Adaptive Optimization Problems under Uncertainty with Limited Feedback

by

Arthur Flajolet

M.S., Ecole Polytechnique (2013)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author . . . . . .
Sloan School of Management
May 15, 2017

Certified by . . . . . .
Patrick Jaillet
Dugald C. Jackson Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . . . .
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center

Adaptive Optimization Problems under Uncertainty with Limited Feedback

by

Arthur Flajolet

Submitted to the Sloan School of Management
on May 15, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis is concerned with the design and analysis of new algorithms for sequential optimization problems with limited feedback on the outcomes of alternatives when the environment is not perfectly known in advance and may react to past decisions. Depending on the setting, we take either a worst-case approach, which protects against a fully adversarial environment, or a hindsight approach, which adapts to the level of adversariality by measuring performance in terms of a quantity known as regret.

First, we study stochastic shortest path problems with a deadline imposed at the destination when the objective is to minimize a risk function of the lateness. To capture distributional ambiguity, we assume that the arc travel times are only known through confidence intervals on some statistics and we design efficient algorithms minimizing the worst-case risk function.

Second, we study the minimax achievable regret in the online convex optimization framework when the loss function is piecewise linear. We show that the curvature of the decision maker's decision set has a major impact on the growth rate of the minimax regret with respect to the time horizon. Specifically, the rate is always square root when the set is a polyhedron while it can be logarithmic when the set is strongly curved.

Third, we study the Bandits with Knapsacks framework, a recent extension to the standard Multi-Armed Bandit framework capturing resource consumption. We extend the methodology developed for the original problem and design algorithms with regret bounds that are logarithmic in the initial endowments of resources in several important cases that cover many practical applications such as bid optimization in online advertising auctions.

Fourth, we study more specifically the problem of repeated bidding in online advertising auctions when some side information (e.g. browser cookies) is available ahead of submitting a bid. Optimizing the bids is modeled as a contextual Bandits with Knapsacks problem with a continuum of arms. We design efficient algorithms with regret bounds that scale as square root of the initial budget.

Thesis Supervisor: Patrick Jaillet

Title: Dugald C. Jackson Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I would like to thank my advisor Patrick Jaillet for his continued and unconditional support and guidance. Patrick is not only a brilliant mind but also a patient and caring professor who is always open to new ideas. I am very grateful for everything he has done for me and in particular for providing me with countless opportunities during my stay at MIT such as attending conferences, interning at a major demand-side platform, and collaborating with brilliant researchers at MIT and elsewhere. It has been a great honor and pleasure to work with him for the past four years.

I would also like to thank my doctoral committee members Vivek Farias and Alexander (Sasha) Rakhlin not only for their valuable feedback on my work but also for giving me the opportunity to TA for them. Vivek and Sasha are exceptional teachers and assisting them in their work has taught me alot. I have also been very fortunate to collaborate with Sasha on open problems in the field of online learning. Sasha is a phenomenal researcher as well as a wonderful person and I have learned a great deal about online learning working with him. I would also like to thank Laura Rose and Andrew Carvalho for all their help over the years.

My PhD would not have been the same without all the great people I met here, starting with my roommates (past and present): Andrew L., Mathieu D., Zach O., Virgile G., Maxime C., and Florent B.. I am grateful for all the good times we have had together. I would also like to make a special mention of the worthy members of the sushi squad: Charles T., Anna P., Sebastien M., Joey H., Max Burq, Max Biggs, Stefano T., Zach S., Colin P., and Elisabeth P.. May the team live on forever. I would also like to thank all the friends I met at MIT and especially: Ludovica R., Cecile C., Sebastien B., Ali A., Anne C., Audren C., Elise D., Rim H., Ilias Z., Alex R., Alex S., Alex B., Rachid N., Alexis T., Pierre B., Maher D., Velibor M., Nishanth M., Mariapaola T., Zeb H., Dan S., Will M., Antoine D., Jean P., Jonathan A., Arthur D., Konstantina M., Nikita K., Chong Yang G., Jehangir A., Swati G., Rajan U., Clark P., and Eli G..

I would also like to thank S. Hunter in a separate paragraph because he has been such a great friend.

Finally, I would like to acknowledge my family, particularly my mother and my father, who have always been there for me. I would not be here if it was not for them.

This research was supported by the National Research Foundation Grant No. 015824-00078 and the Office of Naval Research Grant N00014-15-1-2083.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and General Setting

In many practical applications, the decision making process is fundamentally sequential and takes place in an uncertain, possibly adversarial, environment that is too complex to be fully comprehended. As a result, planning can no longer be regarded as a one-shot procedure, where the entire plan would be laid out ahead of time, but rather as a trial-and-error process, where the decision maker can assess the quality of past actions, adapt to new circumstances, and potentially learn about the surrounding environment. Arguably, we face many such problems in our daily life. Routing vehicles in transportation networks is a good example since traffic conditions are constantly evolving according to dynamics that are difficult to predict. As drivers, we have little control over this exogenous parameter but we are nevertheless free to modify our itinerary at any point in time based on the latest available traffic information. In this particular application, the environment is not cooperative but also not necessarily adversarial. There are, however, other settings, such as dynamic pricing in the airline industry, where the environment is clearly pursuing a conflicting goal. In this last setting, the ticket agent dynamically adjusts prices as a function of the remaining inventory, the selling horizon, and the perceived willingness to pay while potential customers act strategically in order to drive the prices down.

At an abstract level, we consider the following class of sequential optimization problems. At each stage, the decision maker first selects an action out of a set of options, whose

availability depend on the decision maker's current state, based on the information acquired in the past. Next, the environment, defined as everything outside of the decision maker's control (e.g. traffic conditions in the vehicle routing routing application mentioned above), transitions into a new state, possibly in reaction to this choice. Finally, the decision maker transitions into a new state, obtains a scalar reward, and receives some feedback on the outcomes of alternatives as a function of not only the action that was implemented but also the new state of the environment. We then move onto the next stage and this process is repeated until the decision maker reaches a terminal state. His or her goal is to maximize some prescribed objective function of the sum of the rewards earned along the way. The specificity of the sequential optimization problems studied in this thesis lies in two key features:

1. the environment is uncertain. The decision maker knows which states the environment might occupy but he or she has little a priori knowledge on its internal mechanics: its initial state, its transition rules from one state to another, and its objectives;

2. the feedback received at each stage is very limited. As a result, the decision maker can never perform a thorough counterfactual analysis, which limits his or her ability to learn from past mistakes.

In the above description, uncertainty is not used as a placeholder for stochasticity. Stochasticity can be a source of uncertainty but it need not be the only one or the most important one. For instance, even when the environment is governed by a stochastic process, the underlying distributions might be a priori unknown as in Chapter 2. Going further, stochasticity may not even be involved in the modeling. This is the case, for example, when the problem is to optimize an unknown deterministic function $f(\cdot)$, lying in a known set of general functions, by making repeated value queries.

As the environment is uncertain, the optimization problem faced by the decision maker is ill-defined since merely computing the objective function requires knowing the sequence of states occupied by the environment. In Section 1.2, we detail two possible approaches: a worst-case approach, that protects the decision maker against a fully adversarial environment, and a hindsight approach, that adapts to the level of adversariality at the price of

deriving a suboptimal objective function if the environment happens to be fully adversarial. In the next four chapters, depending on the setting, we study one of these approaches with an emphasis on computational tractability and performance analysis. A precise overview of contributions is given in Section 1.3.

The framework considered in this thesis is, in many respects, strongly reminiscent of Reinforcement Learning (RL) but the focus is different. In RL, the environment is often governed by a stochastic process that the decision maker can learn from through repeated interaction. Since there is limited feedback on the outcome of alternatives, this naturally gives rise to an exploration-exploitation trade-off. In this thesis, the focus is on optimization as opposed to learning and we substitute the concept of uncertainty for that of stochasticity. As a consequence, there might be nothing to learn from for the decision maker in general, as in Chapters 2 and 3, but he or she might need to learn something about the environment in order to perform well, as in Chapters 4 and 5. Note that this is merely a choice of focus and this does not imply by any means that there are no connections to learning. Learning and optimization are intrinsically tied topics whose frontiers are becoming increasingly blurred. However, taking a pure optimization standpoint, abstracting away learning considerations, has proved particularly successful in the RL literature. A good example of this trend is given by the growing popularity of the online convex optimization framework, studied in Chapter 3 and originally developed by the machine learning community, which has led to significant advances in RL. As a recent illustration, the authors of [80] show that establishing martingale tail bounds is essentially equivalent to solving online convex optimization problems. Given that concentration inequalities are intrinsic to any kind of learning, this constitutes a clear step towards bringing together optimization and learning theory. Connections between optimization and learning theory will also appear throughout this thesis.

1.2 Mathematical Framework

In this section, we sketch the common mathematical framework underlying the problems studied in this thesis. For the purpose of describing it at a high level in a mathematically

precise way, we need to introduce notations that may differ from the ones used in later chapters.

Stages are discrete and indexed by $t \in \mathbb{N}$. At the beginning of stage t, the environment (resp. the decision maker) is in state $e_t \in E$ (resp. $s_t \in S$). The set of actions available to the decision maker in state $s \in S$ is denoted by $A(s)$. At stage t, the decision maker takes an action $a_t \in A(s_t)$, from which he or she derives a scalar reward r_t and receives some feedback on the outcomes of alternatives in the form of a multidimensional vector o_t. The decision made at stage t can be based upon all the information acquired in the past, namely $((s_\tau, a_\tau, r_\tau, o_\tau))_{\tau \leq t-1}$ along with s_t, which we symbolically denote by $\mathcal{F}_t$. To select these actions, the decision maker uses an algorithm $\mathcal{A}$ in $\mathbb{A}$, a subset of all non-anticipating algorithms that, at any stage t, map $\mathcal{F}_t$ to a (possibly randomized) action in $A(s_t)$. Similarly, the environment uses a decision rule $\mathcal{D}$ in $\mathbb{D}$, a subset of all non-anticipating decision rules that, at any stage t, map the information available to it, namely $((s_\tau, a_\tau, r_\tau, o_\tau, e_\tau))_{\tau \leq t-1}$ along with s_t and a_t, to a (possibly randomized) state $e_{t+1} \in E$. In addition, $\mathcal{D}$ determines the initial state of the environment ahead of stage 1. The reward and feedback provided to the decision maker at the end of stage t are determined by prescribed (possibly stochastic) rules outside of the environment's control. The sequential process ends at stage τ^* when the decision maker reaches a terminal stage. In many settings, there is a deterministic time horizon T, in which case $\tau^* = T$. Given a prescribed scalar function $\phi(\cdot)$, the goal for the decision maker is to maximize the objective function $\phi(\sum_{t=1}^{\tau^*} r_t)$, which we denote by $\mathrm{obj}(\mathcal{A}, \mathcal{D})$ to underline the fact that it is determined by a deep interplay between $\mathcal{A}$ and $\mathcal{D}$. The exact expression of $\phi(\cdot)$ may be complex and involve expectations. In fact, even when the environment is deterministic and perfectly known in advance, maximizing the objective function can be a difficult computational problem.

Since $\mathcal{D}$ is a priori unknown, maximizing $\mathrm{obj}(\mathcal{A}, \mathcal{D})$ over $\mathcal{A}$ in $\mathbb{A}$ is an ill-defined optimization problem. To remove any ambiguity, a natural approach is to infer $\mathcal{D}$ based on the initial knowledge and to take it as an input. However, this approach is not robust in the sense that there are no guarantees on the performance of the optimal solution to $\max_{\mathcal{A} \in \mathbb{A}} \mathrm{obj}(\mathcal{A}, \mathcal{D})$ if $\mathcal{D}$ happens to deviate from what was inferred, even if only so slightly. This is particularly problematic if the decision maker expects the environment to

be governed by a stochastic process while it is, in fact, adversarial. A variety of robust approaches have been developed in the literature, each with its own optimization formulation that comes with provable guarantees on the performance of the solution derived. In this thesis, we focus on two approaches that do not require additional modeling assumptions about the environment: the worst-case approach and the hindsight approach. This effectively rules out the *Bayesian* optimization approach which requires some prior knowledge on $\mathcal{D}$ in the form of a distribution on $\mathbb{D}$, in which case the performance guarantees are measured in terms of this prior. Ideally, the decision maker should pick an approach based on whether it is critical to be prepared for the worst-case scenario for the particular application at hand. However, in many cases, computational tractability and ease of analysis end up being the main criteria.

1.2.1 Worst-Case Approach

In this approach, the decision maker protects himself or herself against the worst-case scenario of facing a fully adversarial environment and solves:

$$\max_{\mathcal{A}\in\mathbb{A}} \min_{\mathcal{D}\in\mathbb{D}} \mathrm{obj}(\mathcal{A},\mathcal{D}). \tag{1.1}$$

Denote by $\mathcal{A}^*$ an optimal solution to (1.1). By construction, the objective function derived from $\mathcal{A}^*$ is always at least no smaller than the optimal value of (1.1), irrespective of which decision rule $\mathcal{D}$ in $\mathbb{D}$ is used by the environment. The main drawback of this approach is that, depending on $\mathcal{D}$, the objective function may be much larger than the optimal value of (1.1) but no better guarantees can be derived. Note that, depending on the modeling, a worst-case analysis is not necessarily over-conservative since $\mathbb{D}$ may be a small set.

1.2.2 Hindsight Approach

In this approach, the goal is to adapt to the level of adversariality of the environment by comparing, uniformly over all $\mathcal{D} \in \mathbb{D}$, the objective function derived by the decision maker with the largest one he or she could have obtained in hindsight, i.e. if $\mathcal{D}$ was initially

known. Several comparison metrics have been proposed in the literature.

Regret metric. When the comparison is done by means of a subtraction, the metric of interest is called regret, often mathematically defined as:

$$\mathcal{R}(\mathcal{A}) = \max_{\mathcal{D}\in\mathbb{D}} \{\max_{\tilde{\mathcal{A}}\in\tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D}) - \mathrm{obj}(\mathcal{A}, \mathcal{D})\}, \tag{1.2}$$

for $\mathcal{A} \in \mathbb{A}$ and where $\tilde{\mathbb{A}}$ is a subclass of non-anticipating algorithms that may be different from $\mathbb{A}$. The computational problem faced by the decision maker is to find an efficient algorithm $\mathcal{A}$ that, at least approximately, minimizes $\mathcal{R}(\mathcal{A})$. If $\min_{\mathcal{A}\in\mathbb{A}} \mathcal{R}(\mathcal{A})$ is small, the decision maker performs almost as well as if he or she had initially been provided with $\mathcal{D}$, as shown by the inequality:

$$\mathrm{obj}(\mathcal{A}, \mathcal{D}) \geq \max_{\tilde{\mathcal{A}}\in\tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D}) - \mathcal{R}(\mathcal{A}) \tag{1.3}$$

that holds for every $\mathcal{A} \in \mathbb{A}$ irrespective of which decision rule $\mathcal{D}$ in $\mathbb{D}$ is used by the environment.

There are many alternative definitions of regret that vary slightly from (1.2). In many frameworks, such as in online convex optimization, the benchmark $\max_{\tilde{\mathcal{A}}\in\tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D})$ differs along two lines. First, $\tilde{\mathcal{A}}$ is not facing the same decision rule $\mathcal{D}$ as $\mathcal{A}$: the environment is assumed to follow a state path identical to the one that would have been followed if the decision maker had implemented $\mathcal{A}$, as opposed to reacting to the actions dictated by $\tilde{\mathcal{A}}$ (which would lead to a different path). Second, $\tilde{\mathbb{A}}$ is a subclass of all-knowing algorithms that are initially provided with the state path followed by the environment. Thus, in stark contrast with frameworks that use (1.2) as a definition for regret, such as stochastic multi-armed bandit problems, the benchmark is not necessarily physically-realizable in online convex optimization. This is a critical difference that is motivated by technical considerations: the framework is completely model-free, i.e. $\mathbb{D}$ is the set of all non-anticipating decision rules, and, as a result, it is usually impossible to get non-trivial bounds on (1.2). This change in definition makes it possible to establish non-trivial regret bounds while still encompassing many settings of interest, such as the the standard i.i.d. statistical learning

framework, see the discussion below. For the purpose of simplifying the presentation, we keep the notations unchanged but they could easily be adapted to include these frameworks.

A new idea that has received considerable attention in the recent years is to derive regret bounds that also adapt to the level of adversariality of the environment, see, for example, [79] and [47]. In this line of work, the goal is to find an algorithm $\mathcal{A} \in \mathbb{A}$ as well as a function $f(\cdot)$ such that:

$$\max_{\mathcal{D}\in\mathbb{D}}\{\max_{\tilde{\mathcal{A}}\in\tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D}) - \mathrm{obj}(\mathcal{A}, \mathcal{D}) - f(e_1, \cdots, e_{\tau^*})\} \leq 0$$

and such that $f(e_1, \cdots, e_{\tau^*})$ is small when the environment is not fully adversarial (e.g. slow-varying). This approach is particularly attractive when $\mathbb{A}$ is very large and it is impossible to establish non-trivial bounds on $\min_{\mathcal{A}\in\mathbb{A}} \mathcal{R}(A)$.

Researchers also now distinguish various notions of regret based on the exact definition of $\tilde{\mathbb{A}}$, some of which turn out to be closely related, see, for example, [26]. For instance, in the online convex optimization framework, researchers distinguish the concept of dynamic regret, when $\tilde{\mathbb{A}}$ is the set of all unconstrained all-knowing algorithms, from that of (static) external regret, when $\tilde{\mathbb{A}}$ is restricted to all-knowing algorithms that are constrained to select the same action across stages. This last notion is motivated by the standard statistical learning setting: if the environment happens to be governed by an i.d.d. stochastic process then bounds on the external regret directly translate into oracle inequalities. Since the minimal external regret can often be shown to be identical, up to constant factors, to the optimal statistical learning rate and since many efficient algorithms have been designed to achieve near-optimal regret, we get optimal statistical learning rates "for free", i.e. without ever assuming that the environment is governed by a stochastic process.

Competitive ratio metric. When $\phi(\cdot)$ is a positive function, the comparison can done by means of a ratio, in which case the metric of interest is called competitive ratio, often mathematically defined as:

$$\mathcal{C}(\mathcal{A}) = \min_{\mathcal{D}\in\mathbb{D}} \frac{\mathrm{obj}(\mathcal{A}, \mathcal{D})}{\max_{\tilde{\mathcal{A}}\in\tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D})}, \tag{1.4}$$

where $\tilde{\mathbb{A}}$ is a subclass of non-anticipating algorithms that may be different from $\mathbb{A}$. Similarly as for the regret metric, the computational problem faced by the decision maker is then to find an efficient algorithm $\mathcal{A}$ that, at least approximately, minimizes $\mathcal{C}(\mathcal{A})$. The resulting performance guarantee is:

$$\mathrm{obj}(\mathcal{A}, \mathcal{D}) \geq \mathcal{C}(\mathcal{A}) \cdot \max_{\tilde{\mathcal{A}} \in \tilde{\mathbb{A}}} \mathrm{obj}(\tilde{\mathcal{A}}, \mathcal{D}) \tag{1.5}$$

and holds for every $\mathcal{A} \in \mathbb{A}$ irrespective of which decision rule $\mathcal{D}$ in $\mathbb{D}$ is used by the environment.

Just like for the regret metric, a number of variants have been proposed in the literature where the exact definition of the competitive ratio may vary slightly from (1.4) and $\tilde{\mathbb{A}}$ is typically a subset of all-knowing algorithms. Since the competitive ratio metric will not be the metric of choice in this thesis, we refer to the literature for a more thorough introduction to competitive analysis, see in particular [28] and [59].

Comparison of metrics. Regret and competitive ratio metrics are known to be incomparable and, in general, incompatible, see [10]. For this reason, the choice is often based on technical considerations, such as whether the metric lends itself well to analysis, which heavily depends on the framework. Competitive analysis is usually well suited for problems where, at each stage, the decision maker receives some information on the state of the environment before selecting an action while regret metrics tend to facilitate the analysis in *Bayesian* settings by linearity of expectation.

1.3 Overview of Thesis

In Chapter 2, we study a vehicle routing problem and take a worst-case approach for which we design efficient algorithms. Chapters 3, 4, and 5 are all dedicated to the hindsight approach. In Chapter 3, we establish a thorough characterization of the minimax achievable regret in online convex optimization when the loss function is piecewise linear. In Chapters 4 and 5, we design efficient algorithms with provable near-optimal regret for Bandits with Knapsacks problems, an extension of Multi-Armed Bandit problems capturing resource

consumption, with a particular emphasis on applications in the online advertising industry. Each chapter is summarized in more detail in the remainder of this section.

Robust Adaptive Routing under Uncertainty. We consider the problem of finding an optimal history-dependent routing strategy on a directed graph weighted by stochastic arc costs when the objective is to minimize the risk of spending more than a prescribed budget. To help mitigate the impact of the lack of information on the arc cost probability distributions, we introduce a worst-case robust counterpart where the distributions are only known through confidence intervals on some statistics such as the mean, the mean absolute deviation, and any quantile. Leveraging recent results in distributionally robust optimization, we develop a general-purpose algorithm to compute an approximate optimal strategy. To illustrate the benefits of the worst-case robust approach, we run numerical experiments with field data from the Singapore road network.

No-Regret Learnability for Piecewise Linear Losses. In the convex optimization approach to online regret minimization, many methods have been developed to guarantee a $O(\sqrt{T})$ bound on regret for linear loss functions. These results carry over to general convex loss functions by a standard reduction to linear ones. This seems to suggest that linear loss functions are the hardest ones to learn against. We investigate this question in a systematic fashion looking at the interplay between the set of possible moves for both the decision maker and the adversarial environment. This allows us to highlight sharp distinctive behaviors about the learnability of piecewise linear loss functions. On the one hand, when the decision set of the decision maker is a polyhedron, we establish $\Omega(\sqrt{T})$ lower bounds on regret for a large class of piecewise linear loss functions with important applications in online linear optimization, repeated Stackelberg games, and online prediction with side information. On the other hand, we exhibit $o(\sqrt{T})$ learning rates, achieved by the Follow-The-Leader algorithm, in online linear optimization when the decision maker's decision set is curved and when 0 does not lie in the convex hull of the environment's decision set. Hence, the curvature of the decision maker's decision set is a determining factor for the optimal rate.

Logarithmic Regret Bounds for Bandits with Knapsacks. Optimal regret bounds for Multi-Armed Bandit problems are now well documented. They can be classified into two categories based on the growth rate with respect to the time horizon T: (i) small, distribution-dependent, bounds of order of magnitude $\ln(T)$ and (ii) robust, distribution-free, bounds of order of magnitude $\sqrt{T}$. The Bandits with Knapsacks model, an extension to the framework allowing to model resource consumption, lacks this clear-cut distinction. While several algorithms have been shown to achieve asymptotically optimal distribution-free bounds on regret, there has been little progress toward the development of small distribution-dependent regret bounds. We partially bridge the gap by designing a general-purpose algorithm with distribution-dependent regret bounds that are logarithmic in the initial endowments of resources in several important cases that cover many practical applications, including dynamic pricing with limited supply, bid optimization in online advertisement auctions, and dynamic procurement.

Real-Time Bidding with Side Information. We consider the problem of repeated bidding in online advertising auctions when some side information (e.g. browser cookies) is available ahead of submitting a bid in the form of a d-dimensional vector. The goal for the advertiser is to maximize the total utility (e.g. the total number of clicks) derived from displaying ads given that a limited budget B is allocated for a given time horizon T. Optimizing the bids is modeled as a linear contextual Multi-Armed Bandit (MAB) problem with a knapsack constraint and a continuum of arms. We develop UCB-type algorithms that combine two streams of literature: the confidence-set approach to linear contextual MABs and the probabilistic bisection search method for stochastic root-finding. Under mild assumptions on the underlying unknown distribution, we establish distribution-independent regret bounds of order $\tilde{O}(d \cdot \sqrt{T})$ when either $B = \infty$ or when B scales linearly with T.

Chapter 2

Robust Adaptive Routing under Uncertainty

2.1 Introduction

2.1.1 Motivation

Stochastic Shortest Path (SSP) problems have emerged as natural extensions to the classical shortest path problem when arc costs are uncertain and modeled as outcomes of random variables. In particular, we consider in this chapter the class of adaptive SSPs, which can be formulated as *Markov Decision Processes* (MDPs), where we optimize over all history-dependent strategies. As standard with MDPs, optimal policies are characterized by dynamic programming equations involving expected values (e.g. [22]). Yet, computing the expected value of a function of a random variable generally requires a full description of its probability distribution, and this can be hard to obtain accurately due to errors and sparsity of measurements. In practice, only finite samples are available and an optimal strategy based on approximated arc cost probability distributions may be suboptimal with respect to the real arc cost probability distributions.

In recent years, *Distributionally Robust Optimization* (DRO) has emerged as a new framework for decision-making under uncertainty when the underlying distributions are only known through some statistics or from collections of samples. DRO was put forth in

an effort to capture both risk (uncertainty on the outcomes) and ambiguity (uncertainty on the probabilities of the outcomes) when optimizing over a set of alternatives. The computational complexity of this approach can vary greatly, depending on the nature of the ambiguity sets and on the structure of the optimization problem, see [100] and [38] for convex problems, and [31] for chance-constraint problems. Even in the absence of decision variables, the theory proves useful to derive either numerical or closed form bounds on expected values using optimization tools, see, for example, [77], [23], and [94].

In the case of limited knowledge of the arc cost probability distributions, we propose to bring DRO to bear on adaptive SSP problems to help mitigate the impact of the lack of information. Our work fits into the literature on Distributionally Robust MDPs (DRMDPs) where the transition probabilities are only known to lie in prescribed ambiguity sets (e.g. [69], [104], and [99]). While the methods developed in the aforementioned literature carry over, adaptive SSPs exhibit a particular structure that allows for a large variety of ambiguity sets and enables the development of faster solution procedures. Specifically, optimal strategies for DRMDPs are characterized by a Bellman recursion on the worst-case expected reward-to-go. While standard approaches focus on computing this quantity for each state independently from one another, closely related problems (e.g. estimating an expected value $\mathbb{E}[f(t - X)]$ where the random variable X is fixed but t varies across states) carry across states for adaptive SSPs. As a result, making the most of previous computations becomes crucial for computational tractability. This entails keeping track of the extreme points of a dynamically changing set efficiently, revealing an interesting connection between DRMDPs and *Dynamic Convex Hull* problems.

2.1.2 Related Work and Contributions

Over the years, many SSP problems have been formulated. They differ along three main features:

- The specific objective function to optimize: in the presence of uncertainty, minimizing the expected costs is a natural approach, see [22], but it is oblivious to risk. [62] proposed earlier to rely on utility functions of statistical moments involving an inher-

ent trade-off, and considered multi-objective criteria. However, Bellman's principle of optimality no longer holds in this case, giving rise to computational hardness. A different approach consists of (1) introducing a budget, set by the user, corresponding to the maximum cost he is willing to pay to reach his terminal node and (2) minimizing either the probability of budget overrun (see [41], [68], and also [105] for probabilistic goal MDPs), more general functions of the budget overrun as in [67], satisficing measures to guarantee good performances with respect to multiple objectives as in [52], or the expected costs while also constraining the probability of budget overrun as in [103].

- The set of strategies over which we are free to optimize: incorporating uncertainty may cause history-dependent strategies to significantly outperform a priori paths depending on the performance index. This is the case when the objective is to maximize the probability of completion within budget for which two types of formulations have been considered: (i) an a priori formulation which consists of finding a path before taking any actions, see [68] and [66]; and (ii) an adaptive formulation which allows to update the path to go based on the remaining budget, see [65] and [85].

- The knowledge on the random arc costs taken as an input: it can range from the full knowledge of the probability distributions to having access to only a few samples drawn from them. In practical settings, the problem of estimating some statistics seems more reasonable than retrieving the full probability distribution. For instance, [52] consider lower-order statistics (minimum, average, and maximum costs) and use closed form bounds derived in the DRO theory. These considerations were extensively investigated in the context of DRMDPs, see [51] and [99] for theoretical developments. The ambiguity sets are parametric in [99], where the parameter lies in the intersection of ellipsoids, are based on likelihood measures in [69], and are defined by linear inequalities in [98].

We give an overview of prior formulations in Table 2.1.

Table 2.1: Literature review of stochastic shortest path problems.

Author(s)	Objective function	Strategy	Uncertainty description	Approach
[62]	utility function	a priori	moments	dominated paths
[68]	probability of budget overrun	a priori	normal distributions	convex optimization
[65] [85]	probability of budget overrun	adaptive	distributions	dynamic programming
[69]	expected cost	adaptive	maximum-likelihood ambiguity sets	dynamic programming
[52] [4]	requirements violation	a priori	distributions or moments	iterative procedure
[74]	monotone risk measure	a priori	distributions	labeling algorithm
Our work	risk function of the budget overrun	adaptive	distributions or confidence intervals on statistics	dynamic programming

Contributions. The main contributions of this chapter can be summarized as follows:

1. We extend the class of adaptive SSP problems to general risk functions of the budget overrun and to the presence of distributional ambiguity.

2. We characterize optimal strategies and identify conditions on the risk function under which infinite cycling is provably suboptimal.

3. For any risk function satisfying these conditions, we provide efficient solution procedures (invoking fast Fourier transforms and dynamic convex hull algorithms as subroutines) to compute ϵ-approximate optimal strategies when the arc cost distributions are either exactly known or only known through confidence intervals on piecewise affine statistics (e.g. the mean, the mean absolute deviation, any quantile...) for any $\epsilon > 0$.

Special cases where (i) the objective is to minimize the probability of budget overrun and (ii) the arc costs are independent discrete random variables can serve as a basis for comparison with prior work on DRMDPs. For this subclass of problems, our formulation can be interpreted as a DRMDP with finite horizon N, finitely many states n (resp. actions m), and a rectangular ambiguity set. Our methodology can be used to compute an ϵ-optimal strategy with complexity $O(m \cdot n \cdot \log(\frac{N}{\epsilon}) \cdot \log(n))$.

The remainder of the chapter is organized as follows. In Section 2.2, we introduce the adaptive SSP problem and its distributionally robust counterpart. Section 2.3 (resp. Section 2.4) is devoted to the theoretical and computational analysis of the nominal (resp. robust) problem. In Section 2.5, we consider a vehicle routing application and present results of numerical experiments run with field data from the Singapore road network. In Section 2.6, we relax some of the assumptions made in Section 2.2 and extend the results presented in Sections 2.3 and 2.4.

Notations. For a function $g(\cdot)$ and a random variable X distributed according to p, we denote the expected value of $g(X)$ by $\mathbb{E}_{X\sim p}[g(X)]$. For a set $S \subset \mathbb{R}^n$, $\bar{S}$ is the closure of S for the standard topology, $\mathrm{conv}(S)$ is the convex hull of S, and $|S|$ is the cardinality of S.

For a set $S \subset \mathbb{R}^2$, $\hat{S}$ denotes the upper convex hull of S, i.e. $\hat{S} = \{(x, y) \in \mathbb{R}^2 \ : \ \exists (a, b) \in \text{conv}(S)$ such that $x = a$ and $y \geq b\}$.

2.2 Problem Formulation

2.2.1 Nominal Problem

Let $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ be a finite directed graph where each arc $(i, j) \in \mathcal{A}$ is assigned a collection of non-negative random costs $(c_{ij}^\tau)_{\tau \geq 0}$. We consider a user traveling through $\mathcal{G}$ leaving from s and wishing to reach d within a prescribed budget T. Having already spent a budget τ and being at node i, choosing to cross arc (i, j) would incur an additional cost c_{ij}^τ, whose value becomes known after the arc is crossed. In vehicle routing applications, c_{ij}^τ typically models the travel time along arc (i, j) at time τ and T is the deadline imposed at the destination. The objective is to find a strategy to reach d maximizing a risk function of the budget overrun, denoted by $f(\cdot)$. Mathematically, this corresponds to solving:

$$\sup_{\pi \in \Pi} \mathbb{E}[f(T - X_\pi)], \tag{2.1}$$

where Π is the set of all history-dependent randomized strategies and X_π is the random cost associated with strategy π when leaving from node s with budget T. Examples of natural risk functions include $f(t) = t \cdot 1_{t \leq 0}$, $f(t) = 1_{t \geq 0}$, and $f(t) = -|t|$ which translate into, respectively, minimizing the expected budget overrun, maximizing the probability of completion within budget, and penalizing the expected deviation from the target budget. We will restrict our attention to risk functions satisfying natural properties meant to prevent infinite cycling in Theorem 2.1 of Section 2.3.1, e.g. maximizing the expected budget overrun is not allowed. Without any additional assumption on the random costs, (2.1) is computationally intractable. To simplify the problem, a common approach in the literature is to assume independence of the arc costs, see for example [40].

Assumption 2.1. $(c_{ij}^\tau)_{(i,j) \in \mathcal{A}, \tau \geq 0}$ *are independent random variables.*

In practice, the costs of neighboring arcs can be highly correlated for some applications

and Assumption 2.1 may then appear unreasonable. Most of the results derived in this chapter can be extended when the experienced costs are modeled as a *Markov* chain of finite order. To simplify the presentation, Assumption 2.1 is used throughout the chapter and this extension is discussed in Section 2.6.1. For the same reason, the arc costs are also assumed to be identically distributed across τ.

Assumption 2.2. *For all arcs $(i,j) \in \mathcal{A}$, the distribution of c_{ij}^{τ} does not depend on τ.*

The extension to τ-dependent arc cost distributions is detailed in Section 2.6.2. For clarity of the exposition, we omit the superscript τ when it is unnecessary and simply denote the costs by $(c_{ij})_{(i,j)\in\mathcal{A}}$, even though the cost of an arc corresponds to an independent realization of its corresponding random variable each time it is crossed. Motivated by computational and theoretical considerations that will become apparent in Section 2.3.2.b, we further assume that the arc cost distributions have compact supports throughout the chapter. This assumption is crucial for the analysis carried out in this chapter but is also perfectly reasonable in many practical settings, such as in transportation networks.

Assumption 2.3. *For all arcs $(i,j) \in \mathcal{A}$, the distribution of c_{ij}, denoted by $\mathrm{p_{ij}}$, has compact support included in $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$ with $\delta_{ij}^{\inf} > 0$ and $\delta_{ij}^{\sup} < \infty$. Thus $\delta^{\inf} = \min_{(i,j)\in\mathcal{A}} \delta_{ij}^{\inf} > 0$ and $\delta^{\sup} = \max_{(i,j)\in\mathcal{A}} \delta_{ij}^{\sup} < \infty$.*

2.2.2 Distributionally Robust Problem

A major limitation of the approach described above is that it requires a full description of the uncertainty, i.e. having access to the arc cost probability distributions. Yet, in practice, we often only have access to a limited number of realizations of the random variables c_{ij}. It is then tempting to estimate empirical arc cost distributions and to take them as input to problem (2.1). However, estimating accurately a distribution usually requires a large sample size, and our experimental evidence suggests that, as a result, the corresponding solutions may perform poorly when only a few samples are available, as we will see in Section 2.5. To address this limitation, we adopt a distributionally robust approach where, for each arc $(i,j) \in \mathcal{A}$, p_{ij} is only assumed to lie in an ambiguity set $\mathcal{P}_{ij}$. We make the following assumption on these ambiguity sets throughout the chapter.

Assumption 2.4. *For all arcs $(i,j) \in \mathcal{A}$, $\mathcal{P}_{ij}$ is not empty, closed for the weak topology, and a subset of $\mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}])$, the set of probability measures on $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$.*

Assumption 2.4 is a natural extension of Assumption 2.3, and is essential for computational tractability, see Section 2.4. The robust counterpart of (2.1) for an ambiguity-averse user is then given by:

$$\sup_{\pi \in \Pi} \inf_{\forall (i,j) \in \mathcal{A},\, p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T - X_\pi)], \tag{2.2}$$

where the notation $\mathbf{p}$ refers to the fact that the costs $(c_{ij})_{(i,j)\in\mathcal{A}}$ are independent and distributed according to $(p_{ij})_{(i,j)\in\mathcal{A}}$. As a byproduct of the results obtained for the nominal problem in Section 2.3.1, (2.2) can be equivalently viewed as a distributionally robust MDP in the extended space state $(i,\tau) \in \mathcal{V} \times \mathbb{R}_+$ where i is the current location and τ is the total cost spent so far and where the transition probabilities from any state (i,τ) to any state (j,τ'), for $j \in \mathcal{V}(i)$ and $\tau' \geq \tau$, are only known to jointly lie in a global ambiguity set. As shown in [99], the tractability of a distributionally robust MDP hinges on the decomposability of the global ambiguity set as a Cartesian product over the space state of individual ambiguity sets, a property coined as *rectangularity*. While the global ambiguity set of (2.2) is rectangular with respect to our original state space $\mathcal{V}$, it is not with respect to the extended space space $\mathcal{V} \times \mathbb{R}_+$. Thus, we are led to enlarge our ambiguity set to make it rectangular and consider a conservative approximation of (2.2). This boils down to allowing the arc cost distributions to vary in their respective ambiguity sets as a function of τ. This approach leads to the following formulation:

$$\sup_{\pi \in \Pi} \inf_{\forall \tau, \forall (i,j) \in \mathcal{A},\, p_{ij}^\tau \in \mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T - X_\pi)], \tag{2.3}$$

where the notation $\mathbf{p}^\tau$ refers to the fact that, for any arc $(i,j) \in \mathcal{A}$, the costs $(c_{ij}^\tau)_{\tau \geq 0}$ are independent and distributed according to $(p_{ij}^\tau)_{\tau \geq 0}$. Note that when Assumption 2.2 is relaxed, we have a different ambiguity set $\mathcal{P}_{ij}^\tau$ for each pair $((i,j),\tau) \in \mathcal{A} \times \mathbb{R}_+$ and (2.3) is precisely the robust counterpart of (2.1) as opposed to a conservative approximation, see Section 2.6.2. Also observe that (2.3) reduces to (2.1) when the ambiguity sets are singletons, i.e. $\mathcal{P}_{ij} = \{p_{ij}\}$. In the sequel, we focus on (2.3), which we refer to as the robust problem. However, we will also investigate the performance of an optimal solution to (2.3)

with respect to the optimization problem (2.2) from a theoretical (resp. practical) standpoint in Section 2.4.2 (resp. Section 2.5). Finally note that we consider general ambiguity sets satisfying Assumption 2.4 when we study the theoretical properties of (2.3). However, for tractability purposes, the solution procedure that we develop in Section 2.4.3.c only applies to ambiguity sets defined by confidence intervals on piecewise affine statistics, such as the mean, the absolute mean deviation, or any quantile. We refer to Section 2.4.3.b for a discussion on the modeling power of these ambiguity sets. Similarly as for the nominal problem, we will also restrict our attention to risk functions satisfying natural properties meant to prevent infinite cycling in Theorem 2.2 of Section 2.4.1.

2.3 Theoretical and Computational Analysis of the Nominal Problem

2.3.1 Characterization of Optimal Policies

Perhaps the most important property of (2.1) is that *Bellman's Principle of Optimality* can be shown to hold irrespective of the choice of the risk function. Specifically, for any history of the previously experienced costs and previously visited nodes, an optimal strategy to (2.1) must also be an optimal strategy to the subproblem of minimizing the risk function given this history. Otherwise, we could modify this strategy for this particular history and take it to be an optimal strategy for this subproblem. This operation could only increase the objective function of the optimization problem (2.1), which would contradict the optimality of the strategy.

Another, less obvious, interesting feature of (2.1) is that, even for perfectly natural risk functions $f(\cdot)$, making decisions according to an optimal strategy may lead to cycle back to a previously visited location. This may happen, for instance, when the objective is to maximize the probability of completion within budget, see [85], and their example can be adapted when the objective is to minimize the expected budget overrun, see Figure 2-1. While counter-intuitive at first, the existence of loops is a direct consequence of the stochasticity of the costs when the decision maker is concerned about the risk of going

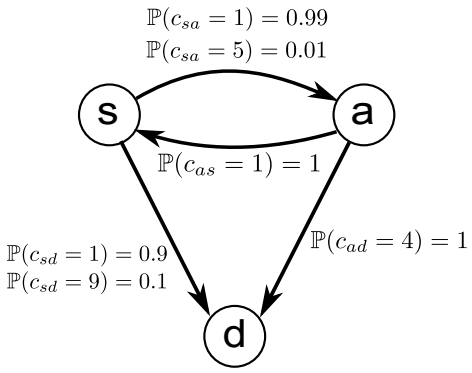
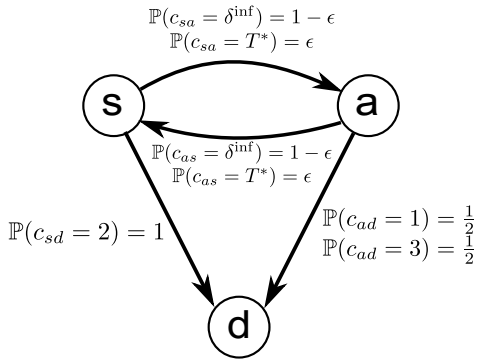


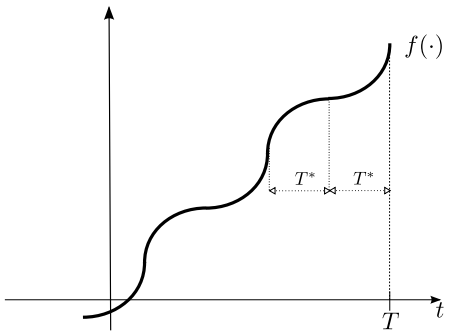
Figure 2-1: Existence of loops. If the initial budget is $T = 8$ and the risk function is $f(t) = t \cdot 1_{t \leq 0}$, the optimal strategy to travel from s to d is to go to a first. This is because going to d directly incurs an expected delay of 0.1, while going to a first and then planning to go to d incurs an expected delay of 0.01. If we end up getting a cost $c_{sa} = 5$ on the way to a, then, performing a similar analysis, the optimal strategy is to go back to s.

over budget, as illustrated in Figure 2-1. On the other hand, the existence of infinitely many loops is particularly troublesome from a modeling perspective as it would imply that a user traveling through $\mathcal{V}$ following the optimal strategy may get at a location $i \neq d$ having already spent an arbitrarily large budget with positive probability. Furthermore, infinite cycling is also problematic from a computational standpoint because describing an optimal strategy would require unlimited storage capacity. We argue that infinite cycling arises only when the risk function is poorly chosen. This is obvious when $f(t) = -t \cdot 1_{t \leq 0}$, which corresponds to maximizing the expected budget overrun, but we stress that it is not merely a matter of monotonicity. Infinite cycling may occur even if $f(\cdot)$ is increasing as we highlight in Example 2.1.

Example 2.1. *Consider the simple directed graph of Figure 2-2a and the risk function $f(\cdot)$ illustrated in Figure 2-2b. $f(\cdot)$ is defined piecewise, alternating between concavity and convexity on intervals of size T^* and the same pattern is repeated every $2T^*$. This means that, for this particular objective, the attitude towards risk keeps fluctuating as the budget decreases, from being risk-averse when $f(\cdot)$ is locally concave to being risk-seeking when $f(\cdot)$ is locally convex. Now take $\delta^{\inf} << 1$, $\epsilon << 1$ and $T^* > 3$ and consider finding a strategy to get to d starting from s with initial budget T which we choose to take at a point where $f(\cdot)$ switches from being concave to being convex, see Figure 2-2b. Going straight to d incurs an expected objective value of $f(T-2) < \frac{1}{2}f(T-1) + \frac{1}{2}f(T-3)$ and we*



(a) Graph, s and d are respectively the source and the destination.



(b) risk function. T is the initial budget, $2T^*$ is the period of $f'(\cdot)$.

Figure 2-2: Existence of infinite cycling from Example 2.1.

can make this gap arbitrarily large by properly defining $f(\cdot)$. Therefore, by taking ϵ and $\delta^{\inf}$ small enough, going to a first is optimal. With probability $\epsilon > 0$, we arrive at a with a remaining budget of $T - T^*$. Afterwards, the situation is reversed as we are willing to take as little risk as possible and the corresponding optimal solution is to go back to s. With probability ϵ, we arrive at s with a budget of $T - 2T^*$ and we are back in the initial situation, showing the existence of infinite cycling.

In light of Example 2.1, we identify a set of sufficient conditions on $f(\cdot)$ ruling out the possibility of infinite cycling.

Theorem 2.1. *Case 1: If there exists T_1 such that either:*

(a) $f(\cdot)$ is increasing, concave, and C^2 on $(-\infty, T_1)$ and such that $\frac{f''}{f'} \to_{-\infty} 0$,

(b) $f(\cdot)$ is C^1 on $(-\infty, T_1)$ and $\lim_{-\infty} f'$ exists, is positive, and is finite,

then there exists T_f such that, for any $T \geq 0$ and as soon as the total cost spent so far is larger than $T - T_f$, any optimal policy to (2.1) *follows the shortest-path tree rooted at d with respect to the mean arc costs, which we denote by $\mathcal{T}$.*
Case 2: If there exists T_f such that the support of $f(\cdot)$ is included in $[T_f, \infty)$, then following $\mathcal{T}$ is optimal as soon as the total cost spent so far is larger than $T - T_f$.

For a node i, $\mathcal{T}(i)$ refers to the set of immediate successors of i in $\mathcal{T}$. The proof is deferred to the Appendix.

Observe that, in addition to not being concave, the choice of $f(\cdot)$ in Example 2.1 does not satisfy property (b) as $f'(\cdot)$ is $2T^*$-periodic. An immediate consequence of Theorem 2.1 is that an optimal strategy to (2.1) does not include any loop as soon as the total cost spent so far is larger than $T - T_f$. Since each arc has a positive minimum cost, this rules out infinite cycling. The parameter T_f can be computed through direct reasoning on the risk function $f(\cdot)$ or by inspecting the proof of Theorem 2.1. Remark that any polynomial of even degree with a negative leading coefficient satisfies condition (a) of Theorem 2.1. Examples of valid objectives include maximization of the probability of completion within budget $f(t) = 1_{t \geq 0}$ with $T_f = 0$, minimization of the budget overrun $f(t) = t \cdot 1_{t \leq 0}$ with $T_f = 0$, and minimization of the squared budget overrun $f(t) = -t^2 \cdot 1_{t \leq 0}$ with

$$T_f = -\frac{|\mathcal{V}| \cdot \delta^{\sup} \cdot \max_{i \in \mathcal{V}} M_i}{2 \cdot \min_{i \neq d} \min_{j \in \mathcal{V}(i), j \notin \mathcal{T}(i)} \{\mathbb{E}[c_{ij}] + M_j - M_i\}},$$

where M_i is the minimum expected cost to go from i to d and with the convention that the minimum of an empty set is equal to ∞. When $f(\cdot)$ is increasing but does not satisfy condition (a) or (b), the optimal strategy may follow a different shortest-path tree. For instance, if $f(t) = -\exp(-t)$, the optimal policy is to follow the shortest path to d with respect to $(\log(\mathbb{E}[\exp(c_{ij})]))_{(i,j) \in \mathcal{A}}$. Conversely, if $f(t) = \exp(t)$, the optimal policy is to follow the shortest path to d with respect to $(-\log(\mathbb{E}[\exp(-c_{ij})]))_{(i,j) \in \mathcal{A}}$. For these reasons, proving that an optimal strategy to (2.1) does not include infinitely many loops when $f(\cdot)$ does not satisfy the assumptions of Theorem 2.1 requires objective-specific (and possibly graph-specific) arguments. To illustrate this last point, observe that the conclusion of Theorem 2.1 always holds for a graph consisting of a single simple path regardless of the definition of $f(\cdot)$, even if this function is decreasing. Hence, the assumptions of Theorem 2.1 are not necessary in general to prevent infinite cycling but restricting our attention to this class of risk functions enables us to study the problem in a generic fashion and to develop a general-purpose algorithm in Section 2.3.2.

Another remarkable property of (2.1) is that it can be equivalently formulated as a MDP in the extended space state $(i, t) \in \mathcal{V} \times (-\infty, T]$ where i is the current location and t is the remaining budget. As a result, standard techniques for MDPs can be applied to show that

there exists an optimal *Markov* policy π_f^* which is a mapping from the current location and the remaining budget to the next node to visit. Furthermore, the optimal *Markov* policies are characterized by the dynamic programming equation:

$$\begin{aligned} u_d(t) &= f(t) && t \leq T \\ u_i(t) &= \max_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T \\ \pi_f^*(i,t) &\in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T, \end{aligned} \tag{2.4}$$

where $\mathcal{V}(i) = \{j \in \mathcal{V} \mid (i,j) \in \mathcal{A}\}$ refers to the set of immediate successors of i in $\mathcal{G}$ and $u_i(t)$ is the expected objective-to-go when leaving $i \in \mathcal{V}$ with remaining budget t. The interpretation of (2.4) is simple. At each node $i \in \mathcal{V}$, and for each potential remaining budget t, the decision maker should pick the outgoing edge (i,j) that yields the maximum expected objective-to-go if acting optimally thereafter.

