{j=1}^\infty\sum_{k=0}^M \lambda_j^2(\Sigma)\frac{\beta\sigma^2\text{tr}\left(\Sigma\right)}{2\lambda_j(\Sigma)}\exp\left(\beta^2\sigma^2\text{tr}\left(\Sigma\right)k + 2\beta\lambda_j(\Sigma)\right) \\ &= \beta\text{tr}\left(\Sigma(I-(I-\beta\Sigma)^M)\right) \\ &+ \beta^3\frac{\sigma^2\text{tr}\left(\Sigma\right)}{2}\sum_{j=1}^\infty \lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\frac{\exp\left(\beta^2\sigma^2\text{tr}\left(\Sigma\right)(M+1)\right)-1}{\exp\left(\beta^2\sigma^2\text{tr}\left(\Sigma\right)\right)-1} \\ &\leq \beta\text{tr}\left(\Sigma\right) + \beta^3\frac{\sigma^2\text{tr}\left(\Sigma\right)}{2\beta^2\sigma^2\text{tr}\left(\Sigma\right)}\sum_{j=1}^\infty \lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\exp\left(\beta^2\sigma^2\text{tr}\left(\Sigma\right)(M+1)\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \beta \mathrm{tr}\left(\Sigma\right) + \frac{\beta}{2}\sum_{j=1}^{\infty} \lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\exp\left(\beta^2\sigma^2\mathrm{tr}\left(\Sigma\right)(M+1)\right) \\
&\leq \beta C + \frac{\beta}{2} C\exp(2\beta\sigma^2)\exp\left(\beta^2\sigma^2 C(M+1)\right) \leq 3,
\end{aligned}
$$

As it was done for the proof of Lemma 19, we are moving to the second term

$$
\begin{aligned}
\beta^2\sum_{k=0}^{M}\sum_{j=k+1}^{M}\mathbb{E}\left[\mathrm{tr}\left(\Sigma A_k\Sigma A_j\right)\right] &= \beta^2\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M}\mathrm{tr}\left(\Sigma A_k \Sigma A_j\right)\middle| B_{\leq k}\right]\right] \\
&= \beta^2\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M}\mathrm{tr}\left(\Sigma A_k\Sigma A_j(I-\beta\Sigma)^{j-k}\right)\middle| B_{\leq k}\right]\right] \\
&= \beta\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma_t A_k\Sigma A_k\Sigma^{-1}\left(I-(I-\beta\Sigma)^{M-k}\right)\right)\middle| B_{\leq k}\right]\right] \\
&\leq \beta\sum_{k=0}^{M}\mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma A_k\Sigma A_k\Sigma^{-1}\right)\middle| B_{\leq k}\right]\right] \\
&\quad \text{(due to } (I-\beta\Sigma)^{M-k} \succcurlyeq 0\text{)} \\
&= \beta\sum_{k=0}^{M}\mathbb{E}\left[\mathrm{tr}\left(A_k\Sigma A_k\right)\right] \leq \beta\sum_{k=0}^{M}\mathrm{tr}\left(\Sigma\left(I-2\beta\Sigma\right)^{k}\right) \\
&+\beta\sum_{j=1}^{\infty}\sum_{k=0}^{M}\lambda_j(\Sigma)\exp\left(\beta^2\sigma^2\mathrm{tr}\left(\Sigma\right)k + 2\beta\lambda_j(\Sigma)\right) \\
&\quad \text{(applying Lemma 20 with } \widetilde{H} = I \text{)} \\
&\leq \mathbb{E}\left[\mathrm{tr}\left(\Sigma\left(\Sigma+\frac{1}{\beta M}I\right)^{-1}\right)\right] + \sum_{j=1}^{\infty}\lambda_j(\Sigma)\exp\left(\beta^2\sigma^2\mathrm{tr}\left(\Sigma\right)(M+1)+2\beta\lambda_j(\Sigma)\right) \\
&\text{(by the same argument as in [35, Proof of Theorem 5])} \\
&\leq d_{eff}\left(\Sigma, \frac{1}{\beta M}\right) + 4C.
\end{aligned}
$$

This results to the final bound. □

As a final step, we need to verify that $\eta\left|\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle\right| < 1$ holds. To this end, notice that

$$\begin{aligned}\eta \cdot \left|\left\langle X_0, \widetilde{\theta}_{t,a,\lambda}\right\rangle\right| &= \eta \cdot \left|X_0^\mathsf{T} \widehat{\Sigma}_{t,a,\lambda}^+ X_t \left\langle X_t, \theta_{t,a}\right\rangle \mathbb{I}_{\{A_t=a\}}\right| \leq \eta \cdot \left|X_0^\mathsf{T} \widehat{\Sigma}_{t,a,\lambda}^+ X_t\right| \\ &\leq \eta\sigma^2 \left\|\widehat{\Sigma}_{t,a,\lambda}^+\right\|_{\text{op}} \leq \eta\sigma^2\beta \left(1 + \sum_{k=1}^{M} \|A_{k,a}\|_{\text{op}}\right) \leq \eta\sigma^2\beta(M+1)/2,\end{aligned}$$

where we used the fact that our choice of β ensures that $\|A_{k,a}\|_{\text{op}} = \left\|\prod_{j=0}^{k}(I - \beta B_{j,a})\right\|_{\text{op}} \leq 1$. Then we have to take $\eta \leq \frac{2}{M+1}\beta\sigma^2$. The expected regret of REALLINEXP3 satisfies

$$R_T \leq 2T\frac{R}{\beta M} + 2\gamma T + \eta\frac{\gamma}{K}\left(3 + 4C + d_{eff}\left(\Sigma, \frac{1}{\beta M}\right)\right)T + \frac{\log K}{\eta}.$$

Chapter 4

Learning in episodic MDP

4.1 Introduction

We study the problem of online learning in episodic Markov Decision Processes (MDP), modelling a sequential decision making problem where the interaction between a learner and its environment is divided into T episodes of fixed length H. At each time step of the episode, the learner observes the current state of the environment, chooses one of the available actions, and earns a reward. Consequently, the state of the environment changes according to the transition function of the underlying MDP, as a function of the previous state and the action taken by the learner. A key distinguishing feature of our setting is that we assume that the reward function can change arbitrarily between episodes, and the learner only has access to bandit feedback: instead of being able to observe the reward function at the end of the episode, the learner only gets to observe the rewards that it actually received. As traditional in this line of work, we aim to design algorithms for the learner with theoretical guarantees on her regret, which is the difference between the total reward accumulated by the learner and the total reward of the best stationary policy fixed in hindsight.

We allow the state space to be very large and aim to prove performance guarantees that do not depend on the size of the state space, bringing theory one step closer to practical scenarios where assuming finite state spaces is unrealistic. To address the challenge of learning in large state spaces, we adopt the classic RL technique of using *linear function approximation* and suppose that we have access to a relatively low-dimensional feature map that can be used to represent policies and value functions. We will assume that the feature map is expressive enough so that all action-value functions can be expressed as linear functions of the features. We study two types of assumptions on the knowledge of the learner about the environment. First, we assume that the learner only has an access to the simulator of the dynamics of the MDP that will allow the learner to generate sample episodes. Then we strength this assumption and consider the case when the learner has full knowledge of the transition function of the MDP. More specifically, our main contributions are:

- Under the assumption that the learner has an access to the simulator of MDP, we design a computationally efficient algorithm called MDP-LinExp3 and prove that its regret is at most $\widetilde{\mathcal{O}}\left(H^2 T^{2/3}(d \log K)^{1/3}\right)$, where K is the number of actions and d is the dimensionality of the feature map. We also show that this bound can be improved to $\widetilde{\mathcal{O}}\left(H^2\sqrt{dT \log K}\right)$ under the additional strong condition that the likelihood ratio between the state distributions generated by any pair of policies is upper bounded by a constant. Notably, both these bounds show only logarithmic dependence on the number of actions K, guaranteeing that they remain meaningful even for very large action spaces.

- Our second contribution is an algorithm called Online Q-REPS, for which, under the assumption that the learner has the perfect knowledge of MDP dynamics, show that the regret is at most $\mathcal{O}\left(\sqrt{dHTD\left(\mu^*\|\mu_0\right)}\right)$, where $D\left(\mu^*\|\mu_0\right)$ is the relative entropy between the state-action distribution μ^* induced by the optimal policy and an initial distribution μ_0 given as input to the algorithm. Notably, our results do not require the likelihood ratio between these distributions to be uniformly bounded, and the bound shows no dependence on the eigenvalues of the feature covariance matrices. Our algorithm itself requires solving a d^2-dimensional convex optimization problem at the beginning of each episode, which can be solved to arbitrary precision ε in time polynomial in d and $1/\varepsilon$, independently of the size of the state-action space.

Our work fits into a long line of research considering online learning in Markov decision processes. The problem of regret minimization in stationary MDPs with a *fixed* reward function has been studied extensively since the work of Burnetas and Katehakis [32], Auer and Ortner [16], Tewari and Bartlett [136], Jaksch et al. [69], with several important advances made in the past decade

[53, 54, 19, 64, 71]. While most of these works considered small finite state spaces, the same techniques have been very recently extended to accommodate infinite state spaces under the assumption of realizable function approximation by Jin et al. [73] and Yang and Wang [149]. In particular, the notion of *linear MDPs* introduced by Jin et al. [73] has become a standard model for linear function approximation and has been used in several recent works (e.g., 115, 146, 6).

Even more relevant is the line of work considering adversarial rewards, initiated by Even-Dar et al. [58], who consider online learning in continuing MDPs with full feedback about the rewards. They proposed a MDP-E algorithm, that achieves $\mathcal{O}(\tau^2\sqrt{T\log K})$ regret, where τ is an upper bound on the mixing time of the MDP. Later, Neu et al. [113] proposed an algorithm which guarantees $\widetilde{\mathcal{O}}\big(\sqrt{\tau^3 KT/\alpha}\big)$ regret with bandit feedback, essentially assuming that all states are reachable with probability $\alpha > 0$ under all policies. In our work, we focus on episodic MDPs with a fixed episode length H. The setting was first considered in the bandit setting by Neu et al. [111], who proposed an algorithm with a regret bound of $\mathcal{O}(H^2\sqrt{TK}/\alpha)$. Although the number of states does not appear explicitly in the bound, the regret scales at least linearly with the size of the state space $\mathcal{S}$, since $|\mathcal{S}| \leq H/\alpha$. Later work by Zimin and Neu [152], Dick et al. [55] eliminated the dependence on α and proposed an algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{TH|\mathcal{S}|K})$ regret. Regret bounds for the full-information case without prior knowledge of the MDP were achieved by Neu et al. [112] and Rosenberg and Mansour [126], of order $\widetilde{\mathcal{O}}(H|\mathcal{S}|K\sqrt{T})$ and $\widetilde{\mathcal{O}}(H|\mathcal{S}|\sqrt{KT})$, respectively. These results were recently extended to handle bandit feedback about the rewards by Jin et al. [72], ultimately resulting in a regret bound of $\widetilde{\mathcal{O}}(H|\mathcal{S}|\sqrt{KT})$.

The adversarial rewards setting with infinite number of states was first considered in the full-information case by Cai et al. [33]. This work proposes the algorithm OPPO, that is guaranteed to achieve $\widetilde{\mathcal{O}}\big(\sqrt{d^3H^3T}\big)$. It is assumed there that the learner has access to d-dimensional features that can perfectly represent all action-value functions, without assuming any prior knowledge of the MDP parameters.

The follow up work of [101] covers the same setting with the results presented in this chapter. It is assumed there that the learner only has access to the simulator of dynamics of the MDP. Under this assumption, they propose an algorithm, with a regret at most $\widetilde{O}(T^{2/3})$. This result is largely based on the analysis of MDP-LinExp3, with an improvement of eliminating the assumption on the lower bound of the covariance matrix of the states induced by an exploratory policy.

Our results are made possible by a careful combination of recently proposed techniques for contextual bandit problems and optimal control in Markov decision processes. In particular, a core component of our algorithm is a regularized linear programming formulation of optimal control in MDPs due to Bas-Serrano et al. [23], which allows us to reduce the task of computing near-optimal policies in linear MDPs to a low-dimensional convex optimization problem. A similar algorithm design has been previously used for tabular MDPs by Zimin and Neu [152], Dick et al. [55], with the purpose of removing factors of $1/\alpha$ from the previous state-of-the-art bounds of Neu et al. [111]. Analogously to this improvement, our methodology enables us to make strong assumptions on problem-dependent constants like likelihood ratios between μ^* and μ_0 or eigenvalues of the feature covariance matrices. Another important building block of our method is a version of the recently proposed Matrix Geometric Resampling procedure of Neu and Olkhovskaya [114] that enables us to efficiently estimate the reward functions. Incorporating these estimators in the algorithmic template of Bas-Serrano et al. [23] is far from straightforward and requires several subtle adjustments.

4.2 Problem setting

We already defined a general framework of episodic Markovian Decision Process, that is denoted as $M = (\mathcal{S}, \mathcal{A}, H, P, r)$, in Section 1.1. In our work, we assume that both $\mathcal{S}$ and $\mathcal{A}$ are finite sets, although we allow the state space $\mathcal{S}$ to be arbitrarily large. Without significant loss of generality, we will assume that the set of available actions is the same $\mathcal{A}$ in each state, with cardinality $|\mathcal{A}| = K$. In this work, we consider an *online learning* problem where the learner interacts with its environment in a sequence of episodes $t = 1, 2, \dots, T$, facing a different *reward functions* $r_{t,1}, \dots r_{t,H+1}$ selected by a (possibly adaptive) adversary at the beginning of each episode t. Oblivious to the reward function chosen by the adversary, at the beginning of each episode the the learner chooses the policy π_t and performs the consecutive steps over the layers of the MDP following the policy π_t. The objective of the learner is selecting a sequence of policies π_t for each episode t in a way that it minimizes the total expected regret, that we defined in 1.1. It follows from standard results that there exists a stationary and deterministic policy π^* that achieves the corresponding supremum [122, Theorem 4.4.2]. Intuitively, the regret measures the gap between the total reward gained by the learner and that of the best stationary policy fixed in hindsight, with full knowledge of the sequence of rewards chosen by the adversary. This performance measure is standard in the related literature on online learning in MDPs, see, for example Neu et al. [111], Zimin and Neu [152], Neu et al. [112], Rosenberg and Mansour [126], Cai et al. [33].

In our work, we focus on MDPs with potentially enormous state spaces, which makes it difficult to design computationally tractable algorithms with nontrivial guarantees, unless we make some assumptions. We particularly focus on the classic technique of relying on *linear function approximation* and assuming that the reward functions occurring during the learning process can be written as a linear function of a low-dimensional feature map. We specify the form of function approximation and the conditions our analysis requires as follows:

Assumption 3 (Linear MDP with adversarial rewards)**.** *There exists a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and a collection of d signed measures $m = (m_1, \ldots, m_d)$ on $\mathcal{S}$, such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$ the transition function can be written as*

$$P(\cdot|x, a) = \langle m(\cdot), \varphi(x, a) \rangle .$$

Furthermore, the reward function chosen by the adversary in each episode t can be written as

$$r_{t,h}(x, a) = \langle \theta_{t,h}, \varphi(x, a) \rangle$$

for some $\theta_{t,h} \in \mathbb{R}^d$. We assume that the features and the parameter vectors satisfy $\|\varphi(x, a)\| \leq \sigma$ and that the first coordinate $\varphi_1(x, a) = 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$. Also we assume that $\|\theta_{t,h}\| \leq R$.

Online learning under this assumption, but with a fixed reward function, has received substantial attention in the recent literature, particularly since the work of Jin et al. [73] who popularized the term "Linear MDP" to refer to this class of MDPs. This has quickly become a common assumption for studying reinforcement learning algorithms (Cai et al. [33], Jin et al. [73], Neu and Pike-Burke [115], Agarwal et al. [6]). This is also a special case of *factored linear models* (Yao et al. [150], Pires and Szepesvári [121]).

4.2.1 Algorithm and reward estimator

This section provides a template of the strategy for the learner, acting in episodic MDP, and an efficient estimator of the reward vectors $\theta_{t,h}$ based on the estimator, presented in Chapter 3. The template of the algorithm is presented as Algorithm 4.1.

Algorithm 4.1 Decision making in episodic MDP

```
**Initialization:** Set $\widehat{\theta}_{1,h} = 0$ for all $h$.
**For** $t = 1, \dots, T$, **repeat:**
    • **For** $h = 1, \dots, H$, **do:**
        – Observe $X_{t,h}$ and compute $\pi_{t,h}(a|X_{t,h})$ for all $a \in \mathcal{A}(X_{t,h})$,
        – Draw $A_{t,h} \sim \pi_{0,h}(\cdot|X_{t,h})$,
        – Observe the reward $\ell_{t,h}(X_{t,h}, A_{t,h})$.
    • Compute $\widehat{\theta}_{t,1}, \dots, \widehat{\theta}_{t,H-1}$.
```

We now turn to describing the reward estimators $\widehat{r}_{t,h}$, which will require several further definitions. Specifically, a concept of key importance will be the following *feature covariance matrix*:

$$\Sigma_{t,h} = \mathbb{E}_{\pi_t}\left[\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h}, A_{t,h})^\mathsf{T}\right].$$

Making sure that $\Sigma_{t,h}$ is invertible, we can define the estimator

$$\widetilde{\theta}_{t,h} = \Sigma_{t,h}^{-1}\varphi(X_{t,h}, A_{t,h})r_{t,h}(X_{t,h}, A_{t,h}). \tag{4.1}$$

This estimate shares many similarities with the estimates that are broadly used in the literature on adversarial linear bandits [103, 17, 52]. It is easy to see that $\widetilde{\theta}_{t,h}$ is an unbiased estimate of $\theta_{t,h}$:

$$\mathbb{E}_t\left[\widetilde{\theta}_{t,h}\right] = \mathbb{E}_t\left[\Sigma_{t,h}^{-1}\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h},, A_{t,h})^\mathsf{T}\theta_{t,h}\right] = \Sigma_{t,h}^{-1}\Sigma_{t,h}\theta_{t,h} = \theta_{t,h}.$$

Unfortunately, exact computation of $\Sigma_{t,h}$ is intractable. To address this issue, we propose a method to directly estimate the inverse of the covariance matrix $\Sigma_{t,h}$ by adapting the Matrix Geometric Resampling method that we already presented in Chapter 3 (which itself is originally inspired by the Geometric Resampling method of 109, 110). This method is using two parameters: $\beta > 0$ and $M \in \mathbb{Z}_+$, and generates an estimate of the inverse covariance matrix through the following procedure

```
Matrix Geometric Resampling
Input: simulator of $P$, policy $\widetilde{\pi}_t = (\widetilde{\pi}_{t,1}, \dots, \widetilde{\pi}_{t,H-1})$.
For $i = 1, \dots, M$, repeat:
    1. Simulate a trajectory
       $\tau(i) = \{(X_1(i), A_1(i)), \dots, (X_{H-1}(i), A_{H-1}(i))\}$,
       following the policy $\widetilde{\pi}_t$ in $P$,
    2. For $h = 1, \dots, H-1$, repeat:
       Compute
           (a) $B_{i,h} = \varphi(X_h(i), A_h(i))\varphi(X_h(i), A_h(i))^{\mathsf{T}}$,
           (b) $C_{i,h} = \prod_{j=1}^{i}(I - \beta B_{j,h})$.
Return $\widehat{\Sigma}^{+}_{t,h} = \beta I + \beta \sum_{i=1}^{M} C_{i,h}$ for all $h \in [H-1]$.
```

Based on the above procedure, we finally define our estimator as

$$\widehat{\theta}_{t,h} = \widehat{\Sigma}^{+}_{t,h}\varphi(X_{t,h}, A_{t,h})r_{t,h}(X_{t,h}, A_{t,h}).$$

The idea of the estimate is based on the truncation of the Neumann-series expansion of the matrix $\Sigma_{t,h}^{-1}$ at the Mth order term. Then, for large enough M, the matrix $\Sigma^{+}_{t,h}$ is a good estimator of the inverse covariance matrix, which will be quantified formally in the analysis.