Proposition 2.1. *Under the same assumptions as in Theorem 2.1, any Markov policy solution to* (2.4) *is an optimal strategy for* (2.1)*.*

The proof is deferred to the Appendix.

2.3.2 Solution Methodology

In order to solve (2.1), we use Proposition 2.1 and compute a *Markov* policy solution to the dynamic program (2.4). We face two main challenges when we carry out this task. First, (2.4) is a continuous dynamic program. To solve this program numerically, we approximate the functions $(u_i(\cdot))_{i \in \mathcal{V}}$ by piecewise constant functions, as detailed in Section 2.3.2.a. Second, as illustrated in Figure 2-1 of Section 2.3.1, an optimal *Markov* strategy solution to (2.4) may contain loops. Hence, in the presence of a cycle in $\mathcal{G}$, say $i \to j \to i$, observe that computing $u_i(t)$ requires to know the value of $u_j(t)$ which in turns depends on $u_i(t)$. As a result, it is a-priori unclear how to solve (2.4) without resorting to value or policy iteration. We explain how to sidestep this difficulty and construct efficient label-setting algorithms in Section 2.3.2.b. In particular, using these algorithms, we can compute:

- an optimal solution to (2.1) in $O(|\mathcal{A}| \cdot \frac{T-T_f}{\Delta t} \cdot \log^2(\frac{\delta^{\sup}}{\Delta t}) + |\mathcal{V}|^2 \cdot \frac{\delta^{\sup}}{\Delta t} \cdot \log(|\mathcal{V}| \cdot \frac{\delta^{\sup}}{\Delta t}))$ computation time when the arc costs only take on values that are multiple of $\Delta t > 0$ and for any risk function $f(\cdot)$ satisfying Theorem 2.1. This simplifies to $O(|\mathcal{A}| \cdot \frac{T}{\Delta t} \cdot \log^2(\frac{\delta^{\sup}}{\Delta t}))$ when the objective is to maximize the probability of completion within budget.

- an ϵ-approximate solution to (2.1) in

$$O(\frac{(|\mathcal{V}| + \frac{T-T_f}{\delta^{\inf}})^2}{\epsilon} \cdot \ |\mathcal{A}| \cdot (T - T_f) \cdot \log^2(\frac{(|\mathcal{V}| + \frac{T-T_f}{\delta^{\inf}}) \cdot \delta^{\sup}}{\epsilon}))$$
$$+ O(\frac{(|\mathcal{V}| + \frac{T-T_f}{\delta^{\inf}})^2}{\epsilon} \cdot |\mathcal{V}|^2 \cdot \delta^{\sup} \cdot \log(\frac{(|\mathcal{V}| + \frac{T-T_f}{\delta^{\inf}}) \cdot |\mathcal{V}| \cdot \delta^{\sup}}{\epsilon}) \ )$$

  computation time when the risk function is Lipschitz on compact sets.

As we explain in Section 2.3.2.b, computing the convolution products arising in (2.4) efficiently (e.g. through fast Fourier transforms) is crucial to get this near-linear dependence on $\frac{1}{\Delta t}$ (or equivalently $\frac{1}{\epsilon}$). A brute-force approach consisting in applying the pointwise definition of convolution products incurs a quadratic dependence.

2.3.2.a Discretization Scheme

For each node $i \in \mathcal{V}$, we approximate $u_i(\cdot)$ by a piecewise constant function $u_i^{\Delta t}(\cdot)$ of uniform stepsize Δt. Under the conditions of Theorem 2.1, we only need to approximate $u_i(\cdot)$ for a remaining budget larger than $k_i^{\min} \cdot \Delta t$, for $k_i^{\min} = \left\lfloor \frac{T_f - (|\mathcal{V}| - \text{level}(i,\mathcal{T}) + 1) \cdot \delta^{\sup}}{\Delta t} \right\rfloor$, where level$(i, \mathcal{T})$ is defined as the level of node i in the rooted tree $\mathcal{T}$, i.e. the number of parent nodes of i in $\mathcal{T}$ plus one. This is because, following the shortest path tree $\mathcal{T}$ once the remaining budget drops below T_f, we can never get to state i with remaining budget less than $k_i^{\min} \cdot \Delta t$. We use the approximation:

$$\begin{aligned} u_i^{\Delta t}(t) &= u_i^{\Delta t}(\left\lfloor \frac{t}{\Delta t} \right\rfloor \cdot \Delta t) && i \in \mathcal{V}, t \in [k_i^{\min} \cdot \Delta t, T] \\ \pi^{\Delta t}(i,t) &= \pi^{\Delta t}(i, \left\lfloor \frac{t}{\Delta t} \right\rfloor \cdot \Delta t) && i \neq d, t \in [k_i^{\min} \cdot \Delta t, T], \end{aligned} \tag{2.5}$$

and the values at the mesh points are determined by the set of equalities:

$$
\begin{aligned}
u_d^{\Delta t}(k\cdot\Delta t) &= f(k\cdot\Delta t) && k = k_d^{\min},...,\left\lfloor\frac{T}{\Delta t}\right\rfloor\\
u_i^{\Delta t}(k\cdot\Delta t) &= \max_{j\in\mathcal{V}(i)}\int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega && i\neq d, k = \left\lfloor\frac{T_f}{\Delta t}\right\rfloor,...,\left\lfloor\frac{T}{\Delta t}\right\rfloor\\
\pi^{\Delta t}(i,k\cdot\Delta t) &\in \operatorname*{argmax}_{j\in\mathcal{V}(i)}\int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega && i\neq d, k = \left\lfloor\frac{T_f}{\Delta t}\right\rfloor,...,\left\lfloor\frac{T}{\Delta t}\right\rfloor\\
u_i^{\Delta t}(k\cdot\Delta t) &= \max_{j\in\mathcal{T}(i)}\int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega && i\neq d, k = k_i^{\min},...,\left\lfloor\frac{T_f}{\Delta t}\right\rfloor-1\\
\pi^{\Delta t}(i,k\cdot\Delta t) &\in \operatorname*{argmax}_{j\in\mathcal{T}(i)}\int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega && i\neq d, k = k_i^{\min},...,\left\lfloor\frac{T_f}{\Delta t}\right\rfloor-1.
\end{aligned}
\tag{2.6}
$$

Notice that for $t \leq T_f$, we rely on Theorem 2.1 and only consider, for each node $i \neq d$, the immediate neighbors of i in $\mathcal{T}$. This is of critical importance to be able to solve (2.6) with a label-setting algorithm, see Section 2.3.2.b. The next result provides insight into the quality of the policy $\pi^{\Delta t}$ as an approximate solution to (2.1).

Proposition 2.2. *Consider a solution to the global discretization scheme* (2.5) *and* (2.6)*,* $(\pi^{\Delta t}, (u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}})$*. We have:*

1. *If $f(\cdot)$ is non-decreasing, the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge pointwise almost everywhere to $(u_i(\cdot))_{i\in\mathcal{V}}$ as $\Delta t \to 0$,*

2. *If $f(\cdot)$ is continuous, the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge uniformly to $(u_i(\cdot))_{i\in\mathcal{V}}$ and $\pi^{\Delta t}$ is a $o(1)$-approximate optimal solution to* (2.1) *as $\Delta t \to 0$,*

3. *If $f(\cdot)$ is Lipschitz on compact sets (e.g. if $f(\cdot)$ is C^1), the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge uniformly to $(u_i(\cdot))_{i\in\mathcal{V}}$ at speed Δt and $\pi^{\Delta t}$ is a $O(\Delta t)$-approximate optimal solution to* (2.1) *as $\Delta t \to 0$,*

4. *If $f(t) = 1_{t\geq 0}$ and the distributions $(p_{ij})_{(i,j)\in\mathcal{A}}$ are continuous, the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge uniformly to $(u_i(\cdot))_{i\in\mathcal{V}}$ and $\pi^{\Delta t}$ is a $o(1)$-approximate optimal solution to* (2.1) *as $\Delta t \to 0$.*

The proof is deferred to the Appendix.

If the distributions $(p_{ij})_{(i,j)\in\mathcal{A}}$ are discrete and $f(\cdot)$ is piecewise constant, an exact optimal solution to (2.1) can be computed by appropriately choosing a different discretization

length for each node. In this chapter, we focus on discretization schemes with a uniform stepsize Δt for mathematical convenience. We stress that choosing adaptively the discretization length can improve the quality of the approximation for the same number of computations, see [50].

2.3.2.b Solution Procedures

The key observation enabling the development of label-setting algorithms to solve (2.4) is made in [85]. They note that, when the risk function is the probability of completion within budget, $u_i(t)$ can be computed for $i \in \mathcal{V}$ and $t \leq T$ as soon as the values taken by $u_j(\cdot)$ on $(-\infty, t-\delta^{\inf}]$ are available for all neighboring nodes $j \in \mathcal{V}(i)$ since $p_{ij}(\omega) = 0$ for $\omega \leq \delta^{\inf}$ under Assumption 2.3. They propose a label-setting algorithm which consists in computing the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ block by block, by interval increments of size $\delta^{\inf}$. After the following straightforward initialization step: $u_i(t) = 0$ for $t \leq 0$ and $i \in \mathcal{V}$, they first compute $(u_i(\cdot)_{[0,\delta^{\inf}]})_{i\in\mathcal{V}}$, then $(u_i(\cdot)_{[0,2\cdot\delta^{\inf}]})_{i\in\mathcal{V}}$ and so on to eventually derive $(u_i(\cdot)_{[0,T]})_{i\in\mathcal{V}}$. While this incremental procedure can still be applied for general risk functions, the initialization step gets tricky if $f(\cdot)$ does not have a one-sided compact support of the type $[a, \infty)$. Theorem 2.1 is crucial in this respect because the shortest-path tree $\mathcal{T}$ induces an ordering of the nodes to initialize the collection of functions $(u_i(\cdot))_{i\in\mathcal{V}}$ for remaining budgets smaller than T_f. The functions can subsequently be computed for larger budgets using the incremental procedure outlined above. To be specific, we solve (2.6) in three steps. First, we compute T_f (defined in Theorem 2.1). Inspecting the proof of Theorem 2.1, observe that T_f only depends on few parameters, namely the risk function $f(\cdot)$, the expected arc costs, and the maximum arc costs. Next, we compute the values $u_i^{\Delta t}(k\cdot\Delta t)$ for $k \in \{k_i^{\min}, \cdots, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor - 1\}$ starting at node $i = d$ and traversing the tree $\mathcal{T}$ in a breadth-first fashion using fast Fourier transforms with complexity $O(|\mathcal{V}|^2 \cdot \frac{\delta^{\sup}}{\Delta t} \cdot \log(|\mathcal{V}| \cdot \frac{\delta^{\sup}}{\Delta t}))$. Note that this step can be made to run significantly faster for specific risk functions, e.g. for the probability of completion within budget where $u_i^{\Delta t}(k\cdot\Delta t) = 0$ for $k < \left\lfloor \frac{T_f}{\Delta t} \right\rfloor$ and any $i \in \mathcal{V}$. Finally, we compute the values $u_i^{\Delta t}(k\cdot\Delta t)$ for $k \in \{\left\lfloor \frac{T_f}{\Delta t} \right\rfloor + m \cdot \left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor, \cdots, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor + (m+1) \cdot \left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor\}$ for all nodes $i \in \mathcal{V}$ by induction on m.

Complexity analysis. The description of the last step of the label-setting approach leaves out one detail that has a dramatic impact on the runtime complexity. We need to specify how to compute the convolution products arising in (2.6) for $k \geq \left\lfloor \frac{T_f}{\Delta t} \right\rfloor$, keeping in mind that, for any node $i \in \mathcal{V}$, the values $u_i^{\Delta t}(k \cdot \Delta t)$ for $k \in \{\left\lfloor \frac{T_f}{\Delta t} \right\rfloor, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor\}$ become available online by chunks of length $\left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor$ as the label-setting algorithm progresses. A naive implementation consisting in applying the pointwise definition of convolution products has a runtime complexity $O(|\mathcal{A}| \cdot \frac{(T-T_f)\cdot(\delta^{\sup}-\delta^{\inf})}{(\Delta t)^2})$. Using fast Fourier transforms for each chunk brings down the complexity to $O(|\mathcal{A}| \cdot \frac{(T-T_f)}{\Delta t} \cdot \frac{\delta^{\sup}}{\delta^{\inf}} \cdot \log(\frac{\delta^{\sup}}{\Delta t}))$. Applying another online scheme developed in [37] and [84], based on the idea of zero-delay convolution, leads to a worst-case complexity $O(|\mathcal{A}| \cdot \frac{(T-T_f)}{\Delta t} \cdot \log^2(\frac{\delta^{\sup}}{\Delta t}))$. Numerical evidence suggest that this last implementation significantly speeds up the computations, see [84].

2.4 Theoretical and Computational Analysis of the Robust Problem

2.4.1 Characterization of Optimal Policies

The properties satisfied by optimal solutions to the nominal problem naturally extend to their robust counterparts, which we recall are defined as optimal solutions to (2.3). In fact, all the results derived in this section are strict generalizations of those obtained in Section 2.3.1 for singleton ambiguity sets. We point out that the rectangularity of the global ambiguity set is essential for the results to carry over to the robust setting as it guarantees that *Bellman's Principle of Optimality* continue to hold, which is an absolute prerequisite for computational tractability.

Similarly as what we have seen for the nominal problem, infinite cycling might occur in the robust setting, depending on the risk function at hand. This difficulty can be shown not to arise under the same conditions on $f(\cdot)$ as for the nominal problem.

Theorem 2.2. *Case 1: If there exists T_1 such that either:*

(a) $f(\cdot)$ is increasing, concave, and C^2 on $(-\infty, T_1)$ and such that $\frac{f''}{f'} \to_{-\infty} 0$,

(b) $f(\cdot)$ is C^1 on $(-\infty, T_1)$ and $\lim_{-\infty} f'$ exists, is positive, and is finite,

then there exists T_f^r such that, for any $T \geq 0$ and as soon as the total cost spent so far is larger than $T - T_f^r$, any optimal policy solution to (2.3) *follows the shortest-path tree rooted at d with respect to the worst-case mean arc costs, which we denote by $\mathcal{T}^r$ (the worst-case mean arc costs are given by $(\max_{p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{X \sim p_{ij}}[X])_{(i,j) \in \mathcal{A}}$).*

Case 2: If there exists T_f such that the support of $f(\cdot)$ is included in $[T_f, \infty)$, then following $\mathcal{T}^r$ is optimal as soon as the total cost spent so far is larger than $T - T_f^r$.

For a node i, $\mathcal{T}^r(i)$ refers to the set of immediate successors of node i in $\mathcal{T}^r$. The proof is deferred to the Appendix.

Interestingly, T_f^r is determined by the exact same procedure as T_f provided the expected arc costs are substituted with the worst-case expected costs. For instance, when $f(t) = -t^2 \cdot 1_{t \leq 0}$, we may take:

$$T_f^r = -\frac{|\mathcal{V}| \cdot \delta^{\sup} \cdot \max_{i \in \mathcal{V}} M_i}{2 \cdot \min_{i \neq d} \min_{j \in \mathcal{V}(i), j \notin \mathcal{T}^r(i)} \{\max_{p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{X \sim p_{ij}}[X] + M_j - M_i\}},$$

where M_i is the worst-case minimum expected cost to go from i to d.

Last but not least, problem (2.3) can be formulated as a distributionally robust MDP in the extended space state $(i, t) \in \mathcal{V} \times (-\infty, T]$. As a result, one can show that there exists an optimal *Markov* policy $\pi^*_{f,\mathcal{P}}$ characterized by the dynamic programming equation:

$$\begin{aligned}
&u_d(t) = f(t) && t \leq T \\
&u_i(t) = \max_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t - \omega) \mathrm{d}\omega && i \neq d, t \leq T \\
&\pi^*_{f,\mathcal{P}}(i, t) \in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t - \omega) \mathrm{d}\omega && i \neq d, t \leq T,
\end{aligned} \tag{2.7}$$

where $u_i(t)$ is the worst-case expected objective-to-go when leaving $i \in \mathcal{V}$ with remaining budget t. Observe that (2.7) only differs from (2.4) through the presence of the infimum over $\mathcal{P}_{ij}$.

Proposition 2.3. *Any Markov policy solution to* (2.7) *is an optimal strategy for* (2.3).

The proof is deferred to the Appendix.

2.4.2 Tightness of the Robust Problem

The optimization problem (2.3) is a conservative approximation of (2.2) in the sense that, for any strategy $\pi \in \Pi$, we have:

$$\inf_{\forall (i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T - X_\pi)] \geq \inf_{\forall \tau, \forall (i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T - X_\pi)].$$

We say that (2.2) and (2.3) are equivalent if they share the same optimal value and if there exists a common optimal strategy. For general risk functions, ambiguity sets, and graphs, (2.2) and (2.3) are not equivalent. In this section, we highlight several situations of interest for which (2.2) and (2.3) happen to be equivalent and we bound the gap between the optimal values of (2.2) and (2.3) for a subclass of risk functions. In this chapter, we solve (2.3) instead of (2.2) for computational tractability, irrespective of whether or not (2.2) and (2.3) are equivalent. Hence, the results presented in this section are included mainly for illustrative purposes, i.e. we do not impose further restrictions on the risk function or the ambiguity sets here.

Equivalence of (2.2) **and** (2.3)**.** As a simple first example, observe that when $f(\cdot)$ is non-decreasing and $\mathcal{P}_{ij} = \mathcal{P}([\delta^{\inf}_{ij}, \delta^{\sup}_{ij}])$, both (2.2) and (2.3) reduce to a standard robust approach where the goal is to find a path minimizing the sum of the worst-case arc costs. The following result identifies conditions of broader applicability when the decision maker is risk-seeking.

Lemma 2.1. *Suppose that $f(\cdot)$ is convex and satisfies property (b) in Case 1 of Theorem 2.2 and that, for any arc $(i,j) \in \mathcal{V}$, either:*

(a) the Dirac distribution supported at $\max_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[X]$ belongs to $\mathcal{P}_{ij}$,

(b) there exist $\mu_{ij} \geq 0$, $\alpha_{ij} \geq 0$, and $\beta_{ij} \in [0,1]$ such that:

$$\mathcal{P}_{ij} = \left\{ p \in \mathcal{P}([\delta^{\inf}_{ij}, \delta^{\sup}_{ij}]) : \begin{array}{r} \mathbb{E}_{X\sim p}[X] = \mu_{ij} \\ \mathbb{E}_{X\sim p}[|X - \mu_{ij}|] = \alpha_{ij} \\ \mathbb{P}[X \geq \mu_{ij}] = \beta_{ij} \end{array} \right\} \tag{2.8}$$

Then, (2.2) *and* (2.3) *are equivalent.*

The proof is deferred to the Appendix.

To illustrate Lemma 2.1, observe that the assumptions are satisfied for $f(t) = \exp(a \cdot t) + b \cdot t$, with a and b taken as positive values, and when the ambiguity sets are defined either through (2.8) or through confidence intervals on the expected costs, i.e.:

$$\mathcal{P}_{ij} = \{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \ \mathbb{E}_{X\sim p}[X] \in [\alpha_{ij}, \beta_{ij}]\}, \tag{2.9}$$

with $\alpha_{ij} \leq \beta_{ij}$. Further note that adding upper bounds on the mean deviation or on higher-order moments in the definition of the ambiguity sets (2.9) does not alter the conclusion of Lemma 2.1. We move on to another situation of interest where (2.2) and (2.3) can be shown to be equivalent.

Lemma 2.2. *Take $K \in \mathbb{N}$. Suppose that:*

- *$\mathcal{G}$ is a single-path graph,*
- *$f(\cdot)$ is C^{K+1} and $f^{(K+1)}(t) > 0 \ \forall t$ or $f^{(K+1)}(t) < 0 \ \forall t$,*
- *For any arc $(i, j) \in \mathcal{A}$:*

$$\mathcal{P}_{ij} = \left\{ p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \begin{array}{c} \mathbb{E}_{X\sim p}(X) = m_{\text{ij}}^1 \\ \vdots \\ \mathbb{E}_{X\sim p}(X^K) = m_{\text{ij}}^K \end{array} \right\}$$

  where $m_{\text{ij}}^1, \cdots, m_{\text{ij}}^K$ are non-negative.

Then (2.2) *and* (2.3) *are equivalent.*

The proof is deferred to the Appendix.

When $\mathcal{G}$ is a single-path graph, the optimal value of (2.2) corresponds to the worst-case risk function when following this path, given that the arc cost distributions are only known to lie in the ambiguity sets. While it is a priori unclear how to compute this quantity,

Proposition 2.4 of Section 2.4.3.a establishes that the optimal value of (2.3) can be determined with arbitrary precision provided the inner optimization problems appearing in the discretization scheme of Section 2.4.3.a can be computed numerically. Hence, even in this seemingly simplistic situation, the equivalence between (2.2) and (2.3) is an important fact to know as it has significant computational implications. Lemma 2.2 shows that, when the risk function is $(K+1)$th order convex or concave and when the arc cost distributions are only known through the first K-order moments, (2.2) and (2.3) are in fact equivalent. For this particular class of ambiguity sets, the inner optimization problems of the discretization scheme of Section 2.4.3.a can be solved using semidefinite programming, see [23].

Bounding the gap between the optimal values of (2.2) **and** (2.3)**.** It turns out that, for a particular subclass of risk functions, we can bound the gap between the optimal values of (2.2) and (2.3) uniformly over all graphs and ambiguity sets.

Lemma 2.3. *Denote the optimal value of* (2.2) *(resp.* (2.3)*) by* v^* *(resp.* v*).*
If there exists $\gamma, a > 0$ *and* β, b *such that one of the following conditions holds:*

- $\gamma \cdot t + \beta \geq f(t) \geq a \cdot t + b \quad \forall t \leq T$,
- $\gamma \cdot \exp(t) + \beta \geq f(t) \geq a \cdot \exp(t) + b \quad \forall t \leq T$,
- $-\gamma \cdot \exp(-t) + \beta \geq f(t) \geq -a \cdot \exp(-t) + b \quad \forall t \leq T$,

then $v^* \geq v \geq \frac{a}{\gamma} \cdot (v^* - \beta) + b$.

The proof is deferred to the Appendix.

2.4.3 Solution Methodology

We proceed as in Section 2.3.2 and compute an approximate *Markov* policy solution to (2.7). The computational challenges faced when solving the nominal problem carry over to the robust counterpart, but with additional difficulties to overcome. Specifically, the continuity of the problem leads us to build a discrete approximation in Section 2.4.3.a similar to the one developed for the nominal approach. We also extend the label-setting algorithm

of Section 2.3.2.b to tackle the potential existence of cycles at the beginning of Section 2.4.3.c. However, the presence of an inner optimization problem in (2.7) is a distinctive feature of the robust problem which poses a new computational challenge. As a result, and in contrast with the situation for the nominal problem where this optimization problem reduces to a convolution product, it is not a priori obvious how to solve the discretization scheme numerically, let alone efficiently. As can be expected, the exact form taken by the ambiguity sets has a major impact on the computational complexity of the inner optimization problem. In an effort to mitigate the computational burden, we restrict our attention to a subclass of ambiguity sets defined by confidence intervals on piecewise affine statistics in Section 2.4.3.b. While this simplification might seem restrictive, we show that this subclass displays significant modeling power. We develop two general-purpose algorithms in Section 2.4.3.c for this particular subclass of ambiguity sets. Using these algorithms, we can compute:

- an ϵ-approximate solution to (2.3) in

$$O\left(\frac{|\mathcal{A}| \cdot (T - T_f^r) + |\mathcal{V}|^2 \cdot \delta^{\sup}}{\Delta t} \cdot \log\left(\frac{\delta^{\sup} - \delta^{\inf}}{\Delta t}\right) \cdot \log\left(\frac{|\mathcal{V}| + \frac{T - T_f^r}{\delta^{\inf}}}{\epsilon}\right)\right)$$

  computation time when the arc costs only take on values that are multiple of $\Delta t > 0$ and for any continuous risk function $f(\cdot)$ satisfying the conditions of Theorem 2.2. This also applies when the objective is to maximize the probability of completion within budget and even simplifies to $O(|\mathcal{A}| \cdot \frac{T}{\Delta t} \cdot \log(\frac{\delta^{\sup} - \delta^{\inf}}{\Delta t}) \cdot \log(\frac{T}{\epsilon \cdot \delta^{\inf}}))$. Note that, in contrast to the nominal problem, we are only able to compute approximate solutions because finding a solution to (2.7) entails solving to optimality optimization programs as opposed to computing convolutions.

- an ϵ-approximate solution to (2.3) in

$$O\left(\frac{(|\mathcal{V}| + \frac{T - T_f^r}{\delta^{\inf}})^2 \cdot (|\mathcal{A}| \cdot (T - T_f^r) + |\mathcal{V}|^2 \cdot \delta^{\sup})}{\epsilon} \cdot \log^2\left(\frac{(|\mathcal{V}| + \frac{T - T_f^r}{\delta^{\inf}}) \cdot \delta^{\sup}}{\epsilon}\right)\right)$$

  computation time when the risk function is Lipschitz on compact sets.

Our methodology can be outlined as follows. We remark that the inner optimization problems arising in (2.7) are conic linear problems whose duals reduce to linear programs with $O(1)$ variables and $O(\frac{1}{\Delta t})$ (or equivalently $O(\frac{1}{\epsilon})$) constraints for the subclass of ambiguity sets defined in Section 2.4.3.b. The computational attractiveness of our approach hinges on the observation that there is a significant overlap between the constraints of these linear programs. This translates into an efficient separation oracle, based on a data structure maintaining the convex hull of a dynamic set of points efficiently, running in amortized time $O(\log(\frac{1}{\Delta t}))$ (or equivalently $O(\log(\frac{1}{\epsilon}))$). As a basis for comparison, a brute-force approach consisting in solving these linear programs independently from one another has runtime complexity polynomial in $\frac{1}{\Delta t}$ (or equivalently $\frac{1}{\epsilon}$). Additionally, constantly recomputing the convex hulls from scratch in a naive fashion would lead to a global running time quadratic in $\frac{1}{\Delta t}$ (or equivalently $\frac{1}{\epsilon}$). The mechanism behind our separation oracle can be regarded as a counterpart of the online fast Fourier scheme for the nominal approach.

2.4.3.a Discretization Scheme

For $i \in \mathcal{V}$, we approximate $u_i(\cdot)$ by a piecewise affine continuous function $u_i^{\Delta t}(\cdot)$ of uniform stepsize Δt. This is in contrast with Section 2.3.2.a where we use a piecewise constant approximation. This change is motivated by computational considerations: the continuity of $u_i^{\Delta t}(\cdot)$ guarantees strong duality for the inner optimization problem appearing in (2.7). Just like for the nominal problem, we only need to approximate $u_i(\cdot)$ for a remaining budget larger than $k_i^{r,\min} \cdot \Delta t$, for $k_i^{r,\min} = \left\lfloor \frac{T_f^r - (|\mathcal{V}| - \text{level}(i,\mathcal{T}^r)+1)\cdot \delta^{\sup}}{\Delta t} \right\rfloor$, where $\text{level}(i,\mathcal{T}^r)$ is the level of node i in $\mathcal{T}^r$. Specifically, we use the approximation:

$$
\begin{aligned}
&u_i^{\Delta t}(t) = (1 - \frac{t}{\Delta t} + \left\lfloor \frac{t}{\Delta t} \right\rfloor) \cdot u_i^{\Delta t}(\left\lfloor \frac{t}{\Delta t} \right\rfloor \cdot \Delta t) + (\frac{t}{\Delta t} - \left\lfloor \frac{t}{\Delta t} \right\rfloor) \cdot u_i^{\Delta t}(\left\lceil \frac{t}{\Delta t} \right\rceil \cdot \Delta t) \\
&\text{for } i \in \mathcal{V}, t \in [k_i^{r,\min} \cdot \Delta t, T] \\
&\pi^{\Delta t}(i,t) = \pi^{\Delta t}(i, \left\lfloor \frac{t}{\Delta t} \right\rfloor \cdot \Delta t) \\
&\text{for } i \neq d, t \in [k_i^{r,\min} \cdot \Delta t, T],
\end{aligned}
\tag{2.10}
$$

and the values at the mesh points are determined by the set of equalities:

$$
\begin{aligned}
&u_d^{\Delta t}(k\cdot\Delta t) = f(k\cdot\Delta t) \quad \text{for } k = k_d^{r,\min}, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor \\
&u_i^{\Delta t}(k\cdot\Delta t) = \max_{j\in\mathcal{V}(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega \\
&\text{for } i\neq d, k = \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor \\
&\pi^{\Delta t}(i,k\cdot\Delta t) \in \operatorname*{argmax}_{j\in\mathcal{V}(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega \\
&\text{for } i\neq d, k = \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor \\
&u_i^{\Delta t}(k\cdot\Delta t) = \max_{j\in\mathcal{T}^r(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega \\
&\text{for } i\neq d, k = k_i^{r,\min}, \cdots, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor - 1 \\
&\pi^{\Delta t}(i,k\cdot\Delta t) \in \operatorname*{argmax}_{j\in\mathcal{T}^r(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega)\cdot u_j^{\Delta t}(k\cdot\Delta t-\omega)\mathrm{d}\omega \\
&\text{for } i\neq d, k = k_i^{r,\min}, \cdots, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor - 1.
\end{aligned}
\tag{2.11}
$$

As we did for the nominal problem, we can quantify the quality of $\pi^{\Delta t}$ as an approximate solution to (2.3) as a function of the regularity of the risk function.

Proposition 2.4. *Consider a solution to the global discretization scheme* (2.10) *and* (2.11)*,* $(\pi^{\Delta t}, (u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}})$*. We have:*

1. *If $f(\cdot)$ is non-decreasing, the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge pointwise almost everywhere to $(u_i(\cdot))_{i\in\mathcal{V}}$ as $\Delta t\to 0$.*

2. *If $f(\cdot)$ is continuous, the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge uniformly to $(u_i(\cdot))_{i\in\mathcal{V}}$ and $\pi^{\Delta t}$ is a $o(1)$-approximate optimal solution to* (2.3) *as $\Delta t\to 0$.*

3. *If $f(\cdot)$ is Lipschitz on compact sets (e.g. if $f(\cdot)$ is C^1), the functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ converge uniformly to $(u_i(\cdot))_{i\in\mathcal{V}}$ at speed Δt and $\pi^{\Delta t}$ is a $O(\Delta t)$-approximate optimal solution to* (2.3) *as $\Delta t\to 0$.*

The proof is deferred to the Appendix.

2.4.3.b Ambiguity Sets

For computational tractability, we restrict our attention to the following subclass of ambiguity sets.

Definition 2.1. *For any arc $(i,j) \in \mathcal{A}$, we have:*

$$\mathcal{P}_{ij} = \{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \ \mathbb{E}_{X\sim p}[g_q^{ij}(X)] \in [\alpha_q^{ij}, \beta_q^{ij}],\ q = 1, \cdots, Q_{ij}\},$$

where:

- $Q_{ij} \in \mathbb{N}$ *denotes the number of statistics used,*
- $-\infty \leq \alpha_q^{ij} \leq \beta_q^{ij} \leq \infty$ *for* $q = 1, \cdots, Q_{ij}$*,*
- *the functions* $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ *are piecewise affine with a finite number of pieces on* $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$ *and such that* $g_q^{ij}(\cdot)$ *is upper (resp. lower) semi-continuous if* $\alpha_q^{ij} > -\infty$ *(resp.* $\beta_q^{ij} < \infty$*), for any* $q = 1, \cdots, Q_{ij}$*.*

The second restriction imposed on the functions $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ is meant to guarantee that $\mathcal{P}_{ij}$ is closed for the weak topology, which is required by Assumption 2.4. Note that Definition 2.1 allows to model one-sided constraints by either taking $\alpha_q^{ij} = -\infty$ or $\beta_q^{ij} = \infty$. For instance, the constraints $\mathbb{E}_{X\sim p}[1_{X\in S}] \leq \beta$ and $\mathbb{E}_{X\sim p}[1_{X\in S'}] \geq \beta$, for S (resp. S') an open (resp. a closed) set, are perfectly valid. In terms of modeling power, Definition 2.1 allows to have constraints on standard statistics, such as the mean value and the mean absolute deviation, but also to capture distributional asymmetry, through constraints on any quantile or of the type $\mathbb{E}_{X\sim p}[X \cdot 1_{X>\theta}] \leq \beta$, and to incorporate higher-order information, e.g. the variance or the skewness, since continuous functions can be approximated arbitrarily well by piecewise affine functions on a compact set. Finally, note that Definition 2.1 allows to model situations where c_{ij} only takes values in a prescribed finite set S through the constraint $\mathbb{E}_{X\sim p}[1_{X\in S}] \geq 1$.

Data-driven ambiguity sets. Ambiguity sets of the form introduced in Definition 2.1 can be built using a combination of prior knowledge and historical data. To illustrate,

suppose that, for any arc $(i,j) \in \mathcal{A}$, we have observed n_{ij} samples $(X_p^{ij})_{p=1,\cdots,n_{ij}}$ drawn from the corresponding arc cost distribution. Setting aside computational aspects, there is an inherent trade-off at play when designing ambiguity sets with this empirical data: using more statistics and/or narrowing the confidence intervals $([\alpha_q^{ij}, \beta_q^{ij}])_{q=1,\cdots,Q_{ij}}$ will shrink the ambiguity sets $\mathcal{P}_{ij}$ with two implications. On one hand, the quality of the guarantee on the risk function provided by the robust approach will improve (i.e. the optimal value of (2.3) will increase). On the other hand, the probability that this guarantee holds will deteriorate. Assuming we want to set this probability value to $1-\epsilon$ and that we are set on which statistics to use $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$, the trade-off is simple to resolve as far as the confidence intervals are concerned. Using Hoeffding's and Boole's inequalities, the confidence interval for statistic q of arc (i,j) should be centered at the empirical mean:

$$\alpha_q^{ij} = \frac{1}{n_{ij}} \sum_{p=1}^{n_{ij}} g_q(X_p^{ij}) - \epsilon_q^{ij}, \quad \beta_q^{ij} = \frac{1}{n_{ij}} \sum_{p=1}^{n_{ij}} g_q(X_p^{ij}) + \epsilon_q^{ij},$$

with half width ϵ_q^{ij} determined by:

$$\epsilon_q^{ij} / (\max_{[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]} g_q^{ij} - \min_{[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]} g_q^{ij}) = \sqrt{\log(\frac{2}{\epsilon} \cdot \sum_{(i,j)\in\mathcal{A}} Q_{ij})/2n_{ij}} \tag{2.12}$$

so that the probability that the true arc cost distribution $\mathbf{p}$ lies in the rectangular ambiguity set $\prod_{(i,j)\in\mathcal{A}} \mathcal{P}_{ij}$ is at least $1-\epsilon$. Choosing how many and which statistics to use is a more complex endeavor as the impact on the size of $\prod_{(i,j)\in\mathcal{A}} \mathcal{P}_{ij}$ is a priori unclear: using more statistics adds constraints in the definition of $\mathcal{P}_{ij}$ which tends to shrink it but at the same time we have to increase all the widths $(\epsilon_q^{ij})_{q=1,\cdots,Q_{ij}}$ to keep the probability that the guarantee holds at the same level $1-\epsilon$ (see the dependence on $\sum_{(i,j)\in\mathcal{A}} Q_{ij}$ in (2.12)). Numerical evidence presented in Section 2.5 suggests that low-order statistics, such as the mean, tend to be more informative when only few samples are available. Conversely, as sample sizes get very large, incorporating higher-order information, for example in the form of piecewise statistics that approximate the variance such as the mean deviation, seems to improve the quality of the strategy derived. In the limit where the statistics can be computed exactly, we should use as many statistics as possible. This observation is supported by the following

lemma.

Lemma 2.4. *For any arc $(i,j) \in \mathcal{A}$, consider $(\mathcal{P}_{ij}^k)_{k\in\mathbb{N}}$, a sequence of nested ambiguity sets satisfying Assumption 2.4. If $f(\cdot)$ is continuous, then the optimal value of the robust problem* (2.3) *when the uncertainty sets are taken as $(\mathcal{P}_{ij}^k)_{(i,j)\in\mathcal{A}}$ monotonically converges to the optimal value of* (2.3) *when the uncertainty sets are taken as $(\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k)_{(i,j)\in\mathcal{A}}$ as $k \to \infty$.*

In particular, if $\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k$ is a singleton for all arcs $(i,j) \in \mathcal{A}$, then the optimal value of the robust problem converges to the value of the nominal problem (2.1).

The proof is deferred to the Appendix.

2.4.3.c Solution Procedures

We develop two general-purpose methods to compute a solution to the discretization scheme (2.11) for the class of ambiguity sets identified in Section 2.4.3.b. The first method, based on the ellipsoid algorithm, computes an $\epsilon-$approximate solution to (2.11) with worst-case complexity:

$$O\left(\frac{|\mathcal{A}| \cdot (T - T_f^r) + |\mathcal{V}|^2 \cdot \delta^{\sup}}{\Delta t} \cdot \log\left(\frac{\delta^{\sup} - \delta^{\inf}}{\Delta t}\right) \cdot \log\left(\frac{|\mathcal{V}| + \frac{T-T_f^r}{\delta^{\inf}}}{\epsilon}\right)\right),$$

provided $f(\cdot)$ is continuous and where the hidden factors are linear in the number of pieces of each statistic and polynomial in the number of statistics. We remind the reader that the complexity of solving the discretization scheme (2.6) for the nominal problem is $O(|\mathcal{A}| \cdot \frac{T-T_f}{\Delta t} \cdot \log^2(\frac{\delta^{\sup}}{\Delta t}) + |\mathcal{V}|^2 \cdot \frac{\delta^{\sup}}{\Delta t} \cdot \log(|\mathcal{V}| \cdot \frac{\delta^{\sup}}{\Delta t}))$ when using zero-delay convolution. While these bounds are not directly comparable because some of the parameters required to specify a robust instance are not relevant for a nominal instance and vice versa, we point out that they share many similarities, including the almost linear dependence on $\frac{1}{\Delta t}$. The second method, based on delayed column generation and warm starting techniques, is more practical but has worst-case complexity exponential in $\frac{1}{\Delta t}$. We stress that none of these approaches can be used to solve the nominal problem as the latter is not a particular case of the robust problem for the restricted class of ambiguity sets defined in Section 2.4.3.b. Indeed, characterizing a single distribution generally requires infinitely many moment constraints.

Label-setting approach. To cope with the potential existence of cycles, we remark that the label-setting approach developed for the nominal approach trivially extends to the robust setting. Similarly as for the nominal problem, we proceed in three steps to solve (2.11). First, we compute T_f^r. Next, we compute the values $u_i^{\Delta t}(k\cdot\Delta t)$ for $k \in \{k_i^{r,\min}, \cdots, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor - 1\}$ starting at node $i = d$ and traversing the tree $\mathcal{T}^r$ in a breadth-first fashion. Finally, we compute the values $u_i^{\Delta t}(k \cdot \Delta t)$ for $k \in \{\left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor + m \cdot \left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor, \cdots, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor + (m+1) \cdot \left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor\}$ for all nodes $i \in \mathcal{V}$ by induction on m. Of course, an efficient procedure solving the inner optimization problem of (2.11) is a prerequisite for carrying out the last two steps. This will be our focus in the remainder of this section.

Solving the Inner Optimization Problem. Consider any arc $(i,j) \in \mathcal{A}$. We need to solve, at each step $k \in \{k_i^{r,\min}, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor\}$, the optimization problem:

$$\begin{array}{ll} \inf\limits_{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}])} & \mathbb{E}_{X\sim p}[u_j^{\Delta t}(k\cdot\Delta t - X)] \\ \text{subject to} & \mathbb{E}_{X\sim p}[g_q^{ij}(X)] \in [\alpha_q^{ij}, \beta_q^{ij}] \quad q = 1, \cdots, Q_{ij}. \end{array} \tag{2.13}$$

Since the set of non-negative measures on $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$ is a cone, (2.13) can be cast as a conic linear problem. As a result, standard conic duality theory applies and the optimal value of (2.13) can be equivalently computed by solving a dual optimization problem which turns out to be easier to study. For a thorough exposition of the duality theory of general conic linear problems, see [87]. To simplify the presentation, we assume that $(\alpha_q^{ij})_{q=1,\cdots,Q_{ij}}$ and $(\beta_q^{ij})_{q=1,\cdots,Q_{ij}}$ are all finite quantities (which implies that the functions $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ are continuous by Definition 2.1) but this is by no means a limitation of our approach.

Lemma 2.5. *The optimization problem* (2.13) *has the same optimal value as the semi-infinite linear program:*

$$\begin{array}{ll} \sup\limits_{\substack{z\in\mathbb{R} \\ y_1,\cdots,y_{Q_{ij}}\in\mathbb{R}_+ \\ x_1,\cdots,x_{Q_{ij}}\in\mathbb{R}_+}} & z + \sum\limits_{q=1}^{Q_{ij}} (\alpha_q^{ij}\cdot x_q - \beta_q^{ij}\cdot y_q) \\ \textit{subject to} & z + \sum\limits_{q=1}^{Q_{ij}} (x_q - y_q)\cdot g_q^{ij}(\omega) \le u_j^{\Delta t}(k\cdot\Delta t - \omega) \quad \forall \omega \in [\delta_{ij}^{\inf}, \delta_{ij}^{\sup}] \end{array} \tag{2.14}$$

The proof is deferred to the Appendix.

Because the functions $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ are all piecewise affine, we can partition the interval $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$ into R_{ij} non-overlapping intervals $(I_r)_{r=1,\cdots,R_{ij}}$ such that the functions $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ are all affine on I_r for any $r \in \{1, \cdots, R_{ij}\}$, i.e.:

$$g_q^{ij}(\omega) = a_{q,r}^{ij} \cdot \omega + b_{q,r}^{ij} \text{ if } \omega \in I_r$$

for any $q \in \{1, \cdots, Q_{ij}\}$ and $\omega \in [\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$. This decomposition enables us to show that the feasible region of (2.14) can be described with finitely many inequalities.

Lemma 2.6. *The semi-infinite linear program* (2.14) *can be reformulated as the following finite linear program:*

$$\begin{aligned}
\sup_{\substack{z\in\mathbb{R}\\ y_1,\cdots,y_{Q_{ij}}\in\mathbb{R}\\ x_1,\cdots,x_{Q_{ij}}\in\mathbb{R}}} \quad & z + \sum_{q=1}^{Q_{ij}} (\alpha_q^{ij} \cdot x_q - \beta_q^{ij} \cdot y_q) \\
\textit{subject to} \quad & z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot l \cdot \Delta t + b_{q,r}^{ij}) \leq u_j^{\Delta t}((k-l)\cdot\Delta t) \\
& \textit{for } l = \left\lceil \frac{\inf(I_r)}{\Delta t} \right\rceil, \cdots, \left\lfloor \frac{\sup(I_r)}{\Delta t} \right\rfloor \textit{ and } r = 1, \cdots, R_{ij} \\
& z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot \sup(I_r) + b_{q,r}^{ij}) \leq u_j^{\Delta t}(k \cdot \Delta t - \sup(I_r)) \\
& \textit{for } r = 1, \cdots, R_{ij} \\
& z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot \inf(I_r) + b_{q,r}^{ij}) \leq u_j^{\Delta t}(k \cdot \Delta t - \inf(I_r)) \\
& \textit{for } r = 1, \cdots, R_{ij} \\
& y_q, x_q \geq 0 \quad q = 1, \cdots, Q_{ij}.
\end{aligned} \tag{2.15}$$

Proof. Take $z, y_1, \cdots, y_{Q_{ij}}, x_1, \cdots, x_{Q_{ij}} \in \mathbb{R}$ and $r \in \{1, \cdots, R_{ij}\}$. Since the function $\omega \to z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot g_q^{ij}(\omega)$ is affine on I_r, this function lies below the continuous piecewise affine function $u_j^{\Delta t}(k \cdot \Delta t - \cdot)$ on I_r if and only if it lies below $u_j^{\Delta t}(k \cdot \Delta t - \cdot)$ at every breakpoint of $u_j^{\Delta t}(k \cdot \Delta t - \cdot)$ on $\bar{I}_r$ and at the boundary points of $\bar{I}_r$. Since the collection of intervals $(I_r)_{r=1,\cdots,R_{ij}}$ forms a partition of $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$, this establishes the

claim. □

While (2.15) is a finite linear program and can thus be solved with an interior point algorithm, the large number of constraints calls for an efficient separation oracle, which we develop next, and the use of the ellipsoid algorithm. The key is to refine the idea of Lemma 2.6. Specifically, for any $r \in \{1, \cdots, R_{ij}\}$ and $l \in \{\left\lceil \frac{\inf(I_r)}{\Delta t} \right\rceil, \cdots, \left\lfloor \frac{\sup(I_r)}{\Delta t} \right\rfloor\}$, the constraint

$$z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot l \cdot \Delta t + b_{q,r}^{ij}) \leq u_j^{\Delta t}((k-l) \cdot \Delta t)$$

does not limit the feasible region if $(l \cdot \Delta t, u_j^{\Delta t}((k-l) \cdot \Delta t))$ is not an extreme point of the upper convex hull of $\{(m \cdot \Delta t, u_j^{\Delta t}((k-m) \cdot \Delta t)),\ m = \left\lfloor \frac{\inf(I_r)}{\Delta t} \right\rfloor, \cdots, \left\lceil \frac{\sup(I_r)}{\Delta t} \right\rceil\}$. Denote by $\mathcal{L}_{ij}^{k,r}$ the subset of integers l such that $(l \cdot \Delta t, u_j^{\Delta t}((k-l) \cdot \Delta t))$ is such an extreme point. Observe that the function

$$l \to u_j^{\Delta t}((k-l) \cdot \Delta t) - [z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot l \cdot \Delta t + b_{q,r}^{ij})]$$

is convex on $\mathcal{L}_{ij}^{k,r}$, therefore a minimizer of this function can be found by binary search. As a result, being able to perform binary search on $\mathcal{L}_{ij}^{k,r}$ efficiently would enable us to separate efficiently the subset of constraints:

$$z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot (a_{q,r}^{ij} \cdot l \cdot \Delta t + b_{q,r}^{ij}) \leq u_j^{\Delta t}((k-l) \cdot \Delta t) \quad l = \left\lceil \frac{\inf(I_r)}{\Delta t} \right\rceil, \cdots, \left\lfloor \frac{\sup(I_r)}{\Delta t} \right\rfloor.$$

We defer the presentation of a data structure designed for this purpose to Section 2.4.3.d and make the following assumption to conclude the computational study.

Assumption 2.5. *For any two integers L, L' such that $\left\lceil \frac{\delta_{ij}^{\inf}}{\Delta t} \right\rceil \leq L < L' \leq \left\lfloor \frac{\delta_{ij}^{\sup}}{\Delta t} \right\rfloor$, there exists a data structure that can maintain, dynamically as k increases from $k = k_i^{r,min}$ to $k = \left\lfloor \frac{T}{\Delta t} \right\rfloor$, a description of the upper convex hull of $\{(l \cdot \Delta t, u_j^{\Delta t}((k-l) \cdot \Delta t)),\ l = L, \cdots, L'\}$ allowing to perform binary search on the first coordinate of the extreme points with a global complexity $O((\frac{T}{\Delta t} - k_i^{r,min}) \cdot \log(\frac{\delta^{\sup} - \delta^{\inf}}{\Delta t}))$.*

Equipped with a data structure satisfying Assumption 2.5, the separation oracle has runtime

complexity $O(\log(\frac{\delta^{\sup}-\delta^{\inf}}{\Delta t}))$ given that there are at most $\left\lfloor \frac{\delta_{ij}^{\sup}}{\Delta t} \right\rfloor - \left\lceil \frac{\delta_{ij}^{\inf}}{\Delta t} \right\rceil$ extreme points at any step k. Using the ellipsoid algorithm, we can compute the optimal value of (2.13) with precision ϵ in $O(\log(\frac{\delta^{\sup}-\delta^{\inf}}{\Delta t}) \cdot \log(\frac{1}{\epsilon}))$ running time, where the hidden factors are polynomial in Q_{ij} and linear in R_{ij}. We point out that relying on a data structure satisfying Assumption 2.5 is critical to achieve this complexity: recomputing the upper convex hull from scratch at every time step k would increase the complexity to $O(\frac{\delta^{\sup}-\delta^{\inf}}{\Delta t} \cdot \log(\frac{1}{\epsilon}))$ (achieved using, for instance, Andrew's monotone chain convex hull algorithm).