The implementation of the MGR procedure presented above requires $O(MKHd + MHd^2)$ time due to the matrix-matrix multiplications involved. Now we explain how to compute $\widehat{\theta}_t$ in $O(MKHd)$ time, exploiting the fact that the matrices $\widehat{\Sigma}_{t,h}$ never actually need to be computed, since the algorithm only works with products of the form $\widehat{\Sigma}_{t,h}\varphi(X_{t,h}, A_{t,h})$ for vectors $X_{t,h}$, $h \in [H]$. This motivates the procedure of Fast Matrix Geometric Resampling. It is easy to see from the procedure of Fast Matrix Geometric Resampling that each iteration k can be computed using $(K+1)Hd$ vector-vector multiplications: sampling each action $A_h(k)$ takes Kd time due to having to compute the products $\left\langle \varphi(X_h(k)), \sum_{s=1}^{t-1} \widehat{\theta}_{s,a,h} \right\rangle$ for each action a, and updating $Y_{k,h}$ can be done by computing the product $\langle Y_{k-1,h}, \varphi(X_h(k)) \rangle$. Overall, this results in a total runtime of order $MKHd$ as promised above.

```
Fast Matrix Geometric Resampling
Input: simulator of transition function $P$, policy $\pi_t$
Initialization: Compute $Y_{0,h} = \varphi(x_h)$ for all $h \in [H]$.
For $k = 1, \dots, M$, repeat:
    1. Generate a path $U(i) = \{(X_1(i), A_1(i)), \dots, (X_H(i), A_H(i))\}$,
       following the policy $\pi_t$ in the simulator of $P$,
    2. For $h = 1, \dots, H$, repeat:
        (a) if $A_h(k) = a_h$, set
            $Y_{k,h} = Y_{k-1,h} - \beta \langle Y_{k-1,h}, \varphi(X_h(k), A_h(k)) \rangle \varphi(X_h(k), A_h(k))$,
        (b) otherwise, set $Y_{k,h} = Y_{k-1,h}$.
Return $q_{t,h} = \beta Y_{0,h} + \beta \sum_{k=1}^{M} Y_{k,h}$ for all $h \in [H]$.
```

4.3 Regret decomposition approach

4.3.1 Preliminaries

Before we state main results, we define some useful concepts that we will use later. First, the value function and action-value function with respect to policy π in episode t are respectively defined as

$$\begin{aligned} Q^{\pi}_{t,h}(x,a) &= \mathbb{E}_{\pi}\left[\sum_{k=h}^{H} \ell_t(\widetilde{X}_k, \widetilde{A}_k) \middle| \widetilde{X}_h = x, \widetilde{A}_h = a \right], \\ V^{\pi}_{t,h}(x) &= \sum_{a} \pi(a|x) Q^{\pi}_{t,h}(x,a), \end{aligned}$$

where the notation $\mathbb{E}_{\pi}[\cdot]$ highlights that the sequence of states $\widetilde{X}_k$ and actions $\widetilde{A}_k$ are generated by following policy π in the MDP. For an action-value function Q, we will sometimes use the notation $Q(x,\pi) = \sum_a \pi(a|x) Q(x,a)$.

Each policy π generates a probability distribution μ^{π}_h over each layer $h \in [H]$, and we will refer to the collection of these distributions in each layer as the *occupancy measure* μ^{π} induced by π. We will use occupancy measures in the analysis later.

We will be interested in developing learning algorithms that select a policy π_t for the learner at the beginning of each episode t. With some abuse of notation, we will use $V_{t,l}(x) = V_{t,l}^{\pi_t}(x)$ and $Q_{t,l}(x,a) = Q_{t,l}^{\pi_t}(x,a)$ to denote the value function and the action-value function of policy π_t in episode t. With this notation, we can reformulate the notion of regret, introduced in (1.1), as following:

$$R_T = \sup_{\pi} \sum_{t=1}^{T} \left(V_{t,1}^{\pi}(x_1) - V_{t,1}(x_1) \right).$$

As in (1.1), the supremum is taken over the set of all stationary policies mapping states to actions.

As shown by Jin et al. [73], an important property of the linear MDPs is that the action-value functions are also linear in the feature map $\varphi(x,a)$:

Lemma 23 (Linear action-value function)**.** *Suppose that Assumption 3 is satisfied. Then, for any t, h and any stationary stochastic policy π, there exists a vector $\zeta_{t,h}^{\pi} \in \mathbb{R}^d$ such that the action-value function can be written as*

$$Q_{t,h}^{\pi}(x,a) = \left\langle \zeta_{t,h}^{\pi}, \varphi(x,a) \right\rangle \tag{4.2}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. Furthermore, $\zeta_{t,h}$ satisfies $\left\| \zeta_{t,h}^{\pi} \right\| \le (H-h)R$.

The statement is a direct consequence of Proposition 2.3 in Jin et al. [73], and is thus omitted. We note that our results in this section do not directly require Assumption 3 to hold, and only make use of the property established in Lemma 4.2.

4.3.2 Algorithm and main results

Our algorithm design is motivated by the following decomposition of the regret first proposed for online MDP problems by Even-Dar et al. [58] and adapted to finite-horizon MDPs by Neu et al. [111]:

Lemma 24. *Let μ^* denote the occupancy measure induced by π^*. Then, for any sequence of policies π_t selected by the learner, the regret satisfies $R_T = \sum_{h=1}^{H} \mathbb{E}_{X_h^* \sim \mu_h^*} \left[\sum_{t=1}^{T} \Big(Q_{t,h}(X_h^*, \pi^*) - V_{t,h}(X_h^*) \Big) \right]$.*

This decomposition is based on the classic performance-difference lemma popularized by Kakade and Langford [74], Kakade [75] (see also 32, 36). As observed in previous work [58, 111], this lemma implies that the global regret minimization problem can be decomposed into a set of local regret minimization problems in each state x, where the reward function associated with each action a is defined as $Q_{t,h}(x, a)$. Indeed, letting $\pi_t(\cdot|x)$ denote the policy played by the local algorithm in state x in round t, we can define the *local regret* against policy π^* as

$$R_{h,T}(x) = \sum_{t=1}^{T} \mathbb{E}\left[Q_{t,h}(x, \pi^*) - V_{t,h}(x)\right].$$

and the regret in layer h as $R_{T,h} = \mathbb{E}_{X_h^* \sim \mu_h^*}\left[R_{h,T}(X_h^*)\right]$. This can be easily seen to be related to the global regret as $R_T = \sum_{h=1}^{H} R_{T,h}$, and thus it is obvious that bounding the local regrets in each state x yields a bound on the global regret.

With this in mind, following the algorithmic template laid out by Even-Dar et al. [58] called MDP-E, we propose an algorithm based on running a variant of the classic EXP3 algorithm of Auer et al. [14] in each state x. The key challenge is constructing the inputs to these local algorithms in a way that yields a computationally tractable algorithm with nontrivial performance bounds, and more specifically to achieve runtime and regret guarantees that are *independent of the size of the state space*. Indeed, instead of the possibly infinite number of states, we prefer to have the *dimensionality of the feature map* appear in our bounds, which is made possible by Assumption 23. Indeed, this assumption allows us to represent each Q-function by its parameter vector, which in turn enables an efficient implementation of the local regret minimization algorithms. Specifically, we design an estimator $\widehat{\zeta}_{t,h}$ of the parameter vector $\zeta_{t,h}$ corresponding to the action-value function $Q_{t,h}(x, a) = \langle \zeta_{t,h}, \varphi(x, a)\rangle$ of policy π_t, and plug the resulting estimates $\langle \widehat{\zeta}_{t,h}, \varphi(x, a)\rangle$ into a local copy of EXP3. The form of the estimator $\widehat{\zeta}_{t,h}$ and overall algorithm design is directly influenced by the LINEXP3 method, presented in Chapter 3, and thus we refer to our algorithm as MDP-LINEXP3. Its pseudocode is presented as Algorithm 4.3.

For stating many of our technical results, we define the filtration $\mathcal{F}_t = \sigma\left(\tau_s, s \leq t\right)$, and the notation $\mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot \left|\mathcal{F}_{t-1}\right.\right]$. Our reward estimator will be based on the observed rewards, and particularly the partial sums $G_{t,h} = \sum_{k=h}^{H} \ell_t(X_{t,k}, A_{t,k})$ for each layer h. Analogically to estimator of the reward function, we can define the estimator of the action-value function

$$\widetilde{\zeta}_{t,h} = \Sigma_{t,h}^{-1} \varphi(X_{t,h}, A_{t,h}) G_{t,h}. \tag{4.3}$$

Algorithm 4.2 MDP-LinExp3

```
Parameters: Learning rate $\eta > 0$, exploration parameter $\gamma \in (0,1)$,
Initialization: Set $\widehat{\zeta}_{1,h} = 0$ for all $h \in [H]$.
For episode $t = 1, \dots, T$, repeat:
  • Draw $Y_t \sim \text{Ber}(\gamma)$,
  • For step $h = 1, \dots, H$, do:
      – Observe $X_{t,h}$ and, for all $a \in \mathcal{A}(X_{t,h})$, set
            $w_t(X_{t,h}, a) = \exp\left(\eta \cdot \sum_{s=1}^{t-1} \langle \varphi(X_{t,h}, a), \widehat{\zeta}_{s,h} \rangle\right),$
      – draw $A_{t,h}$ from the policy defined as
            $\pi_t(a|X_{t,h}) = \frac{w_t(X_{t,h}, a)}{\sum_{a' \in \mathcal{A}(X_{t,h})} w_t(X_{t,h}, a')} \mathbb{I}_{\{Y_t=0\}} + \frac{1}{K} \mathbb{I}_{\{Y_t=1\}},$
      – observe the reward $\ell_t(X_{t,h}, A_{t,h})$.
  • Compute $\widehat{\zeta}_{t,h}$ for all $h = 1, \dots, H$.
```

This is an unbiased estimator of the action-value function:

$$\mathbb{E}_t \left[G_{t,h} | X_{t,h} = x, A_{t,h} = a \right] = Q_{t,h}(x,a) = \langle \varphi(x,a), \zeta_{t,h} \rangle .$$

It is easy to see that $\widetilde{\zeta}_{t,h}$ is an unbiased estimate of vector $\zeta_{t,h}$:

$$\begin{aligned} \mathbb{E}_t \left[\widetilde{\zeta}_{t,h} \right] &= \mathbb{E}_t \left[\Sigma_{t,h}^{-1} \varphi(X_{t,h}, A_{t,h}) \varphi(X_{t,h},, A_{t,h})^\mathsf{T} \zeta_{t,h} \right] \\ &= \Sigma_{t,h}^{-1} \Sigma_{t,h} \zeta_{t,h} = \zeta_{t,h}. \end{aligned}$$

Based on the Matrix geometric resampling procedure, described in Section 4.2.1, we finally define our estimator of value function as

$$\widehat{\zeta}_{t,h} = \widehat{\Sigma}_{t,h}^{+} \varphi(X_{t,h}, A_{t,h}) G_{t,h}.$$

Computing these estimators, MDP-LINEXP3 constructs an EXP3-style policy defined as $\pi_t(a|x) \propto \exp\left(\sum_{k=1}^{t-1} \langle \varphi(x,a), \widehat{\zeta}_{k,h} \rangle\right)$ for each x, a. Notably, the policy only depends on the cumulative parameter vectors and the feature vector $\varphi(x,a)$, and thus does not have to make explicit updates to the individual regret-minimization algorithms acting in the states x. For technical reasons, MDP-LINEXP3 follows *the uniform policy* $\pi_U(a|x) = \frac{1}{K}$ with probability γ in each episode, and follows the above exponential-weights policy otherwise. We will denote the covariance matrix generated by the uniform policy at layer h as Σ_h, and make the following assumption:

Assumption 4. *The eigenvalues of Σ_h for all h are lower bounded by $\lambda_{\min} > 0$.*

Our main result is the following guarantee regarding the performance of MDP-LINEXP3:

Theorem 4.3.1. *Suppose that the MDP satisfies Assumptions 3 and 4 and $\lambda_{\min} > 0$. Then, for $\gamma \in (0,1)$, $M \geq 0$, any positive $\eta \leq \frac{1}{\sigma^2 \beta (M+1) H}$ and any positive $\beta \leq \frac{1}{2\sigma^2 \sqrt{d(M+1)}}$, the expected regret of* MDP-LINEXP3 *over T episodes satisfies*

$$R_T \leq \mathcal{O}\left(\gamma H^2 T + dH^3 T \frac{\eta \sigma^2}{\gamma \lambda_{\min}} + \frac{H \log K}{\eta}\right).$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2 \sqrt{d(M+1)}}$, $M = \left\lceil \frac{\sigma^4 d \log^2(\sqrt{TH}\sigma R)}{\gamma^2 \lambda_{\min}^2} \right\rceil$, $\eta = \frac{(\log K)^{2/3} \lambda_{\min}^{1/3}}{T^{2/3} d^{1/3} H}$, $\gamma = \frac{\sigma (d \log K)^{1/3}}{(T\lambda_{\min})^{1/3}}$ and supposing that T is large enough so that the above constraints on γ, M, η and β are satisfied, we also have

$$R_T = \mathcal{O}\left(\sigma H^2 T^{2/3} \left(\frac{d \log K}{\lambda_{\min}}\right)^{1/3}\right).$$

For the complete regret bound, see Section 4.3.4. The downside of the above result is that it scales with the time horizon as $T^{2/3}$, which is likely to be suboptimal in light of the best known bounds of order $\sqrt{T}$ achieved in the tabular setting [152] in the bandit case and the large-scale setting considered by Cai et al. [33] in the full-information case. The next result shows that this dependence can be improved at the price of making stronger assumptions about the MDP. Specifically, assume that P is such that for any policy π, the occupancy distribution μ_h^π

has a density $f_h^\pi(x)$ on the set of states $\mathcal{S}_h$ with respect to some base measure, and denote the density corresponding to μ_h^* as $f_h^*(x)$. Then, assuming that the likelihood ratio $\frac{f_h^*(x)}{f_h^\pi(x)}$ is uniformly upper bounded, the dependence of our bounds on T can be improved to $\sqrt{T}$:

$$\begin{aligned}R_T \leq & 2T\sigma RH \cdot \exp\left(-\gamma\beta\lambda_{\min}M\right) + \gamma H^2 T \\ & + \eta H^3 d\left(\frac{1}{3} + \rho\right)T + H \cdot \frac{\log K}{\eta}.\end{aligned}$$

Theorem 4.3.2. *Suppose that the MDP satisfies Assumptions 3 and 4 and that the likelihood ratio between the occupancy measures induced by any policy* π *and* π^* *can be bounded uniformly as* $\sup_{\pi,h,x}\frac{f_h^*(x)}{f_h^\pi(x)} \leq \rho$ *for some* $\rho > 0$. *Then, for* $\gamma \in (0,1)$, $M \geq 0$, *any positive* $\eta \leq \frac{1}{\sigma^2\beta(M+1)H}$ *and any positive* $\beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, *the expected regret of* MDP-LINEXP3 *over* T *episodes, satisfies*

$$R_T \leq \mathcal{O}\left(\gamma H^2 T + \eta H^3 d\rho T + H \cdot \frac{\log K}{\eta}\right).$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, $M = \left\lceil\frac{\sigma^4 d\log^2(\sqrt{TH}\sigma R)}{\gamma^2\lambda_{\min}^2}\right\rceil$, $\eta = \frac{1}{H}\sqrt{\frac{\log K}{Td\rho}}$, $\gamma = \sqrt{\frac{d\rho\log K}{T}}$ *and supposing that* T *is large enough so that the above constraints are satisfied, we also have*

$$R_T = \mathcal{O}\left(H^2 \cdot \sqrt{Td\rho\log K} + H^2\sqrt{\frac{Td\log K}{\rho}}\right).$$

For the complete regret bound, see Section 4.3.4.

4.3.3 Analysis

As explained in the previous section, our algorithm and analysis is based on decomposing the overall learning problem to a number of local online learning problems corresponding to each state in the MDP. This approach is closely related to the ghost sample technique used in the Section 3 in the contextual bandit problem, where a similar regret decomposition was suggested. Our analysis in this section will make use of several tools developed to solve the contextual bandit problem, with the added challenge that the feature vectors in the current setting are no longer i.i.d.: in any layer, the distribution of states clearly depends on the learner's policy in the previous layers. Concretely, the main challenge in our analysis comes from the fact that the state distribution μ^* appearing in the regret decomposition of Lemma 24 does not match the actual distribution of states μ_t. In what follows, we highlight the main steps in the analysis. Proofs of the lemmas are given in the Section 4.4.4.

We start by rewriting our reward estimator as $\widehat{\zeta}_{t,h} = \widetilde{\zeta}_{t,h} + b_{t,h}$, where $\widetilde{\zeta}_{t,h}$ is such that $\mathbb{E}_t\left[\widetilde{\zeta}_{t,h}\right] = \zeta_{t,h}$ and $b_{t,h}$ is a bias term. Also, we will use the notation $\widetilde{Q}_t$ to refer to the function taking values $\widetilde{Q}_{t,h}(x,a) = \left\langle \widetilde{\zeta}_{t,h}, \varphi(x,a)\right\rangle$. The bulk of our analysis is based on the following regret decomposition that further refines the decomposition given in Lemma 24:

Lemma 25. *Let X_h^* be sampled from the context distribution generated by μ_h^*. Suppose that $\pi_t \in \mathcal{F}_{t-1}$ and that $\mathbb{E}_t\left[\widetilde{\zeta}_{t,h}\right] = \zeta_{t,h}$ for all t, h. Then, for all h,*

$$R_{T,h} = \sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{E}_{X_h^* \sim \mu_h^*, t}\left[\widetilde{Q}_{t,h}(X_h^*, \pi^*) - \widetilde{Q}_{t,h}(X_h^*, \pi_t)\right].$$

The proof is presented in Section 4.3.4. This suggests that we can define an auxiliary regret minimization game for every layer h and every state x, action a with reward $\left\langle \varphi(x,a), \widehat{\zeta}_{t,h}\right\rangle$ assigned to action a in each round t. The regret in this auxiliary game can be written as

$$\widehat{R}_{T,h}(x) = \sum_{t=1}^{T}\sum_{a=1}^{K}\left(\pi^*(a|x) - \pi_t(a|x)\right)\left\langle \varphi(x,a), \widehat{\zeta}_{t,h}\right\rangle,$$

and the above lemma suggests that the regret in layer h can be simply bounded as

$$R_{T,h} \leq \mathbb{E}\left[\widehat{R}_{T,h}(X_h^*)\right] + 2\sum_{t=1}^{T} \max_a \left|\mathbb{E}\left[\langle \varphi(X_h^*, a), b_{t,h}\rangle\right]\right|.$$

Thus, we are left with the problem of controlling the auxiliary regret in each state, and the bias of our estimators. The following lemma, which is a straightforward application of Lemma 30, gives bounds for the regret in the auxiliary game:

Lemma 26. *Fix any $h \in [H]$, $x \in \mathcal{S}_h$ and suppose that $\widehat{\zeta}_{t,h}$ is such that $\left|\eta\left\langle\varphi(x,a),\widehat{\zeta}_{t,h}\right\rangle\right| < 1$. Then, the regret in the auxiliary game at x satisfies*

$$\begin{aligned}\widehat{R}_{T,h}(x) \leq & \frac{\log K}{\eta} + \eta \sum_{t=1}^{T}\sum_{a=1}^{K} \pi_t(a|x)\left\langle\varphi(x,a),\widehat{\zeta}_{t,h}\right\rangle^2 \\ & + \gamma\sum_{t=1}^{T}\sum_{a=1}^{K}\left(\pi^*(a|x) - \frac{1}{K}\right)\left\langle\varphi(x,a),\widehat{\zeta}_{t,h}\right\rangle.\end{aligned}$$

The main term on the right-hand side of this bound is handled in the next lemma:

Lemma 27. *Suppose that $\varphi(X_{t,h})$ is satisfying $\left\|\varphi(X_{t,h},a)\right\|_2 \leq \sigma$ for any a, $0 < \beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}}$ and $M > 0$. Then for each t and h,*

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_h^*)\left\langle\varphi(X_h^*,a),\widehat{\zeta}_{t,h}\right\rangle^2\right] \leq 3(H-h)^2\left(1 + \frac{d}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right).$$

The proof of this claim is rather complicated and is presented in Section 4.4.4. The main difficulty in the analysis comes from the mismatch of distribution of X_h^* and $X_{t,h}$. To illustrate this difficulty, consider replacing $\widehat{\zeta}_{t,h}$ by the ideal estimator $\widetilde{\zeta}_{t,h}$ defined in Equation (4.3) in the quadratic term bounded in the above lemma. Introducing the notation $\Sigma_{t,h}^* = \mathbb{E}\left[\varphi(X_h^*,\pi_t(X_h^*))\varphi(X_h^*,\pi_t(X_h^*))^\mathsf{T}\right]$, each term in the sum can be bounded as

$$\begin{aligned}&\mathbb{E}_t\left[\pi_t(a|X_h^*)\left\langle\varphi(X_h^*,a),\widetilde{\zeta}_{t,h}\right\rangle^2\right] = \mathbb{E}_t\left[\pi_t(a|X_h^*)\left(\varphi(X_h^*,a)^\mathsf{T}\Sigma_{t,h}^{-1}\varphi(X_t,a)G_{t,h}\right)^2\right] \\ &\quad\leq (H-h)^2\cdot\mathbb{E}_t\left[\text{tr}\left(\pi_t(a|X_h^*)\varphi(X_h^*,a)\varphi(X_h^*,a)^\mathsf{T}\Sigma_{t,h}^{-1}\varphi(X_{t,h},a)\varphi(X_{t,h},a)^\mathsf{T}\Sigma_{t,h}^{-1}\right)\right] \\ &\quad= (H-h)^2\cdot\text{tr}\left(\Sigma_{t,h}^*\Sigma_{t,h}^{-1}\right).\end{aligned}$$

Unfortunately, this latter term cannot be bounded without further assumptions on $\Sigma_{t,h}$ due to the mismatch between the distributions of X_h^* and $X_{t,h}$. We address this issue by mixing the exponential-weights distribution with the uniform policy and appealing to Assumption 5, which together ensure that the smallest eigenvalue of matrix $\Sigma_{t,a,h}$ is at least $\lambda_{\min}\frac{\gamma}{K}$. This yields a bound on the operator norm of the matrix inverse $\Sigma_{t,h}^{-1}$, and eventually the bound of order $H^2Kd/(\gamma\lambda_{\min})$ above. The tighter bounds of Theorem 4.3.2 are derived by using a stronger assumption to bound $\operatorname{tr}\big(\Sigma_{t,h}^{*}\Sigma_{t,h}^{-1}\big)$—the details of these tighter bounds are presented in the Section 4.3.4. The final element in the proof is the following lemma that bounds the bias of the estimator:

Lemma 28. *For $M \geq 0$, $\beta = \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, we have*

$$\left|\mathbb{E}_t\left[\left\langle\varphi(X_h^*, a), \zeta_{t,h} - \widehat{\zeta}_{t,h}\right\rangle\right]\right| \leq \sigma R \exp\left(-\gamma\beta\lambda_{\min}M\right).$$

The proof can be found in the Section 4.4.4. Putting these lemmas together and verifying that the reward estimators indeed satisfy the condition of Lemma 26 (done in Lemma 32 in the Section 4.4.4), we obtain the following bound on the regret in layer h:

$$\begin{aligned} R_{T,h} &\leq 2T\sigma R \cdot \exp\left(-\gamma\beta\lambda_{\min}M\right) + 2\gamma(H-h)T \\ &\quad + 3\eta(H-h)^2\left(1 + d\frac{1}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right)T + \frac{\log K}{\eta}. \end{aligned}$$

Summing up the bound for all $h \in [H]$ proves Theorem 4.4.1.

4.3.4 Proofs

The proof of Lemma 25

By Lemma 24, and since $\widetilde{\zeta}_{t,a,h}$ is unbiased, we have

$$\begin{aligned} R_{T,h} &= \sum_{t=1}^{T}\mathbb{E}_{X_h^*\sim\mu_h^*,t}\left[Q_{t,h}(X_h^*, \pi^*(X_h^*)) - V_{t,h}(X_h^*)\right] \\ &= \sum_{t=1}^{T}\mathbb{E}_{X_h^*\sim\mu_h^*,t}\left[\sum_{a=1}^{K}\left(\pi^*(a|X_h^*) - \pi_t(a|X_h^*)\right)\cdot\left\langle\varphi(X_h^*, a), \zeta_{t,h}\right\rangle\right] \\ &= \sum_{t=1}^{T}\mathbb{E}_{X_h^*\sim\mu_h^*,t}\left[\sum_{a=1}^{K}\left(\pi^*(a|X_h^*) - \pi_t(a|X_h^*)\right)\cdot\left\langle\varphi(X_h^*, a), \widetilde{\zeta}_{t,h}\right\rangle\right]. \end{aligned}$$

■

The proof of Lemma 27

The proof relies on a repeated use of the following identity that holds for any symmetric positive definite matrix S:

$$\sum_{k=0}^{M}(I-S)^k = S^{-1} - (I-S)^M S^{-1}.$$

For ease of readability, we will omit the indices h in this section. We denote the covariance of states, generated by policy π^* as $\Sigma^* = \mathbb{E}\left[\varphi(X^*,\pi_t(X^*))\varphi(X^*,\pi_t(X^*))^\mathsf{T}\right]$. We start by plugging in the definition of $\widehat{\zeta}_t$ and writing

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X^*)\big\langle\varphi(X^*,a),\widehat{\zeta}_t\big\rangle^2\right]\\
&\quad= \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X^*)\left(\varphi(X^*,a)^\mathsf{T}\widehat{\Sigma}_{t,a}^{+}\varphi(X_t,A_t)G_{t,h}\right)^2\right] \qquad\qquad (4.4)\\
&\quad\leq (H-h)^2\cdot\mathbb{E}_t\left[\sum_{a=1}^{K}\mathrm{tr}\left(\pi_t(a|X^*)\varphi(X^*,a)\varphi(X^*,a)^\mathsf{T}\widehat{\Sigma}_{t,a}^{+}\varphi(X_t,A_t)\varphi(X_t,A_t)^\mathsf{T}\widehat{\Sigma}_t^{+}\right)\right],
\end{aligned}$$

where we used $\big\langle X_t,\zeta_{t,a}\big\rangle \leq H-h$ in the inequality. Using the definition of $\Sigma_{t,a}^{+}$ and elementary manipulations, we can get

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{a=1}^{K}\mathrm{tr}\left(\pi_t(a|X^*)\varphi(X^*,a)\varphi(X^*,a)^\mathsf{T}\widehat{\Sigma}_t^{+}\varphi(X_t,A_t)\varphi(X_t,A_t)^\mathsf{T}\widehat{\Sigma}_t^{+}\right)\right]\\
&\quad= \mathbb{E}_t\left[\mathrm{tr}\left(\Sigma^*\Sigma_t^{+}\Sigma_t\Sigma_t^{+}\right)\right] = \beta^2\cdot\mathbb{E}_t\left[\mathrm{tr}\left(\Sigma^*\left(\sum_{k=0}^{M}C_k\right)\Sigma_t\left(\sum_{j=0}^{M}C_j\right)\right)\right]\\
&\quad= \beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\sum_{j=0}^{M}\mathrm{tr}\left(\Sigma^*C_k\Sigma_tC_j\right)\right]\\
&\quad= \beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\mathrm{tr}\left(\Sigma^*C_k\Sigma_tC_k\right)\right] + 2\beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\sum_{j=k+1}^{M}\mathrm{tr}\left(\Sigma^*C_k\Sigma_tC_j\right)\right].
\end{aligned}$$

Let us first address the first term on the right hand side. Let V be a matrix where the columns are from the orthonormal system of eigenvectors of Σ_t. Applying

Lemma 20 with $\widetilde{H} = \Sigma^*$, and recalling the definition $C_k = \prod_{j=1}^k (I - \beta B_j)$, we can obtain

$$
\begin{aligned}
\beta^2 \sum_{k=0}^M \mathrm{tr}\left(\mathbb{E}\left[\Sigma^* C_k \Sigma_t C_k\right]\right) &\le \beta^2 \sum_{k=0}^M \mathrm{tr}\left(\Sigma^* \Sigma_t \left(I - 2\beta\Sigma_t\right)^k\right) \\
&+ \beta^2 \sum_{k=0}^M \mathrm{tr}\left(\Sigma^* \Sigma_t V \mathrm{diag}\left(\frac{\beta\sigma^2 \mathrm{tr}\left(\Sigma_t\right)}{2\lambda_j(\Sigma_t)} \exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma_t\right) k + 2\beta\lambda_j(\Sigma_t)\right)\right) V^\intercal\right) \\
&= \beta \mathrm{tr}\left(\Sigma^*(I - (I - \beta\Sigma_t)^M)\right) \\
&+ \beta^3 \frac{\sigma^2 \mathrm{tr}\left(\Sigma_t\right)}{2} \mathrm{tr}\left(\Sigma^* V \mathrm{diag}\left(\exp(2\beta\lambda_j(\Sigma_t)) \frac{\exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma_t\right)(M+1)\right) - 1}{\exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma_t\right)\right) - 1}\right) V^\intercal\right) \\
&\le \beta \mathrm{tr}\left(\Sigma^*\right) + \frac{\beta}{2} \mathrm{tr}\left(\Sigma^* V \mathrm{diag}\left(\exp(2\beta\lambda_j(\Sigma)) \exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma_t\right)(M+1)\right)\right) V^\intercal\right) \\
&\le \beta\sigma^2 d + \frac{\beta}{2} \sum_{j=1}^d \sigma^2 \exp(2\beta\sigma^2) \exp\left(\beta^2\sigma^4 d(M+1)\right) \le 3,
\end{aligned}
$$

where we used the condition $\beta \le \frac{1}{2\sigma^2 d\sqrt{(M+1)}} \le \frac{1}{2\sigma^2}$ and the fact that $(I - \beta(2 - \beta\sigma^2)\Sigma_t)^M \succcurlyeq 0$ by the same condition. We also used an observation that our assumption on the contexts implies $\mathrm{tr}\left(\Sigma_t\right) \le \mathrm{tr}\left(\sigma^2 I\right) = \sigma^2 d$, so again by our condition on β it implies the final bound.

Moving on to the second term, we first note that for any $j > k$, the conditional expectation of B_j given $B_{\le k} = (B_1, B_2, \dots B_k)$ satisfies $\mathbb{E}\left[C_k \middle| B_{\le k}\right] = C_k(I - \beta\Sigma_t)^{j-k}$ due to conditional independence of all B_j given B_k, for $i > k$. We make use of this equality by writing

$$
\begin{aligned}
\beta^2 \sum_{k=0}^M \sum_{j=k+1}^M \mathbb{E}\left[\mathrm{tr}\left(\Sigma^* C_k \Sigma_t C_j\right)\right] &= \beta^2 \sum_{k=0}^M \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^M \mathrm{tr}\left(\Sigma^* C_k \Sigma_t C_j\right) \middle| B_{\le k}\right]\right] \\
&= \beta^2 \sum_{k=0}^M \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^M \mathrm{tr}\left(\Sigma^* C_k \Sigma_t C_j (I - \beta\Sigma_t)^{j-k}\right) \middle| B_{\le k}\right]\right] \\
&= \beta \sum_{k=0}^M \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma^* C_k \Sigma_t C_k \Sigma_t^{-1}\left(I - (I - \beta\Sigma_t)^{M-k}\right)\right) \middle| B_{\le k}\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \beta \sum_{k=0}^{M} \mathbb{E}\left[\operatorname{tr}\left(\Sigma^* C_k \Sigma_t C_k \Sigma_t^{-1}\right)\right] \\
&\qquad \text{(due to } (I-\beta\Sigma_t)^{M-k} \succcurlyeq 0) \\
&\leq \beta \sum_{k=0}^{M} \operatorname{tr}\left(\Sigma^{-1}\Sigma^*\Sigma\left(I-2\beta\Sigma\right)^k\right) \\
&+ \beta \sum_{k=0}^{M} \operatorname{tr}\left(\Sigma^{-1}\Sigma^*\Sigma_t V \operatorname{diag}\left(\frac{\beta\sigma^2 \operatorname{tr}\left(\Sigma_t\right)}{2\lambda_j(\Sigma_t)} \exp\left(\beta^2\sigma^2\operatorname{tr}\left(\Sigma_t\right)k + 2\beta\lambda_j(\Sigma_t)\right)\right) V^\mathsf{T}\right) \\
&\qquad \text{(applying Lemma 20 with } \widetilde{H} = \Sigma_t^{-1}\Sigma^*) \\
&\leq 3\operatorname{tr}\left(\Sigma^*\Sigma^{-1}\right)\operatorname{tr}\left(\Sigma^*\left(\Sigma_t' + \gamma\Sigma\right)^{-1}\right) \leq \frac{3}{\gamma}\operatorname{tr}\left(\Sigma^*\Sigma^{-1}\right), \\
&\qquad \text{(following the analysis for the first term)}
\end{aligned}
$$

where in the last line we used that Σ_t can be written as $\Sigma_t = (1-\gamma)\Sigma_t' + \gamma\Sigma$ for $\Sigma_t' = \mathbb{E}_t\left[\varphi(X_t, A_t)\varphi(X_t, A_t)^\mathsf{T}\mathbb{I}_{\{Y_t=0\}}\right]$. Now, turning back to the sum over actions in (4.4) and recalling the definition of Σ^*, we can write

$$
\begin{aligned}
\mathbb{E}_t &\left[\sum_{a=1}^{K} \operatorname{tr}\left(\pi_t(a|X^*)\varphi(X^*, a)\varphi(X^*, a)^\mathsf{T}\widehat{\Sigma}_t^+\varphi(X_t, a)\varphi(X_t, a)^\mathsf{T}\widehat{\Sigma}_t^+\right)\right] \\
&\leq 3 + \frac{3}{\gamma}\operatorname{tr}\left(\Sigma^*\Sigma^{-1}\right) \leq 3 + \frac{3}{\gamma}\sqrt{\operatorname{tr}\left((\Sigma^*)^2\right)\operatorname{tr}\left((\Sigma^{-1})^2\right)} \qquad (4.5) \\
&\leq 3 + 3\frac{d}{\gamma}\frac{\sigma^2}{\lambda_{\min}},
\end{aligned}
$$

where we used the Cauchy–Schwarz inequality in the last step. This proves the statement. □

The proof of Lemma 34

We first observe that the bias of $\widehat{\zeta}_{t,h}$ can be easily expressed as

$$
\begin{aligned}
\mathbb{E}_t\left[\widehat{\zeta}_{t,h}\right] &= \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^+\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h}, A_{t,h})^\mathsf{T}\zeta_{t,h}\right] \\
&= \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^+\right]\mathbb{E}_t\left[\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h}, A_{t,h})^\mathsf{T}\right]\zeta_{t,h}
\end{aligned}
$$

$$= \mathbb{E}_t \left[\widehat{\Sigma}_{t,h}^{+}\right] \Sigma_{t,h}\zeta_{t,h} = \zeta_{t,h} - (I - \beta\Sigma_{t,h})^M \zeta_{t,h}.$$

Thus, the bias is bounded as

$$\left|\mathbb{E}_t \left[\varphi(X_0, a)^{\mathsf{T}}(I - \beta\Sigma_{t,h})^M \zeta_{t,h}\right]\right| \leq \|\varphi(X_0, a)\|_2 \cdot \|\zeta_{t,h}\|_2 \left\|(I - \beta\Sigma_{t,h})^M\right\|_{\text{op}}.$$

In order to bound the last factor above, observe that $\Sigma_{t,h} \succcurlyeq \gamma\Sigma_h$ due to the uniform exploration used in the first layer by MDP-LINEXP3, which implies that

$$\left\|(I - \beta\Sigma_{t,h})^M\right\|_{\text{op}} \leq (1 - \gamma\beta\lambda_{\min})^M \leq \exp\left(-\gamma\beta\lambda_{\min}M\right),$$

where the second inequality uses $1 - z \leq e^{-z}$ that holds for all z. This concludes the proof. □

The boundedness of the estimates

Lemma 29. *The loss estimates satisfy* $\eta\left|\left\langle\varphi(X_0, a), \widehat{\zeta}_{t,h}\right\rangle\right| < 1$ *for* $\eta \leq\leq \frac{1}{\sigma^2\beta(M+1)H}$

.