Practical general purpose method. Due to the limited practicability of the ellipsoid algorithm, we have developed another method based on delayed column generation to solve the inner optimization problem. To simplify the presentation, we assume that $(\inf(I_r))_{r=1,\cdots,R_{ij}}$ and $(\sup(I_r))_{r=1,\cdots,R_{ij}}$ are all multiples of Δt. Since (2.15) is a linear program with a non-empty feasible set, we can equivalently compute its value by solving the dual optimization problem given by:

$$\begin{aligned} \inf_{p_0,\cdots,p_L \in \mathbb{R}_+} \quad & \sum_{l=0,\cdots,L} p_l \cdot u_j^{\Delta t}((k-l)\cdot \Delta t - \delta_{ij}^{\inf}) \\ \text{subject to} \quad & \sum_{l=0,\cdots,L} p_l \cdot g_q^{ij}(l\cdot \Delta t + \delta_{ij}^{\inf}) \in [\alpha_q^{ij}, \beta_q^{ij}] \quad q = 1,\cdots,Q_{ij} \\ & \sum_{l=0,\cdots,L} p_l = 1 \end{aligned} \tag{2.16}$$

where $L = \frac{\delta_{ij}^{\sup}-\delta_{ij}^{\inf}}{\Delta t}$. Observe that the feasible set of the linear program (2.16) does not change across steps $k = k_i^{r,\min}, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor$. Hence, we can warm start the primal simplex algorithm with the optimal solution found at the previous step. Furthermore, the separation oracle developed for the dual optimization problem can also be used as a subroutine for delayed column generation.

Faster procedure when the mean is the only statistics. If the ambiguity sets are only defined through a confidence interval on the mean value, i.e.:

$$\mathcal{P}_{ij} = \{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \ \mathbb{E}_{X\sim p}[X] \in [\alpha^{ij}, \beta^{ij}]\},$$

then (2.15) can be solved to optimality in $O(\log(\frac{\delta_{ij}^{\sup}-\delta_{ij}^{\inf}}{\Delta t}))$ computation time without resorting to the ellipsoid algorithm. First observe that (2.15) simplifies to:

$$\begin{aligned}
\sup_{z\in\mathbb{R},y,x\in\mathbb{R}_+} \quad & z+\alpha^{ij}\cdot x-\beta^{ij}\cdot y \\
\text{subject to} \quad & z+(x-y)\cdot l\cdot\Delta t\le u_j^{\Delta t}((k-l)\cdot\Delta t), \quad l=\left\lceil\frac{\delta_{ij}^{\inf}}{\Delta t}\right\rceil,\cdots,\left\lfloor\frac{\delta_{ij}^{\sup}}{\Delta t}\right\rfloor \\
& z+(x-y)\cdot\delta_{ij}^{\sup}\le u_j^{\Delta t}(k\cdot\Delta t-\delta_{ij}^{\sup}) \\
& z+(x-y)\cdot\delta_{ij}^{\inf}\le u_j^{\Delta t}(k\cdot\Delta t-\delta_{ij}^{\inf})
\end{aligned} \tag{2.17}$$

As it turns out, we can identify an optimal feasible basis to (2.17) by direct reasoning.

Lemma 2.7. *An optimal solution to* (2.17) *can be found by performing three binary searches on the first coordinate of the extreme points of the upper convex hull of*

$$\begin{aligned}
&\{(l\cdot\Delta t, u_j^{\Delta t}((k-l)\cdot\Delta t)),\ l=\left\lfloor\frac{\delta_{ij}^{\inf}}{\Delta t}\right\rfloor,\cdots,\left\lceil\frac{\delta_{ij}^{\sup}}{\Delta t}\right\rceil\} \\
&\cup\{(\delta_{ij}^{\sup}, u_j^{\Delta t}(k\cdot\Delta t-\delta_{ij}^{\sup})),(\delta_{ij}^{\inf}, u_j^{\Delta t}(k\cdot\Delta t-\delta_{ij}^{\inf}))\}.
\end{aligned}$$

The proof is deferred to the Appendix.

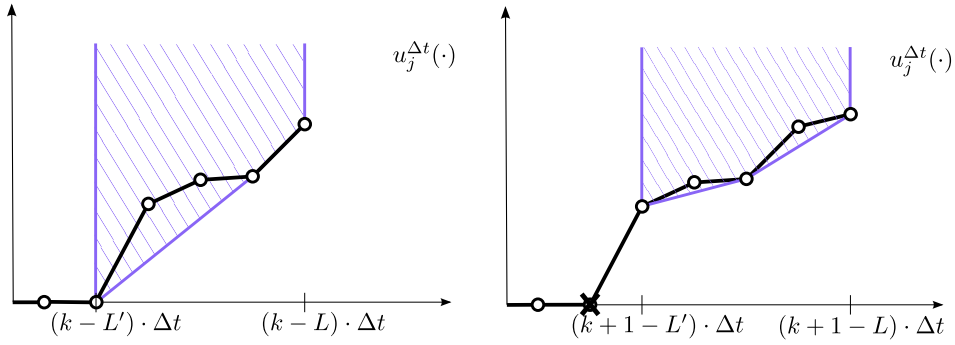
Hence, (2.17) can be solved to optimality in $O(\log(\frac{\delta_{ij}^{\sup}-\delta_{ij}^{\inf}}{\Delta t}))$ running time provided that the extreme points are stored in a data structure satisfying Assumption 2.5.

Faster procedure when the statistics are piecewise constant. When the statistics are piecewise constant, we have:

$$a_{q,r}^{ij}=0 \quad q=1,\cdots,Q_{ij},\ r=1,\cdots,R_{ij}.$$

Hence, for any $r\in\{1,\cdots,R_{ij}\}$, the set of constraints

$$z+\sum_{q=1}^{Q_{ij}}(x_q-y_q)\cdot(a_{q,r}^{ij}\cdot l\cdot\Delta t+b_{q,r}^{ij})\le u_j^{\Delta t}((k-l)\cdot\Delta t) \quad l=\left\lceil\frac{\inf(I_r)}{\Delta t}\right\rceil,\cdots,\left\lfloor\frac{\sup(I_r)}{\Delta t}\right\rfloor$$



(a) $\hat{\mathcal{C}}_k$ is the hatched area. (b) $\hat{\mathcal{C}}_{k+1}$ is the hatched area.

Figure 2-3: The graph of $u_j^{\Delta t}(\cdot)$ is plotted in black. The dot points represent the breakpoints of $u_j^{\Delta t}(\cdot)$.

is equivalent to the single constraint:

$$z + \sum_{q=1}^{Q_{ij}} (x_q - y_q) \cdot b_{q,r}^{ij} \leq \min_{l=\lceil \frac{\inf(I_r)}{\Delta t} \rceil, \cdots, \lfloor \frac{\sup(I_r)}{\Delta t} \rfloor} u_j^{\Delta t}((k-l)\cdot\Delta t),$$

whose right-hand side can be computed by binary search on $\mathcal{L}_{ij}^{k,r}$. As a result, the linear program (2.15) has $2 \cdot Q_{ij} + 1$ variables and $2 \cdot Q_{ij} + 3 \cdot R_{ij}$ constraints and can be solved to precision ϵ with an interior-point algorithm in $O(\log(\frac{1}{\epsilon}))$ computation time. Typically, piecewise constant statistics can be used to bound the probability that a given event occurs, see Section 2.4.3.b.

2.4.3.d Dynamic Convex Hull Algorithm

Fix an arc $(i,j) \in \mathcal{A}$ and two integers $L < L'$ in $\{\left\lceil \frac{\delta_{ij}^{\inf}}{\Delta t} \right\rceil, \cdots, \left\lfloor \frac{\delta_{ij}^{\sup}}{\Delta t} \right\rfloor\}$. We are interested in the extreme points of the upper convex hull of $\{(l\cdot\Delta t, u_j^{\Delta t}((k-l)\cdot\Delta t)),\ l = L, \cdots, L'\}$ for $k \in \{k_i^{r,\min}, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor\}$. To simplify the notations, it is convenient to reverse the x-axis and shift the x-coordinate by $k\cdot\Delta t$ which leads us to equivalently look at the extreme points of the upper convex hull of:

$$\mathcal{C}_k = \{(l\cdot\Delta t, u_j^{\Delta t}(l\cdot\Delta t)),\ l = k-L', \cdots, k-L\},$$

for $k \in \{k_i^{r,\min}, \cdots, \left\lfloor \frac{T}{\Delta t} \right\rfloor\}$. There is a one-to-one mapping between the extreme points of these two sets which consists in applying the reverse transformation. For any k, $\hat{\mathcal{C}}_k$ denotes the upper convex hull of $\mathcal{C}_k$. Note that $\hat{\mathcal{C}}_k$ is a convex set and has a finitely many extreme points, all of which are in $\mathcal{C}_k$. Since the values $(u_j^{\Delta t}(l \cdot \Delta t))_{l=k_j^{r,\min},\cdots,\lfloor \frac{T}{\Delta t} \rfloor}$ become sequentially available in ascending order of l by chunks of size $\left\lfloor \frac{\delta^{\inf}}{\Delta t} \right\rfloor$ as the label-setting algorithm progresses, a search for the extreme points of $\hat{\mathcal{C}}_{k+1}$ begins upon identification of the extreme points of $\hat{\mathcal{C}}_k$. Observe that $\hat{\mathcal{C}}_k$ updates to $\hat{\mathcal{C}}_{k+1}$ by removing the leftmost point $((k-L') \cdot \Delta t, u_j^{\Delta t}((k-L') \cdot \Delta t))$ and appending $((k+1-L) \cdot \Delta t, u_j^{\Delta t}((k+1-L) \cdot \Delta t))$ to the right, see Figure 2-3 for an illustration. In this process, deleting a point is arguably the most challenging operation because it might turn a formerly non-extreme point into one, see Figure 2-3b where this happens to be the case for the third leftmost point. In contrast, inserting a new point can only turn a formerly extreme point into a non-extreme one. Hence, deletions require us to do some bookkeeping other than simply keeping track of the extreme points of $\hat{\mathcal{C}}_k$ as k increases.

Maintaining the extreme points of a dynamically changing set is a well-studied class of problems in computational geometry known as *Dynamic Convex Hull* problems. Specific instances from this class differ along the operations to be performed on the set (e.g. insertions, deletions), the queries to be answered on the extreme points, and the dimensionality of the input data. The authors of [30] design a data structure maintaining a description of the upper convex hull of a finite set of N points in $\mathbb{R}^2$. This data structure satisfies Assumption 2.5 as it allows to insert points, to delete points, and to perform binary search on the first coordinate of the extreme points, all in amortized time $O(\log(N))$ and with $O(N)$ space usage. For the purpose of being self-contained, we design our own data structure in the Appendix to tackle the particular dynamic convex hull problem at hand. Our approach is based on Andrew's monotone chain convex hull algorithm, see [9], and only uses two arrays and a stack. The data structure developed in [30] is more complex than ours but can handle arbitrary dynamic convex hull problems.

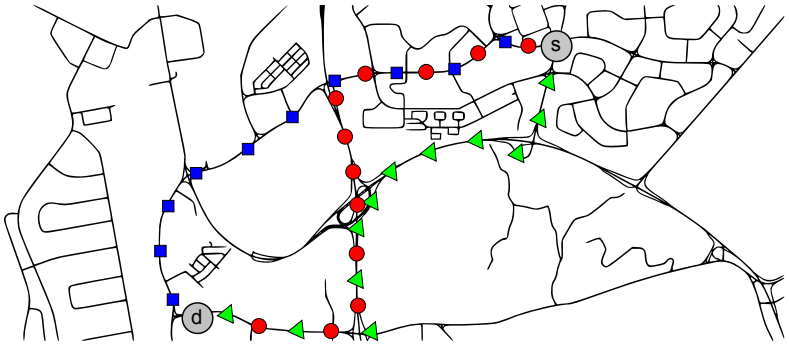


Figure 2-4: Local map. s and d locate the departure and arrival nodes. Three paths are highlighted. The left one (blue) is 5.3-km long and takes 9 minutes to travel. The middle one (red) is 6.4-km long and takes 8 minutes to travel. The rightmost one (green) is 6.1-km long and takes 10 minutes to travel.

2.5 Numerical Experiments

One of the most common applications of SSPs deals with the problem of routing vehicles in transportation networks. Providing driving itineraries is a challenging task as suppliers have to cope simultaneously with limited knowledge about random fluctuations in traffic congestion (e.g. caused by traffic incidents, variability of travel demand) and users' desire to arrive on time. In this section, we compare, using a real-world application with field data from the Singapore road network, the performance of the nominal and robust approaches to vehicle routing when traffic measurements are scarce and uncertain. To benchmark the performance of the robust approach, we propose a realistic framework where both the nominal and robust approaches can be efficiently computed and for which it is up to the user to pick one.

2.5.1 Framework

We work on a network composed of the main roads of Singapore with 20,221 arcs and 11,018 nodes for a total length of 1131 kilometers of roads. The data consists of a 15-day recording of GPS probe vehicle speed samples coming from a combined fleet of over 15,000 taxis. Features of each recording include current location, speed and status (free, waiting for a customer, occupied). We denote by s and d the departure and arrival nodes. Because there is usually only one reasonable route to get from s to d for most pairs (s, d)

in our network, the benefits of using one vehicle routing approach over another would not be apparent if we were to pick (s, d) uniformly at random over $\mathcal{V}^2$. Instead, we choose to hand-pick a pair (s, d) with at least two reasonable routes to get from s to d with similar travel times so that the best driving itinerary depends on the actual traffic conditions. We choose $s =$ "Woodlands avenue 2" and $d =$ "Mandai link", see Figure 2-4, but the results would be similar for other pairs satisfying this property.

Method of performance evaluation. Consider the following real-world situation. A user has to find an itinerary to get from s to d within a given budget T (the deadline) and with an objective to maximize the probability of on-time arrival, but when only a few vehicle speed samples are available in order to assess arc travel time uncertainty.

To model this real-world situation, we assume that the full set of samples of vehicle speed measurement available in our dataset in fact represents the real traffic conditions, characterized by the corresponding travel-time distributions p_{ij}^{real}'s, which are obtained from the full set of samples. Mimicking the fact that the p_{ij}^{real}'s are actually not fully available, we then consider the case where only a fraction of the full set of samples, say $\lambda \in [0, 1]$, is available. Based on this limited data, the challenge is to select an itinerary with a probability of on-time arrival with respect to the real traffic conditions p_{ij}^{real}'s as high as possible. We propose to use the methods listed in Table 2.2 to choose such an itinerary. For each of these methods, the process goes as follows:

1. Estimate the arc-based travel-time parameters required to run the method using the fraction of data available.

2. Run the corresponding algorithm to find an itinerary, depending on the chosen method.

3. Compute the probability of on-time arrival of this itinerary for the **real** traffic conditions ($\lambda = 1$).

The result obtained depends on both λ and the available samples as there are many ways to pick a fraction λ out of the entire dataset. Hence, for each λ in a set Λ, we randomly pick $\lambda \cdot N_{ij}$ samples for each arc (i, j), where N_{ij} is the number of samples collected in

the entire dataset for that particular arc. For each $\lambda \in \Lambda$, and for each method, we store the calculated probability of on-time arrival. We repeat this procedure 100 times.

Table 2.2: Methods considered. I_{ij}^{m} and I_{ij}^{md} are confidence intervals.

Method	Travel-time parameters to estimate from samples
RobustM	$\delta_{ij}^{\inf}$, $\delta_{ij}^{\sup}$, I_{ij}^{m}
RobustMD	$\delta_{ij}^{\inf}$, $\delta_{ij}^{\sup}$, I_{ij}^{m}, I_{ij}^{md}, $m_{ij} = \frac{\max(I_{ij}^{\mathrm{m}})+\min(I_{ij}^{\mathrm{m}})}{2}$
Empirical	empirical distributions p_{ij}
LET	empirical mean m_{ij}

Method	Approach
RobustM	(2.3) with $\mathcal{P}_{ij} = \{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \mathbb{E}_{X\sim p}[X] \in I_{ij}^{\mathrm{m}}\}$
RobustMD	(2.3) with $\mathcal{P}_{ij} = \{p \in \mathcal{P}([\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]) : \begin{array}{c}\mathbb{E}_{X\sim p}[X] \in I_{ij}^{\mathrm{m}} \\ \mathbb{E}_{X\sim p}[\lvert X - m_{ij}\rvert] \in I_{ij}^{\mathrm{md}}\end{array}\}$
Empirical	(2.1) with p_{ij}
LET	deterministic shortest path with the arc costs c_{ij} taken as m_{ij}

A few remarks are in order. We choose $\Lambda = \{0.001, 0.002, 0.005\}$, this corresponds to an average number of samples per arc of $[5.5, 9.4, 25.1]$ respectively (we take at least one sample per arc). The average arc length is 163 meters, hence we set $\Delta t = 0.02$ second to get a good accuracy. This parameter has a significant impact on the running time and it could also be optimized. We include the Least Expected Time (LET) method as it is a reasonably robust approach, although not tailored to the risk function considered, and because it is very fast to solve. The confidence intervals used by the robust approaches are percentile bootstrap 95 % confidence intervals derived from resampling the available data with replacement. When solving the discretization schemes (2.6) and (2.11), ties in the argument of the maximum are broken in favor of the (estimated) least expected travel time to the destination. To solve the robust problems, we use the column generation scheme and the special-purpose procedure described in Section 2.4.3.c while we use the scheme based on fast Fourier transforms described in Section 2.3.2.b for the nominal approach.

2.5.2 Results

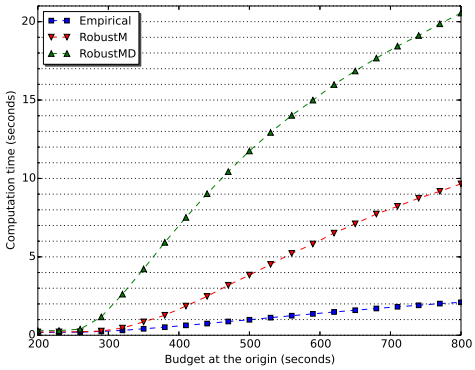
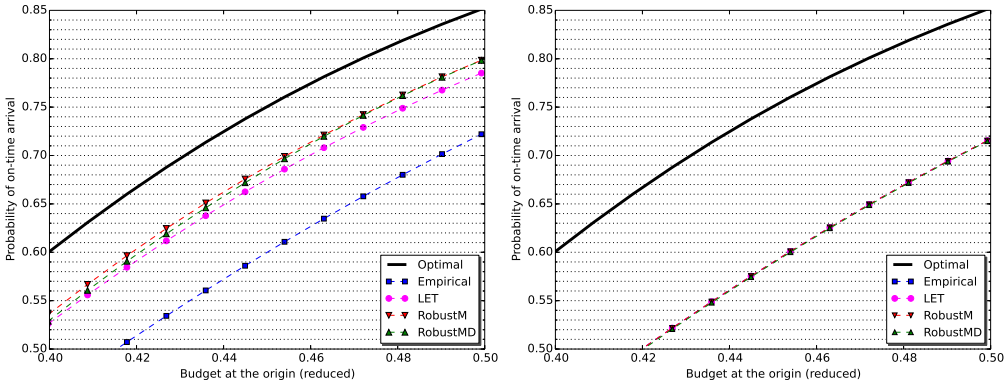


Figure 2-5: Average computation time as a function of the time budget for $\lambda = 0.001$.

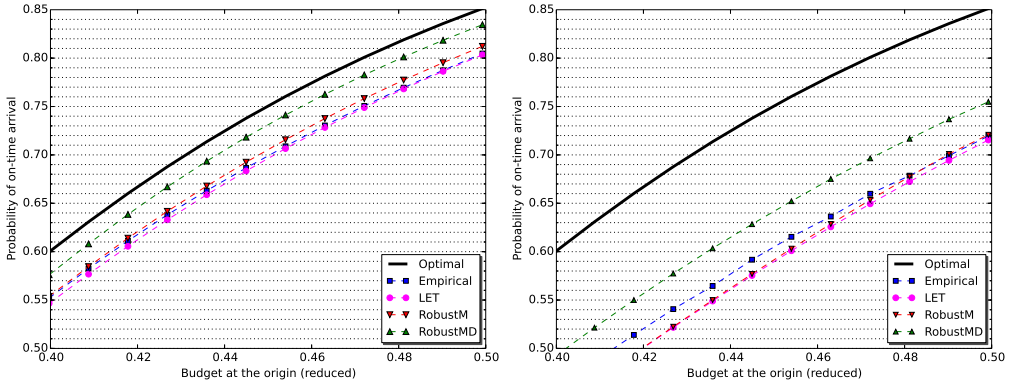


(a) Average probability of on-time arrival.

(b) 5% worst-case probability of on-time arrival.

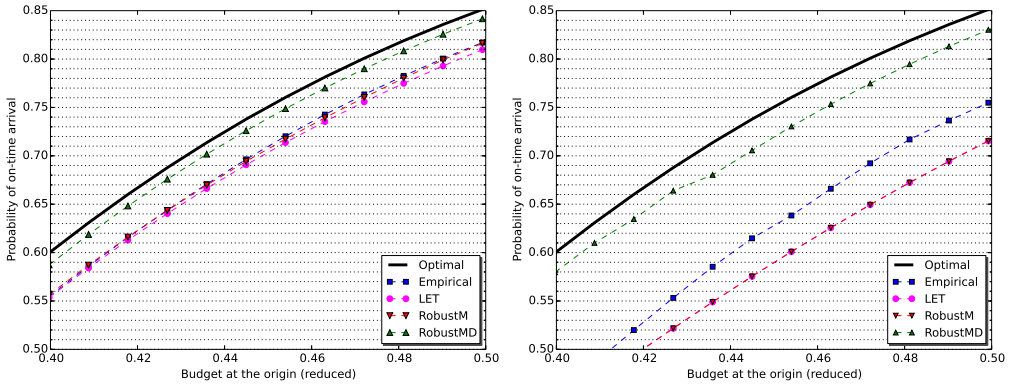
Figure 2-6: $\lambda = 0.001$, average number of samples per link: ~ 5.5.

The results are plotted in Figure 2-6, 2-7, and 2-8. Each of these figures corresponds to one of the fraction $\lambda \in \Lambda$ so as to see the impact of an increasing knowledge. The time budget is "normalized": 0 (resp. 1) corresponds to the minimum (resp. maximum) amount of time it takes to reach d from s. For each λ, for each method in Table 2.2, for each time budget T, and for each of the 100 simulations, we compute the actual probability of on-time arrival of the corresponding strategy. The average (resp. 5 % worst-case) probability of on-time arrival over the simulations is plotted on the figures labeled "a" (resp. "b"). The 5 % worst-case measure, which corresponds to the average over the 5 simulations out 100 that yield the lowest probability of arriving on-time, is particularly relevant as commuters opting for



(a) Average probability of on-time arrival. (b) 5% worst-case probability of on-time arrival.

Figure 2-7: $\lambda = 0.002$, average number of samples per link: ~ 9.4.



(a) Average probability of on-time arrival. (b) 5% worst-case probability of on-time arrival.

Figure 2-8: $\lambda = 0.005$, average number of samples per link: ~ 25.1.

this risk function would expect the approach to have good results even under bad scenarios. We also plot the average runtime for each of the method as a function of the time budget in Figure 2-5.

Conclusions. As can be observed on the figures, Empirical is not competitive when only a few samples are available. To be specific, RobustM slightly outperforms the other methods when there are very few measurements, see Figure 2-6, while RobustMD is a clear winner when more samples are available, in terms of both average and worst-case performances, see Figures 2-7 and 2-8. Observe that, as expected, the performance of Empirical improves as more samples get available and Empirical eventually outperforms RobustM,

see Figure 2-8. Our interpretation of these results is that relying on quantities, either moments or distributions, that cannot be accurately estimated may be misleading even for robust strategies. On the other hand, failure to capture the increasing knowledge on the actual travel-time probability distributions (e.g. by estimating more moments) as the amount of available data increases may lead to poor performances.

2.6 Extensions

In this section, we sketch how to extend the results derived in Sections 2.3 and 2.4 when either Assumption 2.1 or Assumption 2.2 is relaxed. Most of the results also extend when both assumptions are relaxed at the same time but we choose to discuss one assumption at a time to highlight their respective implications.

2.6.1 Relaxing Assumption 2.1: Markovian Costs

We consider here the case where the experienced costs of crossing arcs define a *Markov* chain of finite order m. To simplify the presentation, we provide in details the extensions of our previous results to the case $m = 1$. Adapting these extensions to a general m amounts to augmenting the state space of the underlying MDP by the costs of the last m visited arcs. We emphasize that while *Markov* chains can model the reality of the decision making process more accurately, this comes at a price: this requires an estimation of m-dimensional probability distributions, and the computational time needed to find an optimal strategy grows exponentially with m.

Extension for the nominal problem. A variant of Theorem 2.1 can be shown to hold if the arc cost distributions are discrete. Under this assumption and as soon as the total cost spent so far is larger than $T - T_f$, the optimal strategy coincides with the strategy of minimizing the expected costs, which may no longer be a shortest path but can still be shown to be a solution without cycles. Under the same assumption, Proposition 2.1 remains

valid under the following higher-dimensional dynamic program:

$$
\begin{aligned}
&u_d(t,z,\theta) = f(t) \quad \text{for } t \leq T, z \in \mathcal{A}(d), \theta \in \Theta_{zd} \\
&u_i(t,z,\theta) = \max_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}(\omega | z, \theta) \cdot u_j(t-\omega, i, \omega) \mathrm{d}\omega \\
&\text{for } i \neq d, t \leq T, z \in \mathcal{A}(i), \theta \in \Theta_{zi} \\
&\pi_f^*(i,t,z,\theta) \in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}(\omega | z, \theta) \cdot u_j(t-\omega, i, \omega) \mathrm{d}\omega \\
&\text{for } i \neq d, t \leq T, z \in \mathcal{A}(i), \theta \in \Theta_{zi},
\end{aligned}
\tag{2.18}
$$

where $\mathcal{A}(i)$ denotes the set of immediate antecedents of i in $\mathcal{G}$, Θ_{zi} is the finite set of possible values taken by c_{zi} for $z \in \mathcal{A}(i)$, and $p_{ij}(\cdot \mid z, \theta)$ is the conditional distribution of c_{ij} given that the last visited node is z and that $c_{zi} = \theta$. The discretization scheme of Section 2.3.2.a can be adapted for this new dynamic equation and the approximation guarantees carry over. To solve this new discretization scheme, the label-setting approach from Section 2.3.2.b can be adapted by observing that the functions $(u_i(\cdot, z, \theta))_{i \in \mathcal{V}, z \in \mathcal{A}(i), \theta \in \Theta_{zi}}$ can be computed block by block by interval increments of size $\delta^{\inf}$. However, the schemes based on fast Fourier transforms and the idea of zero-delay convolution do not apply anymore, and we need to use the pointwise definition of convolution products with computational complexity:

$$
O\Big(\max_{(i,j) \in \mathcal{A}} |\Theta_{ij}| \cdot \frac{|\mathcal{A}| \cdot (T - T_f) + |\mathcal{V}|^2 \cdot \delta^{\sup}}{\Delta t}\Big).
$$

Extension for the robust problem. For any $(i,j) \in \mathcal{A}$, $z \in \mathcal{A}(i)$, and $\theta \in \Theta_{zi}$, $p_{ij}(\cdot \mid z, \theta)$ is only known to lie in the ambiguity set $\mathcal{P}_{ij,z,\theta}$. If $\mathcal{P}_{ij,z,\theta}$ is only comprised of discrete distributions with finite support Θ_{ij}, a variant of Theorem 2.2 can be shown to hold. Specifically, as soon as the total cost spent so far is larger than $T - T_f^r$, the optimal strategy coincides with the strategy of minimizing the worst-case expected costs, which can also be shown not to cycle. Under this assumption, Proposition 2.3 remains valid under the

following higher-dimensional dynamic program:

$$
\begin{aligned}
&u_d(t,z,\theta) = f(t) \\
&\text{for } t \leq T, z \in \mathcal{A}(d), \theta \in \Theta_{zd} \\
&u_i(t,z,\theta) = \max_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij,z,\theta}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega, i, \omega) \mathrm{d}\omega \\
&\text{for } i \neq d, t \leq T, z \in \mathcal{A}(i), \theta \in \Theta_{zi} \\
&\pi^*_{f,\mathcal{P}}(i,t,z,\theta) \in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij,z,\theta}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega, i, \omega) \mathrm{d}\omega \\
&\text{for } i \neq d, t \leq T, z \in \mathcal{A}(i), \theta \in \Theta_{zi}.
\end{aligned}
\tag{2.19}
$$

The discretization scheme of Section 2.4.3.a can be adapted for this new set of equations and the approximation guarantees of Proposition 2.4 carry over. Moreover, the label-setting approach can also be adapted along the same lines as for the nominal problem. The ideas underlying the algorithmic developments of Section 2.4.3.c remain valid but we now have to recompute the convex hulls from scratch at each time step using Andrew's monotone chain convex hull algorithm, as opposed to using a dynamic convex hull algorithm, which leads to the computational complexity:

$$
O(\max_{(i,j)\in\mathcal{A}} |\Theta_{ij}| \cdot \log(\max_{(i,j)\in\mathcal{A}} |\Theta_{ij}|) \cdot \frac{|\mathcal{A}| \cdot (T - T_f^r) + |\mathcal{V}|^2 \cdot \delta^{\sup}}{\Delta t} \cdot \log(\frac{|\mathcal{V}| + \frac{T - T_f^r}{\delta^{\inf}}}{\epsilon})),
$$

when we want to compute an ϵ-approximate strategy solution to the discretization scheme (2.11).

2.6.2 Relaxing Assumption 2.2: τ-dependent Arc Cost Probability Distributions

Extension for the nominal problem. For any $\tau \geq 0$ and $(i,j) \in \mathcal{A}$, we denote by p_{ij}^τ the distribution of c_{ij}^τ and by m_{ij}^τ the mean of p_{ij}^τ. Theorem 2.1 remains valid if, for any $(i,j) \in \mathcal{A}$, m_{ij}^τ converges as $\tau \to \infty$, in which case the shortest-path tree mentioned in the statement is defined with respect to the limits of the mean arc costs. For instance, this assumption is satisfied when the distributions are time-varying during a peak period

and stationary anytime thereafter, see [64]. Under this assumption, Proposition 2.1 also remains valid but for the slightly modified dynamic program:

$$\begin{aligned} u_d(t) &= f(t) && t \leq T \\ u_i(t) &= \max_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}^{T-t}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T \\ \pi_f^*(i,t) &\in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \int_0^\infty p_{ij}^{T-t}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T. \end{aligned} \tag{2.20}$$

The discretization scheme of Section 2.3.2.a can be trivially adapted for this new dynamic equation, although we may loose the approximation guarantees provided by Proposition 2.2. For them to carry over, we need additional assumptions. To be specific, one of the following properties must be satisfied:

- the arc cost distributions vary smoothly, in the sense that, for any arc $(i,j) \in \mathcal{A}$, there exists K such that the Kolmogorov distance between $p_{ij}^{\tau_1}$ and $p_{ij}^{\tau_2}$ is smaller than $K \cdot |\tau_1 - \tau_2|$ for any $\tau_1, \tau_2 \geq 0$,
- the arc cost distributions are discrete and the discretization length Δt is chosen appropriately,
- the arc cost distributions change finitely many times and the discretization length Δt is chosen appropriately.

To solve the discretization scheme, the label-setting approach described in Section 2.3.2.b remains relevant but we now have to apply the pointwise definition of convolution products, as opposed to using fast Fourier transforms and zero-delay convolutions, with computational complexity quadratic in $\frac{1}{\Delta t}$:

$$O\Big(\frac{|\mathcal{A}| \cdot (T - T_f) + |\mathcal{V}|^2 \cdot \delta^{\sup}}{\Delta t} \cdot \frac{\delta^{\sup} - \delta^{\inf}}{\Delta t}\Big).$$

Extension for the robust problem. For any $\tau \geq 0$ and $(i,j) \in \mathcal{A}$, p_{ij}^τ is only known to lie in the ambiguity set $\mathcal{P}_{ij}^\tau$. First observe that (2.3) turns into:

$$\sup_{\pi \in \Pi} \inf_{\forall \tau, \forall (i,j) \in \mathcal{A},\, p_{ij}^\tau \in \mathcal{P}_{ij}^\tau} \mathbb{E}_{\mathbf{p}^\tau}[f(T - X_\pi)],$$

which is exactly the robust counterpart of (2.1), as opposed to a conservative approximation when the arc cost distributions are stationary. Theorem 2.2 remains valid if, for any $(i,j) \in \mathcal{A}$, $\max_{p_{ij} \in \mathcal{P}_{ij}^{\tau}} \mathbb{E}_{X \sim p_{ij}}[X]$ converges as $\tau \to \infty$, in which case the shortest-path tree mentioned in the statement is defined with respect to the limits. Again, this assumption is, for instance, satisfied when the ambiguity sets are time-varying during a peak period and stationary anytime thereafter. Under this assumption, Proposition 2.3 also remains valid but for the slightly modified dynamic program:

$$\begin{aligned} u_d(t) &= f(t) && t \leq T \\ u_i(t) &= \max_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}^{T-t}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T \\ \pi^*_{f,\mathcal{P}}(i,t) &\in \operatorname*{argmax}_{j \in \mathcal{V}(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}^{T-t}} \int_0^\infty p_{ij}(\omega) \cdot u_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T. \end{aligned} \tag{2.21}$$

Similarly as for the nominal problem, the discretization scheme can be trivially adapted but we may loose the approximation guarantees provided by Proposition 2.4. For them to carry over, one of the following properties has to be satisfied:

- the ambiguity sets vary smoothly, in the sense that, for any arc $(i,j) \in \mathcal{A}$, there exists K such that the Kolmogorov distance between $\mathcal{P}_{ij}^{\tau_1}$ and $\mathcal{P}_{ij}^{\tau_2}$ is smaller than $K \cdot |\tau_1 - \tau_2|$ for any $\tau_1, \tau_2 \geq 0$,
- the ambiguity sets are only comprised of discrete distributions and the discretization length Δt is chosen appropriately,
- the ambiguity sets change finitely many times and the discretization length Δt is chosen appropriately.

In contrast to the nominal problem, all the algorithms developed in Section 2.4.3.c can still be used to solve the discretization sheme with the same computational complexity as long as the ambiguity sets are defined by confidence intervals on piecewise affine statistics, as precisely defined in Section 2.4.3.b.

Chapter 3

No-Regret Learnability for Piecewise Linear Losses

3.1 Introduction

Online convex optimization has emerged as a popular approach to online learning, bringing together convex optimization methods to tackle problems where repeated decisions need to be made in an unknown, possibly adversarial, environment. A full-information online convex optimization problem is a repeated zero-sum game between a learner (the player) and the environment (the opponent). There are T time periods. At each round t, the player has to choose a point f_t in a convex set $\mathcal{F}$. Subsequent to the choice of f_t, the environment reveals a point $z_t \in \mathcal{Z}$ and the loss incurred to the player is $\ell(z_t, f_t)$, for a loss function ℓ that is convex in its second argument. Both players are aware of all the parameters of the game, namely ℓ, $\mathcal{Z}$, and $\mathcal{F}$, prior to starting the game. Additionally, at the end of each period, the opponent's move is revealed to the player. The performance of the player is measured in terms of a quantity coined regret, defined as the gap between the accumulated losses incurred by the player and the best performance he could have achieved in hindsight with a non-adaptive strategy:

$$r_T((z_t)_{t=1,\cdots,T},(f_t)_{t=1,\cdots,T}) = \sum_{t=1}^{T} \ell(z_t, f_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(z_t, f).$$

In this field, one of the primary focus is to design algorithms, i.e. strategies to select $(f_t)_{t=1,\cdots,T}$ so as to keep the regret as small as possible even when facing an adversarial opponent. Particular emphasis is placed on how the regret scales with T because this dependence relates to a notion of learning rate. If $r_T = o(T)$, the player is, in some sense, learning the game in the long-run since the gap between experienced and best achievable average cumulative payoffs vanishes as $T \to \infty$. Furthermore, the smaller the growth rate of r_T, the faster the learning. A natural question to ask is what is the best learning rate that can be achieved for a given game $(\ell, \mathcal{Z}, \mathcal{F})$. Mathematically, this is equivalent to characterizing the growth rate of the smallest regret that can be achieved by a player against a completely adversarial opponent, expressed as:

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathcal{Z}} \cdots \inf_{f_T \in \mathcal{F}} \sup_{z_T \in \mathcal{Z}} [\sum_{t=1}^{T} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(z_t, f)],$$

which we refer to as the value of the game $(\ell, \mathcal{Z}, \mathcal{F})$. Aside from pure learning considerations, the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$ has important consequences in a variety of fields where no-regret algorithms are used to compute complex quantities, e.g. Nash equilibria in Game Theory [70] or solutions to optimization problems in convex optimization [42], in which case this growth rate translates into the number of iterations required to compute the quantity with a given precision. We investigate this question in a systematic fashion by looking at the interplay between $\mathcal{F}$ and $\mathcal{Z}$ for the following class of piecewise linear loss functions:

$$\ell(z, f) = \max_{x \in \mathcal{X}(z)} (C(z)f + c(z))^{\intercal} x, \tag{3.1}$$

where, for any $z \in \mathcal{Z}$, $C(z)$ is a matrix, $c(z)$ is a vector, and $\mathcal{X}(z) \subset \mathbb{R}^d$ is either a finite set or a polyhedron $\{x \in \mathbb{R}^d \mid A(z)x \leq b(z)\}$ with $A(z)$ a matrix and $b(z)$ a vector. This type of loss functions arises in a number of important contexts such as online linear optimization, repeated zero-sum Stackelberg games, online prediction with side information, and online two-stage optimization, as illustrated in Section 3.1.1. Throughout the chapter, we make the following standard assumption so as to have the game well-defined.

Assumption 3.1. *$\mathcal{Z}$ is a non-empty compact subset of $\mathbb{R}^{d_z}$ and $\mathcal{F}$ is a non-empty, convex,*

and compact subset of $\mathbb{R}^{d_f}$. For any choice of $z \in \mathcal{Z}$, the set $\mathcal{X}(z)$ is not empty. The loss function ℓ is bounded on $\mathcal{Z} \times \mathcal{F}$. Moreover, either $\mathcal{Z}$ has finite cardinality or $\ell(\cdot, f)$ is continuous for any $f \in \mathcal{F}$.

Contributions A number of no-regret algorithms developed in the literature can be used as black boxes for the setting considered in this chapter in order to get $O(\sqrt{T})$ bounds on regret, e.g. Exponential Weights [95], Online Gradient Descent [106], and more generally Online Mirror Descent [48], to cite a few. To get better learning rates, other approaches have been proposed but they usually rely on either the curvature of ℓ, for instance if ℓ is strongly convex in its second argument [46], which is not the case here, or more information about the sequence $(z_t)_{t=1,\cdots,T}$, see for example [79], but which is not available in the fully adversarial setting. Aside from particular instances, e.g. [33] and [3], it is in general unknown how the interplay between ℓ, $\mathcal{Z}$, and $\mathcal{F}$ determines the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$. The main insight from this chapter is that the curvature of the decision maker's set of moves $\mathcal{F}$ is a determining factor for the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$: if $\mathcal{F}$ is a polyhedron then the decision maker is doomed to a rate of $\Theta(\sqrt{T})$, otherwise, if it is curved, the rate can be exponentially smaller. Specifically, we show that:

1. When $\mathcal{F}$ is a polyhedron, either $R_T(\ell, \mathcal{Z}, \mathcal{F}) = 0$ or $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$. To the best of our knowledge, this result constitutes the first systematic $\Omega(\sqrt{T})$ lower bound on regret obtained for a large class of piecewise linear loss functions. This lower bound applies to online combinatorial optimization, also to many experts settings and repeated zero-sum Stackelberg games where the player resorts to a randomized strategy, as well as to many online prediction problems with side information and online two-stage optimization problems.

2. When (i) ℓ is linear, (ii) $\mathcal{F} = \{f \in \mathbb{R}^{d_f} \mid F(f) \leq 0\}$, for F either a strongly convex function or $F(f) = \|f\|_{\mathcal{F}} - C$ where $C \geq 0$ and $\| \|_{\mathcal{F}}$ is a $q-$uniformly convex norm with $q \in [2, 3)$, and (iii) 0 does not lie in the convex hull of $\mathcal{Z}$, we have $R_T(\ell, \mathcal{Z}, \mathcal{F}) = o(\sqrt{T})$, achieved by the Follow-The-Leader algorithm [56]. This result applies to repeated zero-sum games where the player picks a cost vector

Table 3.1: Growth rate of R_T in several settings of interest.

$\ell(z, f)$	$\mathcal{F}$	conditions on $\mathcal{Z}$	R_T
(3.1)	any polyhedron		0 or $\Theta(\sqrt{T})$
$z^\intercal f$	any convex set	$0 \in \text{int}(\text{conv}(\mathcal{Z}))$	0 or $\Theta(\sqrt{T})$
$z^\intercal f$	$B_p(0,1)$ for $p \in (1,3)$	$0 \notin \text{conv}(\mathcal{Z})$	$\begin{cases} O(\log(T)) & \text{if } p \in (1,2) \\ 0 \text{ or } \Theta(\log(T)) & \text{if } p = 2 \\ O(T^{\frac{p-2}{p-1}}) & \text{if } p \in (2,3) \end{cases}$.

(e.g. arc costs) of bounded Euclidean norm and the opponent chooses an element in a combinatorial set (e.g. a path). This also applies to non-linear loss functions when 0 does not lie in the convex hull of the set of subgradients of ℓ with respect to the second-coordinate by a standard reduction to linear loss functions, see [106]. Note that assumption (iii) is required to get $o(\sqrt{T})$ rates as $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$ for linear losses when 0 lies in the interior of the convex hull of $\mathcal{Z}$, see Section 3.2. Also note that, as a byproduct of our results, one can combine the Follow-The-Leader algorithm with Online Gradient Descent using the strategy developed in [86] to get $o(\sqrt{T})$ regret if $0 \notin \text{conv}(\mathcal{Z})$ and the near-optimal rate $O(\sqrt{T \log(T)})$ if $0 \in \text{conv}(\mathcal{Z})$ when (i) and (ii) hold but it is unknown whether $0 \notin \text{conv}(\mathcal{Z})$.

The most interesting cases are summarized in Table 3.1.

Notations For a set $S \subset \mathbb{R}^d$, $\text{conv}(S)$ (resp. $\text{int}(S)$) refers to the convex hull (resp interior) of this set. When S is compact, we define $\mathcal{P}(S)$ as the set of probability measures on S. For $x \in \mathbb{R}^d$, $\|x\|$ refers to the L_2-norm of x while $B_p(x, \epsilon)$ denotes the closed L_p ball centered at x with width ϵ. For a collection of random variables $(Z_1, \cdots, Z_t)$, $\sigma(Z_1, \cdots, Z_t)$ refers to the sigma-field generated by $Z_1, \cdots, Z_t$. For a random variable Z and a probability distribution p, we write $Z \sim p$ if Z is distributed according to p.

3.1.1 Applications

We list examples of situations where losses of the type (3.1) arise.

Online Linear Optimization In this setting, the loss function is given by $\ell(z, f) = z^\intercal f$. In particular, this includes:

- online combinatorial optimization where the opponent picks a cost in $[0, 1]^{d_z}$ and $\mathcal{F}$ is defined as the convex hull of a finite set of elements (e.g. paths, spanning trees, and matchings),
- experts settings where the player picks a distribution over the experts' advice (in which case $\mathcal{F}$ is also a polyhedron) and the opponent reveals a cost for each of the experts.

In online linear optimization, lower bounds on regret are often derived by introducing a randomized zero-mean i.i.d. opponent, see [3]. However, this is possible only if 0 is in the interior of the convex hull of $\mathcal{Z}$, which is typically not the case in online combinatorial optimization. A general feature of online linear optimization that will turn out to be important in the analysis is that there is no loss of generality in assuming that $\mathcal{Z}$ is a convex set in the following sense.

Lemma 3.1. *When $\ell(z, f) = z^\intercal f$, the games $(\ell, \mathcal{Z}, \mathcal{F})$ and $(\ell, \mathrm{conv}(\mathcal{Z}), \mathcal{F})$ are equivalent, i.e.:*

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) = R_T(\ell, \mathrm{conv}(\mathcal{Z}), \mathcal{F}).$$

Repeated Zero-Sum Stackelberg Games A repeated zero-sum Stackelberg game is a repeated zero-sum game with the particularity that one of the players, referred to as the leader, has to commit first to a randomized strategy f without knowing which of the N other players, indexed by z, he is going to face at the next round. The interaction between the leader and player $z \in \{1, \cdots, N\}$ is captured by a payoff matrix $M(z)$. Once the leader is set on a strategy, the identity of the other player is revealed and the latter "best-responds" to the leader's strategy, leading to the following expression for the loss function:

$$\ell(z, f) = \max_{i=1,\cdots,I_z} e_i^\intercal M(z) f,$$

where I_z is the number of possible moves for player z. We illustrate the modeling power of this framework with a network security problem that has applications in urban network

security [53] and fare evasion prevention in transit networks [54]. Consider a directed graph $G = (V, E)$. The leader has a limited number of patrols that can be assigned to arcs in order to intercept the attackers. A configuration $\gamma \in \Gamma$ corresponds to a valid assignment of patrols to arcs and is represented by a vector $(Y_{ij}^{\gamma})_{(i,j)\in E}$ with $Y_{ij}^{\gamma} = 1$ if a patrol is assigned to arc (i, j) and $Y_{ij}^{\gamma} = 0$ otherwise. The leader chooses a mixed strategy f over the set of feasible allocations. Attacker $z \in \{(i_1, j_1), \cdots, (i_N, j_N)\}$ wants to go from z_1 to z_2 while minimizing the probability of being intercepted. This interaction is captured by the loss function:

$$\ell(z, f) = \max_{x\in\mathcal{X}(z)} \sum_{\gamma\in\Gamma} -f_{\gamma}x_{\gamma},$$

with $\mathcal{X}(z) = \{(\max_{(i,j)\in E} X_{ij}^{\pi}Y_{ij}^{\gamma})_{\gamma\in\Gamma} \mid \pi \in \Pi(z)\}$, where $\Pi(z)$ is the set of directed paths joining z_1 to z_2 in G and $X_{ij}^{\pi} = 1$ if $(i, j) \in \pi$ and $X_{ij}^{\pi} = 0$ otherwise. The presentation of repeated Stackelberg games given here follows the model introduced in [18] for general, i.e. not necessarily zero-sum, Stackelberg security games. In this more general setting, the loss function may not be convex and a possible approach, see [18], is to add another layer of randomization which casts the problem back into the realm of online linear optimization.

Online Prediction with Side Information This setting has a slightly different flavor as the opponent first provides some side information x, then the player gets to pick $f \in \mathcal{F}$ and, finally, the opponent reveals the correct prediction y. Nonetheless, the lower bounds established here also apply to this setting through a reduction to the setting without side information, as detailed at the end of Section 3.2. In the standard linear binary prediction problem, where $\mathcal{F}$ is a L_2 ball, $y \in \{-1, 1\}$, and x lies in a L_2 ball, loss functions of the form (3.1) are commonly used, e.g. the absolute loss $\ell((x, y), f) = |y - x^{\mathsf{T}}f|$ and the hinge loss $\ell((x, y), f) = \max(0, 1 - yx^{\mathsf{T}}f)$. This is also true for linear multiclass prediction problems with the multiclass hinge loss:

$$\ell((x, y), f) = \max_{j=1,\cdots,N}(1\{j \neq y\} + f_j^{\mathsf{T}}x - f_y^{\mathsf{T}}x),$$

where N denotes the number of classes, $y \in \{1, \cdots, N\}$, and f is the vector obtained by concatenation of the vectors $f_1, \cdots, f_N$. In the online approach to collaborative filtering,

a typical loss function is $\ell((i,j,y),M) = |M(i,j) - y|$ where M is a (user, item) matrix with bounded trace norm, (i,j) is a (user, item) pair, and y is the rating of item j by user i.

Online Two-Stage Optimization This setting captures situations where the decision making process can be broken down into two consecutive stages. In the first stage, the player makes a decision represented by $f \in \mathcal{F}$. Subsequently, the opponent discloses some information $z \in \mathcal{Z}$, e.g. a demand vector, and then the player chooses another decision vector x in the second stage, taking into account this newly available information to optimize the objective function. The loss function takes on the following form:

$$\ell(z,f) = c_1^\intercal f + \min_{\substack{x \in \mathbb{R}^d \\ Af + Bx \leq z}} c_2^\intercal x,$$

where c_1 and c_2 are cost vectors and A and B are matrices. Using strong duality, this loss function can be expressed in the canonical form (3.1). This framework finds applications in the operation of power grids, where z represents the demand in electricity or the availability of various energy sources. Since z is unknown when it is time to set up conventional generators, the decision maker has to adjust the production or buy additional capacity from a spot market to meed the demand, see for example [57].

Congestion Control We consider a variant of the congestion network game described in [27]. A decision maker has to decide how to ship a given set commodities through a network $G = (V,E)$. This decision can be equivalently represented by a flow vector f. Because the amount of commodities is assumed to be substantial, implementing f will cause congestion which will impact the other users of the network, represented by a flow vector z. The problem faced by the decision maker is to cause as little delay as possible to the other users with the additional difficulty that the traffic pattern z is not known ahead of time. Each arc $e \in E$ has an associated latency function that is convex in the flow on this arc:

$$c_e(f+z) = \max_{k=1,\cdots,K} (c_e^k \cdot (f_e + z_e) + s_e^k).$$

As a result, the total delay incurred to the other users can be expressed as:

$$\ell(z, f) = \sum_{e \in E} z_e \max_{k=1,\cdots,K} (c_e^k \cdot (f_e + z_e) + s_e^k).$$

3.1.2 Related Work

Asymptotically matching lower and upper bounds on $R_T(\ell, \mathcal{Z}, \mathcal{F})$ can be found in the literature for a variety of loss functions although the discussion tends to be restrictive as far as the decision sets $\mathcal{F}$ and $\mathcal{Z}$ are concerned. The value of the game is shown to be $\Theta(\log T)$ for three standard examples of curved loss functions. The first example, studied in [3], is the quadratic loss where $\ell(z, f) = z^\intercal f + \sigma \|f\|_2^2$ for $\sigma > 0$, with $\mathcal{Z}$ and $\mathcal{F}$ bounded L_2 balls. The second, studied in [96], is the online linear regression setting where the opponent plays $z = (x, y) \in \mathcal{Z} = B_\infty(0, 1) \times [-C_y, C_y]$ for $C_y > 0$ ($B_\infty(0, 1)$ denotes the unit L_∞ ball), the loss is $\ell((x, y), f) = (y - x^\intercal f)^2$, and $\mathcal{F}$ is an L_2 ball. The last one, from [71], is the log-loss $\ell(z, f) = -\log(z^\intercal f)$ with $\mathcal{Z}$ any compact set in $\mathbb{R}_+^d$ and $\mathcal{F}$ the simplex of dimension d. For non-curved losses, evidence suggests that the value of the game increases exponentially from $\Theta(\log(T))$ to $\Theta(\sqrt{T})$. Indeed, $\Omega(\sqrt{T})$ lower bounds are proved for several instances involving the absolute loss $\ell(z, f) = |z - f|$ in [32], typically with $\mathcal{Z} = \{0, 1\}$ and $\mathcal{F} = [0, 1]$. For purely linear loss functions, the authors in [3] establish a $\Omega(\sqrt{T})$ lower bound on $R_T(\ell, \mathcal{Z}, \mathcal{F})$ when $\mathcal{Z}$ is an L_2 ball centered at 0 and $\mathcal{F}$ is either an L_2 ball or a bounded rectangle. This result was later generalized in [2] and shown to hold for $\mathcal{F}$ any unit ball and $\mathcal{Z}$ its dual ball. The authors in [33] investigate the experts setting, i.e. $\mathcal{Z} = [0, 1]^d$ and $\mathcal{F}$ is the simplex of dimension d, and proves the same $\Omega(\sqrt{T})$ lower bound (which also holds if $\mathcal{Z}$ is the simplex of dimension d, see [2]). Lower bounds on regret of order $\Omega(\sqrt{T})$ are established in [82] when ℓ is the absolute loss for a prediction with side information setting more general than the one considered in this chapter where the player picks a function $f(\cdot)$, the opponent picks a pair (x, y), and the loss is $\ell((x, y), f(\cdot)) = |f(x) - y|$. The list of results listed above is far from being exhaustive but provides a good picture of the current state of the art. For each loss function, the intrinsic limitations of online algorithms are well understood, usually with the construction of a particular example of $\mathcal{F}$ and $\mathcal{Z}$ for which a lower bound on $R_T(\ell, \mathcal{Z}, \mathcal{F})$

asymptotically matches the best guarantee achieved by one of these algorithms. We aim at studying the value of the game in a more systematic fashion, using tools rooted in duality theory and sensitivity analysis. All the proofs are deferred to the Appendix.

3.2 Lower Bounds

Unless otherwise stated, we assume throughout this section that ℓ can be written in the form (3.1). We build on a powerful result rooted in von Neumann's minimax theorem that enables the derivation of tight lower and upper bounds on $R_T(\ell, \mathcal{Z}, \mathcal{F})$ by recasting the value of the game in a backward order.

Theorem 3.1. *From [2]*

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) = \sup_p \mathbb{E}\Big[\sum_{t=1}^{T} \inf_{f_t \in \mathcal{F}} \mathbb{E}[\ell(Z_t, f_t)|Z_1, \cdots, Z_{t-1}] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)\Big],$$

where the supremum is taken over the distribution p of the random variables $(Z_1, \cdots, Z_T)$ in $\mathcal{Z}^T$.

Any choice for p yields a lower bound on $R_T(\ell, \mathcal{Z}, \mathcal{F})$. The following result identifies a canonical choice for p that leads to $\Omega(\sqrt{T})$ lower bounds on regret.

Lemma 3.2. *Adapted from [2]*
If we can find a distribution p on $\mathcal{Z}$ and two points f_1 and f_2 in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)]$ such that $\ell(Z, f_1) \neq \ell(Z, f_2)$ with positive probability for $Z \sim p$, then $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$.

A distribution p satisfying the requirements of Lemma 3.2 can be viewed as an equalizing strategy for the opponent. This concept, formalized in [81], roughly refers to randomized strategies played by the opponent that cause the player's decisions to be completely irrelevant from a regret standpoint. These strategies are intrinsically hard to play against and often lead to tight lower bounds. To gain some intuition about this result, suppose that the opponent generates an independent copy of Z at each round t, which we denote by Z_t. In the adversarial setting considered in this chapter, the player is aware of the opponent's strategy but does not get to see the realization of Z_t before committing to a decision. For

this reason, at any round, f_1 and f_2 are optimal moves that are completely equivalent from the player's perspective. However, in hindsight, i.e. once all the realizations of the Z_t's have been revealed, f_1 and f_2 are typically not equivalent because $\ell(Z_t, f_1) \neq \ell(Z_t, f_2)$ with positive probability and one of these two moves will turn out to be

$$\max(0, \sum_{t=1}^{T} \ell(Z_t, f_1) - \ell(Z_t, f_2))$$

suboptimal which, in expectation, is of order $\Omega(\sqrt{T})$ by the central limit theorem. Given the conditions imposed on p, it is convenient to work with the following equivalence relation.

Definition 3.1. *We define the equivalence relation $\sim_\ell$ on $\mathcal{F}$ by $f_a \sim_\ell f_b$ for $f_a, f_b \in \mathcal{F}$ if and only if $\ell(z, f_a) = \ell(z, f_b)$ for all $z \in \mathcal{Z}$.*

In Theorem 3.2, we show that we can systematically, with the only exception of trivial games defined below, construct a distribution p satisfying the requirements of Lemma 3.2 whenever $\mathcal{F}$ is a polyhedron.

Definition 3.2. *The game $(\ell, \mathcal{Z}, \mathcal{F})$ is said to be trivial if and only if*

$$\exists f^* \in \mathcal{F} \text{ such that } \forall z \in \mathcal{Z},\ \ell(z, f^*) \leq \min_{f \in \mathcal{F}} \ell(z, f).$$

A simple example of a trivial game is $(\ell(z, f) = zf, [0, 1], [0, 1])$, where $\ell(z, f) \geq 0$ $\forall f \in [0, 1]$ and $\forall z \in [0, 1]$, with $\ell(z, f) = 0$ if $f = 0$ irrespective of z. If the game is trivial, the player will always play f^* irrespective of the time horizon and of the opponent's strategy observed so far to obtain zero regret. As it turns out, this uniquely identifies trivial games as we establish in Lemma 3.3.

Lemma 3.3. *For any $T \in \mathbb{N}$, $R_T(\ell, \mathcal{Z}, \mathcal{F}) \geq 0$. Moreover, in any of the following cases:*

1. *$\mathcal{Z}$ has finite cardinality,*
2. *$\ell(\cdot, f)$ is continuous for any choice of $f \in \mathcal{F}$,*

$R_T(\ell, \mathcal{Z}, \mathcal{F}) = 0$ if and only if the game is trivial.

The following result shows that, in most cases of interest, we can drastically restrict the power of the opponent while still preserving the nature of the game. This enables us to focus on the case where $\mathcal{Z}$ is finite.

Lemma 3.4. *Suppose that $\ell(\cdot, f)$ is continuous for any choice of $f \in \mathcal{F}$. If the game $(\ell, \mathcal{Z}, \mathcal{F})$ is not trivial, there exists a finite subset $\tilde{\mathcal{Z}} \subseteq \mathcal{Z}$ such that the game $(\ell, \tilde{\mathcal{Z}}, \mathcal{F})$ is not trivial.*

We are now ready to derive the $\Omega(\sqrt{T})$ lower bound on regret. The key idea behind the proof is that finding an equalizing strategy amounts to performing a sensitivity analysis for a well-chosen linear program.

Theorem 3.2. *Suppose that $\mathcal{F}$ is a polyhedron. In any of the following cases:*

1. *$\mathcal{Z}$ has finite cardinality,*

2. *$\ell(\cdot, f)$ is continuous for any choice of $f \in \mathcal{F}$,*

either the game is trivial or $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$.

Proof. We sketch the proof, see the Appendix for more details. Without loss of generality, we may assume that the game is not trivial, that $\mathcal{Z}$ is finite by Lemma 3.4, and that $\mathcal{X}(z)$ is finite for any $z \in \mathcal{Z}$ since otherwise, if $\mathcal{X}(z)$ is a polyhedron, the maximum in (3.1) must be attained at one of the finitely many extreme points of $\mathcal{X}(z)$. Write $\mathcal{Z} = \{z_n \mid 1 \leq n \leq N\}$. Without loss of generality, we can construct two probability distributions $(p_0(n))_{1 \leq n \leq N}$ and $(p_1(n))_{1 \leq n \leq N}$ such that: (i) there is a single equivalence class f^* (resp f^{**}) in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_0}[\ell(Z, f)]$ (resp. $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_1}[\ell(Z, f)]$), otherwise we are done by Lemma 3.2, and (ii) f^* and f^{**} do not belong to the same equivalence class (using the fact that the game is not trivial). We move on to show that there must exist $\alpha \in (0, 1)$ such that there are at least two equivalence classes in $\phi(\alpha) = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_\alpha}[\ell(Z, f)]$, where the distribution p_α is defined as $p_\alpha = (1 - \alpha)p_0 + \alpha p_1$. This will conclude the proof by Lemma 3.2. For any $f \in \mathcal{F}$, define $I(f) = \{\alpha \in [0, 1] \mid f \in \phi(\alpha)\}$. Since $\alpha \to \mathbb{E}_{p_\alpha}[\ell(Z, f)]$ is linear in α, $I(f)$ is a closed interval. Moreover, note that

$\min_{f\in\mathcal{F}} \mathbb{E}_{p_\alpha}[\ell(Z,f)]$ is equal to the optimal value of the optimization problem:

$$\begin{aligned}
\min_{q_1,\cdots,q_N,f} \quad & q^\intercal((1-\alpha)p_0+\alpha p_1) \\
\text{subject to} \quad & q=(q_1,\cdots,q_N) \\
& q_n \geq (C(z_n)f+c(z_n))^\intercal x,\ \forall x\in\mathcal{X}(z_n),\forall n \\
& f\in\mathcal{F}, q_1,\cdots,q_N\in\mathbb{R}.
\end{aligned}$$

Since $\mathcal{F}$ is a polyhedron and $\mathcal{X}(z_n)$ is finite for any n, this optimization problem is a linear program. Denoting by $\{f_1,\cdots,f_L\}$ the projections onto the f coordinate of the extreme points of the feasible set, there exists, for any $\alpha\in[0,1]$, $l\in\{1,\cdots,L\}$ such that $f_l\in\phi(\alpha)$. Hence, we can write $[0,1]=\cup_{l=1}^L I(f_l)$. We can further simplify this description by assuming that the f_l's belong to different equivalence classes (because $I(f)=I(f')$ if $f\sim_\ell f'$). Now observe that if $I(f_l)\cap I(f_j)\neq\emptyset$ for some $l\neq j\leq L$, then there are two equivalent classes in $\phi(\alpha)$ for any $\alpha\in I(f_l)\cap I(f_j)$ and we are done by Lemma 3.2. Suppose by contradiction that we cannot find such a pair of indices. Because the only way to partition $[0,1]$ into $L<\infty$ non-overlapping closed intervals is to have $L=1$, we get $[0,1]=I(f_1)$. This implies that $f^*\sim_\ell f^{**}$, a contradiction. □

An immediate consequence of Theorem 3.2 for linear games is the following:

Theorem 3.3. *Suppose that $\mathcal{F}$ is a polyhedron and that $\ell(z,f)=z^\intercal f$. Then, either the game is trivial or $R_T(\ell,\mathcal{Z},\mathcal{F})=\Omega(\sqrt{T})$.*

The proofs rely on Lemma 3.2 which is based on Theorem 3.1 and may, as a result, seem rather obscure. We stress that these lower bounds are derived by means of an equalizing strategy. We present this more intuitive view in the Appendix by exhibiting an equalizing strategy in the online linear optimization setting when $\mathrm{int}(\mathrm{conv}(\mathcal{Z}))\neq\emptyset$.

Note that Theorems 3.2 and 3.3 imply $\Omega(\sqrt{T})$ regret for a number of repeated Stackelberg games and online linear optimization problems as discussed in Section 3.1.1. Furthermore, we stress that Theorem 3.2 can also be used when $\mathcal{F}$ is not a polyhedron but this typically requires a preliminary step which boils down to restricting the opponent's decision set. For instance, the following well-known result is almost a direct consequence of Theorem 3.3.

Lemma 3.5. *Suppose that $\ell(z, f) = z^\mathsf{T} f$, that $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{Z}))$, and that $\mathcal{F}$ contains at least two elements. Then $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$.*

Note that Lemma 3.5 is consistent with Theorem 3.3 as the game $(\ell(z, f) = z^\mathsf{T} f, \mathcal{Z}, \mathcal{F})$ is non-trivial if $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{Z}))$ as soon as $\mathcal{F}$ contains at least two elements. Indeed $\ell(\epsilon \frac{f_2 - f_1}{\|f_2 - f_1\|}, f_2) > \ell(\epsilon \frac{f_2 - f_1}{\|f_2 - f_1\|}, f_1)$ and $\ell(\epsilon \frac{f_1 - f_2}{\|f_2 - f_1\|}, f_1) > \ell(\epsilon \frac{f_1 - f_2}{\|f_2 - f_1\|}, f_2)$, for a small enough $\epsilon > 0$ and any pair $f_1 \neq f_2 \in \mathcal{F}$. When $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{Z}))$, the opponent has some freedom to play, at each time period, a random vector with expected value zero, making every strategy available to the player equally bad. In other words, any i.i.d. zero-mean distribution is an equalizing strategy for the opponent in this case.

A preliminary step is also required to derive $\Omega(\sqrt{T})$ lower bounds on regret for prediction problems with side information where $\mathcal{F}$ is typically not a polyhedron. We sketch this simple argument for the canonical classification problem with the hinge loss, i.e. the game defined by the loss function:

$$\ell((x, y), f) = \max(0, 1 - y x^\mathsf{T} f)$$

along with the decision sets $\mathcal{Z} = B_2(0, 1) \times \{-1, 1\}$ and $\mathcal{F} = B_2(0, 1)$, but the method readily extends to any of the prediction problems described in Section 3.1.1. The idea is to restrict the opponent's decision set by taking a fixed vector x of norm 1 and to impose that, at any round t, the opponent's move is (x, y_t) for $y_t \in \{-1, 1\}$. Since $\ell((x, y), f)$ only depends on f through the scalar product between f and x, the player's decision set can be equivalently described by this value, which lies in $[-1, 1]$. Formally, we define a new loss function $\tilde{\ell}(y, f) = \max(0, 1 - yf)$ with $\tilde{\mathcal{Z}} = \{-1, 1\}$ and $\tilde{\mathcal{F}} = [-1, 1]$ and we have $R_T(\ell, \mathcal{Z}, \mathcal{F}) \geq R_T(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}})$. Observe now that the game $(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}})$ is not trivial, that $\tilde{\mathcal{Z}}$ is discrete, and that:

$$\tilde{\ell}(y, f) = \max_{\alpha \in \{0,1\}} (\alpha, -y\alpha)^\mathsf{T} (1, f).$$

We conclude with Theorem 3.2 that $R_T(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}}) = \Omega(\sqrt{T})$ which implies that:

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T}).$$

Remark about Lemma 3.2 We point out that, in general, it is not possible to weaken the assumptions of Lemma 3.2 (which, in fact, applies to a much more general class of loss functions than the one given by (3.1)). In particular, finding $z \in \mathcal{Z}$ such that there are two equivalence classes f_1 and f_2 in $\operatorname{argmin}_{f\in\mathcal{F}} \ell(z, f)$ does not guarantee that $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$ as we illustrate with a counterexample. This is because the result of Lemma 3.2 is intrinsically tied to the central limit theorem. Consider the following (non-trivial) online linear regression game:

$$(\ell(z, f) = (z^\intercal f)^2, \mathcal{Z} = B_2(z^*, 1), \mathcal{F} = [f_1, f_2]),$$

where $f_1 = (1, 0, \cdots, 0)$, $f_2 = (0, 1, 0, \cdots)$, and $z^* = (1, 1, 0, \cdots, 0)$. Observe that $\forall z \in \mathcal{Z}, \forall f \in \mathcal{F}, z^\intercal f \geq 0$. Hence $\operatorname{argmin}_{f\in\mathcal{F}} \ell(z, f) = \operatorname{argmin}_{f\in\mathcal{F}} f^\intercal z$. Furthermore, $\operatorname{argmin}_{f\in\mathcal{F}} f^\intercal z^* = [f_1, f_2]$ but f_1 and f_2 are clearly not in the same equivalence class. Yet, even though ℓ is not strongly convex in f, there exists an algorithm achieving $O(\log(T))$ regret, see [96].

3.3 Upper Bounds

Looking at Theorem 3.2, Theorem 3.3, and Lemma 3.5, it is tempting to conclude that the existence of pieces where ℓ is linear in its second argument dooms us to $\Omega(\sqrt{T})$ regret bounds. We now argue that the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$ is determined by a more involved interplay between ℓ, $\mathcal{Z}$, and $\mathcal{F}$ so that this assertion requires further examination. In fact, we show that $O(\log(T))$ regret bounds are even possible in online linear optimization. The fundamental reason is that the curvature of $\mathcal{F}$ can make up for the lack of thereof of ℓ. Curvature is key to enforce stability of the player's strategy with respect to perturbations in the opponent's moves. Sometimes, when the predictions are stable, e.g. when ℓ is the square loss $\ell(z, f) = \|z - f\|^2$, a very simple algorithm, known as Follow-The-Leader, yields $O(\log(T))$ regret.

Definition 3.3. *From [56]*

The Follow-The-Leader (FTL) strategy consists in playing:

$$f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{t-1} \sum_{\tau=1}^{t-1} \ell(z_\tau, f).$$

It is well known that FTL fails to yield sublinear regret for online linear optimization in general. However, when $\mathcal{F}$ is strongly curved and $0 \notin \operatorname{conv}(\mathcal{Z})$, the FTL strategy becomes stable, leading to $O(\log(T))$ regret as we next show using sensitivity analysis.

Theorem 3.4. *Suppose that (i) ℓ is linear, (ii) $\mathcal{F} = \{f \in \mathbb{R}^{d_f} \mid F(f) \leq 0\}$ for F a strongly convex and continuously differentiable function, and (iii) $0 \notin \operatorname{conv}(\mathcal{Z})$. Then, FTL yields $O(\log(T))$ regret.*

Proof. Without loss of generality, F is β-strongly convex and $\|\nabla F(f)\| \leq K, \forall f \in \mathcal{F}$. A common inequality on the regret incurred for the FTL strategy is:

$$r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) \leq \sum_{t=1}^{T} \ell(z_t, f_t) - \ell(z_t, f_{t+1}).$$

We use sensitivity analysis to control this last quantity. Specifically, we show that the mapping $\phi : z \to \operatorname{argmin}_{f \mid F(f) \leq 0} z^\mathsf{T} f$ is well-defined (i.e. the minimum is attained at a unique point) and Lipschitz on $\mathcal{Z}$. Using this property, we get:

$$\begin{aligned}
&r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) \\
&\quad \leq \sum_{t=1}^{T} \|z_t\| \, \|f_t - f_{t+1}\| \\
&\quad = O(\sum_{t=1}^{T} \left\| \frac{1}{t-1} \sum_{\tau=1}^{t-1} z_\tau - \frac{1}{t} \sum_{\tau=1}^{t} z_\tau \right\|) \\
&\quad = O(\sum_{t=1}^{T} \frac{1}{t(t-1)} \left\| \sum_{\tau=1}^{t-1} z_\tau \right\| + \frac{1}{t} \|z_t\|) \\
&\quad = O(\sum_{t=1}^{T} \frac{1}{t}) = O(\log(T)),
\end{aligned}$$

since $\mathcal{Z}$ is compact. We move on to show that ϕ is well-defined and Lipschitz. Since $0 \notin \operatorname{conv}(\mathcal{Z})$ and $\operatorname{conv}(\mathcal{Z})$ is closed and convex, there exists $C > 0$ such that $\|z\| \geq C, \forall z \in \mathcal{Z}$. Take $(z_1, z_2) \in \mathcal{Z}^2$ and $(f(z_1), f(z_2)) \in \operatorname{argmin}_{f \mid F(f) \leq 0} z_1^\mathsf{T} f \times \operatorname{argmin}_{f \mid F(f) \leq 0} z_2^\mathsf{T} f$.

Without loss of generality, we have $F(f(z_1)) = F(f(z_2)) = 0$ since the objective is linear. Hence, the constraint qualifications are automatically satisfied at $f(z_1)$ and $f(z_2)$ as ∇F cannot vanish on $\{f \mid F(f) = 0\}$ since F cannot attain its minimum on this set ($\mathcal{F}$ is assumed to contain at least two points). Hence, there exist $\lambda_1, \lambda_2 \geq 0$ such that $z_1 + \lambda_1 \nabla F(f(z_1)) = 0$ and $z_2 + \lambda_2 \nabla F(f(z_2)) = 0$. As $z_1, z_2 \neq 0$, we must have $\lambda_1, \lambda_2 \neq 0$. We obtain $\nabla F(f(z_1)) = -\frac{1}{\lambda_1} z_1$ and $\nabla F(f(z_2)) = -\frac{1}{\lambda_2} z_2$. Since F is β-strongly convex, we get:

$$(\frac{1}{\lambda_2} z_2 - \frac{1}{\lambda_1} z_1)^\intercal (f(z_1) - f(z_2)) \geq \beta \left\| f(z_1) - f(z_2) \right\|^2 .$$

We can break down the last expression in two pieces:

$$\begin{aligned} &\frac{1}{\lambda_2} z_2^\intercal (f(z_1) - f(z_2)) + \frac{1}{\lambda_1} z_1^\intercal (f(z_2) - f(z_1)) \\ &\quad \geq \beta \left\| f(z_1) - f(z_2) \right\|^2 . \end{aligned}$$

Observe that $z_2^\intercal (f(z_1) - f(z_2)) \geq 0$ by definition of $f(z_2)$ and $z_1^\intercal (f(z_2) - f(z_1)) \geq 0$ by definition of $f(z_1)$. Note that $\frac{1}{\lambda_1} = \frac{1}{|\lambda_1|} = \frac{\|\nabla F(f(z_1))\|}{\|z_1\|} \leq \frac{K}{C}$ and the same inequality holds for λ_2. Bringing everything together, we get:

$$\frac{K}{C} (z_2 - z_1)^\intercal (f(z_1) - f(z_2)) \geq \beta \| f(z_1) - f(z_2) \|^2 .$$

Using the Cauchy-Schwarz inequality and simplifying on both sides by $\|f(z_2) - f(z_1)\|$ yields:

$$\frac{K}{\beta C} \| z_2 - z_1 \| \geq \| f(z_1) - f(z_2) \| .$$

□

As an example of application of Theorem 3.4, consider a repeated network game where the player picks the arc costs in a L_2 ball, the opponent picks a path, and the loss incurred to the opponent is the sum of the arc costs along the path. In this setting, FTL yields $O(\log(T))$ regret even though the game is not trivial. Theorem 3.4 also has some implications for non-linear convex loss functions when $\mathcal{F}$ is curved and 0 is not in the convex hull of the set of subgradients of ℓ with respect to the player's moves (i.e. $0 \notin \operatorname{conv}(\{C(z)^\intercal x \mid x \in$

$\mathcal{X}(z), z \in \mathcal{Z}\}$) for the class of loss functions (3.1)). Indeed suppose that, at any time period t, the player follows the FTL strategy as if the loss function were linear and the past moves of the opponents were $y_1, \cdots, y_{t-1}$, i.e.:

$$f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{t-1} \sum_{\tau=1}^{t-1} y_\tau^\intercal f,$$

where, for any $\tau = 1, \cdots, t-1$, y_τ is a subgradient of $\ell(z_t, \cdot)$ at f_t. Then, for any sequence of moves $(z_1, \cdots, z_T)$, we have:

$$\begin{aligned} r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) &\leq \sum_{t=1}^{T} y_t^\intercal f_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} y_t^\intercal f \\ &= O(\log(T)). \end{aligned}$$

It is a priori unclear whether $\log(T)$ is the optimal growth rate for games satisfying the assumptions of Theorem 3.4. Quite surprisingly, i.i.d. opponents appear to be particularly weak for this kind of games, incurring at most a $O(1)$ lower bound on regret as shown in the following lemma. This is in stark contrast with the situations of Section 3.2 where the (tight) $\Omega(\sqrt{T})$ lower bounds are always derived through i.i.d. opponents.

Lemma 3.6. *Consider the game* $(\ell(z, f) = z^\intercal f, \mathcal{Z}, \mathcal{F})$ *with* $0 \notin \operatorname{conv}(\mathcal{Z})$ *and* $\mathcal{F} = B_2(0, 1)$. *Any lower bound derived from Theorem 3.1 with i.i.d. random variables* $Z_1, \cdots, Z_T \sim p$ *is* $O(1)$ *for any choice of* $p \in \mathcal{P}(\mathcal{Z})$.

The authors in [2] remark that restricting the study to i.i.d. sequences is in general not enough to get tight bounds for non-linear losses such as $\ell(z, f) = \|z - f\|^2$. It turns out that this is also true for the game studied in Lemma 3.6 as the value of the game is in fact $\Theta(\log(T))$ when the game is not trivial (e.g. when $\operatorname{int}(\operatorname{conv}(\mathcal{Z})) \neq \emptyset$).

Theorem 3.5. *When* $\ell(z, f) = z^\intercal f$, $0 \notin \operatorname{conv}(\mathcal{Z})$, *and* $\mathcal{F} = B_2(0, 1)$, *either the game is trivial or* $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Theta(\log(T))$.

So far, we have studied two scenarios that are diametrically opposed in terms of the curvature of the decision sets with polyhedra on one side in Section 3.2, with $\Theta(\sqrt{T})$ regret,

and Euclidean balls, with $\Theta(\log(T))$ regret if $0 \notin \operatorname{conv}(\mathcal{Z})$, on the other side of the spectrum. Bridging this gap leads to the rise of intermediate learning rates that can be quantified through the modulus of convexity of $\mathcal{F}$. Precisely, consider any norm $\|\ \|_{\mathcal{F}}$. The modulus of convexity of its unit ball is defined as:

$$\delta_{\mathcal{F}} : \epsilon \to \inf_{\substack{\|f\|_{\mathcal{F}}, \left\|\tilde{f}\right\|_{\mathcal{F}} \leq 1 \\ \left\|f - \tilde{f}\right\|_{\mathcal{F}} \geq \epsilon}} 1 - \left\| \frac{f + \tilde{f}}{2} \right\|_{\mathcal{F}}.$$

The norm $\|\ \|_{\mathcal{F}}$ is said to be uniformly convex if $\delta_{\mathcal{F}}(\epsilon) > 0$ for all $\epsilon \in [0, 2]$. As shown in [76], if $\|\ \|_{\mathcal{F}}$ is uniformly convex, there must exist $q \geq 2$ and $c > 0$ such that $\delta_{\mathcal{F}}(\epsilon) \geq c\epsilon^q$ for all $\epsilon \in [0, 2]$, in which case we say that $\|\ \|_{\mathcal{F}}$ is $q-$uniformly convex. This parameter quantifies how curved $\|\ \|_{\mathcal{F}}$ balls are and determines the growth rate of the value of the game when $\mathcal{F}$ is a $\|\ \|_{\mathcal{F}}$ ball and $0 \notin \operatorname{conv}(\mathcal{Z})$.

Theorem 3.6. *Consider the game $(\ell(z, f) = z^{\mathsf{T}} f, \mathcal{Z}, \mathcal{F})$ with $0 \notin \operatorname{conv}(\mathcal{Z})$ and $\mathcal{F} = \{f \mid \|f\|_{\mathcal{F}} \leq C\}$, where $\|\ \|_{\mathcal{F}}$ is a q-uniformly convex norm and $C \geq 0$. Then, FTL yields regret $O(\log(T))$ if $q = 2$ and regret $O(T^{\frac{q-2}{q-1}})$ if $q \in (2, 3]$.*

In particular, Theorem 3.6 applies to L_p balls since the L_p-norm is p-uniformly convex for $p \in [2, 3)$ and 2-uniformly convex if $p \in (1, 2]$. As a side note, remark that none of Lemma 3.5, Theorem 3.4, Theorem 3.5, or Theorem 3.6 cover the case where 0 lies on the boundary of $\operatorname{conv}(\mathcal{Z})$ and ℓ is linear. We stress that, in general, zero-mean i.i.d. opponents fail to yield $\Omega(\sqrt{T})$ lower bounds on regret in this setting, as we illustrate with an example. Hence, the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$ remains unknown in this setting.

Lemma 3.7. *Define $\mathcal{Z} = \operatorname{conv}(z_1, z_2, z_3, z_4)$ with $z_1 = (-1, 1, 0, 0)$, $z_2 = (1, -1, 0, 0)$, $z_3 = (0, 0, 0, 1)$, and $z_4 = (0, 0, 1, 0)$. Also define $\mathcal{F} = [f^*, f^{**}]$ with $f^* = (0, 0, 0, 0)$ and $f^{**} = (1, 1, -1, 1)$. The game $(\ell(z, f) = z^{\mathsf{T}} f, \mathcal{Z}, \mathcal{F})$ is not trivial and any lower bound derived from Theorem 3.1 with zero-mean random variables is 0.*

3.4 Concluding Remark

An interesting direction for future research is to develop a complete characterization of the growth rate of $R_T(\ell, \mathcal{Z}, \mathcal{F})$ for general loss functions. As shown in this chapter, the curvature of $\mathcal{F}$ is key to get $o(\sqrt{T})$ rates in online linear optimization but it is not enough by itself, as one must also restrict the power of the opponent through the constraint $0 \notin \mathrm{conv}(\mathcal{Z})$. It is unclear how to minimally restrict the power of the opponent to get $o(\sqrt{T})$ rates for general non-curved loss functions.

Chapter 4

Logarithmic Regret Bounds for Bandits with Knapsacks

4.1 Introduction

4.1.1 Motivation

Multi-Armed Bandit (MAB) is a benchmark model for repeated decision making in stochastic environments with very limited feedback on the outcomes of alternatives. In these circumstances, a decision maker must strive to find an overall optimal sequence of decisions while making as few suboptimal ones as possible when exploring the decision space in order to generate as much revenue as possible, a trade-off coined *exploration-exploitation*. The original problem, first formulated in its predominant version in [83], has spurred a new line of research that aims at introducing additional constraints that reflect more accurately the reality of the decision making process. *Bandits with Knapsacks* (BwK), a model formulated in its most general form in [16], fits into this framework and is characterized by the consumption of a limited supply of resources (e.g. time, money, and natural resources) that comes with every decision. This extension is motivated by a number of applications in electronic markets such as dynamic pricing with limited supply, see [25] and [14], online advertising, see [88], online bid optimization for sponsored search auctions, see [92], and crowdsourcing, see [15]. A unifying paradigm of online learning is to evaluate algorithms

based on their regret performance. In the BwK theory, this performance criterion is expressed as the gap between the total payoff of an optimal oracle algorithm aware of how the rewards and the amounts of resource consumption are generated and the total payoff of the algorithm. Many approaches have been proposed to tackle the original MAB problem, where time is the only limited resource with a prescribed time horizon T, and the optimal regret bounds are now well documented. They can be classified into two categories with qualitatively different asymptotic growth rates. Many algorithms, such as UCB1, see [12], Thompson sampling, see [8], and ϵ-greedy, see [12], achieve distribution-dependent, i.e. with constant factors that depend on the underlying unobserved distributions, asymptotic bounds on regret of order $\Theta(\ln(T))$, which is shown to be optimal in [60]. While these results prove very satisfying in many settings, the downside is that the bounds can get arbitrarily large if a malicious opponent was to select the underlying distributions in an adversarial fashion. In contrast, algorithms such as Exp3, designed in [13], achieve distribution-free bounds that can be computed in an online fashion, at the price of a less attractive growth rate $\Theta(\sqrt{T})$. The BwK theory lacks this clear-cut distinction. While provably optimal distribution-free bounds have recently been established, see [6] and [16], there has been little progress toward the development of asymptotically optimal distribution-dependent regret bounds. To bridge the gap, we introduce a template algorithm with proven regret bounds which are asymptotically logarithmic in the initial supply of each resource, in four important cases that cover a wide range of applications:

- Case 1, where there is a single limited resource other than time, which is not limited, and the amount of resource consumed as a result of making a decision is stochastic. Applications in online advertising, see [89], fit in this framework;

- Case 2, where there are arbitrarily many resources and the amounts of resources consumed as a result of making a decision are deterministic. Applications to network revenue management of perishable goods, see [25], shelf optimization of perishable goods, see [44], and wireless sensor networks, see [91], fit in this framework;

- Case 3, where there are two limited resources, one of which is assumed to be time while the consumption of the other is stochastic, under a nondegeneracy condition.

Typical applications include online bid optimization in sponsored search auctions, see [92], dynamic pricing with limited supply, see [14], and dynamic procurement, see [15];

- Case 4, where there are arbitrarily many resources, under a stronger nondegeneracy condition than for Case 3. Typical applications include dynamic ad allocation, see [88], dynamic pricing of multiple products, see [16], and network revenue management, see [25].

In terms of applicability and significance of the results, Case 3 is the most important case. Case 4 is the most general one but the analysis is more involved and requires stronger assumptions which makes it less attractive from a practical standpoint. The analysis is easier for Cases 1 and 2 but their modeling power is more limited.

In fast-paced environments, such as in ad auctions, the stochastic assumptions at the core of the BwK model are only valid for a short period of time but there are typically a large number of actions to be performed per second (e.g. submit a bid for a new ad auction). In these situations, the initial endowments of resources are thus typically large and logarithmic regret bounds can be significantly more attractive than distribution-free ones.

4.1.2 Problem Statement and Contributions

At each time period $t \in \mathbb{N}$, a decision needs to be made among a predefined finite set of actions, represented by arms and labeled $k = 1, \cdots, K$. We denote by a_t the arm pulled at time t. Pulling arm k at time t yields a random reward $r_{k,t} \in [0,1]$ (after scaling) and incurs the consumption of $C \in \mathbb{N}$ different resource types by random amounts $c_{k,t}(1), \cdots, c_{k,t}(C) \in [0,1]^C$ (after scaling). Note that time itself may or may not be a limited resource. At any time t and for any arm k, the vector $(r_{k,t}, c_{k,t}(1), \cdots, c_{k,t}(C))$ is jointly drawn from a fixed probability distribution ν_k independently from the past. The rewards and the amounts of resource consumption can be arbitrarily correlated across arms. We denote by $(\mathcal{F}_t)_{t\in\mathbb{N}}$ the natural filtration generated by the rewards and the amounts of resource consumption revealed to the decision maker, i.e. $((r_{a_t,t}, c_{a_t,t}(1), \cdots, c_{a_t,t}(C)))_{t\in\mathbb{N}}$.

The consumption of any resource $i \in \{1, \cdots, C\}$ is constrained by an initial budget $B(i) \in \mathbb{R}_+$. As a result, the decision maker can keep pulling arms only so long as he does not run out of any of the C resources and the game ends at time period τ^*, defined as:

$$\tau^* = \min\{t \in \mathbb{N} \mid \exists i \in \{1, \cdots, C\}, \sum_{\tau=1}^{t} c_{a_\tau,\tau}(i) > B(i)\}. \tag{4.1}$$

Note that τ^* is a stopping time with respect to $(\mathcal{F}_t)_{t\geq 1}$. When it comes to choosing which arm to pull next, the difficulty for the decision maker lies in the fact that none of the underlying distributions, i.e. $(\nu_k)_{k=1,\cdots,K}$, are initially known. Furthermore, the only feedback provided to the decision maker upon pulling arm a_t (but prior to selecting a_{t+1}) is $(r_{a_t,t}, c_{a_t,t}(1), \cdots, c_{a_t,t}(C))$, i.e. the decision maker does not observe the rewards that would have been obtained and the amounts of resources that would have been consumed as a result of pulling a different arm. The goal is to design a non-anticipating algorithm that, at any time t, selects a_t based on the information acquired in the past so as to keep the pseudo regret defined as:

$$R_{B(1),\cdots,B(C)} = \mathrm{ER}_{\mathrm{OPT}}(B(1), \cdots, B(C)) - \mathbb{E}[\sum_{t=1}^{\tau^*-1} r_{a_t,t}], \tag{4.2}$$

as small as possible, where $\mathrm{ER}_{\mathrm{OPT}}(B(1), \cdots, B(C))$ is the maximum expected sum of rewards that can be obtained by a non-anticipating oracle algorithm that has knowledge of the underlying distributions. Here, an algorithm is said to be non-anticipating if the decision to pull a given arm does not depend on the future observations. We develop algorithms and establish distribution-dependent regret bounds, that hold for any choice of the unobserved underlying distributions $(\nu_k)_{k=1,\cdots,K}$, as well as distribution-independent regret bounds. This entails studying the asymptotic behavior of $R_{B(1),\cdots,B(C)}$ when all the budgets $(B(i))_{i=1,\cdots,C}$ go to infinity. In order to simplify the analysis, it is convenient to assume that the ratios $(B(i)/B(C))_{i=1,\cdots,C}$ are constants independent of any other relevant quantities and to denote $B(C)$ by B.

Assumption 4.1. *For any resource $i \in \{1, \cdots, C\}$, we have $B(i) = b(i) \cdot B$ for some fixed constant $b(i) \in (0, 1]$. Hence $b = \min_{i=1,\cdots,C} b(i)$ is a positive quantity.*

When time is a limited resource, we use the notation T in place of B. Assumption 4.1 is widely used in the dynamic pricing literature where the inventory scales linearly with the time horizon, see [25] and [55]. Assumption 4.1 will only prove useful when deriving distribution-dependent regret bounds and it can largely be relaxed, see Section C.1 of the Appendix.

As the mean turns out to be an important statistics, we denote the mean reward and amounts of resource consumption by $\mu_k^r, \mu_k^c(1), \cdots, \mu_k^c(C)$ and their respective empirical estimates by $\bar{r}_{k,t}$, $\bar{c}_{k,t}(1), \cdots, \bar{c}_{k,t}(C)$. These estimates depend on the number of times each arm has been pulled by the decision maker up to, but not including, time t, which we write $n_{k,t}$. We end with a general assumption, which we use throughout the chapter, meant to have the game end in finite time.

Assumption 4.2. *For any arm $k \in \{1, \cdots, K\}$, we have* $\max_{i=1,\cdots,C} \mu_k^c(i) > 0$.

Note that Assumption 4.2 is automatically satisfied if time is a limited resource.

Contributions. We design an algorithm that runs in time polynomial in K for which we establish $O(K^C \cdot \ln(B)/\Delta)$ (resp. $\sqrt{K^C \cdot B \cdot \ln(B)}$) distribution-dependent (resp. distribution-free) regret bounds, where Δ is a parameter that generalizes the optimality gap for the standard MAB problem. We establish these regret bounds in four cases of increasing difficulty making additional technical assumptions that become stronger as we make progress towards tackling the general case. We choose to present these intermediate cases since: (i) we get improved constant factors under weaker assumptions and (ii) they subsume many practical applications. Note that our distribution-dependent regret bounds scale as a polynomial function of K of degree C, which may be unacceptable when the number of resources is large. We provide evidence that suggests that a linear dependence on K can be achieved by tweaking the algorithm, at least in some particular cases of interest. Finally, we point out that the constant factors hidden in the O notations are not scale-free, in the sense that jointly scaling down the amounts of resources consumed at each round along with their respective initial endowments worsens the bounds. As a consequence, initially scaling down the amounts of resource consumption in order to guarantee

that they lie in $[0, 1]$ should be done with caution: the scaling factors should be as small as possible.

4.1.3 Literature Review

The Bandits with Knapsacks framework was first introduced in its full generality in [16], but special cases had been studied before, see for example [89], [39], and [14]. Since the standard MAB problem fits in the BwK framework, with time being the only scarce resource, the results listed in the introduction tend to suggest that regret bounds with logarithmic growth with respect to the budgets may be possible for BwK problems but very few such results are documented. When there are arbitrarily many resources and a time horizon, the authors of [16] and [6] obtain $\tilde{O}(\sqrt{K \cdot T})$ distribution-free bounds on regret that hold on average as well as with high probability, where the $\tilde{O}$ notation hides logarithmic factors. These results were later extended to the contextual version of the problem in [17] and [7]. The authors of [55] extend Thompson sampling to tackle the general BwK problem and obtain distribution-dependent bounds on regret of order $\tilde{O}(\sqrt{T})$ (with an unspecified dependence on K) when time is a limited resource, under a nondegeneracy condition. The authors of [93] develop algorithms for BwK problems with a continuum of arms and a single limited resource constrained by a budget B and obtain $o(B)$ regret bounds. The authors of [36] consider a closely related framework that allows to model any history-dependent constraint on the number of times any arm can be pulled and obtain $O(K \cdot \ln(T))$ regret bounds when time is a limited resource. However, the benchmark oracle algorithm used in [36] to define the regret is significantly weaker than the one considered here as it only has knowledge of the distributions of the rewards, as opposed to the joint distributions of the rewards and the amounts of resource consumption. The authors of [14] establish a $\Omega(\sqrt{T})$ distribution-dependent lower bound on regret for a dynamic pricing problem which can be cast as a BwK problem with a time horizon, a resource whose consumption is stochastic, and a continuum of arms. This lower bound does not apply here as we are considering finitely many arms and it is well known that the minimax regret can be exponentially smaller when we move from finitely many arms to uncountably many arms

for the standard MAB problem, see [58]. The authors of [90] tackle BwK problems with a single limited resource whose consumption is deterministic and constrained by a global budget B and obtain $O(K \cdot \ln(B))$ regret bounds. This result was later extended to the case of a stochastic resource in [102]. The authors of [101] study a contextual version of the BwK problem when there are two limited resources, one of which is assumed to be time while the consumption of the other is deterministic, and obtain $O(K \cdot \ln(T))$ regret bounds under a nondegeneracy condition. Logarithmic regret bounds are also derived in [88] for a dynamic ad allocation problem that can be cast as a BwK problem.

Organization. The remainder of the chapter is organized as follows. We present applications of the BwK model in Section 4.2. We expose the algorithmic ideas underlying our approach in Section 4.3 and apply theses ideas to Cases (1), (2), (3), and (4) in Sections 4.4, 4.5, 4.6, and 4.7 respectively. We choose to discuss each case separately in a self-contained fashion so that readers can delve into the setting they are most interested in. This comes at the price of some overlap in the analysis. We relax some of the assumptions made in the course of proving the regret bounds and discuss extensions in Section C.1 of the Appendix. To provide as much intuition as possible, the ideas and key technical steps are all included in the main body, sometimes through proof sketches, while the technical details are deferred to the Appendix.

Notations. For a set S, $|S|$ denotes the cardinality of S while $\mathbb{1}_S$ is the indicator function of S. For a vector $x \in \mathbb{R}^n$ and S a subset of $\{1, \cdots, n\}$, x_S refers to the subvector $(x_i)_{i \in S}$. For a square matrix A, $\det(A)$ is the determinant of A while $\mathrm{adj}(A)$ denotes its adjugate. For $x \in \mathbb{R}$, x_+ is the positive part of x. We use standard asymptotic notations such as $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$.

4.2 Applications

A number of applications of the BwK framework are documented in the literature. For the purpose of being self-contained, we review a few popular ones that satisfy our technical

assumptions.

4.2.1 Online Advertising

Bid optimization in repeated second-price auctions. Consider a bidder participating in sealed second-price auctions who is willing to spend a budget B. This budget may be allocated only for a period of time (for the next T auctions) or until it is completely exhausted. Rounds, indexed by $t \in \mathbb{N}$, correspond to auctions the bidder participates in. If the bid submitted by the bidder for auction t is larger than the highest bid submitted by the competitors, denoted by m_t, the bidder wins the auction, derives a private utility $v_t \in [0, 1]$ (whose monetary value is typically difficult to assess), and is charged m_t. Otherwise, m_t is not revealed to the bidder and v_t cannot be assessed. We consider a stochastic setting where the environment and the competitors are not fully adversarial: $((v_t, m_t))_{t\in\mathbb{N}}$ is assumed to be an i.i.d. stochastic process. The goal for the bidder is to design a strategy to maximize the expected total utility derived given that the bidder has selected a grid of bids to choose from $(b_1, \cdots, b_K)$ (e.g. $b_1 = \$0.10, \cdots, b_K = \$1$). This is a BwK problem with two resources: time and money. Pulling arm k at round t corresponds to bidding b_k in auction t, costs $c_{k,t} = m_t \cdot \mathbb{1}_{b_k \geq m_t}$, and yields a reward $r_{k,t} = v_t \cdot \mathbb{1}_{b_k \geq m_t}$. The authors of [97] design bidding strategies for a variant of this problem where the bidder is not limited by a budget and $r_{k,t} = (v_t - m_t) \cdot \mathbb{1}_{b_k \geq m_t}$.

This model was first formalized in [92] in the context of sponsored search auctions. In sponsored search auctions, advertisers can bid on keywords to have ads (typically in the form of a link followed by a text description) displayed alongside the search results of a web search engine. When a user types a search query, a set of relevant ads are selected and an auction is run in order to determine which ones will be displayed. The winning ads are allocated to ad slots based on the outcome of the auction and, in the prevailing cost-per-click pricing scheme, their owners get charged only if the user clicks on their ads. Because the auction is often a variant of a sealed second-price auction (e.g. a generalized second-price auction), very limited feedback is provided to the advertiser if the auction is lost. In addition, both the demand and the supply cannot be predicted ahead of time and are thus

commonly modeled as random variables, see [43]. For these reasons, bidding repeatedly on a keyword can be formulated as a BwK problem. In particular, when the search engine has a single ad slot per query, this problem can be modeled as above: B is the budget the advertiser is willing to spend on a predetermined keyword and rounds correspond to ad auctions the advertiser has been selected to participate in. If the advertiser wins the auction, his or her ad gets displayed and he or she derives a utility $v_t = \mathbb{1}_{A_t}$, where A_t is the event that the ad gets clicked on. The goal is to maximize the expected total number of clicks given the budget constraint. The advertiser may also be interested in optimizing the ad to be displayed, which will affect the probability of a click. In this case, the modeling is similar but arms correspond to pairs of bid values and ads.

Dynamic ad allocation. This problem was first modeled in the BwK framework in [88]. A publisher, i.e. the owner of a collection of websites where ads can be displayed, has previously agreed with K advertisers, indexed by $k \in \{1, \cdots, K\}$, on a predetermined cost-per-click p_k. Additionally, advertiser k is not willing to spend more than a prescribed budget, B_k, for a predetermined period of time (which corresponds to the next T visits or rounds). Denote by A_t^k the event that the ad provided by advertiser k gets clicked on at round t. We consider a stochastic setting where the visitors are not fully adversarial: $(\mathbb{1}_{A_t^k})_{t\in\mathbb{N}}$ is assumed to be an i.d.d. stochastic process for any advertiser k. The goal for the publisher is to maximize the total expected revenues by choosing which ad to display at every round, i.e. every time somebody visits one of the websites. This situation can be modeled as a BwK problem with $K + 1$ resources: time and money for each of the K advertisers. Pulling arm $k \in \{1, \cdots, K\}$ at round t corresponds to displaying the ad owned by advertiser k, incurs the costs $c_{k,t}(i) = p_k \cdot \mathbb{1}_{A_t^k} \cdot \mathbb{1}_{i=k}$ to advertiser $i \in \{1, \cdots, K\}$, and yields a revenue $r_{k,t} = p_k \cdot \mathbb{1}_{A_t^k}$.