Proof. The claim is proven by the following straightforward calculation:

$$\begin{aligned}\eta \cdot \left|\left\langle\varphi(X_0, a), \widehat{\zeta}_{t,h}\right\rangle\right| &= \eta \cdot \left|\varphi(X_0, a)^{\mathsf{T}}\widehat{\Sigma}_{t,h}^{+}\varphi(X_{t,h}, a)\left\langle\varphi(X_{t,h}, a), \zeta_{t,h}\right\rangle\right| \\ &\leq \eta(H - h) \cdot \left|\varphi(X_0, a)^{\mathsf{T}}\widehat{\Sigma}_{t,h}^{+}\varphi(X_{t,h}, a)\right| \leq \eta(H - h)\sigma^2 \left\|\widehat{\Sigma}_{t,h}^{+}\right\|_{\text{op}} \\ &\leq \eta(H - h)\sigma^2\beta\left(1 + \sum_{k=1}^{M}\|C_{k,h}\|_{\text{op}}\right),\end{aligned}$$

where we used the fact that our choice of β ensures $\|C_{k,h}\|_{\text{op}} = \left\|\prod_{j=0}^{k}(I - \beta B_{j,h})\right\|_{\text{op}} \leq 1$. □

Complete regret bound on Theorem 4.4.1.

For $\gamma \in (0, 1)$, $M \geq 0$, any positive $\eta \leq \frac{1}{\sigma^2\beta(M+1)H}$ and any positive $\beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, the expected regret of MDP-LINEXP3 over T episodes satisfies

$$R_T \leq 2T\sigma RH \cdot \exp\left(-\gamma\beta\lambda_{\min}M\right) + \gamma H^2 T + 3\eta H^3\left(1 + \frac{d}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right)T + H \cdot \frac{\log K}{\eta}.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, $M = \left\lceil \frac{\sigma^4 d \log^2(\sqrt{T}H\sigma R)}{\gamma^2\lambda_{\min}^2} \right\rceil$, $\eta = \frac{(\log K)^{2/3}\lambda_{\min}^{1/3}}{T^{2/3}d^{1/3}H}$, $\gamma = \frac{\sigma(d\log K)^{1/3}}{(T\lambda_{\min})^{1/3}}$ and supposing that T is large enough so that the above constraints on γ, M, η and β are satisfied, we also have

$$R_T \leq H^2T^{2/3}\left(\frac{d\log K}{\lambda_{\min}}\right)^{1/3}(2\sigma+1) + 6H^2T^{1/3}(d\log K)^{2/3}\lambda_{\min}^{1/3} + 4H\sqrt{T}.$$

Proof of Theorem 4.3.2

The improvement in the regret bound comes from applying an importance-weighting trick in the proof of Lemma 31 to bound the problematic term $\operatorname{tr}\big(\Sigma_h^*\Sigma_{t,h}^{-1}\big)$. Specifically, we write

$$\begin{aligned}
\operatorname{tr}\big(\Sigma_h^*\Sigma_{t,h}^{-1}\big) &= \operatorname{tr}\Big(\mathbb{E}_t\left[\varphi(X_h^*,\pi_t(X_h^*))\varphi(X_h^*,\pi_t(X_h^*))^\mathsf{T}\right]\Sigma_{t,h}^{-1}\Big)\\
&= \operatorname{tr}\left(\mathbb{E}_t\left[\frac{f_h^*(X_{t,h})}{f_h^{\pi_t}(X_{t,h})}\varphi(X_{t,h},\pi_t(X_{t,h}))\varphi(X_{t,h},\pi_t(X_{t,h}))^\mathsf{T}\right]\Sigma_{t,h}^{-1}\right)\\
&\leq \rho\cdot\operatorname{tr}\Big(\mathbb{E}_t\left[\varphi(X_{t,h},\pi_t(X_{t,h}))\varphi(X_{t,h},\pi_t(X_{t,h}))^\mathsf{T}\right]\Sigma_{t,h}^{-1}\Big) = \rho d,
\end{aligned}$$

where we used our assumption on the likelihood ratio in the inequality. Using this bound instead of the one in Equation (4.5) at the end of the proof of Lemma 31 yields the improved bound

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_h^*)\big\langle\varphi(X_h^*,a),\widehat{\zeta}_{t,h}\big\rangle^2\right] \leq (H-h)^2 d\left(\frac{1}{3}+\rho\right).$$

The proof of Theorem 4.3.2 is then concluded similarly as the proof of Theorem 4.4.1.

□

Complete regret bound on Theorem 4.3.2.

For $\gamma \in (0,1)$, $M \geq 0$, any positive $\eta \leq \frac{2}{(M+1)H}$ and any positive $\beta \leq \frac{1}{2\sigma^2}$, the expected regret of MDP-LINEXP3 over T episodes, satisfies

$$R_T \leq 2T\sigma RH\cdot\exp\left(-\gamma\beta\lambda_{\min}M\right) + \gamma H^2T + \eta H^3 d\left(\frac{1}{3}+\rho\right)T + H\cdot\frac{\log K}{\eta}.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2}$, $M = \left\lceil \frac{\sigma^2 \log(T\sigma^2 R^2)}{\gamma \lambda_{\min}} \right\rceil$, $\eta = \frac{1}{H}\sqrt{\frac{\log K}{Td\rho}}$, $\gamma = \sqrt{\frac{d\rho \log K}{T}}$ and supposing that T is large enough so that the above constraints are satisfied, we also have

$$R_T \leq 3H^2 \cdot \sqrt{Td\rho \log K} + \frac{1}{3}H^2\sqrt{\frac{Td\log K}{\rho}} + 4H\sqrt{T}.$$

4.4 Linear optimization approach

4.4.1 Preliminaries

Analysis in this section is largely based on the concept of occupancy measures. Let $\tau^\pi = ((X_1, A_1), (X_2, A_2), \ldots, (X_H, A_H))$ be a trajectory generated by following the policy π through the MDP. Then, for any $x_h \in \mathcal{S}_h, a_h \in \mathcal{A}$ we define the occupancy measure $\mu_h^\pi(x, a) = \mathbb{P}_\pi\left[(x, a) \in \tau^\pi\right]$. We will refer to the collection of these distributions across all layers h as the occupancy measure induced by π and denote it as $\mu^\pi = (\mu_1^\pi, \mu_2^\pi, \ldots, \mu_H^\pi)$. We will denote the set of all valid occupancy measures by $\mathcal{U}$ and note that this is a convex set, such that for every element $\mu \in \mathcal{U}$ the following set of linear constraints is satisfied:

$$\sum_{a\in\mathcal{A}} \mu_{h+1}(x, a) = \sum_{x',a'\in\mathcal{S}_h\times\mathcal{A}} P(x|x', a')\mu_h(x', a'), \quad \forall x \in \mathcal{S}_{h+1}, h \in [H-1], \tag{4.6}$$

as well as $\sum_a \mu_1(x_1, a) = 1$. From every valid occupancy measure μ, a stationary stochastic policy $\pi = \pi_1, \ldots, \pi_{H-1}$ can be derived as $\pi_{\mu,h}(a|x) = \mu_h(x, a)/\sum_{a'} \mu_h(x, a')$. For each h, introducing the linear operators E and P through their action on a set state-action distribution u_h as $(E^\mathsf{T} u_h)(x) = \sum_{a\in\mathcal{A}} u_h(x, a)$ and $(P_h^\mathsf{T} u_h)(x) = \sum_{x',a'\in\mathcal{S}_h,\mathcal{A}} P(x|x', a')u_h(x', a')$, the constraints can be simply written as $E^\mathsf{T}\mu_{h+1} = P_h^\mathsf{T}\mu_h$ for each h. We will use the inner product notation for the sum over the set of states and actions: $\langle \mu_h, r_h \rangle = \sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})} \mu_h(x, a) r_{t,h}(x, a)$. Using this notation, we can formulate the notion of regret, introduced in (1.1), as

$$\mathfrak{R}_T = \sup_{\pi^*} \sum_{t=1}^{T} \sum_{h=1}^{H} \left(\mathbb{E}_{\pi^*}\left[r_{t,h}(X_h^*, A_h^*)\right] - \mathbb{E}_{\pi_t}\left[r_t(X_{t,h}, A_{t,h})\right]\right)$$

$$= \sup_{\mu^* \in \mathcal{U}} \sum_{t=1}^{T} \sum_{h=1}^{H} \left\langle \mu_h^* - \mu_h^{\pi_t}, r_{t,h} \right\rangle ,$$

where the notations $\mathbb{E}_{\pi^*}[\cdot]$ and $\mathbb{E}_{\pi_t}[\cdot]$ emphasize that the state-action trajectories are generated by following policies π^* and π_t, respectively. As the above expression suggests, we can reformulate our online learning problem as an instance of online linear optimization where in each episode t, the learner selects an occupancy measure $\mu_t \in \mathcal{U}$ (with $\mu_t = \mu^{\pi_t}$) and gains reward $\sum_{h=1}^{H} \langle \mu_{t,h}, r_{t,h} \rangle$.

In this work, we will exploit the useful property shown by Neu and Pike-Burke [115] and Bas-Serrano et al. [23] that all occupancy measures in a linear MDP can be seen to satisfy a relaxed version of the constraints in Equation (4.6). Specifically, for all h, defining the feature matrix $\Phi_h \in \mathbb{R}^{(\mathcal{S}_h \times \mathcal{A}) \times d}$ with its action on the distribution u as $\Phi_h^\mathsf{T} u = \sum_{x,a \in \mathcal{S}_h, \mathcal{A}} u_h(x,a)\varphi(x,a)$, we define $\mathcal{U}_\Phi$ as the set of state-action distributions $(\mu, u) = ((\mu_1, \ldots, \mu_H), (u_1, \ldots, u_H))$ satisfying the following constraints:

$$E^\mathsf{T} u_{h+1} = P_h^\mathsf{T} \mu_h \quad (\forall h), \qquad \Phi_h^\mathsf{T} u_h = \Phi_h^\mathsf{T} \mu_h \quad (\forall h), \qquad E^\mathsf{T} u_1 = 1. \quad (4.7)$$

It is easy to see that for all feasible (μ, u) pairs, u satisfies the original constraints (4.6) if the MDP satisfies Assumption 3: since the transition operator can be written as $P_h = \Phi_h M_h$ for some matrix M_h. In this case, we clearly have

$$E^\mathsf{T} u_{h+1} = P_h^\mathsf{T} \mu_h = M_h^\mathsf{T} \Phi_h^\mathsf{T} \mu_h = M_h^\mathsf{T} \Phi_h^\mathsf{T} u_h = P_h^\mathsf{T} u_h, \qquad (4.8)$$

showing that any feasible u is indeed a valid occupancy measure. Furthermore, due to linearity of the rewards in Φ, we also have $\langle u_h, r_{t,h} \rangle = \langle \mu_h, r_{t,h} \rangle$ for all feasible $(\mu, u) \in \mathcal{U}_\Phi$. While the number of variables and constraints in Equation (4.7) is still very large, it has been recently shown that approximate linear optimization over this set can be performed tractably [115, 23]. Our own algorithm design described in the next section will heavily build on these recent results.

4.4.2 Algorithm and main results

This section presents a new efficient algorithm for the setting described above along with its performance guarantees. Our algorithm design is based on a reduction to online linear optimization, exploiting the structural results established in the previous section. In particular, we will heavily rely on the algorithmic ideas established by Bas-Serrano et al. [23], who proposed an efficient reduction of approximate linear optimization over the high-dimensional set $\mathcal{U}_\Phi$ to a low-dimensional convex optimization problem. Another key component of the algorithm is an estimator of the reward vectors $\theta_{t,h}$, that we already presented in the Section 4.3, which is based on the matrix geometric resampling technique. Despite the fact that the reward estimator is the same as it was used for MDP-LINEXP3, accommodating it into the framework of Bas-Serrano et al. [23] is not straightforward and necessitates some subtle changes.

The policy update rule

Our algorithm is an instantiation of the well-known "Follow the Regularized Leader" (FTRL) template commonly used in the design of modern online learning methods (see, e.g., 118). We will make the following design choices:

- The decision variables will be the vector $(\mu, u) \in \mathbb{R}^{2(\mathcal{S}\times\mathcal{A})}$, with the feasible set $\mathcal{U}_\Phi^2$ defined through the constraints

$$E^\mathsf{T} u_h = P_h^\mathsf{T} \mu_h \quad (\forall h), \qquad \Phi_h^\mathsf{T} \text{diag}(u_h) \Phi_h = \Phi_h^\mathsf{T} \text{diag}(\mu_h) \Phi_h \quad (\forall h). \tag{4.9}$$

  These latter constraints ensure that the feature covariance matrices under u and μ will be identical, which is necessary for technical reasons that will be clarified in Section 3.4. Notice that, due to our assumption that $\varphi_1(x,a) = 1$, we have $\mathcal{U}_\Phi^2 \subseteq \mathcal{U}_\Phi$, so all feasible u's continue to be feasible for the original constraints (4.6).

- The regularization function will be chosen as $\frac{1}{\eta} D(\mu \| \mu_0) + \frac{1}{\alpha} D_C(u \| \mu_0)$ for some positive regularization parameters η and α, where μ_0 is the occupancy measure induced by the uniform π_0 with $\pi_0(a|x) = \frac{1}{K}$ for all x, a, and D and D_C are the marginal and conditional relative entropy functions respectively defined as $D(\mu \| \mu_0) = \sum_{h=1}^H D(\mu_h \| \mu_{0,h})$ and $D_C(\mu \| \mu_0) = \sum_{h=1}^H D_C(\mu_h \| \mu_{0,h})$ with

$$D(\mu_h \| \mu_{0,h}) = \sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})} \mu_h(x,a) \log \frac{\mu_h(x,a)}{\mu_{0,h}(x,a)}, \quad \text{and}$$

$$D_C(\mu_h \| \mu_{0,h}) = \sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})} \mu_h(x,a) \log \frac{\pi_{\mu,h}(a|x)}{\pi_{0,h}(a|x)}.$$

With these choices, the updates of our algorithm in each episode will be given by

$$(\mu_t, u_t) = \arg \max_{(\mu,u)\in\mathcal{U}_\Phi^2} \left\{ \sum_{s=1}^{t-1} \sum_{h=1}^{H-1} \langle \mu_h, \widehat{r}_{s,h} \rangle - \frac{1}{\eta} D(\mu \| \mu_0) - \frac{1}{\alpha} D_C(u \| \mu_0) \right\} \tag{4.10}$$

where $\widehat{r}_{t,h} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ is an estimator of the reward function $r_{t,h}$ that will be defined shortly.

As written above, it is far from obvious if these updates can be calculated efficiently. The following result shows that, despite the apparent intractability of the maximization problem, it is possible to reduce the above problem into a d^2-dimensional unconstrained convex optimization problem:

Proposition 3. *Define for each $h \in [H-1]$, a matrix $Z_h \in \mathbb{R}^{d\times d}$ and let matrix $Z \in \mathbb{R}^{d\times d(H-1)}$ be defined as $Z = (Z_1, \dots, Z_{H-1})$. We will write $h(x) = h$, if $x \in \mathcal{S}_h$. Define the Q-function taking values $Q_Z(x,a) = \varphi(x,a)^\mathsf{T} Z_{h(x)} \varphi(x,a)$ and define the value function*

$$V_Z(x) = \frac{1}{\alpha} \log \left(\sum_{a\in A(x)} \pi_0(a|x) e^{\alpha Q_Z(x,a)} \right)$$

For any $h \in [H-1]$ and for any $x \in \mathcal{X}_h$, $a \in A(x)$, denote $P_{x,a}V_Z = \sum_{x'\in\mathcal{X}_{h(x)+1}} P(x'|x,a)V_Z(x')$ and $\Delta_{t,Z}(x,a) = \sum_{s=1}^{t-1} \widehat{r}_{s,h(x)}(x,a) + P_{x,a}V_Z - Q_Z(x,a)$. Then, the optimal solution of the optimization problem (4.10) is given as

$$\widehat{\pi}_{t,h}(a|x) = \pi_0(a|x) e^{\alpha\left(Q_{Z_t^*}(x,a) - V_{Z_t^*}(x)\right)},$$
$$\widehat{\mu}_{t,h}(x,a) \propto \mu_0(x,a) e^{\eta \Delta_{t,Z_t^*}(x,a)},$$

where $Z_t^* = (Z_{t,1}^*, \dots, Z_{t,H-1}^*)$ is the minimizer of the convex function

$$\mathcal{G}_t(Z) = \frac{1}{\eta} \sum_{h=1}^{H-1} \log \left(\sum_{x\in\mathcal{X}_h, a\in A(x)} \mu_0(x,a) e^{\eta \Delta_{t,Z}(x,a)} \right) + V_Z(x_1). \tag{4.11}$$

A particular merit of this result is that it gives an explicit formula for the policy π_t that induces the optimal occupancy measure u_t, and that $\pi_t(a|x)$ can be evaluated straightforwardly as a function of the features $\varphi(x, a)$ and the parameters Z_t^*. The proof of the result is based on Lagrangian duality, and mainly follows the proof of Proposition 1 in Bas-Serrano et al. [23], with some subtle differences due to the episodic setting we consider and the appearance of the constraints $\Phi_h^\top \text{diag}(u_h)\Phi_h = \Phi_h^\top \text{diag}(\mu_h)\Phi_h$. The proof is presented in Section 4.4.4.