4.2.2 Revenue Management

Dynamic pricing with limited supply. This BwK model was first proposed in [14]. An agent has B identical items to sell to T potential customers that arrive sequentially. Customer $t \in \{1, \cdots, T\}$ is offered a take-it-or-leave-it price p_t and purchases the item only

if p_t is no larger than his or her own valuation v_t, which is never disclosed. Customers are assumed to be non-strategic in the sense that their valuations are assumed to be drawn i.i.d. from a distribution unknown to the agent. The goal for the agent is to maximize the total expected revenues by offering prices among a predetermined list $(p_1, \cdots, p_K)$. This is a BwK problem with two resources: time and item inventory. Pulling arm $k \in \{1, \cdots, K\}$ at round t corresponds to offering the price p_k, depletes the inventory of $c_{k,t} = \mathbb{1}_{p_k \leq v_t}$ unit, and generates a revenue $r_{k,t} = p_k \cdot \mathbb{1}_{p_k \leq v_t}$. The authors of [16] propose an extension where multiple units of M different products may be offered to a customer, which then buys as many as needed of each kind in order to maximize his or her own utility function. In this case, the modeling is similar but arms correspond to vectors of dimension $2M$ specifying the number of items offered along with the price tag for each product and there are $M + 1$ resources: time and item inventory for each of the M products.

Network revenue management.

Non-perishable goods. This is an extension of the dynamic pricing problem developed in [25] which is particularly suited for applications in the online retailer industry, e.g. the online fashion sample sales industry, see [55]. Each product $m = 1, \cdots, M$ is produced from a finite amount of C different kinds of raw materials (which may be products themselves). Producing one unit of product $m \in \{1, \cdots, M\}$ consumes a deterministic amount of resource $i \in \{1, \cdots, C\}$ denoted by $c_m(i)$. Customer $t \in \{1, \cdots, T\}$ is offered a product $m_t \in \{1, \cdots, M\}$ along with a take-it-or-leave-it price $p_t^{m_t}$ and purchases it if his or her valuation $v_t^{m_t}$ is larger than $p_t^{m_t}$. Products are manufactured online as customers order them. We assume that $((v_t^1, \cdots, v_t^M))_{t \in \mathbb{N}}$ is an i.i.d. process with distribution unknown to the agent. This is a BwK problem with $C + 1$ resources: time and the initial endowment of each resource. Given a prescribed list of arms $((m_k, p_k))_{k=1,\cdots,K}$, pulling arm k at round t corresponds to offering product m_k at a price p_k, incurs the consumption of resource i by an amount $c_{k,t}(i) = c_{m_k}(i) \cdot \mathbb{1}_{p_k \leq v_t^{m_k}}$, and generates a revenue $r_{k,t} = p_k \cdot \mathbb{1}_{p_k \leq v_t^{m_k}}$.

Perishable goods. This is a variant of the last model developed for perishable goods, with applications in the food retail industry and the newspaper industry. At each time period

$t \in \{1, \cdots, T\}$, a retailer chooses how many units $\lambda_t^m \in \mathbb{N}$ of product $m \in \{1, \cdots, M\}$ to manufacture along with a price offer for it p_t^m. At time t, the demand for product m sold at the price p is a random quantity denoted by $d_t^m(p)$. We assume that customers are non-strategic: for any vector of prices $(p_1, \cdots, p_m)$, $((d_t^1(p_1), \cdots d_t^M(p_M)))_{t\in\mathbb{N}}$ is an i.i.d. stochastic process with distribution unknown to the agent. Products perish at the end of each round irrespective of whether they have been purchased. Given a predetermined list of arms $((\lambda_k^1, p_k^1, \cdots, \lambda_k^M, p_k^M))_{k=1,\cdots,K}$, pulling arm k at round t corresponds to offering λ_k^m units of product m at the price p_k^m for any $m \in \{1, \cdots, M\}$, incurs the consumption of resource i by a deterministic amount $c_{k,t}(i) = \sum_{m=1}^M \lambda_k^m \cdot c_m(i)$ (where $c_m(i)$ is defined in the previous paragraph), and generates a revenue $r_{k,t} = \sum_{m=1}^M p_k^m \cdot \min(d_t^m(p_k^m), \lambda_k^m)$.

Shelf optimization for perishable goods. This is a variant of the model introduced in [44]. Consider a retailer who has an unlimited supply of M different types of products. At each time period t, the retailer has to decide how many units, λ_t^m, of each product, $m \in \{1, \cdots, M\}$, to allocate to a promotion space given that at most N items fit in the limited promotion space. Moreover, the retailer also has to decide on a price tag p_t^m for each product m. All units of product $m \in \{1, \cdots, M\}$ perish by time period T_m and the retailer is planning the allocation for the next T time periods. At round t, the demand for product m is a random quantity denoted by $d_t^m(p)$. Customers are non-strategic: for any vector of prices $(p_1, \cdots, p_m)$, $((d_t^1(p_1), \cdots d_t^M(p_M)))_{t\in\mathbb{N}}$ is an i.i.d. stochastic process with distribution unknown to the agent. This is a BwK problem with $M + 1$ resources: time horizon and time after which each product perishes. Given a predetermined list of arms $((\lambda_k^1, p_k^1, \cdots, \lambda_k^M, p_k^M))_{k=1,\cdots,K}$ satisfying $\sum_{m=1}^M \lambda_k^m \leq K$ for any $k \in \{1, \cdots, K\}$, pulling arm k at round t corresponds to allocating λ_k^m units of product m to the promotion space for every $m \in \{1, \cdots, M\}$ with the respective price tags $(p_k^1, \cdots, p_k^M)$, incurs the consumption of resource i by a deterministic amount $c_{k,t}(i) = 1$, and generates a revenue $r_{k,t} = \sum_{m=1}^M p_k^m \cdot \min(d_t^m(p_k^m), \lambda_k^m)$.

4.2.3 Dynamic Procurement

This problem was first studied in [15]. Consider a buyer with a budget B facing T agents arriving sequentially, each interested in selling one good. Agent $t \in \{1, \cdots, T\}$ is offered a take-it-or-leave-it price, p_t, and makes a sell only if the value he or she attributes to the item, v_t, is no larger than p_t. We consider a stochastic setting where the sellers are not fully adversarial: $(v_t)_{t\in\mathbb{N}}$ is an i.i.d. stochastic process with distribution unknown to the buyer. The goal for the buyer is to maximize the total expected number of goods purchased by offering prices among a predetermined list $(p_1, \cdots, p_K)$. This is a BwK problem with two resources: time and money. Pulling arm k at round t corresponds to offering the price p_k, incurs a cost $c_{k,t} = p_k \cdot \mathbb{1}_{p_k \geq v_t}$, and yields a reward $r_{k,t} = \mathbb{1}_{p_k \geq v_t}$. It is also possible to model situations where the agents are selling multiple types of products and/or multiple units, in which case arms correspond to vectors specifying the number of units of each product required along with their respective prices, see [16].

Applications of this model to crowdsourcing platforms are described in [15] and [16]. In this setting, agents correspond to workers that are willing to carry out microtasks which are submitted by buyers (called "requesters") using a posted-price mechanism. Requesters are typically submitting large batches of jobs and can thus adjust the posted prices as they learn about the pool of workers.

4.2.4 Wireless Sensor Networks

This is a variant of the model introduced in [91]. Consider an agent collecting information using a network of wireless sensors powered by batteries. Activating sensor $k \in \{1, \cdots, K\}$ consumes some amount of energy, c_k, which is depleted from the sensor's initial battery level, B_k, and triggers a measurement providing a random amount of information (measured in bits), $r_{k,t}$, which is transmitted back to the agent. Sensors cannot harvest energy and the goal for the agent is to maximize the total expected amount of information collected over T actions. This is a BwK problem with $K + 1$ resources: time and the energy stored in the battery of each sensor. Pulling arm $k \in \{1, \cdots, K\}$ corresponds to activating sensor k, incurs the consumption of resource $i \in \{1, \cdots, K\}$ by a deterministic

amount $c_{k,t} = c_k \cdot \mathbb{1}_{k=i}$, and yields a random reward $r_{k,t}$.

4.3 Algorithmic Ideas

4.3.1 Preliminaries

To handle the exploration-exploitation trade-off, an approach that has proved to be particularly successful hinges on the *optimism in the face of uncertainty* paradigm. The idea is to consider all plausible scenarios consistent with the information collected so far and to select the decision that yields the most revenue among all the scenarios identified. Concentration inequalities are intrinsic to the paradigm as they enable the development of systematic closed form confidence intervals on the quantities of interest, which together define a set of plausible scenarios. We make repeated use of the following result.

Lemma 4.1. *Hoeffding's inequality*

Consider $X_1, \cdots, X_n$ n *random variables with support in* $[0, 1]$.

If for every $t \leq n$ $\mathbb{E}[X_t \mid X_1, \cdots, X_{t-1}] \leq \mu$, *then* $\mathbb{P}[X_1 + \cdots + X_n \geq n\mu + a] \leq \exp(-\frac{2a^2}{n})$ *for all* $a \geq 0$.

If for every $t \leq n$ $\mathbb{E}[X_t \mid X_1, \cdots, X_{t-1}] \geq \mu$, *then* $\mathbb{P}[X_1 + \cdots + X_n \leq n\mu - a] \leq \exp(-\frac{2a^2}{n})$ *for all* $a \geq 0$.

The authors of [12] follow the *optimism in the face of uncertainty* paradigm to develop the Upper Confidence Bound algorithm (UCB1). UCB1 is based on the following observations: (i) the optimal strategy always consists in pulling the arm with the highest mean reward when time is the only limited resource, (ii) informally, Lemma 4.1 shows that $\mu_k^r \in [\bar{r}_{k,t} - \epsilon_{k,t}, \bar{r}_{k,t} + \epsilon_{k,t}]$ at time t with probability at least $1 - 2/t^3$ for $\epsilon_{k,t} = \sqrt{2\ln(t)/n_{k,t}}$, irrespective of the number of times arm k has been pulled. Based on these observations, UCB1 always selects the arm with highest UCB index, i.e. $a_t \in \operatorname{argmax}_{k=1,\cdots,K} I_{k,t}$, where the UCB index of arm k at time t is defined as $I_{k,t} = \bar{r}_{k,t} + \epsilon_{k,t}$. The first term can be interpreted as an exploitation term, the ultimate goal being to maximize revenue, while the second term is an exploration term, the smaller $n_{k,t}$, the bigger it is. This fruitful

paradigm go well beyond this special case and many extensions of UCB1 have been designed to tackle variants of the MAB problem, see for example [88]. The authors of [6] embrace the same ideas to tackle BwK problems. The situation is more complex in this all-encompassing framework as the optimal oracle algorithm involves pulling several arms. In fact, finding the optimal pulling strategy given the knowledge of the underlying distributions is already a challenge in its own, see [73] for a study of the computational complexity of similar problems. This raises the question of how to evaluate $\mathrm{ER}_{\mathrm{OPT}}(B(1), \cdots, B(C))$ in (4.2). To overcome this issue, the authors of [16] upper bound the expected payoff of any non-anticipating algorithm by the value of a linear program, which is easier to compute.

Lemma 4.2. *Adapted from [16]*

The total expected payoff of any non-anticipating algorithm is no greater than B times the optimal value of the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K} \in \mathbb{R}_+^K} \quad & \sum_{k=1}^{K} \mu_k^r \cdot \xi_k \\ \textit{subject to} \quad & \sum_{k=1}^{K} \mu_k^c(i) \cdot \xi_k \leq b(i), \quad i = 1, \cdots, C \end{aligned} \tag{4.3}$$

plus the constant term $\max\limits_{\substack{k=1,\cdots,K \\ i=1,\cdots,C| \\ \textit{with } \mu_k^c(i)>0}} \frac{\mu_k^r}{\mu_k^c(i)}$.

The optimization problem (4.3) can be interpreted as follows. For any arm k, $B \cdot \xi_k$ corresponds to the expected number of times arm k is pulled by the optimal algorithm. Hence, assuming we introduce a dummy arm 0 which is equivalent to skipping the current round, ξ_k can be interpreted as the probability of pulling arm k at any round when there is a time horizon T. Observe that the constraints restrict the feasible set of expected number of pulls by imposing that the amounts of resources consumed are no greater than their respective budgets in expectations, as opposed to almost surely which would be a more stringent constraint. This explains why the optimal value of (4.3) is larger than the maximum achievable payoff. In this chapter, we use standard linear programming notions such as the concept of a basis and a basic feasible solution. We refer to [24] for an introduction to linear programming. A pseudo-basis x is described by two subsets $\mathcal{K}_x \subset \{1, \cdots, K\}$

and $\mathcal{C}_x \subset \{1, \cdots, C\}$ such that $|\mathcal{K}_x| = |\mathcal{C}_x|$. A pseudo-basis x is a basis for (4.3) if the matrix $A_x = (\mu_k^c(i))_{(i,k)\in\mathcal{C}_x\times\mathcal{K}_x}$ is invertible. Furthermore, x is said to be a feasible basis for (4.3) if the corresponding basic solution, denoted by $(\xi_k^x)_{k=1,\cdots,K}$ and determined by $\xi_k^x = 0$ for $k \notin \mathcal{K}_x$ and $A_x \xi_{\mathcal{K}_x}^x = b_{\mathcal{C}_x}$ (where $b_{\mathcal{C}_x}$ is the subvector $(b(i))_{i\in\mathcal{C}_x}$), is feasible for (4.3). When x is a feasible basis for (4.3), we denote by $\mathrm{obj}_x = \sum_{k=1}^K \mu_k^r \cdot \xi_k^x$ its objective function. From Lemma 4.2, we derive:

$$R_{B(1),\cdots,B(C)} \leq B \cdot \mathrm{obj}_{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1), \tag{4.4}$$

where x^* is an optimal feasible basis for (4.3). For mathematical convenience, we consider that the game carries on even if one of the resources is already exhausted so that a_t is well defined for any $t \in \mathbb{N}$. Of course, the rewards obtained for $t \geq \tau^*$ are not taken into account in the decision maker's payoff when establishing regret bounds.

4.3.2 Solution Methodology

Lemma 4.2 also provides insight into designing algorithms. The idea is to incorporate confidence intervals on the mean rewards and the mean amounts of resource consumption into the offline optimization problem (4.3) and to base the decision upon the resulting optimal solution. There are several ways to carry out this task, each leading to a different algorithm. When there is a time horizon T, [6] use high-probability lower (resp. upper) bounds on the mean amounts of resource consumption (resp. rewards) in place of the unknown mean values in (4.3) and pull an arm at random according to the resulting optimal distribution. Specifically, at any round t, the authors suggest to compute an optimal solution $(\xi_{k,t}^*)_{k=1,\cdots,K}$ to the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K}\in\mathbb{R}_+^K} \quad & \sum_{k=1}^K (\bar{r}_{k,t} + \epsilon_{k,t}) \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^K (\bar{c}_{k,t}(i) - \epsilon_{k,t}) \cdot \xi_k \leq (1-\gamma) \cdot b(i), \quad i = 1, \cdots, C-1 \\ & \sum_{k=1}^K \xi_k \leq 1 \end{aligned} \tag{4.5}$$

for a well-chosen $\gamma \in (0,1)$, and then to pull arm k with probability $\xi^*_{k,t}$ or skip the round with probability $1-\sum_{k=1}^{K} \xi^*_{k,t}$. If we relate this approach to UCB1, the intuition is clear: the idea is to be optimistic about both the rewards and the amounts of resource consumption. We argue that this approach cannot yield logarithmic regret bounds. First, because γ has to be of order $1/\sqrt{T}$. Second, because, even if we were given an optimal solution to (4.3), $(\xi_k^{x^*})_{k=1,\cdots,K}$, before starting the game, consistently choosing which arm to pull at random according to this distribution at every round would incur regret $\Omega(\sqrt{T})$, as we next show.

Lemma 4.3. *For all the cases treated in this chapter, pulling arm k with probability $\xi_k^{x^*}$ at any round t yields a regret of order $\Omega(\sqrt{T})$ unless pulling any arm in the set $\{k \in \{1,\cdots,K\} \mid \xi_k^{x^*} > 0\}$ incurs the same deterministic amount of resource consumption for all resources in $\mathcal{C}_{x^*}$ and for all rounds $t \in \mathbb{N}$.*

Proof. Define $\mathcal{K}_* = \{k \in \{1,\cdots,K\} \mid \xi_k^{x^*} > 0\}$. For any $i \in \{1,\cdots,C-1\}$, we have:

$$\begin{aligned} T \cdot \mathrm{obj}_{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] &\geq \mathbb{E}[(\sum_{t=\tau^*}^{T} c_{a_t,t}(i) + \sum_{t=1}^{\tau^*-1} c_{a_t,t}(i) - B(i))_+] \cdot \mathrm{obj}_{x^*} + O(1) \\ &= \mathbb{E}[(\sum_{t=1}^{T} \{c_{a_t,t}(i) - b(i)\})_+] \cdot \mathrm{obj}_{x^*} + O(1), \end{aligned}$$

since $\mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] = \mathbb{E}[\tau^*] \cdot \mathrm{obj}_{x^*}$, $c_{a_t,t}(i) \leq 1$ for all t and $\sum_{t=1}^{\tau^*-1} c_{a_t,t}(i) \leq B(i)$. Since, for $i \in \mathcal{C}_{x^*}$, $(c_{a_t,t}(i))_{t\in\mathbb{N}}$ is an i.i.d. bounded stochastic process with mean $b(i)$, we get:

$$\mathbb{E}[(\sum_{t=1}^{T} \{c_{a_t,t}(i) - b(i)\})_+] = \Omega(\sqrt{T}), \tag{4.6}$$

provided that $c_{a_t,t}(i)$ has positive variance, which is true if there exists at least one arm $k \in \mathcal{K}_*$ such that $c_{k,t}(i)$ has positive variance or if there exist two arms $k, l \in \mathcal{K}_*$ such that $c_{k,t}(i)$ and $c_{l,t}(i)$ are not almost surely equal to the same deterministic value. Strictly speaking, this is not enough to conclude that $R_{B(1),\cdots,B(C-1),T} = \Omega(\sqrt{T})$ as $T \cdot \mathrm{obj}_{x^*}$ is only an upper bound on the maximum expected payoff. However, in Sections 4.4, 4.5, 4.6, and 4.7, we show that there exists an algorithm that satisfies $T \cdot \mathrm{obj}_{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] = O(\ln(T))$ for all the cases considered in this thesis. Together with (4.6) and Lemma 4.2, this shows that the regret incurred when pulling arm k with probability $\xi_k^{x^*}$ at any round is $\Omega(\sqrt{T})$. $\square$

The fundamental shortcoming of this approach is that it systematically leads us to plan to consume the same average amount of resource i per round $b(i)$, for any resource $i = 1, \cdots, C-1$, irrespective of whether we have significantly over- or under-consumed in the past. Based on this observation, a natural idea is to solve the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K} \in \mathbb{R}_+^K} \quad & \sum_{k=1}^{K} (\bar{r}_{k,t} + \epsilon_{k,t}) \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^{K} (\bar{c}_{k,t}(i) - \epsilon_{k,t}) \cdot \xi_k \leq (1-\gamma) \cdot b_t(i), \quad i = 1, \cdots, C-1 \\ & \sum_{k=1}^{K} \xi_k \leq 1 \end{aligned} \tag{4.7}$$

instead of (4.5), where $b_t(i)$ denotes the ratio of the remaining amount of resource i at time t to the remaining time horizon, i.e. $T - t + 1$. Bounding the regret incurred by this adaptive algorithm is, however, difficult from a theoretical standpoint. To address this issue, we propose the following family of algorithms, whose behaviors are similar to the adaptive algorithm but lend themselves to an easier analysis.

Algorithm: UCB-Simplex

Take $\lambda \geq 1$ and $(\eta_i)_{i=1,\cdots,C} \geq 0$ (these quantities will need to be carefully chosen). The algorithm is preceded by an initialization phase which consists in pulling each arm a given number of times, to be specified. For each subsequent time period t, proceed as follows.

Step-Simplex: Find an optimal basis x_t to the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K}} \quad & \sum_{k=1}^{K} (\bar{r}_{k,t} + \lambda \cdot \epsilon_{k,t}) \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^{K} (\bar{c}_{k,t}(i) - \eta_i \cdot \epsilon_{k,t}) \cdot \xi_k \leq b(i), \quad i = 1, \cdots, C \\ & \xi_k \geq 0, \quad k = 1, \cdots, K \end{aligned} \tag{4.8}$$

Adapting the notations, x_t is described by two subsets $\mathcal{K}_{x_t} \subset \{1, \cdots, K\}$ and $\mathcal{C}_{x_t} \subset \{1, \cdots, C\}$ such that $|\mathcal{K}_{x_t}| = |\mathcal{C}_{x_t}|$, the matrix $\bar{A}_{x_t,t} = (\bar{c}_{k,t}(i) - \eta_i \cdot \epsilon_{k,t})_{(i,k) \in \mathcal{C}_{x_t} \times \mathcal{K}_{x_t}}$, and the corresponding basic feasible solution $(\xi^{x_t}_{k,t})_{k=1,\cdots,K}$ determined by $\xi^{x_t}_{k,t} = 0$ for $k \notin \mathcal{K}_{x_t}$ and $\bar{A}_{x_t,t} \xi^{x_t}_{\mathcal{K}_x,t} = b_{\mathcal{K}_{x_t}}$.

Step-Load-Balance: Identify the arms involved in the optimal basis, i.e. $\mathcal{K}_{x_t}$. There are at most $\min(K, C)$ such arms. Use a load balancing algorithm $\mathcal{A}_{x_t}$, to be specified, to determine which of these arms to pull.

For all the cases considered in this chapter, (4.8) is always bounded and Step-Simplex is well defined. The Simplex algorithm is an obvious choice to carry out Step-Simplex, especially when $\eta_i = 0$ for any resource $i \in \{1, \cdots, C\}$, because, in this case, we only have to update one column of the constraint matrix per round which makes warm-starting properties attractive. However, note that this can also be done in time polynomial in K and C, see [45]. If we compare (4.8) with (4.5), the idea remains to be overly optimistic but, as we will see, more about the rewards than the amounts of resource consumption through the exploration factor λ which will typically be larger than η_i, thus transferring most of the burden of exploration from the constraints to the objective function. The details of Step-Load-Balance are purposefully left out and will be specified for each of the cases treated in this chapter. When there is a time horizon T, the general idea is to determine, at any time

period t and for each resource $i = 1, \cdots, C$, whether we have over- or under-consumed in the past and to perturb the probability distribution $(\xi^{x_t}_{k,t})_{k=1,\cdots,K}$ accordingly to get back on track.

The algorithm we propose is intrinsically tied to the existence of basic feasible optimal solutions to (4.3) and (4.8). We denote by $\mathcal{B}$ (resp. $\mathcal{B}_t$) the subset of bases of (4.3) (resp. (4.8)) that are feasible for (4.3) (resp. (4.8)). Step-Simplex can be interpreted as an extension of the index-based decision rule of UCB1. Indeed, Step-Simplex assigns an index $I_{x,t}$ to each basis $x \in \mathcal{B}_t$ and outputs $x_t \in \operatorname{argmax}_{x \in \mathcal{B}_t} I_{x,t}$, where $I_{x,t} = \mathrm{obj}_{x,t} + E_{x,t}$ with a clear separation (at least when $\eta_i = 0$ for any resource i) between the exploitation term, $\mathrm{obj}_{x,t} = \sum_{k=1}^{K} \xi^x_{k,t} \cdot \bar{r}_{k,t}$, and the exploration term, $E_{x,t} = \lambda \cdot \sum_{k=1}^{K} \xi^x_{k,t} \cdot \epsilon_{k,t}$. Observe that, for $x \in \mathcal{B}_t$ that is also feasible for (4.3), $(\xi^x_{k,t})_{k=1,\cdots,K}$ and $\mathrm{obj}_{x,t}$ are plug-in estimates of $(\xi^x_k)_{k=1,\cdots,K}$ and obj_x when $\eta_i = 0$ for any resource i. Also note that when $\lambda = 1$ and $\eta_i = 0$ for any resource i and when time is the only limited resource, UCB-Simplex is identical to UCB1 as Step-Load-Balance is unambiguous in this special case, each basis involving a single arm. For any $x \in \mathcal{B}$, we define $\Delta_x = \mathrm{obj}_{x^*} - \mathrm{obj}_x \geq 0$ as the optimality gap. A feasible basis x is said to be suboptimal if $\Delta_x > 0$. At any time t, $n_{x,t}$ denotes the number of times basis x has been selected at Step-Simplex up to time t while $n^x_{k,t}$ denotes the number of times arm k has been pulled up to time t when selecting x at Step-Simplex. For all the cases treated in this chapter, we will show that, under a nondegeneracy assumption, Step-Simplex guarantees that a suboptimal basis cannot be selected more than $O(\ln(B))$ times on average, a result reminiscent of the regret analysis of UCB1 carried out in [12]. However, in stark contrast with the situation of a single limited resource, this is merely a prerequisite to establish a $O(\ln(B))$ bound on regret. Indeed, a low regret algorithm must also balance the load between the arms as closely as possible to optimality. Hence, the choice of the load balancing algorithms $\mathcal{A}_x$ is crucial to obtain logarithmic regret bounds.

4.4 A Single Limited Resource

In this section, we tackle the case of a single resource whose consumption is limited by a global budget B, i.e. $C = 1$ and $b(1) = 1$. To simplify the notations, we omit the indices

identifying the resources as there is only one, i.e. we write μ_k^c, $c_{k,t}$, $\bar{c}_{k,t}$, and η as opposed to $\mu_k^c(1)$, $c_{k,t}(1)$, $\bar{c}_{k,t}(1)$, and η_1. We also use the shorthand $\epsilon = \min_{k=1,\cdots,K} \mu_k^c$. Recall that, under Assumption 4.2, ϵ is positive and a priori unknown to the decision maker. In order to derive logarithmic bounds, we will also need to assume that the decision maker knows an upper bound on the optimal value of (4.3).

Assumption 4.3. *The decision maker knows* $\kappa \geq \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c}$ *ahead of round* 1.

Assumption 4.3 is natural in repeated second-price auctions, as detailed in the last paragraph of this section. Moreover, note that if ϵ happens to be known ahead of round 1 we can take $\kappa = 1/\epsilon$.

Specification of the algorithm. We implement UCB-Simplex with $\lambda = 1 + \kappa$ and $\eta = 0$. The initialization step consists in pulling each arm until the amount of resource consumed as a result of pulling that arm is non-zero. The purpose of this step is to have $\bar{c}_{k,t} > 0$ for all periods to come and for all arms. Step-Load-Balance is unambiguous here as basic feasible solutions involve a single arm. Hence, we identify a basis x such that $\mathcal{K}_x = \{k\}$ and $\mathcal{C}_x = \{1\}$ with the corresponding arm and write $x = k$ to simplify the notations. In particular, $k^* \in \{1, \cdots, K\}$ identifies an optimal arm in the sense defined in Section 4.3. For any arm k, the exploration and exploitation terms defined in Section 4.3 specialize to:

$$\mathrm{obj}_{k,t} = \frac{\bar{r}_{k,t}}{\bar{c}_{k,t}} \text{ and } E_{k,t} = (1+\kappa) \cdot \frac{\epsilon_{k,t}}{\bar{c}_{k,t}},$$

while $\mathrm{obj}_k = \mu_k^r / \mu_k^c$, so that:

$$k^* \in \operatorname*{argmax}_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c},\ a_t \in \operatorname*{argmax}_{k=1,\cdots,K} \frac{\bar{r}_{k,t} + (1+\kappa)\cdot\epsilon_{k,t}}{\bar{c}_{k,t}},\ \text{and } \Delta_k = \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \frac{\mu_k^r}{\mu_k^c}.$$

We point out that, for the particular setting considered in this section, UCB-Simplex is almost identical to the fractional KUBE algorithm proposed in [90] to tackle the case of a single resource whose consumption is deterministic. It only differs by the presence of the scaling factor $1 + \kappa$ to favor exploration over exploitation, which becomes unnecessary when the amounts of resource consumed are deterministic, see Section C.1 of the

Appendix.

Regret analysis. We omit the initialization step in the theoretical analysis because the amount of resource consumed is $O(1)$ and the reward obtained is non-negative and not taken into account in the decision maker's total payoff. Moreover, the initialization step ends in finite time almost surely as a result of Assumption 4.2. First observe that (4.4) specializes to:

$$R_B \leq B \cdot \frac{\mu^r_{k^*}}{\mu^c_{k^*}} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1). \tag{4.9}$$

To bound the right-hand side, we start by estimating the expected time horizon.

Lemma 4.4. *For any non-anticipating algorithm, we have:* $\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}$.

Sketch of proof. By definition of τ^*, we have $\sum_{t=1}^{\tau^*-1} c_{a_t,t} \leq B$. Taking expectations on both sides yields $B \geq \mathbb{E}[\sum_{t=1}^{\tau^*} \mu^c_{a_t}] - 1 \geq \mathbb{E}[\tau^*] \cdot \epsilon - 1$ by Assumption 4.2. Rearranging this last inequality yields the claim. □

The next result is crucial. Used in combination with Lemma 4.4, it shows that any suboptimal arm is pulled at most $O(\ln(B))$ times in expectations, a well-known result for UCB1, see [12]. The proof is along the same lines as for UCB1, namely we assume that arm k has already been pulled more than $\Theta(\ln(\tau^*)/(\Delta_k)^2)$ times and conclude that arm k cannot be pulled more than a few more times, with the additional difficulty of having to deal with the random stopping time and the fact that the amount of resource consumed at each step is stochastic.

Lemma 4.5. *For any suboptimal arm k, we have:*

$$\mathbb{E}[n_{k,\tau^*}] \leq 2^6 (\frac{\lambda}{\mu^c_k})^2 \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + \frac{4\pi^2}{3\epsilon^2}.$$

Sketch of proof. We use the shorthand notation $\beta_k = 2^5(\lambda/\mu^c_k)^2 \cdot (1/\Delta_k)^2$. First observe that if we want to bound $\mathbb{E}[n_{k,\tau^*}]$, we may assume, without loss of generality, that arm k has been pulled at least $\beta_k \cdot \ln(t)$ times at any time t up to an additive term of $2\beta_k \cdot \mathbb{E}[\ln(\tau^*)]$ in the final inequality. We then just have to bound by a constant the probability that k is

selected at any time t given that $n_{k,t} \geq \beta_k \cdot \ln(t)$. If k is selected at time t, it must be that k is optimal for (4.8), which, in particular, implies that $\text{obj}_{k,t} + E_{k,t} \geq \text{obj}_{k^*,t} + E_{k^*,t}$. This can only happen if either: (i) $\text{obj}_{k,t} \geq \text{obj}_k + E_{k,t}$, i.e. the objective value of k is overly optimistic, (ii) $\text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t}$, i.e. the objective value of k^* is overly pessimistic, or (iii) $\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}$, i.e. the optimality gap of arm k is small compared to its exploration factor. The probability of events (i) and (ii) can be bounded by $\sim 1/t^2$ in the same fashion, irrespective of how many times these arms have been pulled in the past. For example for event (i), this is because if $\bar{r}_{k,t}/\bar{c}_{k,t} = \text{obj}_{k,t} \geq \text{obj}_k + E_{k,t} = \mu_k^r/\mu_k^c + E_{k,t}$, then either (a) $\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}$ or (b) $\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}$ and both of these events have probability at most $\sim 1/t^2$ by Lemma 4.1. Indeed, if (a) and (b) do not hold, we have:

$$\begin{aligned}\frac{\bar{r}_{k,t}}{\bar{c}_{k,t}} - \frac{\mu_k^r}{\mu_k^c} &= \frac{(\bar{r}_{k,t} - \mu_k^r)\mu_k^c + (\mu_k^c - \bar{c}_{k,t})\mu_k^r}{\bar{c}_{k,t} \cdot \mu_k^c} \\ &< \frac{\epsilon_{k,t}}{\bar{c}_{k,t}} + \frac{\epsilon_{k,t}}{\bar{c}_{k,t}} \cdot \frac{\mu_k^r}{\mu_k^c} \leq (1+\kappa) \cdot \frac{\epsilon_{k,t}}{\bar{c}_{k,t}} = E_{k,t}.\end{aligned}$$

As for event (iii), observe that if $\text{obj}_{k^*} < \text{obj}_k + 2E_{k,t}$ and $n_{k,t} \geq \beta_k \cdot \ln(t)$ then we have $\bar{c}_{k,t} \leq \mu_k^c/2$, which happens with probability at most $\sim 1/t^2$ by Lemma 4.1 given that arm k has already been pulled at least $\sim \ln(t)/(\mu_c^k)^2$ times. □

Building on the last two results, we derive a distribution-dependent regret bound which improves upon the one derived in [102]: the decision maker is only assumed to know κ, as opposed to a lower bound on ϵ, ahead of round 1. This is more natural in bidding applications as detailed in the last paragraph of this section. This bound generalizes the one obtained by [12] when time is the only scarce resource.

Theorem 4.1. *We have:*

$$R_B \leq 2^6 \lambda^2 \cdot \Big(\sum_{k \in \{1,\cdots,K\} \mid \Delta_k > 0} \frac{1}{\mu_k^c \cdot \Delta_k} \Big) \cdot \ln\Big(\frac{B+1}{\epsilon}\Big) + O(1).$$

Sketch of proof. We build upon (4.9):

$$\begin{aligned} R_B &\leq B \cdot \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\ &= B \cdot \frac{\mu_{k^*}^r}{\mu_{k^*}^c} - \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\ &= \frac{\mu_{k^*}^r}{\mu_{k^*}^c} \cdot (B - \sum_{k \mid \Delta_k = 0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}]) - \sum_{k \mid \Delta_k > 0} \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1). \end{aligned}$$

By definition of τ^*, the resource is exhausted at time τ^*, i.e. $B \leq \sum_{t=1}^{\tau^*} c_{a_t,t}$. Taking expectations on both sides yields $B \leq \sum_{k=1}^{K} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}]$. Plugging this last inequality back into the regret bound, we get:

$$R_B \leq \sum_{k \mid \Delta_k > 0} \mu_k^c \cdot \Delta_k \cdot \mathbb{E}[n_{k,\tau^*}] + O(1).$$

Using the upper bound of Lemma 4.4, the concavity of the logarithmic function, and Lemma 4.5, we derive:

$$R_B \leq 2^6 \lambda^2 \cdot (\sum_{k \mid \Delta_k > 0} \frac{1}{\mu_k^c \cdot \Delta_k}) \cdot \ln(\frac{B+1}{\epsilon}) + \frac{4\pi^2}{3\epsilon^2} \cdot (\sum_{k \mid \Delta_k > 0} \mu_k^c \cdot \Delta_k) + O(1)$$

which yields the claim since $\Delta_k \leq \mu_{k^*}^r / \mu_{k^*}^c \leq \kappa$ and $\mu_k^c \leq 1$ for any arm k. □

Observe that the set of optimal arms, namely $\operatorname{argmax}_{k=1,\cdots,K} \mu_k^r / \mu_k^c$, does not depend on B and that Δ_k is a constant independent of B for any suboptimal arm. We conclude that $R_B = O(K \cdot \ln(B)/\Delta)$ with $\Delta = \min_{k \in \{1,\cdots,K\} \mid \Delta_k > 0} \Delta_k$. Interestingly, the algorithm we propose does not rely on B to achieve this regret bound, much like what happens for UCB1 with the time horizon, see [12]. This result is optimal up to constant factors as the standard MAB problem is a special case of the framework considered in this section, see [60] for a proof of a lower bound in this context. It is possible to improve the constant factors when the consumption of the resource is deterministic as we can take $\lambda = 1$ in this scenario and the resulting regret bound is scale-free, see Section C.1 of the Appendix. Building on Theorem 4.1, we can also derive a near-optimal distribution-free regret bound in the same

fashion as for UCB1.

Theorem 4.2. *We have:*

$$R_B \leq 8\lambda \cdot \sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln(\frac{B+1}{\epsilon})} + O(1).$$

Proof. To get the distribution-free bound, we start from the penultimate inequality derived in the proof sketch of Theorem 4.1 and apply Lemma 4.5 only if Δ_k is big enough, noting that:

$$\sum_{k=1}^{K} \mathbb{E}[n_{k,\tau^*}] = \mathbb{E}[\tau^*] \leq (B+1)/\epsilon.$$

Specifically, we have:

$$\begin{aligned}
R_B &\leq \sup_{\substack{(n_1,\cdots,n_K)\geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \{ \sum_{k \mid \Delta_k > 0} \min(\mu_k^c \cdot \Delta_k \cdot n_k, 2^6\lambda^2 \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\mu_k^c \cdot \Delta_k} + \frac{4\pi^2}{3\epsilon^2} \cdot \mu_k^c \cdot \Delta_k) \} + O(1) \\
&\leq \sup_{\substack{(n_1,\cdots,n_K)\geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \{ \sum_{k \mid \Delta_k > 0} \min(\mu_k^c \cdot \Delta_k \cdot n_k, 2^6\lambda^2 \cdot \frac{\ln(\frac{B+1}{\epsilon})}{\mu_k^c \cdot \Delta_k}) \} + K \cdot \frac{4\pi^2\kappa}{3\epsilon^2} + O(1) \\
&\leq \sup_{\substack{(n_1,\cdots,n_K)\geq 0 \\ \sum_{k=1}^K n_k \leq \frac{B+1}{\epsilon}}} \{ \sum_{k \mid \Delta_k > 0} \sqrt{2^6\lambda^2 \cdot n_k \cdot \ln(\frac{B+1}{\epsilon})} \} + O(1) \\
&\leq 8\lambda \cdot \sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln(\frac{B+1}{\epsilon})} + O(1),
\end{aligned}$$

where the second inequality is obtained with $\Delta_k \leq \mu_{k^*}^r / \mu_{k^*}^c \leq \kappa$ and $\mu_k^c \leq 1$, the third inequality is derived by maximizing on $(\mu_k^c \cdot \Delta_k) \geq 0$ for all arms k, and the last inequality is obtained with the Cauchy$-$Schwarz inequality. □

We conclude that $R_B = O(\sqrt{K \cdot B \cdot \ln(B)})$, where the hidden constant factors are independent of the underlying distributions $(\nu_k)_{k=1,\cdots,K}$.

Applications. Assumption 4.3 is natural for bidding in repeated second-price auctions when the auctioneer sets a reserve price R (this is common practice in sponsored search

auctions). Indeed, then we have:

$$\begin{aligned}\mathbb{E}[c_{k,t}] &= \mathbb{E}[m_t \cdot \mathbb{1}_{b_k \geq m_t}] \\ &\geq R \cdot \mathbb{E}[\mathbb{1}_{b_k \geq m_t}] \\ &\geq R \cdot \mathbb{E}[v_t \cdot \mathbb{1}_{b_k \geq m_t}] = R \cdot \mathbb{E}[r_{k,t}],\end{aligned}$$

for any arm $k \in \{1, \cdots, K\}$ and Assumption 4.3 is satisfied with $\kappa = 1/R$.

4.5 Arbitrarily Many Limited Resources whose Consumptions are Deterministic

In this section, we study the case of multiple limited resources when the amounts of resources consumed as a result of pulling an arm are deterministic and globally constrained by prescribed budgets $(B(i))_{i=1,\cdots,C}$, where C is the number of resources. Note that time need not be a constraint. Because the amounts of resources consumed are deterministic, we can substitute the notation $\mu_k^c(i)$ with $c_k(i)$ for any arm $k \in \{1, \cdots, K\}$ and any resource $i \in \{1, \cdots, C\}$. We point out that the stopping time need not be deterministic as the decision to select an arm at any round is based on the past realizations of the rewards. We define $\rho \leq \min(C, K)$ as the rank of the matrix $(c_k(i))_{1 \leq k \leq K, 1 \leq i \leq C}$.

Specification of the algorithm. We implement UCB-Simplex with an initialization step which consists in pulling each arm ρ times. The motivation behind this step is mainly technical and is simply meant to have:

$$n_{k,t} \geq \rho + \sum_{x \in \mathcal{B} \mid k \in \mathcal{K}_x} n_{k,t}^x \quad \forall t \in \mathbb{N}, \forall k \in \{1, \cdots, K\}. \tag{4.10}$$

Compared to Section 4.4, we choose to take $\lambda = 1$ and $\eta_i = 0$ for any $i \in \{1, \cdots, C\}$. As a result and since the amounts of resource consumption are deterministic, the exploration (resp. exploitation) terms defined in Section 4.3 specialize to $\mathrm{obj}_{x,t} = \sum_{k=1}^K \xi_k^x \cdot \bar{r}_{k,t}$ (resp. $E_{x,t} = \sum_{k=1}^K \xi_k^x \cdot \epsilon_{k,t}$). Compared to the case of a single resource, we are required to specify

the load balancing algorithms involved in Step-Load-Balance of UCB-Simplex as a feasible basis selected at Step-Simplex may involve several arms. Although Step-Simplex will also need to be specified in Sections 4.6 and 4.7, designing good load balancing algorithms is arguably easier here as the optimal load balance is known for each basis from the start. Nonetheless, one challenge remains: we can never identify the (possibly many) optimal bases of (4.3) with absolute certainty. As a result, any basis selected at Step-Simplex should be treated as potentially optimal when balancing the load between the arms involved in this basis, but this inevitably causes some interference issues as an arm may be involved in several bases, and worst, possibly several optimal bases. Therefore, one point that will appear to be of particular importance in the analysis is the use of load balancing algorithms that are decoupled from one another, in the sense that they do not rely on what happened when selecting other bases. More specifically, we use the following class of load balancing algorithms.

Algorithm: Load balancing algorithm $\mathcal{A}_x$ for a feasible basis $x \in \mathcal{B}$

If basis x is selected at time t, pull any arm $k \in \mathcal{K}_x$ such that $n^x_{k,t} \leq n_{x,t} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l}$.

The load balancing algorithms $\mathcal{A}_x$ thus defined are decoupled because, for each basis, the number of times an arm has been pulled when selecting any other basis is not taken into account. The following lemma shows that $\mathcal{A}_x$ is always well defined and guarantees that the ratios $(n^x_{k,t}/n^x_{l,t})_{k,l\in\mathcal{K}_x}$ remain close to the optimal ones $(\xi^x_k/\xi^x_l)_{k,l\in\mathcal{K}_x}$ at all times.

Lemma 4.6. *For any basis $x \in \mathcal{B}$, $\mathcal{A}_x$ is well defined and moreover, at any time t and for any arm $k \in \mathcal{K}_x$, we have:*

$$n_{x,t} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l} - \rho \leq n^x_{k,t} \leq n_{x,t} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l} + 1,$$

while $n^x_{k,t} = 0$ for any arm $k \notin \mathcal{K}_x$.

Proof. We need to show that there always exists an arm $k \in \mathcal{K}_x$ such that $n^x_{k,t} \leq n_{x,t} \cdot \xi^x_k / \sum_{l=1}^K \xi^x_l$. Suppose there is none, we have:

$$n_{x,t} - = \sum_{k\in\mathcal{K}_x} n^x_{k,t} > \sum_{k\in\mathcal{K}_x} n_{x,t} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l} = n_{x,t},$$

a contradiction. Moreover, we have, at any time t and for any arm $k \in \mathcal{K}_x$, $n^x_{k,t} \leq n_{x,t} \cdot \xi^x_k / \sum_{l=1}^K \xi^x_l + 1$. Indeed, suppose otherwise and define $t^* \leq t$ as the last time arm k was pulled, we have:

$$n^x_{k,t^*} = n^x_{k,t} - 1 > n_{x,t} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l} \geq n_{x,t^*} \cdot \frac{\xi^x_k}{\sum_{l=1}^K \xi^x_l},$$

which shows, by definition, that arm k could not have been pulled at time t^*. We also derive as a byproduct that, at any time t and for any arm $k \in \mathcal{K}_x$, $n_{x,t} \cdot \xi^x_k / \sum_{l=1}^K \xi^x_l - \rho \leq n^x_{k,t}$ since $n_{x,t} = \sum_{k \in \mathcal{K}_x} n^x_{k,t}$ and since a basis involves at most ρ arms. □

Observe that the load balancing algorithms $\mathcal{A}_x$ run in time $O(1)$ but may require a memory storage capacity exponential in C and polynomial in K, although always bounded by $O(B)$ (because we do not need to keep track of $n^x_{k,t}$ if x has never been selected). In practice, only a few bases will be selected at Step-Simplex, so that a hash table is an appropriate data structure to store the sequences $(n^x_{k,t})_{k \in \mathcal{K}_x}$. In Section C.1 of the Appendix, we introduce another class of load balancing algorithms that is both time and memory efficient while still guaranteeing logarithmic regret bounds (under an additional assumption) but no distribution-free regret bounds.

Regret Analysis. We use the shorthand notation:

$$\epsilon = \min_{\substack{k=1,\cdots,K \\ i=1,\cdots,C \\ \text{with } c_k(i)>0}} c_k(i).$$

Note that $\epsilon < \infty$ under Assumption 4.2. We discard the initialization step in the theoretical study because the amounts of resources consumed are bounded by a constant and the total reward obtained is non-negative and not taken into account in the decision maker's total payoff. We again start by estimating the expected time horizon.

Lemma 4.7. *For any non-anticipating algorithm, we have:* $\mathbb{E}[\tau^*] \leq \frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1$.

Proof. By definition of τ^*, we have $\sum_{t=1}^{\tau^*-1} c_{a_t,t}(i) \leq B(i)$ for any resource $i \in \{1, \cdots, C\}$. Summing up these inequalities and using Assumption 4.2 and the fact that $(c_{k,t}(i))_{t=1,\cdots,T}$

are deterministic, we get $(\tau^* - 1) \cdot \epsilon \leq \sum_{i=1}^{C} B(i)$. Taking expectations on both sides and using Assumption 4.1 yields the result. □

We follow by bounding the number of times any suboptimal basis can be selected at Step-Simplex in the same spirit as in Section 4.4.

Lemma 4.8. *For any suboptimal basis $x \in \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,\tau^*}] \leq 16\rho \cdot (\sum_{k=1}^{K} \xi_k^x)^2 \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_x)^2} + \rho \cdot \frac{\pi^2}{3}.$$

Sketch of proof. We use the shorthand notation $\beta_x = 8\rho \cdot (\sum_{k=1}^{K} \xi_k^x / \Delta_x)^2$. The proof is along the same lines as for Lemma 4.5. First note that we may assume, without loss of generality, that x has been selected at least $\beta_x \cdot \ln(t)$ times at any time t up to an additive term of $2\beta_x \cdot \mathbb{E}[\ln(\tau^*)]$ in the final inequality. We then just have to bound by a constant the probability that x is selected at any time t given that $n_{x,t} \geq \beta_x \cdot \ln(t)$. If x is selected at time t, x is an optimal basis to (4.8). Since the amounts of resources consumed are deterministic, x^* is feasible to (4.8) at time t, which implies that $\text{obj}_{x,t} + E_{x,t} \geq \text{obj}_{x^*,t} + E_{x^*,t}$. This can only happen if either: (i) $\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}$, (ii) $\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}$, or (iii) $\text{obj}_{x^*} < \text{obj}_x + 2E_{x,t}$. First note that (iii) is impossible because, assuming this is the case, we would have:

$$\begin{aligned}
\frac{\Delta_x}{2} &< \sum_{k \in \mathcal{K}_x} \xi_k^x \cdot \sqrt{\frac{2\ln(t)}{n_{k,t}}} \\
&\leq \sum_{k \in \mathcal{K}_x} \xi_k^x \cdot \sqrt{\frac{2\ln(t)}{\rho + n_{k,t}^x}} \\
&\leq \sqrt{\sum_{k \in \mathcal{K}_x} \xi_k^x} \cdot \sum_{k \in \mathcal{K}_x} \sqrt{\xi_k^x} \cdot \sqrt{\frac{2\ln(t)}{n_{x,t}}} \\
&\leq \sqrt{\rho} \cdot \sum_{k \in \mathcal{K}_x} \xi_k^x \cdot \sqrt{\frac{2}{\beta_x}} = \frac{\Delta_x}{2},
\end{aligned}$$

where we use (4.10), Lemma 4.6 for each $k \in \mathcal{K}_x$, the Cauchy−Schwarz inequality, and the fact that a basis involves at most ρ arms. Along the same lines as for Lemma 4.5, the probability of events (i) and (ii) can be bounded by $\sim \rho/t^2$ in the same fashion, irrespective

of how many times x and x^* have been selected in the past. For example for event (i), this is because if $\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}$, then there must exist $k \in \mathcal{K}_x$ such that $\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}$, but any of these events have individual probability at most $\sim 1/t^2$ by Lemma 4.1. Indeed otherwise, if $\bar{r}_{k,t} < \mu_k^r + \epsilon_{k,t}$ for all $k \in \mathcal{K}_x$, we would reach a contradiction:

$$\mathrm{obj}_{x,t} - \mathrm{obj}_x = \sum_{k \in \mathcal{K}_x} (\bar{r}_{k,t} - \mu_k^r) \cdot \xi_k^x < \sum_{k \in \mathcal{K}_x} \epsilon_{k,t} \cdot \xi_k^x = E_{x,t}.$$

□

Lemma 4.8 used in combination with Lemma 4.7 shows that a suboptimal basis is selected at most $O(\ln(B))$ times. To establish the regret bound, we need to lower bound the expected payoff derived when selecting any of the optimal bases. This is more involved than in Section 4.4 because the load balancing step comes into play at this stage.