The proposition above inspires a very straightforward implementation that is presented as Algorithm 4.3. Due to the direct relation with the algorithm of Bas-Serrano et al. [23], we refer to this method as ONLINE Q-REPS, where Q-REPS stands for "Relative Entropy Policy Search with Q-functions". ONLINE Q-REPS adapts the general idea of Q-REPS to the online setting in a similar way as the O-REPS algorithm of Zimin and Neu [152] adapted the Relative Entropy Policy Search method of **(author?)** [Jan Peters and Altun] to regret minimization in tabular MDPs with adversarial rewards. While O-REPS would in principle be still applicable to the large-scale setting we study in this work and would plausibly achieve similar regret guarantees, its implementation would be nearly impossible due to the lack of the structural properties enjoyed by ONLINE Q-REPS, as established in Proposition 3.

The reward estimator

We already defined the estimator of the reward in Section 4.2.1, which makes use of the covariance matrix $\widehat{\Sigma}_{t,h}^+$. The covariance matrix is computed by the Matrix Geometric Resampling procedure, presented in Section 4.2.1 and there the estimator of the rewards is defined as

$$\widehat{\theta}_{t,h} = \widehat{\Sigma}_{t,h}^+ \varphi(X_{t,h}, A_{t,h}) r_{t,h}(X_{t,h}, A_{t,h}).$$

As was explained, $\widehat{\theta}_{t,h}$ can be computed in $O(MHKd)$ time, using M calls to the simulator.

The regret bound

We are now ready to state our main result: a bound on the expected regret of ONLINE Q-REPS. During the analysis, we will suppose that all the optimization problems solved by the algorithm are solved up to an additive error of $\varepsilon \geq 0$. Furthermore, we will denote the covariance matrix generated by the uniform policy at layer h as $\Sigma_{0,h}$, and make the following assumption:

Algorithm 4.3 ONLINE Q-REPS

```
Parameters: $\eta, \alpha > 0$, exploration parameter $\gamma \in (0,1)$,
Initialization: Set $\widehat{\theta}_{1,h} = 0$ for all $h$, compute $Z_1$.
For $t = 1, \dots, T$, repeat:
    • Draw $Y_t \sim \text{Ber}(\gamma)$,
    • For $h = 1, \dots, H$, do:
        – Observe $X_{t,h}$ and, for all $a \in \mathcal{A}(X_{t,h})$, set
              $\pi_{t,h}(a|X_{t,h}) = \pi_{0,h}(a|X_{t,h}) e^{\alpha\left(Q_{Z_t}(X_{t,h},a) - V_{Z_t}(X_{t,h})\right)}$,
        – if $Y = 0$, draw $A_{t,h} \sim \pi_{t,h}(\cdot|X_{t,h})$, otherwise draw $A_{t,h} \sim \pi_{0,h}(\cdot|X_{t,h})$,
        – observe the reward $\ell_{t,h}(X_{t,h}, A_{t,h})$.
    • Compute $\widehat{\theta}_{t,1}, \dots, \widehat{\theta}_{t,H-1}$, $Z_{t+1}$.
```

Assumption 5. *The eigenvalues of $\Sigma_{0,h}$ for all h are lower bounded by $\lambda_{\min} > 0$.*

Our main result is the following guarantee regarding the performance of ONLINE Q-REPS:

Theorem 4.4.1. *Suppose that the MDP satisfies Assumptions 3 and 5 and $\lambda_{\min} > 0$. Furthermore, suppose that, for all t, Z_t satisfies $\mathcal{G}_t(Z_t) \leq \min_Z \mathcal{G}_t(Z) + \varepsilon$ for some $\varepsilon \geq 0$. Then, for $\gamma \in (0,1)$, $M \geq 0$, positive $\eta \leq \frac{1}{\sigma^2 \beta (M+1) H}$ and any positive $\beta \leq \frac{1}{2\sigma^2 \sqrt{d(M+1)}}$, the expected regret of* ONLINE Q-REPS *over T episodes satisfies*

$$\mathfrak{R}_T \leq 2T\sigma RH \cdot \exp\left(-\gamma\beta\lambda_{\min} M\right) + \gamma HT + 3\eta HdTe^3 + \frac{1}{\eta} D(\mu^* \| \mu_0) + \frac{1}{\alpha} D_C(u^* \| \mu_0) + \sqrt{\alpha\varepsilon}(M+2)HT.$$

Furthermore, letting $\beta = \frac{1}{2\sigma^2\sqrt{d(M+1)}}$, $M = \left\lceil \frac{\sigma^4 d \log^2(\sqrt{TH}\sigma R)}{\gamma^2 \lambda_{\min}^2} \right\rceil$, $\eta = \frac{1}{\sqrt{TdH}}$, $\alpha = \frac{1}{\sqrt{TdH}}$ and $\gamma = \frac{1}{\sqrt{TH}}$ and supposing that T is large enough so that the *above constraints on M, γ, η and β are satisfied, we also have*

$$\mathfrak{R}_T \leq \sqrt{dHT}\left(2 + D(\mu^*\|\mu_0) + D_C(u^*\|\mu_0)\right) + \sqrt{HT} + \sqrt{\varepsilon}T^{7/4}(Hd)^{1/4} + 2.$$

Thus, when all optimization problems are solved up to precision $\varepsilon = T^{-5/2}$, the regret of ONLINE Q-REPS is guaranteed to be of $\mathcal{O}\big(\sqrt{dHTD(\mu^*\|\mu_0)}\big)$.

Implementation

While Proposition 3 establishes the form of the ideal policy updates π_t through the solution of an unconstrained convex optimization problem, it is not obvious that this optimization problem can be solved efficiently. Indeed, one immediate challenge in optimizing $\mathcal{G}_t$ is that its gradient takes the form

$$\nabla \mathcal{G}_t(Z) = \sum_{x,a} \widetilde{\mu}_Z(x,a) \left(\varphi(x,a)\varphi(x,a)^\mathsf{T} - \sum_{x',a'} P(x'|x,a)\pi_Z(a'|x')\varphi(x',a')\varphi(x',a')^\mathsf{T} \right),$$

where $\widetilde{\mu}_Z(x,a) = \frac{\mu_0(x,a)\exp(\eta\Delta_Z(x,a))}{\sum_{x',a'}\mu_0(x',a')\exp(\eta\Delta_Z(x',a'))}$. Sampling from this latter distribution (and thus obtaining unbiased estimators of $\nabla\mathcal{G}_t(Z)$) is problematic due to the intractable normalization constant.

This challenge can be addressed in a variety of ways. First, one can estimate the gradients via weighted importance sampling from the distribution $\widetilde{\mu}_Z$ and using these in a stochastic optimization procedure. This approach has been recently proposed and analyzed for an approximate implementation of REPS by Pacchiano et al. [119], who showed that it results in ε-optimal policy updates given polynomially many samples in $1/\varepsilon$. Alternatively, one can consider an empirical counterpart of the loss function replacing the expectation with respect to μ_0 with an empirical average over a number of i.i.d. samples drawn from the same distribution. The resulting loss function can then be optimized via standard stochastic optimization methods. This approach has been proposed and analyzed by Bas-Serrano et al. [23]. We describe the specifics of this latter approach in Section 4.4.4.

4.4.3 Analysis

This section gives the proof of Theorem 4.4.1 by stating the main technical results as lemmas and putting them together to obtain the final bound. In the first part of the proof, we show the upper bound on the auxiliary regret minimization game with general reward inputs and ideal updates. Then, we relate this quantity to the true expected regret by taking into account the properties of our reward estimates and the optimization errors incurred when calculating the updates. The proofs of all the lemmas are deferred to Section 4.4.4.

We start by defining the idealized updates $(\widehat{\mu}_t, \widehat{u}_t)$ obtained by solving the update steps in Equation (4.10) exactly, and we let u_t be the occupancy measure induced by policy π_t that is based on the near-optimal parameters Z_t satisfying $\mathcal{G}_t(Z_t) \leq \min_Z \mathcal{G}_t(Z) + \varepsilon$. We will also let μ_t be the occupancy measure resulting from mixing u_t with the exploratory distribution μ_0 and note that $\mu_{t,h} = (1-\gamma)u_{t,h} + \gamma\mu_{t,h}$. Using this notation, we will consider an auxiliary online learning problem with the sequence of reward functions given as $\widehat{r}_{t,h}(x,a) = \langle \varphi(x,a), \widehat{\theta}_{t,h} \rangle$, and study the performance of the idealized sequence $(\widehat{\mu}_t, \widehat{u}_t)$ therein:

$$\widehat{\mathfrak{R}}_T = \sum_{t=1}^{T} \sum_{h=1}^{H-1} \langle \mu_h^* - \widehat{u}_{t,h}, \widehat{r}_{t,h} \rangle.$$

Our first lemma bounds the above quantity:

Lemma 30. *Suppose that $\widehat{\theta}_{t,h}$ is such that $\left|\eta \cdot \langle \varphi(x,a), \widehat{\theta}_{t,h} \rangle\right| < 1$ holds for all x, a. Then, the auxiliary regret satisfies*

$$\widehat{\mathfrak{R}}_T \leq \eta \sum_{t=1}^{T} \sum_{h=1}^{H-1} \langle \widehat{\mu}_{t,h}, \widehat{r}_{t,h}^2 \rangle + \frac{1}{\eta} D(\mu^* \| \mu_0) + \frac{1}{\alpha} D_C(u^* \| \mu_0).$$

While the proof makes use of a general potential-based argument commonly used for analyzing FTRL-style algorithms, it involves several nontrivial elements exploiting the structural results concerning ONLINE Q-REPS proved in Proposition 3. In particular, these properties enable us to upper bound the potential differences in a particularly simple way. The main term on contributing to the regret $\widehat{\mathfrak{R}}_T$ can be bounded as follows:

Lemma 31. *Suppose that $\varphi(X_{t,h}, a)$ is satisfying $\|\varphi(X_{t,h}, a)\|_2 \leq \sigma$ for any a, $0 < \beta \leq \frac{1}{2\sigma^2\sqrt{d(M+1)}}$ and $M > 0$. Then for each t and h,*

$$\mathbb{E}_t\left[\langle \widehat{\mu}_{t,h}, \widehat{r}_{t,h}^2 \rangle\right] \leq 3 + 5d + (M+1)^2 \left\|\widehat{u}_{t,h} - u_{t,h}\right\|_1.$$

The proof of this claim makes heavy use of the fact that $\langle \widehat{\mu}_{t,h}, \widehat{r}^2_{t,h} \rangle = \langle \widehat{u}_{t,h}, \widehat{r}^2_{t,h} \rangle$, which is ensured by the construction of the reward estimator $\widehat{r}_{t,h}$ and the constraints on the feature covariance matrices in Equation (4.9). This property is not guaranteed to hold under the first-order constraints (4.7) used in the previous works of Neu and Pike-Burke [115] and Bas-Serrano et al. [23], which eventually justifies the higher complexity of our algorithm.

It remains to relate the auxiliary regret to the actual regret. The main challenge is accounting for the mismatch between μ_t and u_t, and the bias of $\widehat{r}_t$, denoted as $b_{t,h}(x,a) = \mathbb{E}_t \left[\widehat{r}_{t,h}(x,a)\right] - r_{t,h}(x,a)$. To address these issues, we observe that for any t, h, we have

$$\begin{aligned}\langle \mu_{t,h}, r_{t,h} \rangle &= \langle (1-\gamma) u_{t,h} + \gamma \mu_{0,h}, r_{t,h} \rangle \\ &= \langle (1-\gamma) \widehat{u}_{t,h} + \gamma \mu_{0,h}, r_{t,h} \rangle + (1-\gamma) \langle u_{t,h} - \widehat{u}_{t,h}, r_{t,h} \rangle \\ &\geq \mathbb{E}_t \left[\langle (1-\gamma) \widehat{u}_{t,h} + \gamma \mu_{0,h}, \widehat{r}_{t,h} \rangle \right] + \left\| b_{t,h} \right\|_\infty + (1-\gamma) \left\| u_{t,h} - \widehat{u}_{t,h} \right\|_1 ,\end{aligned}$$

where in the last step we used the fact that $\left\| r_{t,h} \right\|_\infty \leq 1$. After straightforward algebraic manipulations, this implies that the regret can be bounded as

$$\mathfrak{R}_T \leq (1-\gamma) \mathbb{E} \left[\widehat{\mathfrak{R}}_T \right] + \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E} \left[\gamma \langle \mu_{0,h} - \mu^*_h, r_{t,h} \rangle + \left\| \widehat{u}_{t,h} - u_{t,h} \right\|_1 + \left\| b_{t,h} \right\|_\infty \right] . \tag{4.12}$$

In order to proceed, we need to verify the condition $\left| \eta \cdot \langle \varphi(x,a), \widehat{\theta}_{t,h} \rangle \right| < 1$ so that we can apply Lemma 30 to bound $\widehat{\mathfrak{R}}_T$. This is done in the following lemma:

Lemma 32. *Suppose that* $\eta \leq \frac{1}{\sigma^2 \beta (M+1) H}$. *Then, for all,* t, h, *the reward estimates satisfy* $\eta \left\| \widehat{r}_{t,h} \right\|_\infty < 1$.

Proceeding under the condition $\eta(M+1)$, we can apply Lemma 30 to bound the first term on the right-hand side of Equation (4.12), giving

$$\begin{aligned}\mathfrak{R}_T \leq & \frac{D(\mu^* \| \mu_0)}{\eta} + \frac{D_C(u^* \| \mu_0)}{\alpha} + (3+5d)\eta HT + \gamma HT \\ & + \sum_{t,h} \mathbb{E} \left[(M+2) \left\| \widehat{u}_{t,h} - u_{t,h} \right\|_1 + \left\| b_{t,h} \right\|_\infty \right] .\end{aligned}$$

It remains to bound the bias of the reward estimators and the effect of the optimization errors that result in the mismatch between u_t and $\widehat{u}_t$. The following lemma shows that this mismatch can be directly controlled as a function of the optimization error:

Lemma 33. *The following bound is satisfied for all t and h: $\left\|\widehat{u}_{t,h} - u_{t,h}\right\|_1 \leq \sqrt{2\alpha\varepsilon}$.*

The final element in the proof is the following lemma that bounds the bias of the estimator:

Lemma 34. *For $M \geq 0$, $\beta = \frac{1}{\sigma^2\beta(M+1)H}$, we have $\left\|b_{t,h}\right\|_\infty \leq \sigma R \exp\left(-\gamma\beta\lambda_{\min}M\right)$.*

Putting these bounds together with the above derivations concludes the proof of Theorem 4.4.1.

4.4.4 Proofs

The proof of Proposition 3

The proof is based on Lagrangian duality: for each $h \in [H-1]$, we introduce a set of multipliers $V_h \in \mathbb{R}^{|X_h|}$ and $Z_h \in \mathbb{R}^{d\times d}$ corresponding to the two sets of constraints connecting $\mu_{t,h}$ and $u_{t,h}$, and $\rho_{t,h}$ for the normalization constraint of $\mu_{t,h}$. Then, we can write the Lagrangian of the constrained optimization problem as

$$\begin{aligned}\mathcal{L}(\mu, u; V, Z, \rho) = &\sum_{h=1}^{H-1}\sum_{s=1}^{t-1}\langle\mu_h, \widehat{r}_{s,h}\rangle + \langle Z_h, \Phi_h^\mathsf{T}(\text{diag}(u_h) - \text{diag}(\mu_h))\Phi_h\rangle \\ &+ \sum_{h=1}^{H-1}\left(\rho_h(1 - \langle\mu_h, \mathbf{1}\rangle) - \frac{1}{\eta}D(\mu_h\|\mu_{0,h}) - \frac{1}{\alpha}D_C(u_h\|\mu_{0,h})\right) \\ &+ V_1(x_1)(1 - E^\mathsf{T}u_1) + \sum_{h=1}^{H-1}\langle V_{h+1}, P^\mathsf{T}\mu_h - E^\mathsf{T}u_{h+1}\rangle.\end{aligned}$$

For any $h \in [H-1]$, for any $x \in \mathcal{X}_h$, $a \in A(x)$, denote $Q_Z(x,a) = \varphi(x,a)^\mathsf{T} Z_{h(x)}\varphi(x,a)$, $P_{x,a}V_{h+1} = \sum_{x'\in\mathcal{X}_{h+1}} P(x'|x,a)V_{h+1}(x')$ and $\Delta_{t,Z}(x,a) = \sum_{s=1}^{t-1}\widehat{r}_{s,h(x)}(x,a) + P_{x,a}V_{h(x)+1} - Q_Z(x,a)$. The above Lagrangian is strictly concave, so the maximum of $\mathcal{L}(\mu, d; V, Z, \rho)$ can be found by setting the derivatives with respect to its parameters to zero. This gives the following expressions for the choices of π and μ:

$$\pi^*_{t,h}(a|x) = \pi_{0,h}(a|x)e^{\alpha(Q_Z(x,a)-V_h(x))},$$

$$\mu^*_{t,h}(x,a) = \mu_0(x,a) e^{\eta(\Delta_{t,Z}(x,a) - \rho_{t,h})},$$

From the constraint $\sum_{x \in \mathcal{X}_h, a \in A(x)} \mu^*_{t,h}(x,a) = 1$ for all h, we get that

$$\rho^*_{t,h} = \frac{1}{\eta} \log \left(\sum_{x \in \mathcal{X}_h, a \in A(x)} \mu_0(x,a) e^{\eta \Delta_{t,Z}(x,a)} \right)$$

and from the constraint $\sum_a \pi^*_t(a|x) = 1$, we get

$$V^*_h(x) = \frac{1}{\alpha} \log \left(\sum_a \pi_0(a|x) e^{\alpha Q_Z(x,a)} \right).$$

We will further use the notation $V_Z(x) := V^*_h(x)$. Then, by plugging $\pi^*_{t,h}$, $\mu^*_{t,h}$, $V_Z(x)$ into the Lagrangian, we get

$$\mathcal{G}_t(Z) = \mathcal{L}(\mu^*, u^*; V^*, Z, \rho^*) = \frac{1}{\eta} \sum_{h=1}^{H-1} \log \left(\sum_{x \in \mathcal{X}_h, a \in A(x)} \mu_0(x,a) e^{\eta \Delta_{t,Z}(x,a)} \right) + V_Z(x_1).$$

Then, the solution of the optimization problem can be written as

$$\begin{aligned} \max_{\mu, u \in U} \min_{V, Z, \rho} \mathcal{L}(\mu, u; V, Z, \rho) &= \min_{V, Z, \rho} \max_{\mu, u \in U} \mathcal{L}(\mu, u; V, Z, \rho) = \min_Z \mathcal{L}(\mu^*, u^*; V^*, Z, \rho^*) \\ &= \min_Z \mathcal{G}_t(Z). \end{aligned}$$

This concludes the proof. ■

The proof of Lemma 30

The proof is based on a variation of the FTRL analysis that studies the evolution of the potential function Ψ_t defined for each t as

$$\Psi_t = \max_{(\mu,u) \in \mathcal{U}^2_\Phi} \left\{ \sum_{s=1}^{t-1} \sum_{h=1}^{H} \langle \mu_h, \widehat{r}_{s,h} \rangle - \frac{1}{\eta} D(\mu \| \mu_0) - \frac{1}{\alpha} D_C(u \| u_0) \right\}.$$

This definition immediately implies the following bound:

$$\Psi_{T+1} \geq \sum_{s=1}^{T} \sum_{h=1}^{H-1} \langle \mu^*_h, \widehat{r}_{s,h} \rangle - \frac{1}{\eta} D(\mu^* \| \mu_0) - \frac{1}{\alpha} D_C(u^* \| u_0). \tag{4.13}$$

To proceed, we will heavily exploit the fact that, by Proposition 3, the potential satisfies $\Psi_t = \min_Z \mathcal{G}_t$. Introducing the notation $Z_t^* = \arg\min_Z \mathcal{G}_t(Z)$, we have

$$\begin{aligned}
&\Psi_{t+1} - \Psi_t = \mathcal{G}_{t+1}(Z_{t+1}^*) - \mathcal{G}_t(Z_t^*) \leq \mathcal{G}_{t+1}(Z_t^*) - \mathcal{G}_t(Z_t^*) \\
&= \frac{1}{\eta}\sum_{h=1}^{H-1}\Bigg(\log\Bigg(\sum_{x\in\mathcal{S}_h, a\in\mathcal{A}} \mu_{0,h}(x,a)\exp\Bigg(\eta\Bigg(\sum_{s=1}^{t}\widehat{r}_{s,h}(x,a) + P_{x,a}V_{Z_t^*} - Q_{Z_t^*}(x,a)\Bigg)\Bigg)\Bigg) \\
&- \log\Bigg(\sum_{x'\in\mathcal{S}_h, a'\in\mathcal{A}} \mu_{0,h}(x',a')\exp\Bigg(\eta\Bigg(\sum_{s=1}^{t-1}\widehat{r}_{s,h}(x,a) + P_{x',a'}V_{Z_t^*} - Q_{Z_t^*}(x',a')\Bigg)\Bigg)\Bigg)\Bigg) \\
&= \frac{1}{\eta}\sum_{h=1}^{H-1}\log\left(\sum_{x\in\mathcal{X}_h, a\in\mathcal{A}} \mu_{t,h}(x,a)\exp\left(\eta\widehat{r}_{t,h}(x,a)\right)\right) \\
&\qquad\qquad \text{(using the expression of } \mu_{t,h}(x,a) \text{ obtained in Proposition 3)} \\
&\leq \frac{1}{\eta}\sum_{h=1}^{H-1}\log\left(1 + \sum_{x\in\mathcal{X}_h, a\in\mathcal{A}} \mu_{t,h}(x,a)\eta\left(\widehat{r}_{t,h}(x,a) + \eta\widehat{r}^2_{t,h}(x,a)\right)\right) \\
&\leq \sum_{h=1}^{H-1}\left(\langle\mu_{t,h}, \widehat{r}_{t,h}\rangle + \eta\langle\mu_{t,h}, \widehat{r}_t^2\rangle\right),
\end{aligned}$$

where in the last two lines we have used the inequalities $e^z \leq 1 + z + z^2$, which holds for $z \leq 1$ and $\log(1+z) \leq z$, which holds for all $z > -1$, which conditions are verified due to our constraint on η. Summing up both sides for all t and combining the result with the inequality (4.13), we obtain

$$\begin{aligned}
\widehat{\mathfrak{R}}_T = \sum_{s=1}^{T}\sum_{h=1}^{H-1}\langle\mu_h^*, \widehat{r}_{s,h}\rangle - \sum_{t=1}^{T}\langle\mu_t, \widehat{r}_t\rangle \leq \eta\sum_{t=1}^{T}\sum_{h=1}^{H-1}\langle\mu_{t,h}, \widehat{r}^2_{t,h}\rangle + \frac{1}{\eta}D(\mu^*||\mu_0) \\
+ \frac{1}{\alpha}D_C(u^*||u_0),
\end{aligned}$$

concluding the proof. ■

The proof of Lemma 31

We start by using the the definition of $\widehat{\theta}_{t,h}$ to obtain

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}\widehat{\mu}_{t,h}(x,a)\left\langle\varphi(x,a),\widehat{\theta}_{t,h}\right\rangle^2\right]\\
&\quad=\mathbb{E}_t\left[\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}\widehat{\mu}_{t,h}(x,a)\text{tr}\left(\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\theta}_{t,h}\widehat{\theta}_{t,h}^\mathsf{T}\right)\right]\\
&\quad=\mathbb{E}_t\left[\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}\widehat{u}_{t,h}(x,a)\text{tr}\left(\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\theta}_{t,h}\widehat{\theta}_{t,h}^\mathsf{T}\right)\right]\\
&\qquad\qquad\qquad\text{(by the constraint }\Phi_h^\mathsf{T}\text{diag}(\widehat{\mu}_t)\Phi_h=\Phi_h^\mathsf{T}\text{diag}(\widehat{u}_t)\Phi_h)\\
&\quad=\mathbb{E}_t\left[\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}u_{t,h}(x,a)\text{tr}\left(\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\theta}_{t,h}\widehat{\theta}_{t,h}^\mathsf{T}\right)\right]\\
&\quad+\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}\left(u_{t,h}(x,a)-\widehat{u}_{t,h}(x,a)\right)\mathbb{E}_t\left[\left\langle\varphi(x,a),\widehat{\theta}_{t,h}\right\rangle^2\right]\\
&\quad\leq\mathbb{E}_t\left[\sum_{x\in\mathcal{X}_h,a\in\mathcal{A}}u_{t,h}(x,a)\text{tr}\left(\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\theta}_{t,h}\widehat{\theta}_{t,h}^\mathsf{T}\right)\right]\\
&\quad+\left\|u_{t,h}-\widehat{u}_{t,h}\right\|_1\cdot\left\|\mathbb{E}_t\left[\widehat{r}_{t,h}^2\right]\right\|_\infty.
\end{aligned}$$

The second term can be bounded straightforwardly by $\left\|u_{t,h}-\widehat{u}_{t,h}\right\|_1(M+1)^2$, using Lemma 32 to bound $\left\|\widehat{r}_{t,h}\right\|_\infty\leq(M+1)$. As for the first term, we have

$$\begin{aligned}
&(1-\gamma)\mathbb{E}_t\left[\sum_{x,a}u_{t,h}(x,a)\text{tr}\left(\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\theta}_{t,h}\widehat{\theta}_{t,h}^\mathsf{T}\right)\right]\\
&\leq(1-\gamma)\mathbb{E}_t\left[\sum_{x,a}\text{tr}\left(u_{t,h}(x,a)\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\Sigma}_{t,h}^+\varphi(X_{t,h},A_{t,h})\varphi(X_{t,h},A_{t,h})^\mathsf{T}\widehat{\Sigma}_{t,h}^+\right)\right]\\
&\leq(1-\gamma)\mathbb{E}_t\left[\sum_{x,a}\text{tr}\left(u_{t,h}(x,a)\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\Sigma}_{t,h}^+\varphi(X_{t,h},A_{t,h})\varphi(X_{t,h},A_{t,h})^\mathsf{T}\widehat{\Sigma}_{t,h}^+\right)\right]
\end{aligned}$$

$$+\gamma\mathbb{E}_t\left[\sum_{x,a}\operatorname{tr}\left(u(x,a)\varphi(x,a)\varphi(x,a)^\mathsf{T}\widehat{\Sigma}^+_{t,h}\varphi(X_{t,h},A_{t,h})\varphi(X_{t,h},A_{t,h})^\mathsf{T}\widehat{\Sigma}^+_{t,h}\right)\right]$$
$$=\mathbb{E}_t\left[\operatorname{tr}\left(\Sigma_{t,h}\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\widehat{\Sigma}^+_{t,h}\right)\right],$$

where we used $|r_{t,h}(X_{t,h},A_{t,h})|\leq 1$ in the first inequality. For ease of readability, we will omit the indices h in the rest of the proof. Using the definition of Σ_t^+ and elementary manipulations, we get

$$\mathbb{E}_t\left[\operatorname{tr}\left(\Sigma_t\Sigma_t^+\Sigma_t\Sigma_t^+\right)\right]=\beta^2\cdot\mathbb{E}_t\left[\operatorname{tr}\left(\Sigma^*\left(\sum_{k=0}^M C_k\right)\Sigma_t\left(\sum_{j=0}^M C_j\right)\right)\right]$$
$$=\beta^2\mathbb{E}_t\left[\sum_{k=0}^M\sum_{j=0}^M\operatorname{tr}\left(\Sigma_tC_k\Sigma_tC_j\right)\right]$$
$$=\beta^2\mathbb{E}_t\left[\sum_{k=0}^M\operatorname{tr}\left(\Sigma_tC_k\Sigma_tC_k\right)\right]+2\beta^2\mathbb{E}_t\left[\sum_{k=0}^M\sum_{j=k+1}^M\operatorname{tr}\left(\Sigma_tC_k\Sigma_tC_j\right)\right].$$

Let us first address the first term on the right hand side. Applying Lemma 20 with $\widetilde{H}=\Sigma_t$, we get

$$\beta^2\sum_{k=0}^M\operatorname{tr}\left(\mathbb{E}\left[\Sigma_tC_k\Sigma_tC_k\right]\right)\leq\beta^2\sum_{k=0}^M\operatorname{tr}\left(\Sigma^2\left(I-2\beta\Sigma\right)^k\right)$$
$$+\beta^2\sum_{j=1}^d\sum_{k=0}^M\lambda_j^2(\Sigma)\frac{\beta\sigma^2\operatorname{tr}(\Sigma)}{2\lambda_j(\Sigma)}\exp\left(\beta^2\sigma^2\operatorname{tr}(\Sigma)k+2\beta\lambda_j(\Sigma)\right)$$
$$=\beta\operatorname{tr}\left(\Sigma(I-(I-\beta\Sigma)^M)\right)$$
$$+\beta^3\frac{\sigma^2\operatorname{tr}(\Sigma)}{2}\sum_{j=1}^d\lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\frac{\exp\left(\beta^2\sigma^2\operatorname{tr}(\Sigma)(M+1)\right)-1}{\exp\left(\beta^2\sigma^2\operatorname{tr}(\Sigma)\right)-1}$$
$$\leq\beta\operatorname{tr}(\Sigma)+\beta^3\frac{\sigma^2\operatorname{tr}(\Sigma)}{2\beta^2\sigma^2\operatorname{tr}(\Sigma)}\sum_{j=1}^d\lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\exp\left(\beta^2\sigma^2\operatorname{tr}(\Sigma)(M+1)\right)$$
$$\leq\beta\operatorname{tr}(\Sigma)+\frac{\beta}{2}\sum_{j=1}^d\lambda_j(\Sigma)\exp(2\beta\lambda_j(\Sigma))\exp\left(\beta^2\sigma^2\operatorname{tr}(\Sigma)(M+1)\right)$$

$$\leq \beta\sigma^2 d + \frac{\beta}{2}\sum_{j=1}^{d} \sigma^2 \exp(2\beta\sigma^2) \exp\left(\beta^2\sigma^4 d(M+1)\right) \leq 3,$$

where we used the condition $\beta \leq \frac{1}{2\sigma^2 d\sqrt{(M+1)}} \leq \frac{1}{2\sigma^2}$ and the fact that $(I - \beta(2-\beta\sigma^2)\Sigma)^M \succcurlyeq 0$ by the same condition. We also used an observation that our assumption on the contexts implies $\mathrm{tr}\left(\Sigma\right) \leq \mathrm{tr}\left(\sigma^2 I\right) = \sigma^2 d$, so again by our condition on β it implies the final bound.

Moving on to the second term, we first note that for any $j > k$, the conditional expectation of B_j given $B_{\leq k} = (B_1, B_2, \dots B_k)$ satisfies $\mathbb{E}\left[C_k \middle| B_{\leq k}\right] = C_k(I - \beta\Sigma)^{j-k}$ due to conditional independence of all B_j given B_k, for $i > k$. We make use of this equality by writing

$$\beta^2 \sum_{k=0}^{M}\sum_{j=k+1}^{M} \mathbb{E}\left[\mathrm{tr}\left(\Sigma_t C_k \Sigma_t C_j\right)\right] = \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \mathrm{tr}\left(\Sigma_t C_k \Sigma_t C_j\right) \middle| B_{\leq k}\right]\right]$$

$$= \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \mathrm{tr}\left(\Sigma_t C_k \Sigma_t C_j (I - \beta\Sigma_t)^{j-k}\right) \middle| B_{\leq k}\right]\right]$$

$$= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma_t C_k \Sigma_t C_k \Sigma_t^{-1}\left(I - (I - \beta\Sigma_t)^{M-k}\right)\right) \middle| B_{\leq k}\right]\right]$$

$$\leq \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathrm{tr}\left(\Sigma_t C_k \Sigma_t C_k \Sigma_t^{-1}\right) \middle| B_{\leq k}\right]\right]$$

(due to $(I - \beta\Sigma_t)^{M-k} \succcurlyeq 0$)

$$= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathrm{tr}\left(C_k \Sigma C_k\right)\right]$$

$$\leq \beta \sum_{k=0}^{M} \mathrm{tr}\left(\Sigma\left(I - 2\beta\Sigma\right)^k\right)$$

$$+ \beta \sum_{j=1}^{d}\sum_{k=0}^{M} \lambda_j(\Sigma) \frac{\beta\sigma^2 \mathrm{tr}\left(\Sigma\right)}{2\lambda_j(\Sigma)} \exp\left(\beta^2\sigma^2 \mathrm{tr}\left(\Sigma\right) k + 2\beta\lambda_j(\Sigma)\right)$$

(applying Lemma 20 with $\widetilde{H} = I$)

$$
\begin{aligned}
&\leq d + \frac{\beta^2\sigma^2 \mathrm{tr}\,(\Sigma)}{2} \sum_{j=1}^{d} \frac{\exp\big(\beta^2\sigma^2\mathrm{tr}\,(\Sigma)\,(M+1) + 2\beta\lambda_j(\Sigma)\big) - 1}{\exp\left(\beta^2\sigma^2\mathrm{tr}\,(\Sigma) + 2\beta\lambda_j(\Sigma)\right) - 1} \\
&\leq d + \frac{1}{2}\sum_{j=1}^{d} \exp\big(\beta^2\sigma^2\mathrm{tr}\,(\Sigma)\,(M+1) + 2\beta\lambda_j(\Sigma)\big) \\
&\leq d + \frac{1}{2}\sum_{j=1}^{d} \exp\big(\beta^2\sigma^4 d(M+1) + 2\beta\lambda_j(\Sigma)\big) \leq 5d.
\end{aligned}
$$

The proof of the theorem is concluded by putting everything together. ■

The proof of Lemma 34

We first observe that the bias of $\widehat{\theta}_{t,h}$ can be easily expressed as

$$
\begin{aligned}
\mathbb{E}_t\big[\widehat{\theta}_{t,h}\big] &= \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h}, A_{t,h})^\mathsf{T}\theta_{t,h}\right] \\
&= \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\mathbb{E}_t\left[\varphi(X_{t,h}, A_{t,h})\varphi(X_{t,h}, A_{t,h})^\mathsf{T}\right]\theta_{t,h} \\
&= \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\Sigma_{t,h}\theta_{t,h} = \theta_{t,h} - (I - \beta\Sigma_{t,h})^M\theta_{t,h}.
\end{aligned}
$$

Thus, the bias is bounded as

$$
\left|\mathbb{E}_t\left[\varphi(X_{t,h}, a)^\mathsf{T}(I - \beta\Sigma_{t,h})^M\theta_{t,h}\right]\right| \leq \left\|\varphi(X_{t,h}, a)\right\|_2 \cdot \left\|\theta_{t,h}\right\|_2 \left\|(I - \beta\Sigma_{t,h})^M\right\|_{\mathrm{op}}.
$$

In order to bound the last factor above, observe that $\Sigma_{t,h} \succcurlyeq \gamma\Sigma_h$ due to the uniform exploration used in the first layer by MDP-LINEXP3, which implies that

$$
\left\|(I - \beta\Sigma_{t,h})^M\right\|_{\mathrm{op}} \leq (1 - \gamma\beta\lambda_{\min})^M \leq \exp\left(-\gamma\beta\lambda_{\min}M\right),
$$

where the second inequality uses $1 - z \leq e^{-z}$ that holds for all z. This concludes the proof. ■

The proof of Lemma 33

The proof consists of two main components: proving that the conditional relative entropy between u_t and $\widehat{u}_t$ can be bounded in terms of the optimization error ε, and then using this quantity to bound the total variation distance between these occupancy measures. For ease of readability, we state these results as separate lemmas.

We will first need the following statement:

Lemma 35. $D_C(\widehat{u}_t \| u_t) \leq \alpha\varepsilon$.