Theorem 4.3. *We have:*

$$R_{B(1),\cdots,B(C)} \leq 16 \frac{\rho \cdot \sum_{i=1}^C b(i)}{\epsilon} \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \Big) \cdot \ln\Big(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1 \Big) + O(1).$$

Sketch of proof. The proof proceeds along the same lines as for Theorem 4.1. We build upon (4.4):

$$\begin{aligned} R_{B(1),\cdots,B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\ &= B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}^x] + O(1). \end{aligned}$$

Using the properties of the load balancing algorithm established in Lemma 4.6, we derive:

$$\begin{aligned} R_{B(1),\cdots,B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \{ \frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^K \xi_k^x} \cdot (\sum_{k=1}^K \mu_k^r \cdot \xi_k^x) \} + O(1) \\ &= (\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*}) \cdot (B - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^K \xi_k^x}) \\ &\quad - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \{ (\sum_{k=1}^K \mu_k^r \cdot \xi_k^x) \cdot \frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^K \xi_k^x} \} + O(1). \end{aligned}$$

Now observe that, by definition, at least one resource is exhausted at time τ^*. Hence, there exists $i \in \{1, \cdots, C\}$ such that:

$$\begin{aligned} B(i) &\leq \sum_{x \in \mathcal{B}} \sum_{k \in \mathcal{K}_x} c_k(i) \cdot n^x_{k,\tau^*} \\ &\leq O(1) + \sum_{x \in \mathcal{B}} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi^x_k} \cdot \sum_{k \in \mathcal{K}_x} c_k(i) \cdot \xi^x_k \\ &\leq O(1) + b(i) \cdot \sum_{x \in \mathcal{B}} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi^x_k}, \end{aligned}$$

where we use Lemma 4.6 again and the fact that any basis $x \in \mathcal{B}$ satisfies all the constraints of (4.3). We conclude that:

$$\sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi^x_k} \geq B - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi^x_k} + O(1).$$

Taking expectations on both sides and plugging the result back into the regret bound yields:

$$R_{B(1),\cdots,B(C)} \leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\Delta_x}{\sum_{k=1}^K \xi^x_k} \cdot \mathbb{E}[n_{x,\tau^*}] + O(1).$$

Using Lemma 4.7, Lemma 4.8, and the concavity of the logarithmic function, we obtain:

$$\begin{aligned} R_{B(1),\cdots,B(C)} \leq\ & 16\rho \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\sum_{k=1}^K \xi^x_k}{\Delta_x} \Big) \cdot \ln\Big(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1 \Big) \\ &+ \frac{\pi^2}{3} \rho \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\Delta_x}{\sum_{k=1}^K \xi^x_k} \Big) + O(1) \end{aligned}$$

which yields the claim since $\Delta_x \leq \sum_{k=1}^K \mu^r_k \cdot \xi^{x^*}_k \leq \sum_{i=1}^C b(i)/\epsilon$, $\sum_{k=1}^K \xi^x_k \geq b$, and $\sum_{k=1}^K \xi^x_k \leq \sum_{i=1}^C b(i)/\epsilon$. □

We point out that, if time is a limited resource with a time horizon T, we can also derive the (possibly better) regret bound:

$$R_{B(1),\cdots,B(C)} \leq 16\rho \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \Big) \cdot \ln(T) + O(1).$$

Since the number of feasible bases to (4.3) is at most $\binom{K+\rho}{K} \leq 2K^\rho$, we get the distribution-dependent regret bound $O(\rho \cdot K^\rho \cdot \ln(B)/\Delta)$ where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$. In Section C.1 of the Appendix, we introduce an alternative class of load balancing algorithms which yields a better dependence on K and C with a regret bound of order $O(\rho^3 \cdot K \cdot \ln(B)/\Delta^2)$ provided that there is a unique optimal basis to (4.3). Along the sames lines as in Section 4.4, the distribution-dependent bound of Theorem 4.3 almost immediately implies a distribution-free one.

Theorem 4.4. *We have:*

$$R_{B(1),\cdots,B(C)} \leq 4\sqrt{\rho \cdot |\mathcal{B}| \cdot (\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1) \cdot \ln(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1)} + O(1).$$

Sketch of proof. The proof is along the same lines as for the case of a single limited resource, we start from the penultimate inequality derived in the proof sketch of Theorem 4.3 and apply Lemma 4.8 only if Δ_x is big enough, taking into account the fact that $\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x,\tau^*}] \leq \mathbb{E}[\tau^*] \leq \sum_{i=1}^C b(i) \cdot B/\epsilon + 1$. □

We conclude that $R_{B(1),\cdots,B(C)} = O(\sqrt{\rho \cdot K^\rho \cdot B \cdot \ln(B)})$, where the hidden constant factors are independent of the underlying distributions $(\nu_k)_{k=1,\cdots,K}$. If time is a limited resource, we can also derive the (possibly better) regret bound:

$$R_{B(1),\cdots,B(C)} \leq 4\sqrt{\rho \cdot |\mathcal{B}| \cdot T \cdot \ln(T)} + O(1).$$

In any case, we stress that the dependence on K and C is not optimal since the authors of [16] and [6] obtain a $\tilde{O}(\sqrt{K \cdot B})$ bound on regret, where the $\tilde{O}$ notation hides factors logarithmic in B. Observe that the regret bounds derived in this section do not vanish with b. This can be remedied by strengthening Assumption 4.2, additionally assuming that $c_{k,t}(i) > 0$ for any arm $k \in \{1, \cdots, K\}$ and resource $i \in \{1, \cdots, C\}$. In this situation, we can refine the analysis and substitute $\sum_{i=1}^C b(i)$ with b in the regret bounds of Propositions 4.3 and 4.4 which become scale-free.

Applications. BwK problems where the amounts of resources consumed as a result of pulling an arm are deterministic find applications in network revenue management of perishable goods, shelf optimization of perishable goods, and wireless sensor networks, as detailed in Section 4.2.

4.6 A Time Horizon and Another Limited Resource

In this section, we investigate the case of two limited resources, one of which is assumed to be time, with a time horizon T, while the consumption of the other is stochastic and constrained by a global budget B. To simplify the notations, we omit the indices identifying the resources since the second limited resource is time and we write μ_k^c, $c_{k,t}$, $\bar{c}_{k,t}$, B, and T as opposed to $\mu_k^c(1)$, $c_{k,t}(1)$, $\bar{c}_{k,t}(1)$, $B(1)$, and $B(2)$. Moreover, we refer to resource $i = 1$ as "the" resource. Observe that, in the particular setting considered in this section, $\tau^* = \min(\tau(B), T+1)$ with $\tau(B) = \min\{t \in \mathbb{N} \mid \sum_{\tau=1}^{t} c_{a_\tau,\tau} > B\}$. Note that the budget constraint is not limiting if $B \geq T$, in which case the problem reduces to the standard MAB problem. Hence, without loss of generality under Assumption 4.1, we assume that the budget scales linearly with time, i.e. $B = b \cdot T$ for a fixed constant $b \in (0, 1)$, and we study the asymptotic regime $T \to \infty$. Motivated by technical considerations, we make two additional assumptions for the particular setting considered in this section that are perfectly reasonable in many applications, such as in repeated second-price auctions, dynamic procurement, and dynamic pricing, as detailed in the last paragraph of this section.

Assumption 4.4. *There exists $\sigma > 0$ such that $\mu_k^r \leq \sigma \cdot \mu_k^c$ for any arm $k \in \{1, \cdots, K\}$.*

Assumption 4.5. *The decision maker knows $\kappa > 0$ such that:*

$$|\mu_k^r - \mu_l^r| \leq \kappa \cdot |\mu_k^c - \mu_l^c| \quad \forall (k, l) \in \{1, \cdots, K\}^2,$$

ahead of round 1.

Note that σ, as opposed to κ, is not assumed to be known to the decision maker. Assumption 4.4 is relatively weak and is always satisfied in practical applications. In particular, note that if $\mu_k^c > 0$ for all arms $k \in \{1, \cdots, K\}$, we can take $\sigma = 1/\min_{k=1,\cdots,K} \mu_k^c$.

Specification of the algorithm. We implement UCB-Simplex with $\lambda = 1 + 2\kappa$, $\eta_1 = 1$, $\eta_2 = 0$, and an initialization step which consists in pulling each arm once. Because the amount of resource consumed at each round is a random variable, a feasible basis for (4.8) may not be feasible for (4.3) and conversely. In particular, x^* may not be feasible for (4.8), thus effectively preventing it from being selected at Step-Simplex, and an infeasible basis for (4.3) may be selected instead. This is in contrast to the situation studied in Section 4.5 and this motivates the choice $\eta_1 > 0$ to guarantee that any feasible solution to (4.3) will be feasible to (4.8) with high probability at any round t.

Just like in Section 4.5, we need to specify Step-Load-Balance because a basis selected at Step-Simplex may involve up to two arms. To simplify the presentation, we introduce a dummy arm $k = 0$ with reward 0 and resource consumption 0 (pulling this arm corresponds to skipping the round) and K dummy arms $k = K + 1, \cdots, 2K$ with reward identical to arm $K - k$ but resource consumption 1 so that any basis involving a single arm can be mapped to an "artificial" one involving two arms. Note, however, that we do not introduce new variables ξ_k in (4.8) for these arms as they are only used for mathematical convenience in Step-Load-Balance once a basis has been selected at Step-Simplex. Specifically, if a basis x_t involving a single arm determined by $\mathcal{K}_{x_t} = \{k_t\}$ and $\mathcal{C}_{x_t} = \{1\}$ (resp. $\mathcal{C}_{x_t} = \{2\}$) is selected at Step-Simplex, we map it to the basis x'_t determined by $\mathcal{K}_{x'_t} = \{0, k_t\}$ (resp. $\{k_t, K + k_t\}$) and $\mathcal{C}_{x'_t} = \{1, 2\}$. We then use a load balancing algorithm specific to this basis, denoted by $\mathcal{A}_{x_t}$, to determine which of the two arms in $\mathcal{K}_{x'_t}$ to pull. Similarly as in Section 4.5, using load balancing algorithms that are decoupled from one another is crucial because the decision maker can never identify the optimal bases with absolute certainty. This implies that each basis should be treated as potentially optimal when balancing the load between the arms, but this inevitably causes interference issues as an arm may be involved in several bases. Compared to Section 4.5, we face an additional challenge when designing the load balancing algorithms: the optimal load balances are initially unknown to the decision maker. It turns out that we can still approximately achieve the unknown optimal load balances by enforcing that, at any round t, the total amount of resource consumed remains close to the pacing target $b \cdot t$ with high probability, as precisely described below.

Algorithm: Load balancing algorithm $\mathcal{A}_x$ for any basis x

For any time period t, define $b_{x,t}$ as the total amount of resource consumed when selecting x in the past $t-1$ rounds. Suppose that x is selected at time t. Without loss of generality, write $\mathcal{K}_x = \{k, l\}$ with $\bar{c}_{k,t} - \epsilon_{k,t} \geq \bar{c}_{l,t} - \epsilon_{l,t}$. Pull arm k if $b_{x,t} \leq n_{x,t} \cdot b$ and pull arm l otherwise.

Observe that a basis x with $\mathcal{K}_x = \{k, l\}$ is feasible for (4.3) if either $\mu_k^c > b > \mu_l^c$ or $\mu_l^c > b > \mu_k^c$. Assuming we are in the first situation, the exploration and exploitation terms defined in Section 4.3 specialize to:

$$\mathrm{obj}_{x,t} = \xi_{l,t}^x \cdot \bar{r}_{l,t} + \xi_{k,t}^x \cdot \bar{r}_{k,t} \text{ and } E_{x,t} = \lambda \cdot (\xi_{l,t}^x \cdot \epsilon_{l,t} + \xi_{k,t}^x \cdot \epsilon_{k,t})$$

with:

$$\xi_{l,t}^x = \frac{(\bar{c}_{k,t} - \epsilon_{k,t}) - b}{(\bar{c}_{k,t} - \epsilon_{k,t}) - (\bar{c}_{l,t} - \epsilon_{l,t})} \text{ and } \xi_{k,t}^x = \frac{b - (\bar{c}_{l,t} - \epsilon_{l,t})}{(\bar{c}_{k,t} - \epsilon_{k,t}) - (\bar{c}_{l,t} - \epsilon_{l,t})},$$

provided that $\bar{c}_{k,t} - \epsilon_{k,t} > b > \bar{c}_{l,t} - \epsilon_{l,t}$. Moreover, their offline counterparts are given by:

$$\mathrm{obj}_x = \xi_l^x \cdot \mu_l^r + \xi_k^x \cdot \mu_k^r,\ \xi_l^x = \frac{\mu_k^c - b}{\mu_k^c - \mu_l^c},\ \text{and } \xi_k^x = \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c}.$$

Regret Analysis. We start by pointing out that, in degenerate scenarios, using the linear relaxation (4.3) as an upper bound on $\mathrm{ER}_{\mathrm{OPT}}(B, T)$ already dooms us to $\Omega(\sqrt{T})$ regret bounds. Precisely, if there exists a unique optimal basis x^* to (4.3) that happens to be degenerate, i.e. $\mathcal{K}_{x^*} = \{k^*\}$ (pre-mapping) with $\mu_{k^*}^c = b$, then, in most cases, $T \cdot \mathrm{obj}_{x^*} \geq \mathrm{ER}_{\mathrm{OPT}}(B, T) + \Omega(\sqrt{T})$ as shown below.

Lemma 4.9. *If there exists $k^* \in \{1, \cdots, K\}$ such that: (i) the i.i.d. process $(c_{k^*,t})_{t\in\mathbb{N}}$ has positive variance, (ii) $\mu_{k^*}^c = b$, and (iii) $(\xi_k)_{k=1,\cdots,K}$ determined by $\xi_{k^*} = 1$ and $\xi_k = 0$ for $k \neq k^*$ is the unique optimal solution to* (4.3)*, then there exists a subsequence of $(\frac{T\cdot \mathrm{obj}_{x^*} - \mathrm{ER}_{\mathrm{OPT}}(B,T)}{\sqrt{T}})_{T\in\mathbb{N}}$ that does not converge to* 0.

Sketch of proof. For any time horizon $T \in \mathbb{N}$ and any arm $k \in \{1, \cdots, K\}$, we denote by $n_{k,T}^{\mathrm{opt}}$ the expected number of times arm k is pulled by the optimal non-anticipating algorithm when the time horizon is T and the budget is $B = b \cdot T$. We expect that consistently

pulling arm k^* is near-optimal. Unfortunately, this is also nothing more than an i.i.d. strategy which implies, along the same lines as in Lemma 4.3, that $\mathbb{E}[\tau^*] = T - \Omega(\sqrt{T})$ so that the total expected payoff is $\mathbb{E}[\tau^*] \cdot \mu^r_{k^*} = T \cdot \mathrm{obj}_{x^*} - \Omega(\sqrt{T})$. To formalize these ideas, we study two cases: $T - n^{\mathrm{opt}}_{k^*,T} = \Omega(\sqrt{T})$ (Case A) and $T - n^{\mathrm{opt}}_{k^*,T} = o(\sqrt{T})$ (Case B) and we show that $\mathrm{ER}_{\mathrm{OPT}}(B,T) = T \cdot \mathrm{obj}_{x^*} - \Omega(\sqrt{T})$ in both cases. In Case A, this is because the optimal value of (4.3) remains an upper bound on the maximum total expected payoff if we add the constraint $\xi_{k^*} \leq n^{\mathrm{opt}}_{k^*,T}/T$ to the linear program (4.3) by definition of $n^{\mathrm{opt}}_{k^*,T}$. Since the constraint $\xi_{k^*} \leq 1$ is binding for (4.3), the optimal value of this new linear program can be shown to be smaller than $\mathrm{obj}_{x^*} - \Omega((T - n^{\mathrm{opt}}_{k^*,T})/T)$ (by strong duality and strict complementary slackness). In Case B, up to an additive term of order $o(\sqrt{T})$ in the final bound, the optimal non-anticipating algorithm is equivalent to consistently pulling arm k^*, which is an i.i.d. strategy so the study is very similar to that of Lemma 4.3. □

Dealing with these degenerate scenarios thus calls for a completely different approach than the one taken on in the BwK literature and we choose instead to rule them out in such a way that there can be no degenerate optimal basis to (4.3).

Assumption 4.6. *We have* $|\mu^c_k - b| > 0$ *for any arm* $k \in \{1, \cdots, K\}$.

We use the shorthand notation $\epsilon = \min_{k=1,\cdots,K} |\mu^c_k - b|$. Assumption 4.6 is equivalent to assuming that any basis for (4.3) is non-degenerate. This assumption can be relaxed to some extent at the price of more technicalities. However, in light of Lemma 4.9, the minimal assumption is that there is no degenerate optimal basis to (4.3). As a final remark, we stress that Assumption 4.6 is only necessary to carry out the analysis but Step-Simplex can be implemented in any case as ϵ is not assumed to be known to the decision maker.

We are now ready to establish regret bounds. Without loss of generality, we can assume that any pseudo-basis for (4.3) involves two arms, one of which may be a dummy arm introduced in the specification of the algorithm detailed above. As stressed at the beginning of this section, UCB-Simplex may sometimes select an infeasible basis or even a pseudo-basis x with $\det(A_x) = 0$ (i.e. such that $\mu^c_k = \mu^c_l$ assuming $\mathcal{K}_x = \{k, l\}$). Interestingly the load balancing algorithm plays a crucial role to guarantee that this does not happen very often.

Lemma 4.10. *For any basis $x \notin \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,T}] \leq \frac{2^6}{\epsilon^3} \cdot \ln(T) + \frac{10\pi^2}{3\epsilon^2}.$$

The same inequality holds if x is a pseudo-basis but not a basis for (4.3).

Proof. We use the shorthand notation $\beta_x = 2^5/\epsilon^3$. Without loss of generality, we can assume that $\mathcal{K}_x = \{k, l\}$ with $\mu_k^c, \mu_l^c > b$ (the situation is symmetric if the reverse inequality holds). Along the same lines as in Lemma 4.5, we only have to bound by a constant the probability that x is selected at any round t given that x has already been selected at least $\beta_x \cdot \ln(t)$ times. If x is selected at round t and $n_{x,t} \geq \beta_x \cdot \ln(t)$, then $b_{x,t}$ must be larger than $n_{x,t} \cdot b$ by at least a margin of $\sim 1/\epsilon^2 \cdot \ln(t)$ with high probability given that $\mu_k^c, \mu_l^c > b$. Moreover, at least one arm, say k, has been pulled at least $\sim 1/\epsilon^3 \cdot \ln(t)$ times and, as a result, $\bar{c}_{k,\tau} - \epsilon_{k,\tau} \geq b$ with high probability for the last $s \sim 1/\epsilon^2 \cdot \ln(t)$ rounds $\tau = \tau_1, \cdots, \tau_s$ where x was selected. This implies that arm l must have been pulled at least $\sim 1/\epsilon^2 \cdot \ln(t)$ times already by definition of the load balancing algorithm but then we have $\bar{c}_{l,t} - \epsilon_{l,t} \geq b$ with high probability and x cannot be feasible for (4.8) at time t with high probability. □

What remains to be done is to: (i) show that suboptimal bases are selected at most $O(\ln(T))$ times and (ii) lower bound the expected total payoff derived when selecting any of the optimal bases. The major difficulty lies in the fact that the amounts of resource consumed, the rewards obtained, and the stopping time are correlated in a non-trivial way through the budget constraint and the decisions made in the past. This makes it difficult to study the expected total payoff derived when selecting optimal bases independently from the amounts of resource consumed and the rewards obtained when selecting suboptimal ones. However, a key point is that, by design, the pulling decision made at Step-Load-Balance is based solely on the past history associated with the basis selected at Step-Simplex because the load balancing algorithms are decoupled. For this reason, the analysis proceeds in two steps irrespective of the number of optimal bases. In a first step, we show that, for any basis x for (4.3), the amount of resource consumed per round when selecting x remains close to the pacing target b with high probability. This enables us to show that the ra-

tios $(\mathbb{E}[n^x_{k,T}]/\mathbb{E}[n^x_{l,T}])_{k,l\in\mathcal{K}_x}$ are close to the optimal ones $(\xi^x_k/\xi^x_l)_{k,l\in\mathcal{K}_x}$, as precisely stated below.

Lemma 4.11. *For any basis $x \in \mathcal{B}$ and time period t, we have:*

$$\mathbb{P}[|b_{x,t} - n_{x,t}\cdot b| \geq u + (\frac{4}{\epsilon})^2\cdot\ln(t)] \leq \frac{4}{\epsilon^2}\cdot\exp(-\epsilon^2\cdot u) + \frac{8}{\epsilon^2\cdot t^2} \quad \forall u \geq 1,$$

which, in particular, implies that:

$$\begin{aligned}\mathbb{E}[n^x_{k,T}] &\geq \xi^x_k\cdot\mathbb{E}[n_{x,T}] - 13/\epsilon^5 - 16/\epsilon^3\cdot\ln(T)\\ \mathbb{E}[n^x_{l,T}] &\geq \xi^x_l\cdot\mathbb{E}[n_{x,T}] - 13/\epsilon^5 - 16/\epsilon^3\cdot\ln(T).\end{aligned} \tag{4.11}$$

Sketch of proof. Without loss of generality, we can assume that $\mathcal{K}_x = \{k,l\}$ with $\mu^c_k > b > \mu^c_l$. Observe that, if the decision maker knew that $\mu^c_k > b > \mu^c_l$ ahead of round 1, he would always pull the "correct" arm in order not to deviate from the pacing target $n_{x,t}\cdot b$ and $|b_{x,t} - n_{x,t}\cdot b|$ would remain small with high probability given Assumption 4.6. However, because this information is not available ahead of round 1, the decision maker is led to pull the incorrect arm when arm k and l are swapped, in the sense that $\bar{c}_{k,t} - \epsilon_{k,t} \leq \bar{c}_{l,t} - \epsilon_{l,t}$. Fortunately, at any time t, there could have been at most $1/\epsilon^2\cdot\ln(t)$ swaps with probability at least $\sim 1 - 1/t^2$ given Assumption 4.6. To derive (4.11), we use: $|\mathbb{E}[b_{x,T} - n_{x,T}\cdot b]| \leq \int_0^T \mathbb{P}[|b_{x,T} - n_{x,T}\cdot b| \geq u]\mathrm{d}u$ and $\mathbb{E}[b_{x,T}] = \mu^k_c\cdot\mathbb{E}[n^x_{k,T}] + \mu^l_c\cdot\mathbb{E}[n^x_{l,T}]$. □

The next step is to show, just like in Section 4.5, that any suboptimal feasible basis is selected at most $O(\ln(T))$ times on average. Interestingly, the choice of the load balancing algorithm plays a minor role in the proof. Any load balancing algorithm that pulls each arm involved in a basis at least a constant fraction of the time this basis is selected does enforce this property.

Lemma 4.12. *For any suboptimal basis $x \in \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,T}] \leq 2^9\frac{\lambda^2}{\epsilon^3}\cdot\frac{\ln(T)}{(\Delta_x)^2} + \frac{10\pi^2}{\epsilon^2}.$$

Sketch of proof. We use the shorthand notation $\beta_x = 2^8/\epsilon^3\cdot(\lambda/\Delta_x)^2$. Without loss of

generality, we can assume that $\mathcal{K}_{x^*} = \{k^*, l^*\}$ with $\mu^c_{k^*} > b > \mu^c_{l^*}$ and $\mathcal{K}_x = \{k, l\}$ with $\mu^c_k > b > \mu^c_l$. Along the same lines as in Lemma 4.5, we only have to bound by a constant the probability that x is selected at any round t given that x has already been selected at least $\beta_x \cdot \ln(t)$ times. If x is selected at time t, x is optimal for (4.8). Observe that $(\xi^{x^*}_k)_{k=1,\cdots,K}$ is a feasible solution to (4.8) when $\bar{c}_{k^*,t} - \epsilon_{k^*,t} \leq \mu^c_{k^*}$ and $\bar{c}_{l^*,t} - \epsilon_{l^*,t} \leq \mu^c_{l^*}$, which happens with probability at least $\sim 1 - 1/t^2$. As a result, $\mathrm{obj}_{x,t} + E_{x,t} \geq \mathrm{obj}_{x^*}$ when additionally $\bar{r}_{k^*,t} + \epsilon_{k^*,t} \geq \mu^r_{k^*}$ and $\bar{r}_{l^*,t} + \epsilon_{l^*,t} \geq \mu^r_{l^*}$, which also happens with probability at least $\sim 1 - 1/t^2$. If $\mathrm{obj}_{x,t} + E_{x,t} \geq \mathrm{obj}_{x^*}$ then we have either (i) $\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}$ or (ii) $\mathrm{obj}_{x^*} < \mathrm{obj}_x + 2E_{x,t}$. Observe that (ii) can only happen with probability at most $\sim 1/t^2$ given that $n_{x,t} \geq \beta_x \cdot \ln(t)$ because (ii) implies that either $n_{l,t} \leq 8(\lambda/\Delta_x)^2 \cdot \ln(t)$ or $n_{k,t} \leq 8(\lambda/\Delta_x)^2 \cdot \ln(t)$ but the load balancing algorithm guarantees that each arm is pulled a fraction of the time x is selected (using Lemma 4.11). As for (i), if $\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}$, then, using Assumption 4.4, either $\bar{r}_{k,t} \geq \mu^r_k + \epsilon_{k,t}$, $\bar{c}_{k,t} \notin [\mu^c_k - \epsilon_{k,t}, \mu^c_k + \epsilon_{k,t}]$, $\bar{r}_{l,t} \geq \mu^r_l + \epsilon_{l,t}$, or $\bar{c}_{l,t} \notin [\mu^c_l - \epsilon_{l,t}, \mu^c_l + \epsilon_{l,t}]$ but all of these events have individual probability at most $\sim 1/t^2$ by Lemma 4.1. □

In a last step, we show, using Lemma 4.11, that, at the cost of an additive logarithmic term in the regret bound, we may assume that the game lasts exactly T rounds. This enables us to combine Lemmas 4.10, 4.11, and 4.12 to establish a distribution-dependent regret bound.

Theorem 4.5. *We have:*

$$R_{B,T} \leq 2^9 \frac{\lambda^2}{\epsilon^3} \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \Big) \cdot \ln(T) + O\Big(\frac{K^2 \cdot \sigma}{\epsilon^3} \cdot \ln(T)\Big),$$

where the O notation hides universal constant factors.

Sketch of proof. We build upon (4.4):

$$\begin{aligned} R_{B,T} &\leq T \cdot \sum_{k=1}^{K} \mu^r_k \cdot \xi^{x^*}_k - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\ &\leq T \cdot \sum_{k=1}^{K} \mu^r_k \cdot \xi^{x^*}_k - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \sigma \cdot \mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t} - B)_+] + O(1), \end{aligned}$$

where we use Assumption 4.4 for the second inequality. Moreover:

$$\mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t} - B)_+] \leq \sum_{x \in \mathcal{B}} \mathbb{E}[|b_{x,T} - n_{x,T} \cdot b|] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}] = O(\frac{K^2}{\epsilon^3} \ln(T)),$$

using Lemmas 4.10 and 4.11. Plugging this last inequality back into the regret bound yields:

$$\begin{aligned}
R_{B,T} &\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,T}^x] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&= \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (T - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \mathbb{E}[n_{x,T}]) - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] \\
&+ O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&= \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \mathbb{E}[n_{x,T}] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}]) \\
&- \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq 2^9 \frac{\lambda^2}{\epsilon^3} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(T) + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)),
\end{aligned}$$

where we use Lemma 4.11 for the third inequality, Lemma 4.10 along with:

$$\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^{K} \xi_k^{x^*} \leq 1$$

for the fourth inequality, and Lemma 4.12 for the last inequality. □

Since there are at most $2K^2$ feasible bases, we get the regret bound $O(K^2 \cdot (1/\Delta + \sigma/\epsilon^3) \cdot \ln(T))$, where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$. Along the sames lines as in Sections 4.4 and 4.5, pushing the analysis further almost immediately yields a distribution-free regret bound.

Theorem 4.6. *We have:*

$$R_{B,T} \leq 2^5 \frac{\lambda}{\epsilon^{3/2}} \cdot \sqrt{|\mathcal{B}| \cdot T \cdot \ln(T)} + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)),$$

where the O notation hides universal constant factors.

Sketch of proof. The proof is along the same lines as for Theorems 4.2 and 4.4, we start from the penultimate inequality derived in the proof sketch of Theorem 4.5 and apply Lemma 4.12 only if Δ_x is big enough, taking into account that $\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x,T}] \leq T$. □

We conclude that $R_{B,T} = O(\sqrt{K^2 \cdot T \cdot \ln(T)})$, where the hidden factors are independent of the underlying distributions $(\nu_k)_{k=1,\cdots,K}$. Just like in Section 4.5, we stress that the dependence on K is not optimal since the authors of [16] and [6] obtain a $\tilde{O}(\sqrt{K \cdot T})$ bound on regret, where the $\tilde{O}$ notation hides factors logarithmic in T. Observe that the regret bounds derived in Theorems 4.5 and 4.6 do not vanish with b, which is not the expected behavior. This is a shortcoming of the analysis that can easily be remedied when $\min_{k=1,\cdots,K} \mu_k^c > 0$ provided that instead of pulling the dummy arm 0 we always pull the other arm involved in the basis (i.e. we never skip rounds). Note that not skipping rounds can only improve the regret bounds derived in Theorems 4.5 and 4.6: arm 0 was introduced only in order to harmonize the notations for mathematical convenience.

Theorem 4.7. *Relax Assumption 4.6 and redefine $\epsilon = \min_{k=1,\cdots,K} \mu_k^c$. Suppose that $b \leq \epsilon/2$ and that we never skip rounds, then we have:*

$$R_{B,T} \leq 2^{12} \frac{\lambda^2}{\epsilon^3} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(\frac{B+1}{\epsilon}) + O(\frac{K^2 \cdot \kappa}{\epsilon^3} \cdot \ln(\frac{B+1}{\epsilon}))$$

and

$$R_{B,T} \leq 2^6 \frac{\lambda}{\epsilon^{3/2}} \cdot \sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln(\frac{B+1}{\epsilon})} + O(\frac{K^2 \cdot \kappa}{\epsilon^3} \cdot \ln(\frac{B+1}{\epsilon})),$$

where the O notations hide universal constant factors.

Applications. Similarly, as in the case of a single resource, Assumptions 4.4 and 4.5 are natural when bidding in repeated second-price auctions if the auctioneer sets a reserve price R (which is common practice in sponsored search auctions). Indeed, we have:

$$\begin{aligned}|\mathbb{E}[c_{k,t}] - \mathbb{E}[c_{l,t}]| &= \mathbb{E}[m_t \cdot \mathbb{1}_{b_k \geq m_t > b_l}] \\ &\geq R \cdot \mathbb{E}[\mathbb{1}_{b_k \geq m_t > b_l}] \\ &\geq R \cdot \mathbb{E}[v_t \cdot \mathbb{1}_{b_k \geq m_t > b_l}] = R \cdot |\mathbb{E}[r_{k,t}] - \mathbb{E}[r_{l,t}]|,\end{aligned}$$

for any pair of arms $(k, l) \in \{1, \cdots, K\}$ with $b_k \geq b_l$. Hence, Assumption 4.4 (resp. 4.5) is satisfied with $\sigma = 1/R$ (resp. $\kappa = 1/R$).

In dynamic procurement, Assumptions 4.4 and 4.5 are satisfied provided that the agents are not willing to sell their goods for less than a known price P. Indeed, in this case, pulling any arm k associated with a price $p_k \leq P$ is always suboptimal and we have:

$$\begin{aligned}|\mathbb{E}[c_{k,t}] - \mathbb{E}[c_{l,t}]| &= p_k \cdot \mathbb{P}[p_k \geq v_t] - p_l \cdot \mathbb{P}[p_l \geq v_t] \\ &\geq p_k \cdot \mathbb{P}[p_k \geq v_t > p_l] \\ &\geq P \cdot \mathbb{P}[p_k \geq v_t > p_l] = P \cdot |\mathbb{E}[r_{k,t}] - \mathbb{E}[r_{l,t}]|,\end{aligned}$$

for any pair of arms $(k, l) \in \{1, \cdots, K\}$ with $p_k \geq p_l \geq P$. Hence, Assumption 4.4 (resp. 4.5) is satisfied with $\sigma = 1/P$ (resp. $\kappa = 1/P$).

In dynamic pricing, Assumptions 4.4 and 4.5 are satisfied if the distribution of valuations has a positive probability density function $f(\cdot)$. Indeed, in this case, we have:

$$\begin{aligned}|\mathbb{E}[r_{k,t}] - \mathbb{E}[r_{l,t}]| &= |p_l \cdot \mathbb{P}[p_l \leq v_t] - p_k \cdot \mathbb{P}[p_k \leq v_t]| \\ &= |p_l \cdot \mathbb{P}[p_l \leq v_t < p_k] + (p_l - p_k) \cdot \mathbb{P}[p_k \leq v_t]| \\ &\leq \max_{r=1,\cdots,K} p_r \cdot \mathbb{P}[p_l \leq v_t < p_k] + |p_k - p_l| \\ &\leq (\max_{r=1,\cdots,K} p_r + \frac{1}{\inf f(\cdot)}) \cdot \mathbb{P}[p_l \leq v_t < p_k] \\ &= (\max_{r=1,\cdots,K} p_r + \frac{1}{\inf f(\cdot)}) \cdot |\mathbb{E}[r_{k,t}] - \mathbb{E}[r_{l,t}]|,\end{aligned}$$

for any pair of arms $(k, l) \in \{1, \cdots, K\}$ with $k \geq l$. Hence, Assumption 4.4 (resp. 4.5) is satisfied with $\sigma = \max_{k=1,\cdots,K} p_k + 1/\inf f(\cdot)$ (resp. $\kappa = \max_{k=1,\cdots,K} p_k + 1/\inf f(\cdot)$).

4.7 Arbitrarily Many Limited Resources

In this section, we tackle the general case of arbitrarily many limited resources. Additionally, we assume that one of them is time, with index $i = C$, but this assumption is almost without loss of generality, as detailed at the end of this section. To simplify the presentation, we consider the regime $K \geq C$, which is the most common in applications. This implies that $|\mathcal{K}_x| = |\mathcal{C}_x| \leq C$ for any pseudo-basis x. We also use the shorthand notation $\bar{A}_t = (\bar{c}_{k,t}(i))_{(i,k)\in\{1,\cdots,C\}\times\{1,\cdots,K\}}$ at any round t. For similar reasons as in Section 4.6, we are led to make two additional assumptions which are discussed in the last paragraph of this section.

Assumption 4.7. *There exists $\sigma > 0$ such that $r_{k,t} \leq \sigma \cdot \min_{i=1,\cdots,C} c_{k,t}(i)$ for any arm $k \in \{1, \cdots, K\}$ and for any round $t \in \mathbb{N}$.*

Note that Assumption 4.7 is stronger than Assumption 4.4 given that the amounts of resources consumed at each round have to dominate the rewards almost surely, as opposed to on average. Assumption 4.7 is not necessarily satisfied in all applications but it simplifies the analysis and can be relaxed at the price of an additive term of order $O(\ln^2(T))$ in the final regret bounds, see the last paragraph of this section.

Assumption 4.8. *There exists $\epsilon > 0$, known to the decision maker ahead of round 1, such that every basis x for* (4.3) *is ϵ-non-degenerate for* (4.3) *and satisfy $|\det(A_x)| \geq \epsilon$.*

Without loss of generality, we assume that $\epsilon \leq 1$. Observe that Assumption 4.8 generalizes Assumption 4.6 but is more restrictive because ϵ is assumed to be known to the decision maker initially. Just like in Section 4.6, this assumption can be relaxed to a large extent. For instance, if ϵ is initially unknown, taking ϵ as a vanishing function of T yields the same asymptotic regret bounds. However, note that Lemma 4.9 carries over to this more general setting and, as a result, the minimal assumption we need to get logarithmic rates is that any optimal basis for (4.3) is non-degenerate.

Specification of the algorithm. We implement UCB-Simplex with $\lambda = 1+2(C+1)!^2/\epsilon$, $\eta_i = 0$ for any $i \in \{1, \cdots, C\}$, and an initialization step which consists in pulling each arm $2^8(\mathrm{C}+2)!^4/\epsilon^6 \cdot \ln(T)$ times in order to get i.i.d. samples. Hence, Step-Simplex is run for the first time after round $\mathrm{t_{ini}} = K \cdot 2^8(\mathrm{C}+2)!^4/\epsilon^6 \cdot \ln(T)$. Compared to Section 4.6, the initialization step plays as a substitute for the choice $\eta_i > 0$ which was meant to incentivize exploration. This significantly simplifies the analysis but the downside is that ϵ has to be known initially. Similarly, as in Section 4.6, we introduce a dummy arm which corresponds to skipping the round (i.e. pulling this arm yields a reward 0 and does not consume any resource) so that any basis can be mapped to one for which the time constraint is always binding, i.e. w.l.o.g. we assume that $C \in \mathcal{C}_x$ for any pseudo-basis x. Following the ideas developed in Section 4.6, we design load balancing algorithms for any basis x that pull arms in order to guarantee that, at any round t, the total amount of resource i consumed remains close to the target $t \cdot b(i)$ with high probability for any resource $i \in \mathcal{C}_x$. This is more involved that in Section 4.6 since we need to enforce this property for many resources. This can be done by perturbing the probability distribution solution to (4.8) taking into account whether we have over- or under-consumed in the past for each binding resource $i \in \mathcal{C}_{x_t}$.

Algorithm: Load balancing algorithm $\mathcal{A}_x$ for any basis x

---

For any time period $t > \mathrm{t_{ini}}$ and $i \in \mathcal{C}_x - \{C\}$, define $b_{x,t}(i)$ as the total amount of resource i consumed when selecting basis x in the past $t-1$ rounds. Suppose that basis x is selected at time t and define the vector e_t^x by $e_{C,t}^x = 0$ and $e_{i,t}^x = -1$ (resp. $e_{i,t}^x = 1$) if $b_{x,t}(i) \geq n_{x,t} \cdot b(i)$ (resp. $b_{x,t}(i) < n_{x,t} \cdot b(i)$) for any $i \in \mathcal{C}_x - \{C\}$. Since x is selected at round t, $\bar{A}_{x,t}$ is invertible and we can define, for any $\delta \geq 0$, $p_{k,t}^x(\delta) = (\bar{A}_{x,t}^{-1}(b_{\mathcal{C}_x} + \delta \cdot e_t^x))_k$ for $k \in \mathcal{K}_x$ and $p_{k,t}^x(\delta) = 0$ otherwise, which together define the probability distribution $p_t^x(\delta) = (p_{k,t}^x(\delta))_{k \in \{1,\cdots,K\}}$. Define

$$\delta_{x,t}^* = \max_{\substack{\delta \geq 0 \\ (\bar{A}_t p_t^x(\delta))_i \leq b(i), i \notin \mathcal{C}_x \\ p_t^x(\delta) \geq 0}} \delta$$

and $p_t^x = p_t^x(\delta_{x,t}^*)$. Note that $\delta_{x,t}^*$ is well defined as x must be feasible for (4.8) if it is selected at Step-Simplex. Pull an arm at random according to the distribution p_t^x.

Observe that the load balancing algorithms generalize the ones designed in Section 4.6 (up to the change $\eta_i = 0$). Indeed, when there is a single limited resource other than time, the probability distribution p_t^x is a Dirac supported at the arm with smallest (resp. largest) empirical cost when $b_{x,t} \geq n_{x,t} \cdot b$ (resp. $b_{x,t} < n_{x,t} \cdot b$). Similarly, as in Section 4.5, the load balancing algorithms $\mathcal{A}_x$ may require a memory storage capacity exponential in C and polynomial in K, but, in practice, we expect that only a few bases will be selected at Step-Simplex, so that a hash table is an appropriate data structure to store the sequences $(b_{x,t}(i))_{i \in \mathcal{C}_x}$. Note, however, that the load balancing algorithms are computationally efficient because p_t^x can be computed in $O(C^2)$ running time if $\bar{A}_{x,t}^{-1}$ is available once we have computed an optimal basic feasible solution to (4.8), which is the case if we use the revised simplex algorithm.

Regret analysis. The regret analysis follows the same recipe as in Section 4.6 but the proofs are more technical and are thus deferred to Section C.6 of the Appendix. First, we show that the initialization step guarantees that infeasible bases or pseudo-bases x with $\det(A_x) = 0$ cannot be selected more than $O(\ln(T))$ times on average at Step-Simplex.

Lemma 4.13. *For any basis $x \notin \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,T}] \leq 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6}\,.$$

The same inequality holds if x is a pseudo-basis but not a basis for (4.3)*.*

The next step is to show that the load balancing algorithms guarantee that, for any basis x, the amount of resource $i \in \mathcal{C}_x$ (resp. $i \notin \mathcal{C}_x$) consumed per round when selecting x remains close to (resp. below) the pacing target $b(i)$ with high probability. This enables us to show that the ratios $(\mathbb{E}[n_{k,T}^x]/\mathbb{E}[n_{l,T}^x])_{k,l \in \mathcal{K}_x}$ are close to the optimal ones $(\xi_k^x/\xi_l^x)_{k,l \in \mathcal{K}_x}$.

Lemma 4.14. *For any feasible basis x and time period t, we have:*

$$\mathbb{P}[|b_{x,t}(i) - n_{x,t} \cdot b(i)| \geq u] \leq 2^5 \frac{(C+1)!^2}{\epsilon^4} \cdot \exp(-u \cdot (\frac{\epsilon^2}{4 \cdot (\mathrm{C}+1)!})^2) + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6} \cdot \frac{1}{T}, \tag{4.12}$$

for all $u \geq 1$ and for any resource $i \in \mathcal{C}_x$ while

$$\mathbb{P}[b_{x,t}(i) - n_{x,t} \cdot b(i) \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)] \leq 2^{10} \frac{(C+4)!^4}{\epsilon^6 \cdot T}, \tag{4.13}$$

for any resource $i \notin \mathcal{C}_x$. In particular, this implies that:

$$\mathbb{E}[n^x_{k,T}] \geq \xi^x_k \cdot \mathbb{E}[n_{x,T}] - 2^{10} \frac{(C+3)!^4}{\epsilon^9}, \tag{4.14}$$

for any arm $k \in \mathcal{K}_x$.

Next, we show that a suboptimal basis cannot be selected more than $O(\ln(T))$ times on average at Step-Simplex. Just like in Section 4.6, the exact definition of the load balancing algorithms has little impact on the result: we only need to know that, for any feasible basis x, each arm $k \in \mathcal{K}_x$ is pulled at least a fraction of the time x is selected with high probability.

Lemma 4.15. *For any suboptimal basis $x \in \mathcal{B}$, we have:*

$$\mathbb{E}[n_{x,T}] \leq 2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot \frac{\ln(T)}{(\Delta_x)^2} + 2^{11} \frac{(C+4)!^4}{\epsilon^6}.$$

We are now ready to derive both distribution-dependent and distribution-independent regret bounds.

Theorem 4.8. *We have:*

$$R_{B(1),\cdots,B(C-1),T} \leq 2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(T) + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)),$$

where the O notation hides universal constant factors.

Theorem 4.9. *We have:*

$$R_{B(1),\cdots,B(C-1),T} \leq 2^5 \frac{(C+3)!^2 \cdot \lambda}{\epsilon^3} \cdot \sqrt{|\mathcal{B}| \cdot T \cdot \ln(T)} + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)),$$

where the O notation hides universal constant factors.

Since the number of feasible bases is at most $2K^C$, we get the distribution-dependent regret bound

$$R_{B(1),\cdots,B(C-1),T} = O(K^C \cdot (C+3)!^4/\epsilon^6 \cdot (\lambda^2/\Delta + \sigma) \cdot \ln(T)),$$

where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$, and the distribution-independent bound

$$R_{B(1),\cdots,B(C-1),T} = O((C+3)!^2 \cdot \lambda/\epsilon^3 \cdot \sqrt{K^C \cdot T \cdot \ln(T)}).$$

We stress that the dependence on K and C is not optimal since the authors of [6] obtain a $\tilde{O}(\sqrt{K \cdot T})$ distribution-independent bound on regret, where the $\tilde{O}$ notation hides factors logarithmic in T. Just like in Section 4.6, we can also derive regret bounds that vanish with b under the assumption that pulling any arm incurs some positive amount of resource consumption in expectations for all resources, but this requires a minor tweak of the algorithm.

Theorem 4.10. *Suppose that:*

$$\epsilon \leq \min_{\substack{i=1,\cdots,C-1 \\ k=1,\cdots,K}} \mu_k^c(i)$$

and that $b \leq \epsilon$. If the decision maker artificially constrains himself or herself to a time horizon $\tilde{T} = b \cdot T/\epsilon \leq T$, then the regret bounds derived in Theorems 4.8 and 4.9 hold with T substituted with $\tilde{T}$.

Similarly, if the decision maker is not constrained by a time horizon, artificially constraining himself or herself to a time horizon $\tilde{T} = \min_{i=1,\cdots,C} B(i)/\epsilon$ yields the regret bounds derived in Theorems 4.8 and 4.9 with T substituted with $\tilde{T}$.

Applications. In dynamic pricing and online advertising applications, Assumption 4.7 is usually not satisfied as pulling an arm typically incurs the consumption of only a few resources. We can relax this assumption but this comes at the price of an additive term of order $O(\ln^2(T))$ in the final regret bounds.

Theorem 4.11. *If Assumption 4.7 is not satisfied, the regret bounds derived in Theorems 4.8 and 4.9 hold with $\sigma = 0$ up to an additive term of order $O(\frac{(C+4)!^4 \cdot |\mathcal{B}|^2}{b \cdot \epsilon^6} \cdot \ln^2(T))$.*

As for Assumption 4.8, the existence of degenerate optimal bases to (4.3) is determined by a complex interplay between the mean rewards and the mean amounts of resource consumption. However, we stress that the set of parameters $(\mu_k^r)_{k=1,\cdots,K}$ and $(\mu_k^c(i))_{k=1,\cdots,K,i=1,\cdots,C}$ that satisfy these conditions has Lebesgue measure 0, hence such an event is unlikely to occur in practice. Additionally, while ϵ is typically not known in applications, taking ϵ as a vanishing function of T yields the same asymptotic regret bounds.

4.8 Concluding Remark

In this chapter, we develop an algorithm with a $O(K^C \cdot \ln(B)/\Delta)$ distribution-dependent bound on regret, where Δ is a parameter that generalizes the optimality gap for the standard MAB problem. It is however unclear whether the dependence on K is optimal. Extensions discussed in Section C.1 of the Appendix suggest that it may be possible to achieve a linear dependence on K but this calls for the development of more efficient load balancing algorithms.

Chapter 5

Real-Time Bidding with Side Information

5.1 Introduction

On the internet, advertisers and publishers now interact through real-time marketplaces called ad exchanges. Through them, any publisher can sell the opportunity to display an ad when somebody is visiting a webpage he or she owns. Conversely, any advertiser interested in such an opportunity can pay to have his or her ad displayed. In order to match publishers with advertisers and to determine prices, ad exchanges commonly use a variant of second-price auctions which typically runs as follows. Each participant is initially provided with some information about the person that will be targeted by the ad (e.g. browser cookies, IP address, and operating system) along with some information about the webpage (e.g. theme) and the ad slot (e.g. width and visibility). Based on this limited knowledge, advertisers must submit a bid in a timely fashion if they deem the opportunity worthwhile. Subsequently, the highest bidder gets his or her ad displayed and is charged the second-highest bid. Moreover, the winner can usually track the customer's interaction with the ad (e.g. clicks). Because the auction is sealed, very limited feedback is provided to the advertiser if the auction is lost. In particular, the advertiser does not receive any customer feedback in this scenario. In addition, the demand for ad slots, the supply of ad slots, and the websurfers' profiles cannot be predicted ahead of time and are thus commonly modeled

as random variables, see [43]. These last two features contribute to making the problem of bid optimization in ad auctions particularly challenging for advertisers.