The proof follows along similar lines as the proof of Lemma 1 in Bas-Serrano et al. [23]. To preserve clarity, we delegate its proof to Section 4.4.4 below. The second lemma lemma bounds the relative entropy between two occupancy measures in terms of their *conditional* relative entropies:

Lemma 36. *For any two occupancy measures u and u' and any h, we have*

$$D\left(u_h \| u'_h\right) \leq \sum_{k=1}^{h} D_C(u_k \| u'_k).$$

Proof. The proof follows from exploiting some basic properties of the relative entropy. Specifically, the result follows from the following chain of inequalities:

$$\begin{aligned}
D(u_h \| u'_h) &= D(E^\mathsf{T} u_h \| E^\mathsf{T} u'_h) + D_C(u_h \| u'_h) \\
&\qquad \text{(by the chain rule of the relative entropy)} \\
&= D(P^\mathsf{T} u_{h-1} \| P^\mathsf{T} u'_{h-1}) + D_C(u_h \| u'_h) \\
&\qquad \text{(by the fact that } u \text{ and } u' \text{ are valid occupancy measures)} \\
&\leq D(u_{h-1} \| u'_{h-1}) + D_C(u_h \| u'_h) \\
&\qquad \text{(by the data processing inequality)} \\
&\leq \cdots \leq \sum_{k=1}^{h} D_C(u_k \| u'_k),
\end{aligned}$$

where the last step follows from iterating the same argument for all layers. □

Putting the above two lemmas together and using Pinsker's inequality, we obtain

$$\|\widehat{u}_{t,h} - u_{t,h}\|_1 \leq \sqrt{2D\big(\widehat{u}_{t,h} \big\| u_{t,h}\big)} \leq \sqrt{2\sum_{k=1}^{h} D_C\big(\widehat{u}_{t,k} \big\| u_{t,k}\big)} \leq \sqrt{2D_C\big(\widehat{u}_t \big\| u_t\big)} \leq \sqrt{2\alpha\varepsilon},$$

concluding the proof of Lemma 33. ■

The proof of Lemma 35

For the proof, let us introduce the notation $\widetilde{\mu}_{t,h}$ with

$$\widetilde{\mu}_{t,h}(x,a) = \frac{\mu_{0,h}(x,a)e^{\eta\Delta_{t,Z_t}(x,a)}}{\sum_{(x',a')\in(\mathcal{S}_h\times\mathcal{A})}\mu_{0,h}(x,a)e^{\eta\Delta_{t,Z_t}(x,a)}}.$$

and also $\mathcal{G}_{t,h}(Z) = \frac{1}{\eta}\log\left(\sum_{x\in\mathcal{X}_h,a\in A(x)}\mu_0(x,a)e^{\eta\Delta_{t,Z}(x,a)}\right)$ and $Z_t^* = \arg\min_Z \mathcal{G}_t(Z)$. Then, observe that

$$\begin{aligned}
D(\widehat{\mu}_{t,h}\|\widetilde{\mu}_{t,h}) &= \sum_{x,a\in\mathcal{S}_h\times\mathcal{A}}\widehat{\mu}_{t,h}(x,a)\log\frac{\widehat{\mu}_{t,h}(x,a)}{\widetilde{\mu}_{t,h}(x,a)}\\
&= \eta\left\langle\widehat{\mu}_{t,h},\Delta_{t,Z_t^*} - \mathcal{G}_{t,h}(Z_t^*)\mathbf{1} - \Delta_{t,Z_t} + \mathcal{G}_{t,h}(Z_t)\mathbf{1}\right\rangle\\
&= \eta\left\langle\widehat{\mu}_{t,h},P_hV_{Z_t^*} - Q_{Z_t^*} - P_hV_{Z_t} + Q_{Z_t}\right\rangle + \eta\left(\mathcal{G}_{t,h}(Z_t^*) - \mathcal{G}_{t,h}(Z_t)\right)\\
&= \eta\sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\sum_{x'\in\mathcal{S}_{h+1}}\widehat{\mu}_{t,h}(x,a)P(x'|x,a)(V_{Z_t^*}(x') - V_{Z_t}(x'))\\
&\quad + \eta\left(\mathcal{G}_{t,h}(Z_t^*) - \mathcal{G}_{t,h}(Z_t)\right) + \eta\sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\widehat{\mu}_{t,h}(x,a)\varphi(x,a)^\mathsf{T}(Z_{t,h} - Z_{t,h}^*)\varphi(x,a)\\
&= \eta\sum_{(x',a')\in(\mathcal{S}_{h+1}\times\mathcal{A})}\widehat{u}_{t,h+1}(x',a')(V_{Z_t^*}(x') - V_{Z_t}(x')) + \eta(\mathcal{G}_{t,h}(Z_t) - \mathcal{G}_{t,h}(Z_t^*)).\\
&\quad + \eta\sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\widehat{u}_{t,h}(x,a)\varphi(x,a)^\mathsf{T}(Z_{t,h} - Z_{t,h}^*)\varphi(x,a).
\end{aligned}$$

Here, the last equality follows from the fact that $(\widehat{\mu}_t,\widehat{u}_t)$ satisfy the constraints of the optimization problem (4.10). On the other hand, we have

$$\begin{aligned}
D_C(\widehat{u}_{t,h}\|u_{t,h}) &= \sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\widehat{u}_{t,h}(x,a)\log\frac{\widehat{\pi}_{t,h}(a|x)}{\pi_{t,h}(a|x)}\\
&=\alpha\sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\widehat{u}_{t,h}(x,a)\sum_{x'\in\mathcal{S}_{h+1}}P(x'|x,a)(V_{Z_t^*}(x') - V_{Z_t}(x'))\\
&\quad + \alpha\sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})}\widehat{u}_{t,h}(x,a)\varphi(x,a)^\mathsf{T}(Z_{t,h} - Z_{t,h}^*)\varphi(x,a)\\
&=\alpha\sum_{(x',a')\in(\mathcal{S}_{h+1}\times\mathcal{A})}\widehat{u}_{t,h+1}(x,a)(V_{Z_t^*}(x') - V_{Z_t}(x'))
\end{aligned}$$

$$+\alpha \sum_{(x,a)\in(\mathcal{S}_h\times\mathcal{A})} \widehat{u}_{t,h}(x,a)\varphi(x,a)^\mathsf{T}(Z_{t,h}-Z^*_{t,h})\varphi(x,a),$$

where the last equality follows from the fact that $\widehat{u}_t$ is a valid occupancy measure, as shown in Equation (4.8). Putting the two equalities together, we get

$$\frac{D(\widehat{\mu}_{t,h}\|\mu_{t,h})}{\eta} - \frac{D_C(\widehat{u}_{t,h}\|u_{t,h})}{\alpha} = \mathcal{G}_{t,h}(Z_t) - \mathcal{G}_{t,h}(Z^*_t).$$

Then, summing up over all h gives

$$\frac{D(\widehat{\mu}_t\|\mu_t)}{\eta} - \frac{D_C(\widehat{u}_t\|u_t)}{\alpha} = \sum_{h=1}^{H}\left(\mathcal{G}_{t,h}(Z_t) - \mathcal{G}_{t,h}(Z^*_t)\right) = \mathcal{G}_t(Z_t) - \mathcal{G}_t(Z^*_t) \leq \varepsilon.$$

Reordering gives the result. ■

The proof of Lemma 32

The claim is proven by the following straightforward calculation:

$$\begin{aligned}
\eta\cdot\left|\left\langle\varphi(X_{t,h},a),\widehat{\theta}_t\right\rangle\right| &= \eta\cdot\left|\varphi(X_{t,h},a)^\mathsf{T}\widehat{\Sigma}^+_{t,h}\varphi(X_{t,h},a)\left\langle\varphi(X_{t,h},a),\theta_t\right\rangle\right| \\
&\leq \eta\left|\varphi(X_{t,h},a)^\mathsf{T}\widehat{\Sigma}^+_{t,h}\varphi(X_{t,h},a)\right| \leq \eta\sigma^2\left\|\widehat{\Sigma}^+_{t,h}\right\|_{\text{op}} \\
&\leq \eta\sigma^2\beta\left(1+\sum_{k=1}^{M}\|C_{k,h}\|_{\text{op}}\right) \leq \eta(M+1)/2,
\end{aligned}$$

where we used the fact that our choice of β ensures $\|C_{k,h}\|_{\text{op}} = \left\|\prod_{j=0}^{k}(I-\beta B_{j,h})\right\|_{\text{op}} \leq 1$. ■

Implementation by optimizing the empirical loss

This section outlines a possible implementation of the policy update steps based on approximate minimization of an empirical counterpart of the loss function $\mathcal{G}_t$. To this end, we define

$$\mathcal{G}_{t,h}(Z) = \frac{1}{\eta}\log\left(\sum_{x,a}\mu_0(x,a)e^{\eta\Delta_{Z,t,h}(x,a)}\right)$$

and its empirical counterpart that replaces the expectation by an empirical mean over state-action pairs sampled from μ_0. Concretely, for all h, we let $(X_h(i), A_h(i))_{i=1}^N$ be N independent samples from μ_0 that can be obtained by running policy π_0 in the transition model P. Using these samples, we define

$$\widehat{\mathcal{G}}_{t,h}(Z) = \frac{1}{\eta} \log \left(\sum_{n=1}^{N} e^{\eta \Delta_{Z,t,h}(X_h(i), A_h(i))} \right). \tag{4.14}$$

This objective function has several desirable properties: it is convex in Z, has bounded gradients, and is $(\alpha + \eta)$-smooth. Furthermore, its gradients can be evaluated efficiently in $\mathcal{O}(N)$ time, given that we can efficiently evaluate expectations of the form $\sum_{x'} P(x'|x, a)V(x')$. As a result, it can be optimized up to arbitrary precision ε in time polynomial in $1/\varepsilon$ and N.

The downside of this estimator is that it is potentially biased. Nevertheless, as the following lemma shows, it is well-concentrated around the true objective function, under some reasonable conditions:

Lemma 37. *Fix Z and suppose that $|\Delta_Z(x, a)| \leq B$ for all x, a. Then, with probability at least $1 - \delta$, the following holds:*

$$\left| \widehat{\mathcal{G}}_{t,h}(Z) - \mathcal{G}_{t,h}(Z) \right| \leq 56 \sqrt{\frac{\log(1/\delta)}{N}}.$$

This statement is a variant of Theorem 1 from Bas-Serrano et al. [23], with the key difference being that being able to exactly calculate expectations with respect to $P(\cdot|x, a)$ enables us to prove a tighter bound.

Proof. Let us start by defining the shorthand notations $\widehat{S}_i = \Delta_{Z,t}(X_h(i), A_h(i))$ and $\overline{\mathcal{W}} = \frac{1}{N} \sum_{i=1}^{N} e^{\eta S_i}$. Furthermore, we define the function

$$f(s_1, s_2, \ldots, s_N) = \frac{1}{N} \sum_{i=1}^{N} e^{\eta s_i}$$

and notice that it satisfies the bounded-differences property

$$f(s_1, s_2, \ldots, s_i, \ldots, s_N) - f(s_1, s_2, \ldots, s_i', \ldots, s_N) = \frac{1}{N} \left(e^{\eta s_i} - e^{\eta s_i'} \right) \leq \frac{\eta e^{2\eta B}}{N}.$$

Here, the last step follows from Taylor's theorem that implies that there exists a $\chi \in (0, 1)$ such that

$$e^{\eta s'_i} = e^{\eta s_i} + \eta e^{\eta \chi \left(s'_i - s_i\right)}$$

holds, so that $e^{\eta s'_i} - e^{\eta s_i} = \eta e^{\eta \chi \left(s'_i - s_i\right)} \leq \eta e^{2\eta B}$, where we used the assumption that $|s_i - s'_i| \leq 2B$ in the last step. Notice that our assumption $\eta B \leq 1$ further implies that $e^{2\eta B} \leq e^2$. Thus, also noticing that $W = f(S_1, \ldots, S_N)$, we can apply McDiarmid's inequality that to show that the following holds with probability at least $1 - \delta'$:

$$|W - \mathbb{E}\left[W\right]| \leq \eta e^2 \sqrt{\frac{\log(2/\delta')}{N}}. \tag{4.15}$$

Thus, we can write

$$\begin{aligned} \widehat{\mathcal{G}}_{t,h}(\theta) - \mathcal{G}_{t,h}(\theta) &= \frac{1}{\eta} \log\left(W\right) - \frac{1}{\eta} \log\left(\mathbb{E}\left[\overline{\mathcal{W}}\right]\right) = \frac{1}{\eta} \log\left(\frac{W}{\mathbb{E}\left[\overline{\mathcal{W}}\right]}\right) \\ &= \frac{1}{\eta} \log\left(1 + \frac{W - \mathbb{E}\left[\overline{\mathcal{W}}\right]}{\mathbb{E}\left[\overline{\mathcal{W}}\right]}\right) \leq \frac{W - \mathbb{E}\left[\overline{\mathcal{W}}\right]}{\eta \mathbb{E}\left[\overline{\mathcal{W}}\right]} \leq e^4 \sqrt{\frac{\log(2/\delta')}{N}}, \end{aligned}$$

where the last line follows from the inequality $\log(1 + u) \leq u$ that holds for $u > -1$ and our assumption on η that implies $\overline{\mathcal{W}} \geq e^{-2}$. Similarly, we can show

$$\mathcal{G}_{t,h}(\theta) - \widehat{\mathcal{G}}_{t,h}(\theta) = \frac{1}{\eta} \log\left(1 + \frac{\mathbb{E}\left[\overline{\mathcal{W}}\right] - W}{W}\right) \leq \frac{\mathbb{E}\left[W\right] - W}{\eta W} \leq e^4 \sqrt{\frac{\log(2/\delta')}{N}}.$$

This concludes the proof. □

Chapter 5

Conclusions

5.1 Sequential influence maximization

In this section we highlight some features of our results and discuss directions for future work. Our main result is showing for all considered graph models both instance-dependent bounds of order $O(\log T)$ and worst-case bounds of order $O(\sqrt{T})$ on the quantile regret of our algorithm. Notably, our bounds hold for both the subcritical and supercritical regimes of the random-graph models considered, and show no explicit dependence on the number of nodes n.

Previous work. Related online influence maximization algorithms consider more general classes of networks, but make more restrictive assumptions about the interplay between rewards and feedback. One line of work explored by Wen et al. [147], Wang and Chen [145] assumes that the algorithm receives *full feedback* on where the information reached in the previous trials (i.e., not only the number of influenced nodes, but their exact identities and influence paths, too). Clearly, such detailed measurements are nearly impossible to obtain in practice, as opposed to the local observations that our algorithm requires.

Another related setup was considered by Carpentier and Valko [38], whose algorithm only receives feedback about the nodes that were directly influenced by the chosen node, but the model does not assume that neighbours in the graph share the information to further neighbours and counts the reward only by the nodes directly connected to the selected one. That is, in contrast to our work, this work does not attempt to show any relation between local and global influence maximization. One downside to all the above works is that they all provide rather conservative performance guarantees: On one hand, Wen et al. [147] and Carpentier and Valko [38] are concerned with worst-case regret bounds that uniformly hold for all problem instances for a fixed time horizon T. On the other hand, the bounds of Wang and Chen [145] depend on topological (rather than probabilistic) characteristics of the underlying graph structure, which inevitably leads to conservative results. For example, their bounds instantiated in our graph model lead to a regret bound of order $n^3 \log T$, which is virtually void of meaning in our regime of interest where n is very large (e.g, in the order of billions). In contrast, our bounds do not show explicit dependence on n. In this light, our work can be seen as the first one that takes advantage of specific probabilistic characteristics of the mechanism of information spreading to obtain strong instance-dependent global performance guarantees, all while having access to only local observations.

Other related framework is stochastic multi-armed bandits with partial monitoring [4, 22, 82]. In this setting the loss is not directly observed by the learner, which makes this setting applicable to a wider range of practical problem. Works on the partial monitoring would not capture the graph structure, so that results cannot be applied directly, and in addition the regret would scale with n^2.

Tightness of our bounds. In terms of dependence on T, both our instance-dependent and worst-case bounds are near-optimal in their respective settings: even in the simpler stochastic multi-armed bandit problem, the best possible regret bounds are $\Omega_T(\log T)$ and $\Omega_T(\sqrt{T})$ in the respective settings [12, 15, 30]. The optimality of our bounds with respect to other parameters such as c^*, μ^* and n is less clear, but we believe that these factors cannot be improved substantially for the models that we studied in this work. As for the subproblem of identifying nodes with the highest degrees, we believe that our bounds on the number of suboptimal draws is essentially tight, closely matching the classic lower bounds by Lai and Robbins [86].

Our assumptions. One may wonder how far our argument connecting local and global influence maximization can be stretched. Clearly, not every random graph model enables establishing such a strong connection. Finally, let us comment on our condition that the number of vertices n needs to be "sufficiently large". We regard this condition as a technical artifact due to our proofs relying on asymptotic analysis. We expect that the required monotonicity property holds for small values of n under mild conditions. Whenever this is the case, the regret bounds remain valid.

Graph model. The degree of a given vertex has a binomial distribution. This is a strongly concentrated distribution with an exponentially decreasing tail. Many graphs from "real life" has much larger tails, for example power-law tails, and it is therefore important to study also random graph models with such behaviour. An example of such a random graph model is the configuration model (for a survey see [141]). This model is different from the family of the models where the nodes are connected with some probability. In the configuration model, nodes are associated with those degrees, and the random graph is generated by uniformly at random selecting a pair of stubs and connecting them. The model for the sequential spread of the influence can be the following: first, fix a realisation of a stochastic block model, and then assume that the information transmits between connected nodes with fixed probability, equal for all edges. The goal of the learner is then to minimize the Bayesian regret by taking the men of the expected regret over all realisations of the configuration model.

5.2 Adversarial contextual bandits

Our work is the first to address the natural adversarial variant of the widely popular framework of linear contextual bandits, thus filling an important gap in the literature. Our algorithm REALLINEXP3 achieves the optimal regret bound of of order $\sqrt{KdT}$ and runs in time polynomial in the relevant problem parameters. To our knowledge, REALLINEXP3 is the first computationally efficient algorithm to achieve near-optimal regret bounds in an adversarial contextual bandit problem, and is among the first ones to achieve any regret guarantees at all for an infinite set of policies (besides results on learning with surrogate losses, cf. 60). In the case of misspecified loss functions, our algorithm ROBUSTLINEXP3 achieves a regret guarantee of order $(Kd)^{1/3}T^{2/3} + \varepsilon\sqrt{d}T$.

Whether or not the overhead of $\varepsilon\sqrt{d}T$ can be improved is presently unclear: while Lattimore et al. [91] proved that the dependence on $\sqrt{d}$ is inevitable even in the stochastic linear bandit setting when K is large (say, order of T), the very recent work of Foster and Rakhlin [62] shows that the overhead can be reduced to $\varepsilon\sqrt{K}T$ in the same setting. These results together suggest that the regret bound $\sqrt{KdT}+\varepsilon\sqrt{\min\{K,d\}}T$ is achievable in for stochastic linear contextual bandits. Whether such guarantees can be achieved in the more challenging adversarial setting we considered here remains an interesting open problem.

The reader may be curious if it is possible to remove the i.i.d. assumption that we make about the contexts. Unfortunately, it can be easily shown that no learning algorithm can achieve sublinear regret if the contexts and losses are both allowed to be chosen by an adversary. To see this, we observe that one can embed the problem of online learning of one-dimensional threshold classifiers into our setting, which is known to be impossible to learn with sublinear regret [Ben-David et al., 134]. While one can conceive other assumptions on the contexts that make the problem tractable, such as assuming that the entire sequence of contexts is known ahead of time (the so-called *transductive setting* studied by 134), such assumptions may end up being a lot more artificial than our natural i.i.d. condition. In addition, it is unclear what the best achievable performance bounds in such alternative frameworks actually are. In contrast, the regret bounds we prove for REALLINEXP3 are essentially minimax optimal.

Our algorithm design and analysis introduces a couple of new techniques that could be of more general interest. First, a key element in our analysis is introducing a set of auxiliary bandit problems for each context x and relating the regrets in these problems to the expected regret in the contextual bandit problem (Lemma 25). While this lemma is stated in terms of linear losses, it can be easily seen to hold for general losses as long as one can construct unbiased estimates of the entire loss function. In this view, our algorithms can be seen as the first instances of a new family of contextual bandit methods that are based on estimating the loss functions rather than working with a policy class. An immediate extension of our approach is to assume that the loss functions belong to a reproducing kernel Hilbert space and define suitable kernel-based estimators analogously to our estimators—a widely considered setting in the literature on stochastic contextual bandits [131, 31, 34]. We also remark that our technique used to prove Lemma 25 is similar in nature to the reduction of stochastic sleeping bandit problems to static bandit problems used by **(author?)** [Kanade et al.], Neu and Valko [117].

A second potentially interesting algorithmic trick we introduce is the Matrix Geometric Resampling for estimating inverse covariance matrices. While such matrices are broadly used for loss estimation in the literature on adversarial linear bandits [104, 18, 51, 10], the complexity of computing them never seems to be discussed in the literature. Our MGR method provides a viable option for tackling this problem. For the curious reader, we remark that the relation between the iterations defining MGR and the dynamics of gradient descent for linear least-squares estimation is well-known in the stochastic optimization literature, where SGD is known to implement a spectral filter function approximating the inverse covariance matrix [125, 66, 20, 116].

Besides the most important question of whether or not our guarantees for the misspecified setting can be improved, we leave a few more questions open for further investigation. One limitation of our methods is that they require prior knowledge of the context distribution $\mathcal{D}$. We conjecture that it may be possible to overcome this limitation by designing slightly more sophisticated algorithms that estimate this distribution from data. Second, it appears to be an interesting challenge to prove versions of our performance guarantees that hold with high probability by using optimistically estimators as done by **(author?)** [Beygelzimer et al.], Neu [107], or if data-dependent bounds depending on the total loss of the best expert rather than T can be achieved in our setting [7, 9, 108]. We find it likely that such improvements are possible at the expense of a significantly more involved analysis.

5.3 Learning in episodic MDP

We merge two important lines of work on online learning in MDPs concerned with linear function approximation [73, 33] and bandit feedback with adversarial rewards [111, 113, 152]. Our results are the first in this setting and not directly comparable with any previous work, although some favorable comparisons can be made with previous results in related settings. In the tabular setting where $d = |\mathcal{S}||\mathcal{A}|$, our bounds exactly recover the minimax optimal guarantees first achieved by the O-REPS algorithm of Zimin and Neu [152]. For realizable linear function approximation, the work closest to ours is that of Cai et al. [33], who prove bounds of order $\sqrt{d^2 H^3 T}$, which is worse by a factor of $\sqrt{d}H$ than our result. Their setting, however, is not exactly comparable to ours due to the different assumptions about the feedback about the rewards and the knowledge of the transition function.

One particular strength of our work is providing a complete analysis of the propagation of optimization errors incurred while performing the updates. This is indeed a unique contribution in the related literature, where the effect of such errors typically go unaddressed. Specifically, the algorithms of Zimin and Neu [152], Rosenberg and Mansour [126], and Jin et al. [72] are all based on solving convex optimization problems similar to ours, the effect of optimization errors or potential methods for solving the optimization problems are not discussed at all. That said, we believe that the methods for calculating the updates discussed in Section 4.4.2 are far from perfect, and more research will be necessary to find truly practical optimization methods to solve this problem.

The most important open question we leave behind concerns the requirement to have full prior knowledge of P. In the tabular case, this challenge has been successfully addressed in the adversarial MDP problem recently by Jin et al. [72], whose technique is based on adjusting the constraints (4.6) with a confidence set over the transition functions, to account for the uncertainty about the dynamics. We find it plausible that a similar extension of ONLINE Q-REPS is possible by incorporating a confidence set for linear MDPs, as has been done in the case of i.i.d. rewards by Neu and Pike-Burke [115]. Nevertheless, the details of such an extension remain highly non-trivial, and we leave the challenge of working them out open for future work.

Bibliography

[1] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, Cs. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320.

[2] Abe, N. and Long, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 3–11.

[3] Abe, N. and Nakamura, A. (1999). Learning to optimally schedule internet banner advertisements. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 12–21, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[4] Agarwal, A., Bartlett, P., and Dama, M. (2010). Optimal allocation strategies for the dark pool problem. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 9–16, Chia Laguna Resort, Sardinia, Italy. PMLR.

[5] Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646.

[6] Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020). FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Agarwal, A., Krishnamurthy, A., Langford, J., Luo, H., and E., S. R. (2017).

Open problem: First-order regret bounds for contextual bandits. In *Proceedings of the 30th Conference on Learning Theory*, pages 4–7.

[8] Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

[9] Allen-Zhu, Z., Bubeck, S., and Li, Y. (2018). Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, pages 186–194.

[10] Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45.

[11] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422.

[12] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

[13] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (1998). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Foundations of Computer Science, 1975., 16th Annual Symposium on*.

[14] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77.

[15] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002c). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77.

[16] Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19, pages 49–56. MIT Press.

[17] Awerbuch, B. and Kleinberg, R. D. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. pages 45–53.

[18] Awerbuch, B. and Kleinberg, R. D. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. pages 45–53.

[19] Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272.

[20] Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781.

[21] Banos, A. (1968). On Pseudo-Games. *The Annals of Mathematical Statistics*, 39(6):1932 – 1945.

[22] Bartók, G., Zolghadr, N., and Szepesvári, C. (2012). An adaptive algorithm for finite stochastic partial monitoring.

[23] Bas-Serrano, J., Curi, S., Krause, A., and Neu, G. (2021). Logistic Q-learning. In *AI & Statistics*, pages 3610–3618.

[Ben-David et al.] Ben-David, S., Pál, D., and Shalev-Shwartz, S. Agnostic online learning.

[Beygelzimer et al.] Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandit algorithms with supervised learning guarantees. pages 19–26.

[26] Beygelzimer, A., Orabona, F., and Zhang, C. (2017). Efficient online bandit multiclass learning with $\widetilde{O}(\sqrt{T})$ regret. In *International Conference on Machine Learning*, pages 488–497.

[27] Blackwell, D. (1956). Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume III, pages 336–338. E. P. Noordhoff, Groningen. (Amsterdam, 2 September–9 September 1954). MR:0085141. Zbl:0073.13204.

[28] Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms*, 31(1):3–122.

[29] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities:A Nonasymptotic Theory of Independence*. Oxford University Press.

[30] Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc.

[31] Bubeck, S., Lee, Y. T., and Eldan, R. (2017). Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85.

[32] Burnetas, A. N. and Katehakis, M. N. (1997). Optimal adaptive policies for Markov Decision Processes. *Mathematics of Operations Research*, 22(1):222–255.

[33] Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv e-prints*, page arXiv:1912.05830.

[34] Calandriello, D., Carratino, L., Lazaric, A., Valko, M., and Rosasco, L. (2019). Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557.

[35] Camoriano, R., Angles, T., Rudi, A., and Rosasco, L. (2016). Nytro: When subsampling meets early stopping. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1403–1411, Cadiz, Spain. PMLR.

[36] Cao, X. R. (2007). *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, New York.

[37] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation. *Annals of Statistics*, 41(3):1516–1541.

[38] Carpentier, A. and Valko, M. (2016). Revealing graph bandits for maximizing local influence. In *Artificial Intelligence and Statistics*, pages 10–18.

[39] Cesa-Bianchi, N. and Lugosi, G. (2006a). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA.

[40] Cesa-Bianchi, N. and Lugosi, G. (2006b). *Prediction, Learning, and Games*. Cambridge University Press, USA.

[41] Chaudhuri, K., Freund, Y., and Hsu, D. J. (2009). A parameter-free hedging algorithm. In *Advances in neural information processing systems*, pages 297–305.

[42] Chen, W., Lakshmanan, L. V., and Castillo, C. (2013a). Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177.

[43] Chen, W., Wang, C., and Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1029–1038, New York, NY, USA. ACM.

[44] Chen, W., Wang, Y., and Yuan, Y. (2013b). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159.

[45] Chernov, A. and Vovk, V. (2010). Prediction with advice of unknown number of experts. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 117–125. AUAI Press.

[46] Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087.

[47] Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR.

[48] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

[49] Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882.

[50] Chung, F. and Lu, L. (2006). *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. American Mathematical Society, USA.

[51] Dani, V., Hayes, T., and Kakade, S. (2008a). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, volume 20, pages 345–352.

[52] Dani, V., Kakade, S. M., and Hayes, T. P. (2008b). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, pages 345–352.

[53] Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

[54] Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pages 5713–5723.

[55] Dick, T., György, A., and Szepesvári, Cs. (2014). Online learning in Markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pages 512–520.

[56] Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*.

[57] Dudík, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178.

[58] Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online Markov decision processes. *Math. Oper. Res.*, 34(3):726–736.

[59] Filippi, S., Cappé, O., Garivier, A., and Szepesvári, Cs. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

[60] Foster, D. J. and Krishnamurthy, A. (2018). Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In *Advances in Neural Information Processing Systems*, pages 2621–2632.

[61] Foster, D. J., Krishnamurthy, A., and Luo, H. (2019). Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14714–14725.

[62] Foster, D. J. and Rakhlin, A. (2020). Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*.

[63] Frieze, A. and Karonski, M. (2015). *Introduction to Random Graphs*. Cambridge University Press, New York. Hardcover.

[64] Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1573–1581.

[65] Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376.

[66] Györfi, L. and Walk, H. (1996). On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61.

[67] Hannan, J. (1957). Approximation to bayes risk in repeated plays. dresher m, tucker a, wolfe p, eds. contributions to the theory of games, vol. iii.

[68] Huaming, W. (2012). On total progeny of multitype Galton-Watson process and the first passage time of random walk with bounded jumps. *ArXiv e-prints*.

[69] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600.

[Jan Peters and Altun] Jan Peters, K. M. and Altun, Y. Relative entropy policy search. *AAAI2010*, pages 1607–1612.

[71] Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

[72] Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020a). Learning adversarial MDPs with bandit feedback and unknown transition. In *International Conference on Machine Learning*.

[73] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020b). Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT 2020)*, pages 2137–2143.

[74] Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274.

[75] Kakade, S. M. (2003). On the sample complexity of reinforcement learning.

[76] Kakade, S. M., Kalai, A. T., and Ligett, K. (2009). Playing games with approximation algorithms. *SIAM Journal on Computing*, 39(3):1088–1106.

[77] Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *International Conference on Machine Learning*, pages 440–447. ACM.

[Kanade et al.] Kanade, V., McMahan, H. B., and Bryan, B. Sleeping experts and bandits with stochastic action availability and adversarial rewards. pages 272–279.

[79] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

[80] Khim, J., Jog, V., and Loh, P.-L. (2019). Adversarial influence maximization. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1–5. IEEE.

[81] Kim, B. and Tewari, A. (2019). Near-optimal oracle-efficient algorithms for stationary and non-stationary stochastic linear bandits. *arXiv preprint arXiv:1912.05695*.

[82] Komiyama, J., Honda, J., and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1792–1800, Cambridge, MA, USA. MIT Press.

[83] Koolen, W. M. and Van Erven, T. (2015). Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175.

[84] Kovalenko, I. (1971). Theory of random graphs. *Cybernetics*, 7(4):575–579.

[85] Krause, A. and Ong, C. (2011). Contextual gaussian process bandit optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[86] Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

[87] Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114.

[88] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[89] Lattimore, T. and Szepesvári, Cs. (2017). The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737.

[90] Lattimore, T. and Szepesvári, Cs. (2019). Bandit algorithms. *book draft*.

[91] Lattimore, T., Szepesvári, Cs., and Weisz, G. (2020). Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*.

[92] Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

[93] Leskovec, J. (2008). *Dynamics of large networks*. PhD thesis, Carnegie Mellon University, School of Computer Science, Machine Learning ....

[94] Leskovec, J., Chakrabarti, D., Kleinberg, J., and Faloutsos, C. (2005). Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In Jorge, A. M., Torgo, L., Brazdil, P., Camacho, R., and Gama, J., editors, *Knowledge Discovery in Databases: PKDD 2005*, pages 133–145, Berlin, Heidelberg. Springer Berlin Heidelberg.

[95] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. (2010). Kronecker graphs: an approach to modeling networks. *Journal of Machine Learning Research*, 11(2).

[96] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010a). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web - WWW '10*.

[97] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010b). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

[98] Littlestone, N. and Warmuth, M. (1994a). The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

[99] Littlestone, N. and Warmuth, M. (1994b). The weighted majority algorithm. *Information and Computation*, 108:212–261.

[100] Luo, H. and Schapire, R. E. (2014). A drifting-games analysis for online learning and applications to boosting. In *Advances in Neural Information Processing Systems*, pages 1368–1376.

[101] Luo, H., Wei, C.-Y., and Lee, C.-W. (2021). Policy optimization in adversarial mdps: Improved exploration via dilated bonuses.

[102] Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback–Leibler divergences. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 497–514.

[103] McMahan, H. B. and Blum, A. Online geometric optimization in the bandit setting against an adaptive adversary. pages 109–123.

[104] McMahan, H. B. and Blum, A. Online geometric optimization in the bandit setting against an adaptive adversary. pages 109–123.

[105] Megiddo, N. and Avivl, T. (1980). On repeated games with incomplete information played by non-bayesian players. *International Journal of Game Theory*, pages 157–167.

[106] Ménard, P. and Garivier, A. (2017). A minimax and asymptotically optimal algorithm for stochastic bandits. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, pages 223–237.

[107] Neu, G. (2015a). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3150–3158.

[108] Neu, G. (2015b). First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375.

[109] Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248.

[110] Neu, G. and Bartók, G. (2016). Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research*, 17:1–21.

[111] Neu, G., György, A., and Szepesvári, Cs. (2010). The online loop-free stochastic shortest-path problem. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT 2010)*, pages 231–243.

[112] Neu, G., György, A., and Szepesvári, Cs. (2012). The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 805–813.

[113] Neu, G., György, A., Szepesvári, Cs., and Antos, A. (2013). Online Markov decision processes under bandit feedback. volume 59, pages 1804–1812.

[114] Neu, G. and Olkhovskaya, J. (2020). Efficient and robust algorithms for adversarial linear contextual bandits. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT 2020)*, pages 3049–3068.

[115] Neu, G. and Pike-Burke, C. (2020). A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[116] Neu, G. and Rosasco, L. (2018). Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, pages 3222–3242.

[117] Neu, G. and Valko, M. (2014). Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pages 2780–2788.

[118] Orabona, F. (2019). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

[119] Pacchiano, A., Lee, J., Bartlett, P., and Nachum, O. (2021). Near optimal policy optimization via REPS. *arXiv preprint arXiv:2103.09756*.

[120] Perrault, P., Healey, J., Wen, Z., and Valko, M. (2020). Budgeted online influence maximization. In *ICML2020*.

[121] Pires, B. Á. and Szepesvári, Cs. (2016). Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151.

[122] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., USA, 1st edition.

[123] Rakhlin, A. and Sridharan, K. (2016). BISTRO: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, pages 1977–1985.

[124] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

[125] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

[126] Rosenberg, A. and Mansour, Y. (2019). Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5478–5486.

[127] Rusmevichientong, P. and Tsitsiklis, J. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35:395–411.

[128] Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026.

[129] Scarlett, J., Bogunovic, I., and Cevher, V. (2017). Lower bounds on regret for noisy Gaussian process bandit optimization. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1723–1742. PMLR.

[130] Shekhar, S. and Javidi, T. (2018). Gaussian process bandits with adaptive discretization.

[131] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022.

[132] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2009). Gaussian process bandits without regret: An experimental design approach. *CoRR*, abs/0912.3995.

[133] Streeter, M. and Golovin, D. (2009). An online algorithm for maximizing submodular functions. In *Advances in Neural Information Processing Systems*, pages 1577–1584.

[134] Syrgkanis, V., Krishnamurthy, A., and Schapire, R. (2016a). Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, pages 2159–2168.

[135] Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. (2016b). Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3135–3143.

[136] Tewari, A. and Bartlett, P. (2020). Optimistic linear programming gives logarithmic regret for irreducible mdps.

[137] Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, pages 495–517.

[138] Thomson, W. R. (1933). ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES. *Biometrika*, 25(3-4):285–294.

[139] Vakili, S., Khezeli, K., and Picheny, V. (2021). On information gain and regret bounds in gaussian process bandits. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 82–90. PMLR.

[140] Valko, M. (2016). *Bandits on graphs and structures.* Habilitation à diriger des recherches, École normale supérieure de Cachan - ENS Cachan.

[141] van der Hofstad, R. (2016). *Random Graphs and Complex Networks*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

[142] Van Roy, B. and Dong, S. (2019). Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910*.

[143] Vaswani, S., Lakshmanan, L., and Schmidt, M. (2015). Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*.

[144] Vovk, V. (1990). Aggregating strategies. In Fulk, M. and Case, J., editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann.

[145] Wang, Q. and Chen, W. (2017). Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171.

[146] Wei, C.-Y., Jafarnia Jahromi, M., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward MDPs with linear function approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3007–3015.

[147] Wen, Z., Kveton, B., Valko, M., and Vaswani, S. (2017). Online influence maximization under independent cascade model with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 3026–3036.

[148] Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806.

[149] Yang, L. F. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 36th International Conference on Machine Learning*.

[150] Yao, H., Szepesvári, Cs., Pires, B., and Zhang, X. (2014). Pseudo-MDPs and factored linear action models.

[151] Zhang, T. (2003). Effective dimension and generalization of kernel learning. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

[152] Zimin, A. and Neu, G. (2013). Online learning in episodic markovian decision processes by relative entropy policy search. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1583–1591. Curran Associates, Inc.