5.1.1 Problem Statement and Contributions

We consider an advertiser interested in purchasing ad impressions through an ad exchange. As standard practice in the online advertising industry, we suppose that the advertiser has allocated a limited budget B for a limited period of time, which corresponds to the next T ad auctions. Rounds, indexed by $t \in \mathbb{N}$, correspond to ad auctions the advertiser participates in. At the beginning of round $t \in \mathbb{N}$, some contextual information about the ad slot and the person that will be targeted is revealed to the advertiser in the form of a multidimensional vector $x_t \in \mathcal{X}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^d$. Without loss of generality, the coordinates of x_t are assumed to be normalized in such a way that $\|x\|_\infty \leq 1$ for all $x \in \mathcal{X}$. Given x_t, the advertiser must submit a bid b_t in a timely fashion. If b_t is larger than the highest bid submitted by the competitors, denoted by p_t and also referred to as the market price, the advertiser wins the auction, is charged p_t, and gets his or her ad displayed, from which he or she derives a utility v_t. Monetary amounts and utility values are assumed to be normalized in such a way that $b_t, p_t, v_t \in [0, 1]$. In this modeling, one of the competitors is the publisher himself who submits a reserve price so that $p_t > 0$. No one wins the auction if no bid is larger than the reserve price. For the purpose of modeling, we suppose that ties are broken in favor of the advertiser but this choice is arbitrary and by no means a limitation of the approach. Hence, the advertiser collects a reward $r_t = v_t \cdot \mathbb{1}_{b_t \geq p_t}$ and is charged $c_t = p_t \cdot \mathbb{1}_{b_t \geq p_t}$ at the end of round t. Since the monetary value of getting an ad displayed is typically difficult to assess, v_t and c_t may be expressed in different units and thus cannot be compared directly in general, which makes the problem two-dimensional. This is the case, for example, when the goal of the advertiser is to maximize the number of clicks, in which case $v_t = 1$ if the ad was clicked on and $v_t = 0$ otherwise. We consider a stochastic setting where the environment and the competitors are not fully adversarial. Specifically, we assume that, at any round $t \in \mathbb{N}$, the vector (x_t, v_t, p_t) is jointly drawn from a fixed probability distribution ν independently from the past. This is motivated by

two observations. First, ad auctions are sealed: the identity of the competitors are never revealed and they value ad slots differently based on undisclosed rules that are specific to them. Second, websurfers connect to websites with no a priori knowledge of which advertisers will participate in the resulting ad auctions. Moreover, while the assumption that the distribution of (x_t, v_t, p_t) is stationary may only be valid for a short period of time, advertisers tend to participate in a large number of ad auctions per second so that T and B are typically large values, which motivates an asymptotic study. We generically denote by (X, V, P) a vector of random variables distributed according to ν. We make a structural assumption about ν, which we use throughout the paper.

Assumption 5.1. *The random variables V and P are conditionally independent given X. Moreover, there exists $\theta_* \in \mathbb{R}^d$ such that $\mathbb{E}[V \mid X] = X^\intercal \theta_*$ and $\|\theta_*\|_\infty \leq 1$.*

The first part of Assumption 5.1 is motivated by the fact that websurfers are oblivious to the ad auctions that take place behind the scenes to determine which ad they will be presented with. The second part of Assumption 5.1 is standard in the literature on linear contextual MABs, see [1] and [34], and is arguably the simplest model capturing a dependence between x_t and v_t. When the advertiser's objective is to maximize the number of clicks, this assumption translates into a linear Click-Through Rate (CTR) model.

We denote by $(\mathcal{F}_t)_{t\in\mathbb{N}}$ (resp. $(\tilde{\mathcal{F}}_t)_{t\in\mathbb{N}}$) the natural filtration generated by $((x_t, v_t, p_t))_{t\in\mathbb{N}}$ (resp. $((x_{t+1}, v_t, p_t))_{t\in\mathbb{N}}$). Since the advertiser can keep bidding only so long as he or she does not run out of money or time, he or she can no longer participate in ad auctions at round τ^*, mathematically defined by:

$$\tau^* = \min(T + 1, \min\{t \in \mathbb{N} \mid \sum_{\tau=1}^{t} c_\tau > B\}). \tag{5.1}$$

Note that τ^* is a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. The difficulty for the advertiser when it comes to determining how much to bid at each round lies in the fact that the underlying distribution ν is initially unknown. This task is further complicated by the fact that the feedback provided to the advertiser upon bidding b_t is partially censored: p_t and v_t are only revealed if the advertiser wins the auction, i.e. if $b_t \geq p_t$. In particular when $b_t < p_t$, the advertiser can never evaluate how much reward would have been obtained and

what price would have been charged if he or she had submitted a higher bid. The goal for the advertiser is to design a non-anticipating algorithm that, at any round t, selects b_t based on the information acquired in the past so as to keep the pseudo-regret defined as:

$$R_{B,T} = \mathrm{ER}_{\mathrm{OPT}}(B,T) - \mathbb{E}[\sum_{t=1}^{\tau^*-1} r_t] \tag{5.2}$$

as small as possible, where $\mathrm{ER}_{\mathrm{OPT}}(B,T)$ is the maximum expected sum of rewards that can be obtained by a non-anticipating oracle algorithm that has knowledge of the underlying distribution. Here, an algorithm is said to be non-anticipating if the bid selection process does not depend on the future observations. We develop algorithms with bounds on the pseudo-regret that do not depend on the underlying distribution ν, which are referred to as distribution-independent regret bounds. This entails studying the asymptotic behavior of $R_{B,T}$ when B and T go to infinity. For mathematical convenience, we consider that the advertiser keeps bidding even if he or she has run out of time or money so that all quantities (including b_t) are well-defined for any $t \in \mathbb{N}$. Of course, the rewards obtained for $t \geq \tau^*$ are not taken into account in the advertiser's total reward when establishing regret bounds.

Contributions. We develop UCB-type algorithms that combine the ellipsoidal confidence set approach to linear contextual MAB problems with a special-purpose stochastic binary search procedure. When the budget is unlimited or when it scales linearly with time, we show that, under additional technical assumptions on the underlying distribution ν, our algorithms incur a regret $R_{B,T} = \tilde{O}(d \cdot \sqrt{T})$, where the $\tilde{O}$ notation hides logarithmic factors in d and T. A key feature of our approach is that overbidding is not only essential to incentivize exploration in order to estimate θ_*, but also crucial to find the optimal bidding strategy given θ_* because bidding higher always provide more feedback in real-time bidding.

5.1.2 Literature Review

To handle the exploration-exploitation trade-off inherent to MAB problems, an approach that has proved to be particularly successful hinges on the *optimism in the face of uncer-*

tainty paradigm. The idea is to consider all plausible scenarios consistent with the information collected so far and to select the decision that yields the largest reward among all identified scenarios. The authors of [12] use this idea to solve the standard MAB problem where decisions are represented by $K \in \mathbb{N}$ arms and pulling arm $k \in \{1, \cdots, K\}$ at round $t \in \{1, \cdots, T\}$ yields a random reward drawn from an unknown distribution specific to this arm independently from the past. Specifically, the authors of [12] develop the Upper Confidence Bound algorithm (UCB1), which consists in systematically selecting the arm with the current largest upper confidence bound on its mean reward, and establish near-optimal regret bounds. This approach has since been successfully extended to a number of more general settings. Of most notable interest to us are: (i) linear contextual MAB problems, where, for each arm k and at each round t, some context x_t^k is provided to the decision maker ahead of pulling any arm and the expected reward of arm k is $\theta_*^\intercal x_t^k$ for some unknown $\theta_* \in \mathbb{R}^d$, (ii) the Bandits with Knapsacks (BwK) framework, an extension to the standard MAB problem allowing to model resource consumption, and (iii) the linear contextual BwK framework, a combination of the two aforementioned extensions.

UCB-type algorithms for linear contextual MAB problems were first developed in [11] and later extended and improved upon in [1] and [34]. In this line of work, the key idea is to build, at any round t, an ellipsoidal confidence set $\mathcal{C}_t$ on the unknown parameter θ_* and to pull the arm k that maximizes $\max_{\theta \in \mathcal{C}_t} \theta^\intercal x_t^k$. Using this idea, the authors of [34] derive $\tilde{O}(\sqrt{d \cdot T})$ upper bounds on regret that hold with high probability, where the $\tilde{O}$ notations hides logarithmic factors in d and T. While this result is not directly applicable in our setting, partly because the knapsack constraint is not taken into account, we rely on this technique to estimate θ_*.

The real-time bidding problem considered in this work can be formulated as a BwK problem with contextual information and a continuum of arms. This framework, first introduced in its full generality in [16] and later extended to incorporate contextual information in [17], [7], and [5], captures resource consumption by assuming that pulling any arm incurs the consumption of possibly many different limited resource types by random amounts. The authors of [5] consider a particular case where the expected rewards and the expected amounts of resource consumption are linear in the context and derive, in partic-

ular, $\tilde{O}(\sqrt{d \cdot T})$ bounds on regret when the initial endowments of resources scale linearly with the time horizon T. These results do not carry over to our setting because the expected costs, and in fact also the expected rewards, are not linear in the context. The authors of [7] and [17] develop algorithms to tackle more general settings where the expected rewards and amounts of resource consumption are not necessarily linear in the context when there is a finite number of arms K. They derive regret bounds that scale as $\tilde{O}(\sqrt{K \cdot T \cdot \ln(\Pi)})$, where Π is the size of the set of benchmark policies. To some extent, at least when θ_* is known, it is possible to apply these results but this requires to discretize $[0, 1]$ with an $\epsilon(T)$ -additive mesh for a well-chosen function $\epsilon(\cdot)$. However, the regret bounds thus derived would scale as $\sim \sqrt{T^{2/3}}$, see the analysis in [16], which would be suboptimal.

On the modeling side, the most closely related prior works studying repeated auctions under the lens of online learning are [97], [92], [19], and [35]. The authors of [97] develop algorithms to solve the problem considered in this work when no contextual information is available and when there is no budget constraint, in which case the rewards are defined as $r_t = (v_t - p_t) \cdot \mathbb{1}_{b_t \geq p_t}$, but in a more general adversarial setting where few assumptions are made concerning the sequence $((v_t, p_t))_{t \in \mathbb{N}}$. They obtain $\tilde{O}(\sqrt{T})$ regret bounds with an improved rate $O(\ln(T))$ in some favorable settings of interest. The authors of [92] study a particular case of the problem considered in this work when no contextual information is available and when the goal is to maximize the number of impressions. They derive $\tilde{O}(\sqrt{T})$ regret bounds using a dynamic programming approach. The authors of [19] study repeated multi-commodity auctions without contextual information when the goal is to maximize revenues subject to a budget constraint that has to be enforced for each individual period, as opposed to globally in the setting considered in this work. They develop an algorithm based on a dynamic program and derive $O(\sqrt{T \cdot \ln(T)})$ regret bounds. Finally, the authors of [35] take the point of view of the publisher whose goal is to price ad impressions, as opposed to purchasing them, in order to maximize revenues with no knapsack constraint. They derive $O(\ln(d^2 \cdot \ln(T/d)))$ bounds on regret with high probability.

On the technical side, our work builds upon and contributes to the stream of literature on probabilistic bisection search algorithms. This class of algorithms was originally developed for solving stochastic root finding problems, see [75] for an overview, but has also

recently appeared in the MAB literature, see [35] and [61]. The authors of [35] propose a binary search approach based on the ellipsoid method for linear programming to obtain $O(\ln(d^2 \cdot \ln(T/d)))$ regret bounds in the MAB setting described above. However, the binary search problem that arises in [35] differs from ours along two key features that makes their approach inapplicable here: the feedback provided to the decision maker is deterministic and the contextual information available at each round is adversarially chosen, as opposed to being stochastically generated in our work. Our approach is largely inspired by the work of the authors of [61] who develop a stochastic binary search algorithm to solve a dynamic pricing problem with limited supply but no contextual information, which can be modeled as a BwK problem with a continuum of arms. The technical challenge in [61] differs from ours in one key aspect: the feedback provided to the decision maker is completely censored in dynamic pricing problems, since the customers' valuations are never revealed, while it is only partially censored in real-time bidding, since the market price is revealed if the auction is won. Making the most of this additional feature enables us to develop a stochastic binary search procedure that can be compounded with the ellipsoidal confidence set approach to linear contextual bandits in order to incorporate contextual information.

Organization. The remainder of the paper is organized as follows. In order to increase the level of difficulty progressively, we start by studying the situation of an advertiser with unlimited budget, i.e. $B = \infty$, in Section 5.2. Given that second-price auctions induce truthful bidding when the bidder has no budget constraint, this setting is easier since the optimal bidding strategy is to bid $b_t = x_t^\intercal \theta_*$ at any round $t \in \mathbb{N}$. This drives us to focus on the problem of estimating θ_*, which we do by means of ellipsoidal confidence sets. Next, in Section 5.3, we study the setting where B is finite and scales linearly with the time horizon T. We show that a near-optimal strategy is to bid $b_t = x_t^\intercal \theta_* / \lambda_*$ at any round $t \in \mathbb{N}$, where $\lambda_* \geq 0$ is a scalar factor whose purpose is to spread the budget as evenly as possible, i.e. $\mathbb{E}[P \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P}] = B/T$. Given this characterization, we first assume that θ_* is known a priori to focus instead on the problem of computing an approximate solution $\lambda \geq 0$ to $\mathbb{E}[P \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda \cdot P}] = B/T$ in Section 5.3.1. We develop a stochastic binary search algorithm for this purpose which is shown to incur $\tilde{O}(\sqrt{T})$ regret under mild assumptions

on the underlying distribution ν. In Section 5.3.2, we bring the stochastic binary search algorithm together with the estimation method based on ellipsoidal confidence sets to tackle the general problem and derive $\tilde{O}(d \cdot \sqrt{T})$ regret bounds. All the proofs are deferred to the Appendix.

Notations. For a vector $x \in \mathbb{R}^d$, $\|x\|_\infty$ (resp. $\|x\|_2$) refers to the L_∞-norm (resp. L_2-norm) of x. For a positive definite matrix $M \in \mathbb{R}^{d \times d}$ and a vector $x \in \mathbb{R}^d$, we define the norm $\|x\|_M$ as $\|x\|_M = \sqrt{x^\intercal M x}$. When M is the identity matrix, which we denote by I_d, $\|x\|_M = \|x\|_2$. For $x, y \in \mathbb{R}^d$, it is well known that the following Cauchy-Schwarz inequality holds: $|x^\intercal y| \leq \|x\|_M \cdot \|y\|_{M^{-1}}$. We use the standard asymptotic notation $O(\cdot)$ when T, B, and d go to infinity. While $O(\cdot)$ only hides universal constant factors, we also use the notation $\tilde{O}(\cdot)$ that hides logarithmic factors in d, T, and B. For $x \in \mathbb{R}$, $(x)_+$ refers to the positive part of x. For a finite set S, $|S|$ denotes the cardinality of S. Additionally, as a slight abuse of notation, we also use $|I|$ to denote the length of a compact interval $I \subset \mathbb{R}$. For a set S, $\mathcal{P}(S)$ denotes the set of all subsets of S. Finally, for a real-valued function $f(\cdot)$, $\operatorname{supp} f(\cdot)$ denotes the support of $f(\cdot)$.

5.2 Unlimited Budget

In this section, we suppose that the budget is unlimited, i.e. $B = \infty$, which implies that the rewards have to be redefined in order to directly incorporate the costs. For this purpose, we assume in this section that v_t is expressed in monetary value and we redefine the rewards as $r_t = (v_t - p_t) \cdot \mathbb{1}_{b_t \geq p_t}$. Since the budget constraint is irrelevant when $B = \infty$, we use the notations R_T and $\mathrm{ER}_{\mathrm{OPT}}(T)$ in place of $R_{B,T}$ and $\mathrm{ER}_{\mathrm{OPT}}(B, T)$. As standard in the literature on MAB problems, we start by analyzing the optimal oracle strategy that has knowledge of the underlying distribution. This will not only guide the design of algorithms when ν is unknown but this will also facilitate the regret analysis. The algorithm developed in this section as well as the regret analysis are extensions of the work of [97] to the contextual setting.

Benchmark analysis. It is well known that second-price auctions induce truthful bidding in the sense that any participant whose only objective is to maximize the immediate payoff should always bid what he or she thinks the good being auctioned is worth. The following result should thus come at no surprise in the context of real-time bidding given Assumption 5.1 and the fact that each participant is provided with the contextual information x_t before the t-th auction takes place.

Lemma 5.1. *The optimal non-anticipating strategy is to bid $b_t = x_t^\intercal \theta_*$ at any time period $t \in \mathbb{N}$ and we have:*

$$\mathrm{ER}_{\mathrm{OPT}}(T) = \sum_{t=1}^{T} \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+]. \tag{5.3}$$

Lemma 5.1 shows that the problem faced by the advertiser essentially boils down to estimating θ_*. Since the bidder only gets to observe v_t if the auction is won, this gives advertisers a clear incentive to overbid early on so that they can progressively refine their estimates downward as they collect more data points.

Specification of the algorithm. Following the approach developed in [11] for linear contextual MAB problems, we define, at any round t, the regularized least square estimate of θ_* given all the feedback acquired in the past $\hat{\theta}_t = M_t^{-1} \sum_{\tau=1}^{t-1} \mathbb{1}_{b_\tau \geq p_\tau} \cdot v_\tau \cdot x_\tau$, where $M_t = I_d + \sum_{\tau=1}^{t-1} \mathbb{1}_{b_\tau \geq p_\tau} \cdot x_\tau x_\tau^\intercal$, as well as the corresponding ellipsoidal confidence set:

$$\mathcal{C}_t = \{\theta \in \mathbb{R}^d \mid \left\|\theta - \hat{\theta}_t\right\|_{M_t} \leq \delta_T\}, \tag{5.4}$$

where $\delta_T = 2\sqrt{d \cdot \ln((1 + d \cdot T) \cdot T)}$. For the reasons mentioned above, we take the *optimism in the face of uncertainty approach* and bid:

$$b_t = \max(0, \min(1, \max_{\theta \in \mathcal{C}_t} \theta^\intercal x_t)) \tag{5.5}$$

at any round t. Since $\mathcal{C}_t$ was designed with the objective of guaranteeing that $\theta_* \in \mathcal{C}_t$ with high probability at any round t, irrespective of the number of auctions won in the past, b_t is larger than the optimal bid $x_t^\intercal \theta_*$ in general, i.e. we tend to overbid. Note that b_t can be computed in closed form since $\max_{\theta \in \mathcal{C}_t} \theta^\intercal x_t = \hat{\theta}_t^\intercal x_t + \delta_T \cdot \sqrt{x_t^\intercal M_t^{-1} x_t}$.

Regret analysis. Concentration inequalities are intrinsic to any kind of learning and are thus key to derive regret bounds in online learning. We start with the following lemma, which is a consequence of the results derived in [1] for linear contextual MABs, that shows that θ_* lies in all the ellipsoidal confidence sets with high probability.

Lemma 5.2. *We have:*

$$\mathbb{P}[\theta^* \notin \cap_{t=1}^{T} \mathcal{C}_t] \leq \frac{1}{T}.$$

Equipped with Lemma 5.2 along with some standard results for linear contextual bandits, we are now ready to extend the analysis of [97] to the contextual setting.

Theorem 5.1. *Bidding according to* (5.5) *at any round* t *incurs a regret* $R_T = \tilde{O}(d \cdot \sqrt{T})$.

Alternative algorithm with lazy updates. As first pointed out by [1] in the context of linear bandits, updating the confidence set $\mathcal{C}_t$ at every round is not only inefficient but also unnecessary from a performance standpoint. Instead, we can perform batch updates, only updating $\mathcal{C}_t$ using all the feedback collected in the past at rounds t for which $\det(M_t)$ has increased by a factor at least $(1 + A)$ compared to the last time there was an update, for some constant $A > 0$ of our choosing. This leads to an interesting trade-off between computational efficiency and deterioration of the regret bound captured in our next result. For mathematical convenience, we keep the same notations as when we were updating the confidence sets at every round. The only difference lies in the fact that the bid submitted at time t is now defined as:

$$b_t = \max(0, \min(1, \max_{\theta \in \mathcal{C}_{\tau_t}} \theta^\intercal x_t)), \tag{5.6}$$

where τ_t is the last round before round t where the last batch update happened.

Theorem 5.2. *Bidding according to* (5.6) *at any round* t *incurs a regret* $R_T = \tilde{O}(d \cdot \sqrt{A \cdot T})$.

The fact that we can afford lazy updates will turn out to be important to tackle the general case in Section 5.3.2 since we will only be able to update the confidence sets at most $O(\ln(T))$ times.

5.3 Limited Budget

In this section, we consider the setting where B is finite and scales linearly with the time horizon T. We will need the following assumptions for the remainder of the paper.

Assumption 5.2. *(a) The ratio B/T is a constant independent of any other relevant quantities, denoted by $\beta > 0$.*

(b) There exists a reserve price $r > 0$, known to the advertiser, such that $p_t \geq r$ for all $t \in \mathbb{N}$.

(c) We have $\mathbb{E}[1/(X^\intercal \theta_*)^3] < \infty$.

(d) The random variable P has a continuous conditional probability density function given the occurrence of the value x of X, denoted by $f_x(\cdot)$, that is upper bounded by $\bar{L} < \infty$.

(e) There exists $K \geq 0$ such that $w \in \mathbb{R}_+ \to w \cdot f_X(w)$ is K-Lipschitz on $\operatorname{supp} f_X(\cdot)$ almost surely.

Condition (b) is very natural in real-time bidding where r corresponds to the minimum reserve price across ad auctions. Conditions (c), (d), and (e) are motivated by technical considerations that will appear clear in the analysis. Note that $\bar{L}$ and K are not assumed to be known to the advertiser.

In order to increase the level of difficulty progressively and to prepare for the integration of the ellipsoidal confidence sets, we first look at an artificial setting in Section 5.3.1 where we assume that there exists a known set $\mathcal{C} \subset \mathbb{R}^d$ such that $\mathbb{E}[V|X] = \min(1, \max_{\theta \in \mathcal{C}} X^\intercal \theta)$ (as opposed to $\mathbb{E}[V|X] = X^\intercal \theta_*$) and such that $\theta_* \in \mathcal{C}$. This is to sidestep the estimation problem in a first step in order to focus on determining an optimal bidding strategy given θ_*. Next, in Section 5.3.2, we bring together the methods developed in Section 5.2 and Section 5.3.1 to tackle the general setting.

5.3.1 Preliminary Work

In this section, we make the following modeling assumption in lieu of $\mathbb{E}[V|X] = X^\intercal \theta_*$.

Assumption 5.3. *There exists* $\mathcal{C} \subset \mathbb{R}^d$ *such that* $\mathbb{E}[V|X] = \min(1, \max_{\theta \in \mathcal{C}} X^\intercal \theta)$ *and* $\theta_* \in \mathcal{C}$.

Furthermore, we assume that $\mathcal{C}$ is known to the advertiser initially. Of course, we recover the original setting introduced in Section 5.1 when $\mathcal{C} = \{\theta_*\}$ (since $V \in [0,1]$ implies $\mathbb{E}[V|X] \in [0,1]$) and θ_* is known but the level of generality considered here will prove useful to tackle the general case in Section 5.3.2 when we define $\mathcal{C}$ as an ellipsoidal confidence set on θ_*. As in Section 5.2, we start by identifying a near-optimal oracle bidding strategy that has knowledge of the underlying distribution. This will not only guide the design of algorithms when ν is unknown but this will also facilitate the regret analysis. We use the shorthand $g(X) = \min(1, \max_{\theta \in \mathcal{C}} X^\intercal \theta)$ throughout this section.

Benchmark analysis. To bound the performance of any non-anticipating strategy, we will be interested in the mappings:

$$\phi : \lambda, \mathcal{C} \in [0, 2/r] \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{E}[P \cdot \mathbb{1}_{g(X) \geq \lambda \cdot P}], \tag{5.7}$$

and:

$$R : \lambda, \mathcal{C} \in [0, 2/r] \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{E}[g(X) \cdot \mathbb{1}_{g(X) \geq \lambda \cdot P}]. \tag{5.8}$$

Note that $\phi(\cdot, \mathcal{C})$ is non-increasing and that, without loss of generality, we can restrict λ to be no larger than $2/r$ because $\phi(\lambda, \mathcal{C}) = \phi(2/r, \mathcal{C}) = 0$ for $\lambda \geq 2/r$ since $P \geq r$. Exploiting the structure of the MAB problem at hand, we can bound the sum of rewards obtained by any non-anticipating strategy by the value of a knapsack problem where the weights and the values of the items are drawn in an i.i.d. fashion from a fixed distribution. Since characterizing the expected optimal value of a knapsack problem is a well studied problem, see [63], we can derive a simple upper bound on $\mathrm{ER}_{\mathrm{OPT}}(B,T)$ through this reduction, as we next show.

Lemma 5.3. *We have:*

$$\mathrm{ER}_{\mathrm{OPT}}(B,T) \leq T \cdot R(\lambda_*, \mathcal{C}) + O(1),$$

where $\lambda_* \geq 0$ *satisfies* $\phi(\lambda_*, \mathcal{C}) = \beta$ *or* $\lambda_* = 0$ *if no such solution exists (i.e. if* $\mathbb{E}[P] < \beta$*) in which case* $\phi(\lambda_*, \mathcal{C}) \leq \beta$*.*

Lemma 5.3 suggests that, given $\mathcal{C}$, a good strategy is to bid:

$$b_t = \min(1, \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta)/\lambda_*),$$

at any round t. The following result shows that we can actually afford to settle for an approximate solution $\lambda \geq 0$ to $\phi(\lambda, \mathcal{C}) = \beta$.

Lemma 5.4. *For any* $\lambda_1, \lambda_2 \geq 0$*, we have:* $|R(\lambda_1, \mathcal{C}) - R(\lambda_2, \mathcal{C})| \leq 1/r \cdot |\phi(\lambda_1, \mathcal{C}) - \phi(\lambda_2, \mathcal{C})|$.

Lemma 5.3 combined with Lemma 5.4 suggests that the problem of computing a near-optimal bidding strategy essentially reduces to a stochastic root-finding problem for the function $|\phi(\cdot, \mathcal{C}) - \beta|$. As it turns out, the fact that the feedback is only partially censored makes a stochastic bisection search possible with minimal assumptions on $\phi(\cdot, \mathcal{C})$. Specifically, we only need that $\phi(\cdot, \mathcal{C})$ be Lipschitz, while the technique developed in [61] for a dynamic pricing problem requires $\phi(\cdot, \mathcal{C})$ to be bi-Lipschitz. This is a significant improvement because this last condition is not necessarily satisfied uniformly for all confidence sets $\mathcal{C}$, which will be important when we use a varying ellipsoidal confidence set instead of $\mathcal{C} = \{\theta_*\}$ in Section 5.3.2. Note, however, that Assumption 5.2 guarantees that $\phi(\cdot, \mathcal{C})$ is always Lipschitz, as we next show.

Lemma 5.5. $\phi(\cdot, \mathcal{C})$ *is* $\bar{L} \cdot \mathbb{E}[1/X^\intercal \theta_*]$*-Lipschitz.*

Specification of the algorithm. At any round $t \in \mathbb{N}$, we bid:

$$b_t = \min(1, \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta)/\lambda_t), \tag{5.9}$$

where $\lambda_t \geq 0$ is the current proxy for λ_*. We perform a binary search on λ_* by repeatedly using the same value of λ_t for consecutive rounds forming phases, indexed by $k \in \mathbb{N}$, and by keeping track of an interval, denoted by $I_k = [\underline{\lambda}_k, \bar{\lambda}_k]$. We start with phase $k = 0$ and

we initially set $\underline{\lambda}_0 = 0$ and $\bar{\lambda}_0 = 2/r$. The length of the interval is shrunk by half at the end of every phase so that $|I_k| = (2/r)/2^k$ for any k. Phase k lasts for $N_k = 3 \cdot 4^k \cdot \ln^2(T)$ rounds during which we set the value of λ_t to $\underline{\lambda}_k$. Since $\underline{\lambda}_k$ will be no larger than λ_* with high probability, this means that we tend to overbid. Note that there are at most $\bar{k}_T = \inf\{n \in \mathbb{N} \mid \sum_{k=0}^{n} N_k \geq T\}$ phases overall. The key observation enabling a bisection search approach is that, since the feedback is only partially censored, we can build, at the end of any phase k, an empirical estimate of $\phi(\lambda, \mathcal{C})$, which we denote by $\hat{\phi}_k(\lambda, \mathcal{C})$, for any $\lambda \geq \underline{\lambda}_k$ using all of the N_k samples obtained during phase k. The decision rule used to update I_k at the end of phase of k is specified next.

Algorithm: Interval updating procedure at the end of phase k

Data: $\bar{\lambda}_k$, $\underline{\lambda}_k$, $\Delta_k = 3\sqrt{2\ln(2T)/N_k}$, and $\hat{\phi}_k(\lambda, \mathcal{C})$ for any $\lambda \geq \underline{\lambda}_k$

Result: $\bar{\lambda}_{k+1}$ and $\underline{\lambda}_{k+1}$

$\bar{\gamma}_k = \bar{\lambda}_k, \underline{\gamma}_k = \underline{\lambda}_k$;

while $\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) > \beta + \Delta_k$ **do**

  $\bar{\gamma}_k = \bar{\gamma}_k + |I_k|$;

  $\underline{\gamma}_k = \underline{\gamma}_k + |I_k|$;

end

if $\hat{\phi}_k(1/2\bar{\gamma}_k + 1/2\underline{\gamma}_k, \mathcal{C}) \leq \beta + \Delta_k$ **then**

  $\bar{\lambda}_{k+1} = 1/2\bar{\gamma}_k + 1/2\underline{\gamma}_k$;

  $\underline{\lambda}_{k+1} = \underline{\gamma}_k$;

else

  $\bar{\lambda}_{k+1} = \bar{\gamma}_k$;

  $\underline{\lambda}_{k+1} = 1/2\bar{\gamma}_k + 1/2\underline{\gamma}_k$;

end

The splitting decision is trivial when $|\phi_k(1/2\bar{\gamma}_k + 1/2\underline{\gamma}_k, \mathcal{C}) - \beta| > \Delta_k$ because we get a clear signal that dominates the stochastic noise to either increase or decrease the current proxy for λ_*. The tricky situation is when $|\hat{\phi}_k(1/2\bar{\gamma}_k + 1/2\underline{\gamma}_k, \mathcal{C}) - \beta| \leq \Delta_k$, in which case the level of noise is too high to draw any conclusion. In this situation, we always favor a smaller value for $\underline{\lambda}_k$ even if that means shifting the interval upwards later on if we realize that we have made a mistake (which is the purpose of the while loop). This is

because we can always recover from underestimating λ_* since the feedback is only partially censored. Finally, note that the while loop of Algorithm 5 always ends after a finite number of iterations since $\hat{\phi}_k(2/r, \mathcal{C}) = 0 \leq \beta + \Delta_k$.

Regret analysis. Just like in Section 5.2, using concentration inequalities is essential to establish regret bounds but this time we need uniform concentration inequalities. We use the Rademacher complexity approach to concentration inequalities to control the deviations of $\hat{\phi}_k(\cdot, \mathcal{C})$ uniformly.

Lemma 5.6. *For any $k \in \{0, \cdots, \bar{k}_T\}$, we have:*

$$\mathbb{P}[\sup_{\lambda \in [\underline{\lambda}_k, 2/r]} |\hat{\phi}_k(\lambda, \mathcal{C}) - \phi(\lambda, \mathcal{C})| \leq \Delta_k] \geq 1 - 1/T.$$

Next, we bound the number of phases as a function of the time horizon.

Lemma 5.7. *For $T \geq 3$, we have:*

$$\bar{k}_T \leq \ln(T+1) \text{ and } 4^{\bar{k}_T} \leq \frac{T}{\ln^2(T)} + 1.$$

Using Lemma 5.6, we next show that the stochastic bisection search procedure correctly identifies $\lambda \geq 0$ such that $|\phi(\lambda, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})|$ is small with high probability, which is all we really need to lower bound the rewards accumulated in all rounds given Lemma 5.4

Proposition 5.1. *We have:*

$$\mathbb{P}[\cap_{k=0}^{\bar{k}_T} \{|\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq 4C \cdot |I_k|,\ |\phi(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq 3C \cdot |I_k|\}] \geq 1 - \frac{2\ln^2(T)}{T}$$

where $C = \bar{L} \cdot \mathbb{E}[1/X^\intercal \theta_*]$ and provided that $T \geq \exp(8r^2/C^2)$.

In a last step, we show, using the above result and at the cost of an additive logarithmic term in the regret bound, that we may assume that the advertiser participates in exactly T auctions. This enables us to combine Lemma 5.4, Lemma 5.7, and Proposition 5.1 to establish a distribution-independent regret bound.

Theorem 5.3. *Bidding according to* (5.9) *at any round* t *yields a regret:*

$$R_{B,T} = \tilde{O}(\frac{\bar{L} \cdot \mathbb{E}[1/X^\intercal \theta^*]}{r^2} \cdot \sqrt{T} \cdot \ln(T)).$$

Observe that Theorem 5.3 applies in particular when θ_* is known to the advertiser initially, in which case we can take $\mathcal{C} = \{\theta_*\}$. Furthermore, note that the regret bound derived is independent of d.

5.3.2 General Case

In this section, we combine the methods developed in Sections 5.2 and 5.3.1 to tackle the general case.

Specification of the algorithm. At any round $t \in \mathbb{N}$, we bid:

$$b_t = \min(1, \min(1, \max_{\theta \in \mathcal{C}_{\tau_t}} x_t^\intercal \theta)/\lambda_t), \tag{5.10}$$

where τ_t is defined in the last paragraph of Section 5.2 and $\lambda_t \geq 0$ is specified below. We use the bisection search method developed in Section 5.3.1 as a subroutine in a master algorithm that also runs in phases. Master phases are indexed by $q = 0, \cdots, Q$ and a new master phase starts whenever $\det(M_t)$ has increased by a factor at least $(1 + A)$ compared to the last time there was an update, for some $A > 0$ of our choosing. By construction, the ellipsoidal confidence set used during the q-th master phase is fixed so that we can denote it by $\mathcal{C}_q$. During the q-th master phase, we run the bisection search method described in Section 5.3.1 from scratch for the choice $\mathcal{C} = \mathcal{C}_q$ in order to identify a solution $\lambda_{q,*} \geq 0$ to $\phi(\lambda_{q,*}, \mathcal{C}_q) = \beta$ (or $\lambda_{q,*} = 0$ if no solution exists). Thus, λ_t is a proxy for $\lambda_{q,*}$ during the q-th master phase. This bisection search lasts for $\bar{k}_q$ phases and stops as soon as we move on to a new master phase. Hence, there are at most $\bar{k}_q \leq \bar{k}_T = \inf\{n \in \mathbb{N} \mid \sum_{k=0}^{n} N_k \geq T\}$ phases during the q-th master phase. We denote by $\underline{\lambda}_{q,k}$ the lower end of the interval used at the k-th phase of the bisection search run during the q-th master phase.

Regret analysis. While Q is, in general, a random variable, we can always guarantee that there are at most $O(d \cdot \ln(T \cdot d))$ master phases overall.

Lemma 5.8. *We have* $Q \leq \bar{Q} = d \cdot \ln(T \cdot d)/\ln(1 + A)$ *almost surely.*

Lemma 5.8 is important because it implies that the bisection searches run long enough to be able to identify sufficiently good approximate values for $\lambda_{q,*}$. Note that our approach is "doubly" optimistic since both $\underline{\lambda}_{q,k} \leq \lambda_{q,*}$ and $\theta_* \in \mathcal{C}_q$ hold with high probability at any point in time. At a high level, the regret analysis goes as follows. First, just like in Section 5.3.1, we show, using Proposition 5.1 and at the cost of an additive logarithmic term in the final regret bound, that we may assume that the advertiser participates in exactly T auctions. Second, we show, using the analysis of Theorem 5.2, that we may assume that the expected per-round reward obtained during phase q is $\mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_q} x_t^\intercal \theta)]$ (as opposed to $x_t^\intercal \theta_*$) at any round t, up to an additive term of order $\tilde{O}(d \cdot \sqrt{T})$ in the final regret bound. Third, we note that Theorem 5.3 essentially shows that the expected per-round reward obtained during phase q is $R(\lambda_{q,*}, \mathcal{C}_q)$, up to an additive term of order $\tilde{O}(\sqrt{T})$ in the final regret bound. Finally, what remains to be done is to compare $R(\lambda_{q,*}, \mathcal{C}_q)$ with $R(\lambda_*, \{\theta_*\})$, which we do using Lemmas 5.2 and 5.3.

Theorem 5.4. *Bidding according to* (5.10) *at any round* t *yields a regret:*

$$R_{B,T} = \tilde{O}(d \cdot \frac{\bar{L} \cdot \mathbb{E}[1/X^\intercal \theta^*]}{r^2} \cdot (\frac{1}{\ln(1 + A)} + \sqrt{1 + A}) \cdot \sqrt{T}).$$

5.4 Concluding Remark

An interesting direction for future research is to develop a complete characterization of the achievable regret bounds, in particular through the derivation of lower bounds on regret. When there is no budget limit and no contextual information, the authors of [97] provide a very thorough characterization with rates ranging from $\Theta(\ln(T))$, when a margin condition on the underlying distribution is satisfied, to $\Theta(\sqrt{T})$ when this condition is not satisfied.

Chapter 6

Concluding Remarks

We conclude this thesis with a summary of contributions, with an emphasis on technical contributions, and a discussion of future research directions.

6.1 Summary

In Chapter 1, we introduce the class of adaptive optimization problems considered in this thesis and we describe two possible approaches to deal with the fact that the environment is uncertain: a worst-case approach, which protects against a fully adversarial environment, and a hindsight approach, which adapts to the level of adversariality by measuring performance in terms of a quantity known as regret, defined as the gap between the objective function derived and the best objective function that could have been obtained in hindsight.

In Chapter 2, we take the worst-case approach and study adaptive stochastic shortest path problems with a deadline imposed at the destination under distributional ambiguity. We develop an efficient algorithm based on a trick to avoid recomputing the value function from scratch at each state when solving the underlying dynamic program. Our algorithm provably solves the problem to near-optimality with a runtime complexity that is comparable, up to logarithmic factors in the parameters, to the runtime complexity of the optimal algorithm solving the nominal version of the problem (i.e. when the distributions are known).

In Chapter 3, we establish, using tools rooted in duality theory and sensitivity analy-

sis, that the minimax achievable regret in the online convex optimization framework when the loss function is piecewise linear always scales as square root of the time horizon when the decision maker's decision set is strongly curved. In stark contrast, we show, using sensitivity analysis, that the Follow-The-Leader algorithm incurs a regret that scales as a logarithmic function of the time horizon when the decision maker's decision set is curved, the loss function is linear, and 0 does not lie in the convex hull of the environment's decision set.

In Chapter 4, we study the Bandits with Knapsacks framework and we develop UCB-type algorithms with distribution-dependent bounds on regret that scale as a logarithmic function of the initial endowments of each resource, thus generalizing the result of [12] for the standard Multi-Armed Bandit problem. The key idea, borrowed from the literature, to design algorithms is to upper bound the performance of the optimal algorithm that knows the underlying distributions by a simple linear program whose decision variables are the anytime probabilities of pulling each arm. This linear program only involves a few unknown quantities (namely the mean rewards and costs). Following the optimism in the face of uncertainty, we plug optimistic empirical estimates for these quantities into the linear program and pull an arm at random according to a slightly perturbed version of the optimal distribution.

In Chapter 5, we model the problem of repeated bidding in online advertisement auctions as a contextual Bandits with Knapsack problem with a continuum of arm. We develop UCB-type algorithms that combine the ellipsoid confidence set-based approach to linear contextual Multi-Armed Bandit problems with a special-purpose stochastic binary search procedure. We establish distribution-independent bounds on regret that scale as square root of the initial budget. Similarly as in Chapter 4, the key to design algorithms is to upper bound the performance of the optimal algorithm that knows the underlying distributions by a simple optimization problem which only involves a few unknown quantities. Following the optimism in the face of uncertainty, we plug optimistic empirical estimates for these quantities into the optimization problem and use the optimal solution derived as a basis for the next bid submission.

6.2 Future Research Directions

This thesis focuses on optimization problems where decisions can be made and implemented globally. In fast-paced environments, such as in real-time bidding, this is often not physically possible given: (i) the volume of decisions and (ii) the fact that decisions have to be made in a timely fashion (advertisers have about 100 milliseconds to submit a bid once an ad auction is posted). As a consequence, computations have to be distributed across machines and decisions have to be made locally and concurrently. This raises several questions for the decision maker. First, how should the global optimization problem be broken down? For instance, each machine could have its own local optimization problem to solve which would be defined online by a master machine collecting the feedback from all machines. Since communications between machines are subject to delays and failures, this implies that the feedback on the outcomes of alternatives is no longer immediately available, which is an assumption underlying the work presented in this thesis. Second, in settings where there is an inherent exploration-exploitation trade-off, how should the exploration component be handled? Anecdotal evidence, see, for example, [49], suggests that even minimal collaboration between machines can increase the performance tremendously compared to the situation where each machine is managing the trade-off locally. Developing provably near-optimal algorithms in the distributed setting remains an active area of research under the general areas of collaborative distributed learning and distributed optimization.

Bibliography

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Adv. Neural Inform. Processing Systems*, pages 2312–2320, 2011.

[2] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proc. 22nd Annual Conf. Learning Theory*, 2009.

[3] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proc. 21st Annual Conf. Learning Theory*, pages 415–424, 2008.

[4] Y. Adulyasak and P. Jaillet. Models and algorithms for stochastic and robust vehicle routing with deadlines. *Transportation Sci.*, 50(2):608–626, 2015.

[5] S. Agrawal and N. Devanur. Linear contextual bandits with knapsacks. In *Adv. Neural Inform. Processing Systems*, pages 3450–3458, 2016.

[6] S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proc. 15th ACM Conf. Economics and Comput.*, pages 989–1006, 2014.

[7] S. Agrawal, N. R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proc. 29th Annual Conf. Learning Theory*, pages 4–18, 2016.

[8] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proc. 25th Annual Conf. Learning Theory*, volume 23, 2012.

[9] A. Andrew. Another efficient algorithm for convex hulls in two dimensions. *Inform. Processing Lett.*, 9(5):216–219, 1979.

[10] L. Andrew, S. Barman, K. Ligett, M. Lin, A. Meyerson, A. Roytman, and A. Wierman. A tale of two metrics: Simultaneous bounds on competitiveness and regret. In *Proc. 2013 ACM SIGMETRICS Int. Conf. Measurement Modeling Computer Systems*, pages 329–330, 2013.

[11] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learning Res.*, 3(Nov):397–422, 2002.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[14] M. Babaioff, S. Dughmi, R. Kleinberg, and A. Slivkins. Dynamic pricing with limited supply. In *Proc. 13th ACM Conf. Electronic Commerce*, pages 74–91, 2012.

[15] A. Badanidiyuru, R. Kleinberg, and Y. Singer. Learning on a budget: posted price mechanisms for online procurement. In *Proc. 13th ACM Conf. Electronic Commerce*, pages 128–145, 2012.

[16] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *Proc. 54th IEEE Annual Symp. Foundations of Comput. Sci.*, pages 207–216, 2013.

[17] A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Proc. 27th Annual Conf. Learning Theory*, volume 35, pages 1109–1134, 2014.

[18] M. Balcan, A. Blum, N. Haghtalab, and A. D. Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proc. 16th ACM Conf. Economics and Comput.*, pages 61–78, 2015.

[19] S. Baltaoglu, L. Tong, and Q. Zhao. Online learning of optimal bidding strategy in repeated multi-commodity auctions. *Working Paper*, 2017.

[20] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Machine Learning Res.*, 3(Nov):463–482, 2002.

[21] A. Ben-Tal and E. Hochman. More bounds on the expectation of a convex function of a random variable. *J. Applied Probability*, pages 803–812, 1972.

[22] D. P. Bertsekas and J. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16(3):580–595, 1991.

[23] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.*, 15(3):780–804, 2005.

[24] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific, 1997.

[25] O. Besbes and A. Zeevi. Blind network revenue management. *Oper. Res.*, 60(6):1537–1550, 2012.

[26] A. Blum and Y. Mansour. From external to internal regret. *J. Machine Learning Res.*, 8:1307–1324, 2007.

[27] V. Bonifaci, T. Harks, and G. Schäfer. Stackelberg routing in arbitrary networks. *Math. Oper. Res.*, 35(2):330–346, 2010.

[28] A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 2005.

[29] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statist.*, 9:323–375, 2005.

[30] G. S. Brodal and R. Jacob. Dynamic planar convex hull. In *Proc. 43rd IEEE Annual Symp. Foundations Comput. Sci.*, pages 617–626, 2002.

[31] C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *J. Optim. Theory and Applications*, 130(1):1–22, 2006.

[32] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals Statist.*, 27(6):1865–1895, 1999.

[33] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[34] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *J. Machine Learning Res. - Proc.*, volume 15, pages 208–214, 2011.

[35] M. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. In *Proc. 17th ACM Conf. Economics and Comput.*, pages 817–817, 2016.

[36] R. Combes, C. Jian, and R. Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. In *Proc. 2015 ACM SIGMETRICS Int. Conf. Measurement Modeling Computer Systems*, pages 245–257, 2015.

[37] B. Dean. Speeding up stochastic dynamic programming with zero-delay convolution. *Algorithmic Oper. Res.*, 5(2):96–104, 2010.

[38] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.*, 58(3):595–612, 2010.

[39] W. Ding, T. Qin, X. Zhang, and T. Liu. Multi armed bandit with budget constraint and variable costs. In *Proc. 27th AAAI Conf. Artificial Intelligence*, pages 232–238, 2013.

[40] Y. Fan, R. Kalaba, and I. Moore. Arriving on time. *J. Optim. Theory and Applications*, 127(3):497–513, 2005.

[41] H. Frank. Shortest paths in probabilistic graphs. *Oper. Res.*, 17(4):583–599, 1969.

[42] Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

[43] A. Ghosh, B. I. P. Rubinstein, S. Vassilvitskii, and M. Zinkevich. Adaptive bidding for display advertising. In *Proc. 18th Int. Conf. World Wide Web*, pages 251–260, 2009.

[44] D. Graczová and P. Jacko. Generalized restless bandits and the knapsack problem for perishable inventories. *Oper. Res.*, 62(3):696–711, 2014.

[45] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer, 2012.

[46] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[47] E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010.

[48] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

[49] E. Hillel, Z. Karnin, T. Koren, R. Lempel, and O. Somekh. Distributed exploration in multi-armed bandits. In *Adv. Neural Inform. Processing Systems*, pages 854–862, 2013.

[50] D. Hoy and E. Nikolova. Approximately optimal risk-averse routing policies via adaptive discretization. In *Proc. 29th Int. Conf. Artificial Intelligence*, pages 3533–3539, 2015.

[51] G. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005.

[52] P. Jaillet, J. Qi, and M. Sim. Routing optimization under uncertainty. *Oper. Res.*, 64(1):186–200, 2016.

[53] M. Jain, V. Conitzer, and M. Tambe. Security scheduling for real-world networks. In *Proc. 2013 Int. Conf. Autonomous Agents and MultiAgent Systems*, pages 215–222, 2013.

[54] A. X. Jiang, Z. Yin, M. P. Johnson, M. Tambe, C. Kiekintveld, K. Leyton-Brown, and T. Sandholm. Towards optimal patrol strategies for fare inspection in transit systems. In *AAAI Spring Symposium: Game Theory Security, Sustainability, and Health*, 2012.

[55] K. Johnson, D. Simchi-Levi, and H. Wang. Online network revenue management using thompson sampling. *Working Paper*, 2015.

[56] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *J. Comput. System Sci.*, 71(3):291–307, 2005.

[57] S. Kim, G. B. Giannakis, and K. Y. Lee. Online optimal power flow with renewables. In *48th Asilomar Conf. Signals, Systems and Comput.*, pages 355–360. IEEE, 2014.

[58] R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proc. 44th IEEE Annual Symp. Foundations of Comput. Sci.*, pages 594–605, 2003.

[59] S. Krumke. *Competitive analysis and beyond*. PhD thesis, Habilitationsschrift, Technische Universität Berlin, 2002.

[60] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Applied Math.*, 6(1):4–22, 1985.

[61] Y. Lei, S. Jasin, and A. Sinha. Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. *Working Paper*, 2015.

[62] R. P. Loui. Optimal paths in graphs with stochastic or multidimensional weights. *Comm. ACM*, 26(9):670–676, 1983.

[63] G. Lueker. Average-case analysis of off-line and on-line knapsack problems. *Journal of Algorithms*, 29(2):277–305, 1998.

[64] E. Miller-Hooks and H. Mahmassani. Least expected time paths in stochastic, time-varying transportation networks. *Transportation Sci.*, 34(2):198–215, 2000.

[65] Y. Nie and Y. Fan. Arriving-on-time problem: discrete algorithm that ensures convergence. *Transportation Res. Record*, 1964:193–200, 2006.

[66] Y. Nie and X. Wu. Shortest path problem considering on-time arrival probability. *Transportation Res. B*, 43(6):597–613, 2009.

[67] E. Nikolova, M. Brand, and D. R. Karger. Optimal route planning under uncertainty. In *Proc. Int. Conf. Automated Planning Scheduling*, pages 131–140, 2006.

[68] E. Nikolova, J. A. Kelner, M. Brand, and M. Mitzenmacher. Stochastic shortest paths via quasi-convex maximization. In *Proc. 14th Annual Eur. Sympos. Algorithms*, pages 552–563, 2006.

[69] A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, 2005.

[70] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press, 2007.

[71] E. Ordentlich and T. M. Cover. The cost of achieving the best portfolio in hindsight. *Math. Oper. Res.*, 23(4):960–982, 1998.

[72] M. H. Overmars and J. Van Leeuwen. Maintenance of configurations in the plane. *J. Comput. and System Sci.*, 23(2):166–204, 1981.

[73] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24(2):293–305, 1999.

[74] A. Parmentier and F. Meunier. Stochastic shortest paths and risk measures. *arXiv preprint arXiv:1408.0272*, 2014.

[75] R. Pasupathy and S. Kim. The stochastic root-finding problem: Overview, solutions, and open questions. *ACM Trans. Modeling and Comput. Simulation*, 21(3):19, 2011.

[76] Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel J. Math.*, 20(3-4):326–350, 1975.

[77] A. Prékopa. The discrete moment problem and linear programming. *Discrete Applied Math.*, 27(3):235–254, 1990.

[78] M. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[79] A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Adv. Neural Inform. Processing Systems*, pages 3066–3074, 2013.

[80] A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. *arXiv preprint arXiv:1510.03925*, 2015.

[81] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *Proc. 24th Annual Conf. Learning Theory*, pages 559–594, 2011.

[82] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *J. Machine Learning Res.*, 16(1):155–186, 2015.

[83] H. Robbins. Some aspects of the sequential design of experiments. *Bulleting American Math. Society*, 58(5):527–535, 1952.

[84] S. Samaranayake, S. Blandin, and A. Bayen. Speedup techniques for the stochastic on-time arrival problem. In *12th Workshop Algorithmic Approaches Transportation Model. Optim. Systems*, volume 25, pages 83–96, 2012.

[85] S. Samaranayake, S. Blandin, and A. Bayen. A tractable class of algorithms for reliable routing in stochastic networks. *Transportation Res. C*, 20(1):199–217, 2012.

[86] A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In *Adv. Neural Inform. Processing Systems*, pages 810–818, 2014.

[87] A. Shapiro. On duality theory of conic linear problems. In *Semi-Infinite Programming*, volume 57, pages 135–165. Kluwer Academic Publishers, 2001.

[88] A. Slivkins. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*, 2013.

[89] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. ϵ-first policies for budget-limited multi-armed bandits. In *Proc. 24th AAAI Conf. Artificial Intelligence*, pages 1211–1216, 2010.

[90] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proc. 26th AAAI Conf. Artificial Intelligence*, pages 1134–1140, 2012.

[91] L. Tran-Thanh, A. Rogers, and N. R. Jennings. Long-term information collection with energy harvesting wireless sensors: a multi-armed bandit based approach. *Autonomous Agents and Multi-Agent Systems*, 25(2):352–394, 2012.

[92] L. Tran-Thanh, C. Stavrogiannis, V. Naroditskiy, V. Robu, N. R. Jennings, and P. Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *Proc. 30th Conf. Uncertainty in Artificial Intelligence*, pages 809–818, 2014.

[93] F. Trovo, S. Paladino, M. Restelli, and N. Gatti. Budgeted multi–armed bandit in continuous action space. *Working Paper*, 2016.

[94] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyschev bounds via semidefinite programming. *SIAM Rev.*, 49(1):52–64, 2007.

[95] V. Vovk. Aggregating strategies. In *Proc. 3rd Workshop Comput. Learning Theory*, pages 371–383, 1990.

[96] V. Vovk. Competitive on-line linear regression. In *Adv. Neural Inform. Processing Systems*, pages 364–370, 1998.

[97] J. Weed, V. Perchet, and P. Rigollet. Online learning in repeated auctions. In *Proc. 29th Annual Conf. Learning Theory*, volume 49, pages 1562–1583, 2016.

[98] C. C. White III and H. K. Eldeib. Markov decision processes with imprecise transition probabilities. *Oper. Res.*, 42(4):739–749, 1994.

[99] W. Wiesemann, D. Kuhn, and B. Rustem. Robust markov decision processes. *Math. Oper. Res.*, 38(1):153–183, 2013.

[100] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Oper. Res.*, 62(6):1358–1376, 2014.

[101] H. Wu, S. Srikant, X. Liu, and C. Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Adv. Neural Inform. Processing Systems*, pages 433–441, 2015.

[102] Y. Xia, W. Ding, X. Zhang, N. Yu, and T. Qin. Budgeted bandit problems with continuous random costs. In *Proc. 7th Asian Conf. Machine Learning*, pages 317–332, 2015.

[103] H. Xu, C. Caramanis, and S. Mannor. Optimization under probabilistic envelope constraints. *Oper. Res.*, 60(3):682–699, 2012.

[104] H. Xu and S. Mannor. Distributionally robust markov decision processes. In *Adv. Neural Inform. Processing Systems*, pages 2505–2513, 2010.

[105] H. Xu and S. Mannor. Probabilistic goal Markov decision processes. In *Proc. 22th Int. Joint Conf. Artificial Intelligence*, pages 2046–2052, 2011.

[106] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Int. Conf. Machine Learning*, pages 928–936, 2003.

Appendix A

Appendix for Chapter 2

A.1 Tailored Dynamic Convex Hull Algorithm

The fact that deletions and insertions always occur on the same side of the set allows us to deal with deletions in an indirect way, by building and merging upper convex hulls of partial input data. The only downside is that this requires an efficient merging procedure. In this respect, we state without proof a result derived from [72].

Lemma A.1. *Consider a set S of N points in $\mathbb{R}^2$ partitioned into two sets of points S_1 and S_2 such that, for any two points $(x_1, y_1) \in S_1$ and $(x_2, y_2) \in S_2$ we have $x_1 < x_2$. Suppose that the extreme points of $\hat{S}_1$ (resp. $\hat{S}_2$) are stored in an array A_1 (resp. A_2) of size N in ascending order of their first coordinates. We can find two indices l_1 and l_2 in $O(\log(N))$ time such that the set comprised of the points contained in A_1 with index smaller than l_1 and the points contained in A_2 with index larger than l_2 is precisely the set of extreme points of $\hat{S}$.*

Algorithm. We use two arrays A_{left} and A_{right} along with a stack $\mathcal{S}$. The arrays A_{left} and A_{right} are of size $L' - L + 1$, indexed from 0 to $L' - L$, and store points in $\mathbb{R}^2$ in ascending order of their first coordinates. The stack $\mathcal{S}$ stores stacks of points in $\mathbb{R}^2$. We keep track of two indices l_{left} and l_{right} such that, at any step $k = k_i^{r,\min} + p \cdot (L' - L + 2) + r$ for some $p \in \mathbb{N}$ and $0 \leq r \leq L' - L + 1$, the following invariant holds:

- $\{A_{\text{left}}[l], l = l_{\text{left}} + 1, \cdots, L' - L\}$ is the set of extreme points of the upper convex hull of $\{(l \cdot \Delta t, u_j^{\Delta t}(l \cdot \Delta t)),\ l = k - L', \cdots, k - L - r\}$,

- $\{A_{\text{right}}[l], l = 0, \cdots, l_{\text{right}} - 1\}$ is the set of extreme points of the upper convex hull of $\{(l \cdot \Delta t, u_j^{\Delta t}(l \cdot \Delta t)),\ l = k - L - r + 1, \cdots, k - L\}$.

Using the procedure of Lemma A.1 and this invariant, we can find a pair of indices (l_1, l_2) in $O(\log(L' - L))$ time such that $\{A_{\text{left}}[l], l = l_{\text{left}} + 1, \cdots, l_1\} \cup \{A_{\text{right}}[l], l = l_2, \cdots, l_{\text{right}} - 1\}$ is the set of extreme points of $\hat{\mathcal{C}}_k$. Hence, all we have left to do is to provide a procedure to maintain A_{left}, A_{right}, l_{left} and l_{right}, which we do next.

A_{left}, A_{right}, and $\mathcal{S}$ are initially empty. The algorithm proceeds in two phases and loops back to the first one every $L' - L + 2$ steps. For convenience, we define "cross" as the function taking as an input three points a, b, c in $\mathbb{R}^2$ and returning the cross product of the vector $\vec{ab}$ and $\vec{ac}$.

Phase 1: Suppose that the current step is k. Hence, the values $u_j^{\Delta t}(k - L'), \cdots, u_j^{\Delta t}(k - L)$ are available. This phase is based on Andrew's monotone chain convex hull algorithm to find the extreme points of $\hat{\mathcal{C}}_k$ with the difference that we store the points removed along the process in stacks for future use. Specifically, set $l_{\text{left}} = L' - L$ and $l_{\text{right}} = 0$ and for l decreasing from $k - L$ to $k - L'$, do the following:

(a) Initialize a new stack $\mathcal{S}'$,

(b) While $l_{\text{left}} \leq L' - L - 2$ and $\text{cross}(A_{\text{left}}[l_{\text{left}} + 2], A_{\text{left}}[l_{\text{left}} + 1], (l \cdot \Delta t, u_j^{\Delta t}(l \cdot \Delta t))) \geq 0$:

- Push $A_{\text{left}}[l_{\text{left}} + 1]$ to $\mathcal{S}'$,
- Increment l_{left},

(c) Push $\mathcal{S}'$ to $\mathcal{S}$,

(d) Set $A_{\text{left}}[l_{\text{left}}] = (l \cdot \Delta t, u_j^{\Delta t}(l \cdot \Delta t))$ and decrement l_{left}. At this point, $\{A_{\text{left}}[l], l = l_{\text{left}} + 1, \cdots, L' - L\}$ is the set of extreme points of the upper convex hull of $\{(m \cdot \Delta t, u_j^{\Delta t}(m \cdot \Delta t)),\ m = l, \cdots, k - L\}$.

Phase 2: At step $k + l$, for l increasing from 1 to $L' - L$, observe that the value $u_j^{\Delta t}(k + l - L)$ becomes available. To maintain A_{left} and l_{left}, we remove the leftmost point (x, y) and

reinsert the points, stored in the topmost stack of $\mathcal{S}$, that were previously removed from A_{left} when appending (x, y) to A_{left} in the course of running Andrew's monotone chain convex hull algorithm. Specifically:

(a) Increment l_{left},

(b) Pop the topmost stack $\mathcal{S}'$ out of $\mathcal{S}$,

(c) While $\mathcal{S}'$ is not empty:

- Pop the topmost point (x, y) of $\mathcal{S}'$,
- Set $A_{\text{left}}[l_{\text{left}}] = (x, y)$,
- Decrement l_{left}.

To maintain A_{right} and l_{right}, we run an iteration of Andrew's monotone chain convex hull algorithm. Specifically:

(a) While $l_{\text{right}} \geq 2$ and $\text{cross}(A_{\text{right}}[l_{\text{right}} - 2], A_{\text{right}}[l_{\text{right}} - 1], ((k + l - L) \cdot \Delta t, u_j^{\Delta t}((k + l - L) \cdot \Delta t))) \leq 0$:

- Decrement l_{right},

(b) Set $A_{\text{right}}[l_{\text{right}}] = ((k + l - L) \cdot \Delta t, u_j^{\Delta t}((k + l - L) \cdot \Delta t))$ and increment l_{right}.

Complexity Analysis. Observe that any point added to A_{left} can only be removed once, and the same holds for A_{right}. This means that Phase 1 and Phase 2 take $O(L' - L)$ computation time. These two phases are repeated $\left\lceil \frac{\left\lfloor \frac{T}{\Delta t} \right\rfloor - k_i^{r,\min}}{L' - L + 2} \right\rceil$ times leading to an overall complexity of $O(\left\lfloor \frac{T}{\Delta t} \right\rfloor - k_i^{r,\min})$. Since the merging procedure outlined in Lemma A.1 takes $O(\log(L' - L))$ computation time at each step, the global complexity is $O((\left\lfloor \frac{T}{\Delta t} \right\rfloor - k_i^{r,\min}) \cdot \log(L' - L))$.

A.2 Omitted Proofs

A.2.1 Proof of Theorem 2.1

Proof. Proof of Theorem 2.1. We denote by $\mathcal{H}$ the set of all possible histories of the previously experienced costs and previously visited nodes. Let us start with the last part of the theorem. If the support of $f(\cdot)$ is included in $[T_f, \infty)$, any strategy is optimal when having already spent a budget of $T - T_f$ with an optimal objective function of 0.

Let us now focus on the first part of the theorem. Consider an optimal strategy π_f^* solution to (2.1). For a given history $h \in \mathcal{H}$, we define t_h as the remaining budget, i.e. T minus the total cost spent so far, and i_h as the current location. The policy π_f^* maps $h \in \mathcal{H}$ to a probability distribution over $\mathcal{V}(i_h)$. Observe that randomizing does not help because the costs are independent across time and arcs so that, without loss of generality, we can assume that π_f^* actually maps h to the node in $\mathcal{V}(i_h)$ minimizing the objective function given h. For $h \in \mathcal{H}$, we denote by $X^h_{\pi_f^*}$ the random cost-to-go incurred by following strategy π_f^*, i.e. not including the total cost spent up to this point of the history $T - t_h$. We define $(m_{ij})_{(i,j)\in\mathcal{A}}$ as the expected arc costs and M_i as the minimum expected cost incurred to reach d from $i \in \mathcal{V}$. We also define π_s as a policy associated with an arbitrary shortest path from i to d with respect to the expected costs. Specifically, π_s maps the current location i_h to a node in $\mathcal{T}(i_h)$, irrespective of the history of the process. Similarly as for π_f^*, we denote by $X^h_{\pi_s}$ the random cost-to-go incurred by following strategy π_s for $h \in \mathcal{H}$. We first show that there exists T_f such that, for both cases (a) and (b):

$$\mathbb{E}[X^h_{\pi_f^*}] - M_{i_h} < \min_{i\neq d} \min_{j\in\mathcal{V}(i), j\notin\mathcal{T}(i)} \{m_{ij} + M_j - M_i\} \quad \forall h \in \mathcal{H} \text{ such that } t_h \leq T_f, \text{ (A.1)}$$

with the convention that the minimum of an empty set is equal to infinity. Note that the right-hand side is always positive. Let $\alpha = |\mathcal{V}| \cdot \delta^{\sup}$.

(a) For $h \in \mathcal{H}$ such that $t_h < T_1$, we have, using a Taylor's series expansion:

$$f(t_h - X^h_{\pi_f^*}) = f(t_h - \alpha) + f'(t_h - \alpha) \cdot (\alpha - X^h_{\pi_f^*}) + \frac{1}{2} \cdot f''(\xi_h) \cdot (\alpha - X^h_{\pi_f^*})^2,$$

where $\xi_h \in [\min(t_h - \alpha, t_h - X^h_{\pi^*_f}), \max(t_h - \alpha, t_h - X^h_{\pi^*_f})]$, and:

$$f(t_h - X^h_{\pi_s}) = f(t_h - \alpha) + f'(t_h - \alpha) \cdot (\alpha - X^h_{\pi_s}) + \frac{1}{2} \cdot f''(\zeta_h) \cdot (\alpha - X^h_{\pi_s})^2,$$

where $\zeta_h \in [\min(t_h - \alpha, t_h - X^h_{\pi_s}), \max(t_h - \alpha, t_h - X^h_{\pi_s})]$. Using ***Bellman's Principle of Optimality*** for π^*_f, we have:

$$\mathbb{E}[f(t_h - X^h_{\pi^*_f})] = \mathbb{E}[f(T - ((T - t_h) + X^h_{\pi^*_f})] \geq \mathbb{E}[f(T - ((T - t_h) + X^h_{\pi_s})] \geq \mathbb{E}[f(t_h - X^h_{\pi_s})].$$

Expanding and rearranging yields:

$$-f'(t_h - \alpha) \cdot (\mathbb{E}[X^h_{\pi^*_f}] - \mathbb{E}[X^h_{\pi_s}]) \geq \frac{1}{2} \cdot (\mathbb{E}[-f''(\xi_h) \cdot (\alpha - X^h_{\pi^*_f})^2] + \mathbb{E}[f''(\zeta_h) \cdot (\alpha - X^h_{\pi_s})^2]).$$

Since the costs are independent across time and arcs:

$$\mathbb{E}[X^h_{\pi_s}] = M_{i_h}.$$

Concavity of $f(\cdot)$ implies that $f''(\xi_h) \cdot (\alpha - X^h_{\pi^*_f})^2 \leq 0$ almost surely. Since $f(\cdot)_{(-\infty, T_1)}$ is increasing, we obtain $\mathbb{E}[X^h_{\pi^*_f}] - M_{i_h} \leq \frac{\mathbb{E}[-f''(\zeta_h) \cdot (\alpha - X^h_{\pi_s})^2]}{2 \cdot f'(t_h - \alpha)}$. As $X^h_{\pi_s}$ is the cost of a path, Assumption 2.3 implies $0 \leq X^h_{\pi_s} \leq \alpha$. We get that $\zeta_h \in [t_h - \alpha, t_h]$ and:

$$\mathbb{E}[X^h_{\pi^*_f}] - M_{i_h} \leq -\alpha^2 \cdot \frac{\inf\limits_{[t_h - \alpha, t_h]} f''}{2 \cdot f'(t_h - \alpha)}.$$

As $f''(\cdot)$ is continuous, there exists $\alpha_{t_h} \in [0, \alpha]$ such that $\inf\limits_{[t_h - \alpha, t_h]} f'' = f''(t_h - \alpha_{t_h})$. Since $f'(\cdot)$ is non-increasing on $(-\infty, T_1)$, we derive:

$$\mathbb{E}[X^h_{\pi^*_f}] - M_{i_h} \leq -\alpha^2 \cdot \frac{f''(t_h - \alpha_{t_h})}{2 \cdot f'(t_h - \alpha_{t_h})}.$$

By assumption $\frac{f''}{f'}(\cdot)$ vanishes at $-\infty$ therefore we can pick T_f small enough to get the desired inequality.

(b) As $f' \to_{-\infty} a > 0$, we can find $T_f < T_1$ small enough such that:

$$|f'(t) - a| < \epsilon \quad \forall t \leq T_f,$$

with $\epsilon = a \cdot \frac{\beta}{2\alpha+\beta}$ and where β is the right-hand side of the desired inequality. Consider $h \in \mathcal{H}$ such that $t_h \leq T_f$. Using *Bellman's Principle of Optimality* for π_f^*, we have:

$$\mathbb{E}[f(t_h - X^h_{\pi_f^*})] = \mathbb{E}[f(T-((T-t_h)+X^h_{\pi_f^*})] \geq \mathbb{E}[f(T-((T-t_h)+X^h_{\pi_s})] \geq \mathbb{E}[f(t_h - X^h_{\pi_s})].$$

Since f is C^1 on $(-\infty, T_f)$, this yields:

$$\begin{aligned}
0 &\leq \mathbb{E}[f(t_h - X^h_{\pi_f^*}) - f(t_h - X^h_{\pi_s})] \\
&\leq \mathbb{E}[f(t_h - X^h_{\pi_f^*}) - f(t_h) + f(t_h) - f(t_h - X^h_{\pi_s})] \\
&\leq \mathbb{E}[-\int_{t_h - X^h_{\pi_f^*}}^{t_h} f' + \int_{t_h - X^h_{\pi_s}}^{t_h} f'] \\
&\leq \mathbb{E}[-(a-\epsilon) \cdot X^h_{\pi_f^*} + (a+\epsilon) \cdot X^h_{\pi_s}].
\end{aligned}$$

Since the costs are independent across time and arcs:

$$\mathbb{E}[X^h_{\pi_s}] = M_{i_h}.$$

Rearranging the last inequality, we derive:

$$\begin{aligned}
\mathbb{E}[X^h_{\pi_f^*}] - M_{i_h} &\leq \frac{2\epsilon}{a-\epsilon} \cdot M_{i_h} \\
&\leq \frac{2\epsilon}{a-\epsilon} \cdot \alpha \\
&< \beta,
\end{aligned}$$

where we use the fact that $M_{i_h} \leq \alpha$ and the definition of ϵ.

Starting from (A.1), consider $h \in \mathcal{H}$ such that $t_h \leq T_f$ and suppose by contradiction that $\pi_f^*(h) = j_h \notin \mathcal{T}(i_h)$. Even though the overall policy can be fairly complicated (history-dependent), the first action is deterministic and incurs an expected cost of $m_{i_h j_h}$ because the

costs are independent across time and arcs. Moreover, when the objective is to minimize the average cost, the optimal strategy among all history-dependent rules is to follow the shortest path with respect to the mean arc costs (once again because the costs are independent across time and arcs). As a result:

$$\mathbb{E}[X^h_{\pi^*_f}] \geq m_{i_h j_h} + M_{j_h},$$

which implies:

$$\mathbb{E}[X^h_{\pi^*_f}] - M_{i_h} \geq m_{i_h j_h} + M_{j_h} - M_{i_h},$$

a contradiction.

□

A.2.2 Proof of Proposition 2.1

Proof. Proof of Proposition 2.1. Using Theorem 2.1, the optimization problem (2.1) can be equivalently formulated as a discrete-time finite-horizon MDP in the extended space state $(i,t) \in \mathcal{V} \times [T - \delta^{\sup} \cdot \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil, T]$ where i is the current location and t is the, possibly negative, remaining budget. Specifically:

- The time horizon is $\left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil$.
- The initial state is (s,T).
- The set of available actions at state (i,t), for $i \neq d$, is taken as $\mathcal{V}(i)$. Picking $j \in \mathcal{V}(i)$ corresponds to crossing link (i,j) and results in a transition to state $(j, t-\omega)$ with probability $p_{ij}(\omega)\mathrm{d}\omega$.
- The only available action at a state (d,t) is to remain in this state.
- The transition rewards are all equal to 0.
- The final reward at the epoch $\left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil$ for any state (i,t) is equal to $f_i(t)$, which is the optimal expected objective-to-go when following the shortest path tree $\mathcal{T}$ starting at node i with remaining budget t. Specifically, the collection of functions $(f_i(\cdot))_{i\in\mathcal{V}}$ is

a solution to the following program:

$$\begin{aligned} f_d(t) &= f(t), && t \leq T_f, \\ f_i(t) &= \max_{j \in \mathcal{T}(i)} \int_0^\infty p_{ij}(\omega) \cdot f_j(t-\omega) \mathrm{d}\omega && i \neq d, t \leq T_f. \end{aligned}$$

Observe that Theorem 2.1 is crucial to be able to define the final rewards. Proposition 4.4.3 of [78] shows that any *Markov* policy solution to (2.4) is an optimal solution to (2.1).

□

A.2.3 Proof of Proposition 2.2

Proof. Proof of Proposition 2.2. For any node $i \in \mathcal{V}$, and $t \leq T$, we denote by $u_i^{\pi^{\Delta t}}(t)$ the expected risk function when following policy $\pi^{\Delta t}$ starting at i with remaining budget t. We deal with each case separately.

Case 1. We use the following useful facts:

- The functions $(u_i(\cdot))_{i \in \mathcal{V}}$ are non-decreasing,
- The functions $(u_i^{\Delta t}(\cdot))_{i \in \mathcal{V}}$ are non-decreasing,
- The functions $(u_i^{\Delta t}(\cdot))_{i \in \mathcal{V}}$ lower bound the functions $(u_i(\cdot))_{i \in \mathcal{V}}$.

The main difficulty in proving convergence lies in the fact that the approximation $u_i^{\Delta t}(t)$ may not necessarily improve as Δt decreases. However, this is the case for regular mesh size sequences such as $(\Delta t_p = \frac{1}{2^p})_{p \in \mathbb{N}}$. Hence, we first demonstrate convergence in that particular case in Lemma A.2 and rely on this last result to prove pointwise convergence in general in Lemma A.3.

Lemma A.2. *For the regular mesh $(\Delta t_p = \frac{1}{2^p})_{p \in \mathbb{N}}$, the sequence $(u_i^{\Delta t_p}(t))_{p \in \mathbb{N}}$ converges to $u_i(t)$ for almost every point t in $[k_i^{min} \cdot \Delta t, T]$.*

Proof. First observe that, for any t, the sequence $(u_i^{\Delta t_p}(t))_{p \in \mathbb{N}}$ is non-decreasing since (i) the discretization mesh used at step $p+1$ is strictly contained in the discretization mesh

used at step p and (ii) the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are non-decreasing. This shows that the functions $(u_i^{\Delta t_p}(\cdot))_{i\in\mathcal{V}}$ converge pointwise to some limits $(f_i(\cdot))_{i\in\mathcal{V}}$. Using the preliminary remarks, we get:

$$f_i(t) \leq u_i(t) \quad \forall t \in [k_i^{\min} \cdot \Delta t, T], \forall i \in \mathcal{V}.$$

Next, we establish that for any $i \in \mathcal{V}, t \in [k_i^{\min} \cdot \Delta t, T]$ and $\epsilon > 0$, $f_i(t) \geq u_i(t-\epsilon)$. This will enable us to squeeze $f_i(t)$ to finally derive $f_i(t) = u_i(t)$. We start with node d. Observe that, by construction of the approximation, $u_d^{\Delta t}(\cdot)$ converges pointwise to $f(\cdot)$ at every point of continuity of $f(\cdot)$. Furthermore, since $f_d(\cdot)$ and $u_d(\cdot)$ are non-decreasing, we have $f_d(t) \geq u_d(t-\epsilon)$ for all $t \in [k_d^{\min} \cdot \Delta t, T]$ and for all $\epsilon > 0$. Consider $\epsilon > 0$ and a large enough p such that $\epsilon > \frac{1}{2^p}$ which implies $\Delta t_p \cdot \lfloor \frac{t}{\Delta t_p} \rfloor \geq t - \epsilon$. We first show by induction on the level of the nodes in $\mathcal{T}$ that:

$$f_i(t) \geq u_i(t - \text{level}(i, \mathcal{T}) \cdot \epsilon) \quad \forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t), \forall i \in \mathcal{V}.$$

The base case follows from the discussion above. Assume that the induction property holds for all nodes of level less than l and consider a node $i \in \mathcal{V}$ of level $l+1$. We have, for $t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t)$:

$$\begin{aligned} u_i^{\Delta t_p}(t) &= u_i^{\Delta t_p}(\left\lfloor \frac{t}{\Delta t_p} \right\rfloor \cdot \Delta t_p) \\ &\geq \max_{j \in \mathcal{T}(i)} \mathbb{E}[u_j^{\Delta t_p}(\left\lfloor \frac{t}{\Delta t_p} \right\rfloor \cdot \Delta t_p - c_{ij})] \\ &\geq \max_{j \in \mathcal{T}(i)} \mathbb{E}[u_j^{\Delta t_p}(t - \epsilon - c_{ij})]. \end{aligned}$$

To take the limit $p \to \infty$ in the previous inequality, note that, for any $j \in \mathcal{T}(i)$:

$$u_j^{\Delta t_p}(t - \epsilon - c_{ij}) \geq u_j^{\Delta t_1}(t - \epsilon - \delta^{\sup}),$$

while $(u_j^{\Delta t_p}(t-\epsilon-c_{ij}))_{p\in\mathbb{N}}$ is non-decreasing and converges almost surely to $f_j(t-\epsilon-c_{ij})$

as $p \to \infty$. Therefore, we can apply the monotone convergence theorem and derive:

$$f_i(t) \geq \mathbb{E}[f_j(t - \epsilon - c_{ij})].$$

As the last inequality holds for any $j \in \mathcal{T}(i)$, we finally obtain:

$$f_i(t) \geq \max_{j \in \mathcal{T}(i)} \mathbb{E}[f_j(t - \epsilon - c_{ij})] \quad \forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t).$$

Using the induction property along with Theorem 2.1, we get:

$$\begin{aligned} f_i(t) &\geq \max_{j \in \mathcal{T}(i)} \mathbb{E}[u_j(t - \text{level}(i, \mathcal{T}) \cdot \epsilon - c_{ij})] \\ &\geq u_i(t - \text{level}(i, \mathcal{T}) \cdot \epsilon), \end{aligned}$$

for all $t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t)$, which concludes the induction. We can now prove by induction on m, along the same lines as above, that:

$$f_i(t) \geq u_i(t - (|\mathcal{V}| + m) \cdot \epsilon) \quad \forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t + m \cdot \delta^{\inf}), \forall i \in \mathcal{V},$$

for all $m \in \mathbb{N}$. This last result can be reformulated as:

$$f_i(t) \geq u_i(t - \epsilon) \quad \forall \epsilon > 0, \forall t \in [k_i^{\min} \cdot \Delta t, T], \forall i \in \mathcal{V}.$$

Combining this lower bound with the upper bound previously derived, we get:

$$u_i(t) \geq f_i(t) \geq u_i(t^-) \quad \forall t \in [k_i^{\min} \cdot \Delta t, T], \forall i \in \mathcal{V},$$

where $u_i(t^-)$ refers to the left one-sided limit of $u_i(\cdot)$ at t. Since, $u_i(\cdot)$ is non-decreasing, it has countably many discontinuity points and the last inequality shows that $f_i(\cdot) = u_i(\cdot)$ almost everywhere on $[k_i^{\min} \cdot \Delta t, T]$.

□

Lemma A.3. *For any sequence $(\Delta t_p)_{p \in \mathbb{N}}$ converging to 0, the sequence $(u_i^{\Delta t_p}(t))_{p \in \mathbb{N}}$ con-*

verges to $u_i(t)$ *for almost every point* t *in* $[k_i^{min} \cdot \Delta t, T]$.

Proof. In contrast to the particular case handled by Lemma A.2, our approximation of $u_i(t)$ may not improve as p increases. For that reason, there is no straightforward comparison between $(u_i^{\Delta t_p}(t))_p$ and $u_i(t)$. However, for a given $i \in \mathcal{V}$, $t \in [k^{\min} \cdot \Delta t, T]$, $\epsilon > 0$ and a large enough p, $(u_i^{\Delta t_p}(t))_p$ can be shown to be lower bounded by a subsequence of $(u_i^{\frac{1}{2^p}}(t-\epsilon))_p$. This is how we proceed to establish convergence.

Consider $i \in \mathcal{V}$, $t \in [k^{\min} \cdot \Delta t, T]$, $\epsilon > 0$ and $p \in \mathbb{N}$. Define $\sigma(p) \in \mathbb{N}$ as the unique integer satisfying $\frac{1}{2^{\sigma(p)-1}} < \Delta t_p \leq \frac{1}{2^{\sigma(p)}}$. Since $\lim_{p\to\infty} \Delta t_p = 0$, we necessarily have $\lim_{p\to\infty} \sigma(p) = \infty$. Remark that $u_i^{\frac{1}{2^{\sigma(p)}}}(\cdot)$ has steps of size $\frac{1}{2^{\sigma(p)}} \geq \Delta t_p$, i.e. $u_i^{\Delta t_p}(\cdot)$ is expected to be a tighter approximation of $u_i(\cdot)$ than $u_i^{\frac{1}{2^{\sigma(p)}}}(\cdot)$ is. However, the time steps do not overlap (multiples of either Δt_p or $\frac{1}{2^p}$) making the two sequences impossible to compare. Nevertheless, the time steps differ by no more than Δt_p. Thus, if p is large enough so that $\Delta t_p < \epsilon$, for each update needed to calculate $u_i^{\frac{1}{2^{\sigma(p)}}}(t-\epsilon)$, there is a corresponding update for a larger budget to compute $u_i^{\Delta t_p}(t)$. As a consequence, the sequence $(u_i^{\frac{1}{2^{\sigma(p)}}}(t-\epsilon))_p$ constitutes a lower bound on the sequence of interest $(u_i^{\Delta t_p}(t))_p$. Using the preliminary remarks, we are able to squeeze $(u_i^{\Delta t_p}(t))_p$:

$$u_i^{\frac{1}{2^{\sigma(p)}}}(t-\epsilon) \leq u_i^{\Delta t_p}(t) \leq u_i(t),$$

for all $i \in \mathcal{V}$, $t \in [k^{\min} \cdot \Delta t, T]$, $\epsilon > 0$ and for p large enough. This can be proved first by induction on the level of the nodes in $\mathcal{T}$ and then by interval increments of size $\delta^{\inf}$ along the same lines as what is done in Lemma A.2. Yet, Lemma A.2 shows that:

$$\lim_{p\to\infty} u_i^{\frac{1}{2^{\sigma(p)}}}(t-\epsilon) = u_i(t-\epsilon),$$

provided $t-\epsilon$ is a point of continuity for $u_i(\cdot)$. As $u_i(\cdot)$ has countably many discontinuity points (it is non-decreasing), the last inequality shows, by taking p large enough and ϵ small enough, that $u_i^{\Delta t_p}(t) \to_{p\to\infty} u_i(t)$ for t a point of continuity of $u_i(\cdot)$.

□

Case 2. The first step consists in proving that the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $(-\infty, T]$. By induction on l, we start by proving that $u_i(\cdot)$ is continuous on $(-\infty, T_f)$ for all nodes i of level l in $\mathcal{T}$. The base case follows from the continuity of $f(\cdot)$. Assuming the property holds for some $l \geq 1$, we consider a node i of level $l+1$ in $\mathcal{T}$, $t < T_f$ and a sequence $t_n \to_{n\to\infty} t$. Using Theorem 2.1, we have:

$$|u_i(t) - u_i(t_n)| \leq \max_{j\in\mathcal{T}(i)} \mathbb{E}[|u_j(t - c_{ij}) - u_j(t_n - c_{ij})|].$$

For any $j \in \mathcal{T}(i)$, we can use the uniform continuity of $u_j(\cdot)$ on $[t - 2\cdot\delta^{\sup}, t]$ to prove that this last term converges to 0 as $n \to \infty$. We conclude that all the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $(-\infty, T_f)$. By induction on m, we can then show that the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $(-\infty, T_f + m\cdot\delta^{\inf})$, to finally conclude that they are continuous on $(-\infty, T]$. We are now able to prove uniform convergence. Since $[T_f - |\mathcal{V}|\cdot\delta^{\sup}, T]$ is a compact set, the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are also uniformly continuous on this set. Take $\epsilon > 0$, there exists $\alpha > 0$ such that:

$$\forall i \in \mathcal{V}, |u_i(\omega) - u_i(\omega')| \leq \epsilon,\ \forall(\omega,\omega') \in [T_f - |\mathcal{V}|\cdot\delta^{\sup}, T]^2 \text{ with } |\omega - \omega'| \leq \alpha.$$

Building on this, we can show, by induction on the level of the nodes in $\mathcal{T}$, that:

$$\sup_{\omega\in[k_i^{\min}\cdot\Delta t, \left\lfloor\frac{T_f}{\Delta t}\right\rfloor\cdot\Delta t)} |u_i(\omega) - u_i^{\Delta t}(\omega)| \leq \text{level}(i,\mathcal{T})\cdot\epsilon, \quad \forall i \in \mathcal{V}.$$

This follows from the sequence of inequalities:

$$\begin{aligned}
&\sup_{\omega \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t)} |u_i^{\Delta t}(\omega) - u_i(\omega)| \\
&\qquad \leq \sup_{k \in \{k_i^{\min}, \cdots, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor - 1\}} |u_i^{\Delta t}(k \cdot \Delta t) - u_i(k \cdot \Delta t)| \\
&\qquad + \sup_{\omega \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t]} |u_i(\omega) - u_i(\left\lfloor \frac{\omega}{\Delta t} \right\rfloor \cdot \Delta t)| \\
&\qquad \leq \sup_{k \in \{k_i^{\min}, \cdots, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor - 1\}} \max_{j \in \mathcal{T}(i)} \mathbb{E}[|u_j^{\Delta t}(k \cdot \Delta t - c_{ij}) - u_j(k \cdot \Delta t - c_{ij})|] + \epsilon \\
&\qquad \leq (\text{level}(i, \mathcal{T}) - 1) \cdot \epsilon + \epsilon = \text{level}(i, \mathcal{T}) \cdot \epsilon.
\end{aligned}$$

We conclude that:

$$\sup_{\omega \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t)} |u_i(\omega) - u_i^{\Delta t}(\omega)| \leq |\mathcal{V}| \cdot \epsilon, \forall i \in \mathcal{V}.$$

Along the same lines, we can show by induction on m that:

$$\sup_{\omega \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t + m \cdot \delta^{\inf})} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq (|\mathcal{V}| + m) \cdot \epsilon, \quad \forall i \in \mathcal{V}.$$

This implies:

$$\sup_{\omega \in [k_i^{\min} \cdot \Delta t, T]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq (|\mathcal{V}| + \left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 1) \cdot \epsilon, \quad \forall i \in \mathcal{V},$$

assuming $\Delta t \leq \delta^{\inf}$. In particular, this shows uniform convergence. To conclude the proof of Case 2, we show that $\pi^{\Delta t}$ is a $o(1)$-approximate optimal solution to (2.1) as $\Delta t \to 0$. Using the last set of inequalities derived in combination with the uniform continuity of the functions $(u_i(\cdot))_{i \in \mathcal{V}}$, we can show that:

$$\forall i \in \mathcal{V}, |u_i^{\Delta t}(\omega) - u_i^{\Delta t}(\omega')| \leq (2 \cdot |\mathcal{V}| + 2 \cdot \left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 3) \cdot \epsilon, \tag{A.2}$$

$\forall(\omega,\omega') \in [k_i^{\min}\cdot\Delta t, T]^2$ with $|\omega-\omega'| \leq \alpha$, and:

$$\forall i \in \mathcal{V}, |u_i^{\Delta t}(\omega) - u_i(\omega')| \leq (|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon, \tag{A.3}$$

$\forall(\omega,\omega') \in [k_i^{\min}\cdot\Delta t, T]^2$ with $|\omega-\omega'| \leq \alpha$. We can now prove, by induction on the level of the nodes in $\mathcal{T}$, that:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 3\cdot\text{level}(i,\mathcal{T})\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon,\ \forall t \in [k_i^{\min}\cdot\Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t), \forall i \in \mathcal{V}.$$

This follows from the sequence of inequalities:

$$\begin{aligned}
u_i^{\pi^{\Delta t}}(t) &= \int_0^\infty p_{i\pi^{\Delta t}(i,t)}(\omega)\cdot u^{\pi^{\Delta t}}_{\pi^{\Delta t}(i,t)}(t-w)\mathrm{d}\omega \\
&\geq \int_0^\infty p_{i\pi^{\Delta t}(i,t)}(\omega)\cdot u_{\pi^{\Delta t}(i,t)}(t-w)\mathrm{d}\omega \\
&- 3\cdot(\text{level}(i,\mathcal{T})-1)\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon \\
&\geq \int_0^\infty p_{i\pi^{\Delta t}(i,t)}(\omega)\cdot u^{\Delta t}_{\pi^{\Delta t}(i,t)}(t-w)\mathrm{d}\omega - \epsilon \\
&- 3\cdot(\text{level}(i,\mathcal{T})-1)\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon \\
&\geq \int_0^\infty p_{i\pi^{\Delta t}(i,t)}(\omega)\cdot u^{\Delta t}_{\pi^{\Delta t}(i,\lfloor \frac{t}{\Delta t} \rfloor\cdot\Delta t)}(\left\lfloor \frac{t}{\Delta t} \right\rfloor\cdot\Delta t - w)\mathrm{d}\omega - \epsilon \\
&- (2\cdot|\mathcal{V}| + 2\cdot\left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 3)\cdot\epsilon \\
&- 3\cdot(\text{level}(i,\mathcal{T})-1)\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon \\
&\geq u_i^{\Delta t}(\left\lfloor \frac{t}{\Delta t} \right\rfloor\cdot\Delta t) - 3\cdot\text{level}(i,\mathcal{T})\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon,
\end{aligned}$$

where we use the induction property for the first inequality, the uniform convergence for the second, (A.2) for the third, the definition of $\pi^{\Delta t}(i,t)$ for the fourth and finally (A.3). We conclude that:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 3\cdot|\mathcal{V}|\cdot(|\mathcal{V}| + \left\lceil \frac{T-T_f}{\delta^{\inf}} \right\rceil + 2)\cdot\epsilon, \quad \forall t \in [k_i^{\min}\cdot\Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t), \forall i \in \mathcal{V}.$$

We can then prove by induction on m, in the same fashion as above, that, for all m:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 3 \cdot (m + |\mathcal{V}|) \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 2) \cdot \epsilon,$$

$\forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor \frac{T_f}{\Delta t} \right\rfloor \cdot \Delta t + m \cdot \delta^{\inf})$ and $\forall i \in \mathcal{V}$. We conclude that

$$u_s^{\pi^{\Delta t}}(T) \geq u_s(T) - 3 \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 2)^2 \cdot \epsilon,$$

which establishes the claim.

Case 3. The first step consists in showing that the functions $(u_i(\cdot))_{i \in \mathcal{V}}$ are Lipschitz on $[T_f - |\mathcal{V}| \cdot \delta^{\sup}, T]$. Take K to be a Lipschitz constant for $f(\cdot)$ on $[T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 2 \cdot |\mathcal{V}| - 1) \cdot \delta^{\sup}, T]$. We first show by induction on l that $u_i(\cdot)$ is K-Lipschitz on $[T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 2 \cdot |\mathcal{V}| - l) \cdot \delta^{\sup}, T_f]$ for all nodes i of level l in $\mathcal{T}$. The base case follows from the definition of K. Assuming the property holds for some $l \geq 1$, we consider a node i of level $l + 1$ in $\mathcal{T}$. Using Theorem 2.1, we have, for $(t, t') \in [T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + 2 \cdot |\mathcal{V}| - l - 1) \cdot \delta^{\sup}, T_f]^2$:

$$\begin{aligned} |u_i(t) - u_i(t')| &\leq \max_{j \in \mathcal{T}(i)} \mathbb{E}[|u_j(t - \omega) - u_j(t' - \omega)|] \\ &\leq K \cdot |t - t'|, \end{aligned}$$

where we use the induction property for l (recall that $p_{ij}(\omega) = 0$ for $\omega \geq \delta^{\sup}$). We conclude that the functions $(u_i(\cdot))_{i \in \mathcal{V}}$ are all $K-$Lipschitz on $[T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil + |\mathcal{V}|) \cdot \delta^{\sup}, T_f]$. We now prove, by induction on m, that the functions $(u_i(\cdot))_{i \in \mathcal{V}}$ are all K-Lipschitz on $[T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil - m + |\mathcal{V}|) \cdot \delta^{\sup}, T_f + m \cdot \delta^{\inf}]$. The base case follows from the previous induction. Assuming the property holds for some m, we have for $i \in \mathcal{V}$ and for $(t, t') \in [T_f - (\left\lceil \frac{T - T_f}{\delta^{\inf}} \right\rceil - m - 1 + |\mathcal{V}|) \cdot \delta^{\sup}, T_f + (m + 1) \cdot \delta^{\inf}]^2$:

$$\begin{aligned} |u_i(t) - u_i(t')| &\leq \max_{j \in \mathcal{V}(i)} \mathbb{E}[|u_j(t - \omega) - u_j(t' - \omega)|] \\ &\leq K \cdot |t - t'|, \end{aligned}$$

where we use the fact that $p_{ij}(\omega) = 0$ for $\omega \leq \delta^{\inf}$ or $\omega \geq \delta^{\sup}$ and the induction property. We conclude that the function $(u_i(\cdot))_{i\in\mathcal{V}}$ are all K-Lipschitz on $[T_f - |\mathcal{V}| \cdot \delta^{\sup}, T]$. Using this last fact, we can prove, by induction on the level of the nodes in $\mathcal{T}$, in a similar fashion as done for Case 2, that:

$$\sup_{\omega\in[k_i^{\min}\cdot\Delta t,\left\lfloor\frac{T_f}{\Delta t}\right\rfloor\cdot\Delta t)} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq \text{level}(i,\mathcal{T}) \cdot K \cdot \Delta t, \quad \forall i \in \mathcal{V}.$$

By induction on m, we can then show that:

$$\sup_{\omega\in[k_i^{\min}\cdot\Delta t,\left\lfloor\frac{T_f}{\Delta t}\right\rfloor\cdot\Delta t+m\cdot\delta^{\inf})} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq (|\mathcal{V}| + m) \cdot K \cdot \Delta t, \quad \forall i \in \mathcal{V}, \forall m \in \mathbb{N}.$$

This implies:

$$\sup_{\omega\in[k_i^{\min}\cdot\Delta t,T]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq (|\mathcal{V}| + \left\lceil\frac{T - T_f}{\delta^{\inf}}\right\rceil + 1) \cdot K \cdot \Delta t, \quad \forall i \in \mathcal{V},$$

assuming $\Delta t \leq \delta^{\inf}$. This shows uniform convergence at speed Δt. To conclude the proof of Case 3, we show that $\pi^{\Delta t}$ is a $O(\Delta t)$-approximate optimal solution to (2.1) as $\Delta t \to 0$. We can show, using the last inequality derived along with the same sequence of inequalities as in Case 2, by induction on the level of the nodes in $\mathcal{T}$ that:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 6 \cdot \text{level}(i,\mathcal{T}) \cdot (|\mathcal{V}| + \left\lceil\frac{T - T_f}{\delta^{\inf}}\right\rceil + 1) \cdot K \cdot \Delta t,$$

$\forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor\frac{T_f}{\Delta t}\right\rfloor \cdot \Delta t)$ and $\forall i \in \mathcal{V}$. We can then prove by induction on m, in the same fashion as in Case 2, that, for all $m \in \mathbb{N}$:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 6 \cdot (m + |\mathcal{V}|) \cdot (|\mathcal{V}| + \left\lceil\frac{T - T_f}{\delta^{\inf}}\right\rceil + 1) \cdot K \cdot \Delta t,$$

$\forall t \in [k_i^{\min} \cdot \Delta t, \left\lfloor\frac{T_f}{\Delta t}\right\rfloor \cdot \Delta t + m \cdot \delta^{\inf})$ and $\forall i \in \mathcal{V}$. We conclude that:

$$u_s^{\pi^{\Delta t}}(T) \geq u_s(T) - 6 \cdot (|\mathcal{V}| + \left\lceil\frac{T - T_f}{\delta^{\inf}}\right\rceil + 1)^2 \cdot K \cdot \Delta t,$$

which establishes the claim.

Case 4. In this situation, we can show by induction on m that the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $[0, m \cdot \delta^{\inf}]$ and conclude that the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $[0, T]$. Since $[0, T]$ is a compact set, these functions are also uniformly continuous on $[0, T]$. Moreover, $u_i^{\Delta t}(t) = u_i(t) = 0$ for all $t \leq 0$ and $i \in \mathcal{V}$. Using these two observations, we can apply the same techniques as in Case 2 to obtain the same results.

□

A.2.4 Proof of Theorem 2.2

Proof. Proof of Theorem 2.2. We denote by $\mathcal{H}$ the set of all possible histories of the previously experienced costs and previously visited nodes. As in Theorem 2.1, the last part of the theorem is trivial because any strategy is optimal when having already spent a budget of $T - T_f^r$.

The proof for the first part is an extension of Theorem 2.1 and follows the same steps. We denote by $(m_{ij})_{(i,j)\in\mathcal{A}}$ the worst-case expected costs, i.e.:

$$m_{ij} = \sup_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[X] \quad \forall (i,j) \in \mathcal{A}.$$

Observe that these quantities are well-defined as $\mathcal{P}_{ij}$ is not empty and $\mathbb{E}_{X\sim p_{ij}}[X] \leq \delta^{\sup}$ for any $p_{ij} \in \mathcal{P}_{ij}$. Furthermore, there exists $p_{ij}^* \in \mathcal{P}_{ij}$ such that $m_{ij} = \mathbb{E}_{X\sim p_{ij}^*}[X]$ as $\mathcal{P}_{ij}$ is compact for the weak topology. For any node $i \neq d$, we define M_i as the length of a shortest path from i to d in $\mathcal{G}$ when the arc costs are taken as $(m_{ij})_{(i,j)\in\mathcal{A}}$. Just like in Theorem 2.1, we consider an optimal strategy $\pi_{f,\mathcal{P}}^*$ solution to (2.3). For a given history $h \in \mathcal{H}$, we define t_h as the remaining budget, i.e. T minus the total cost spent so far, and i_h as the current location. The policy $\pi_{f,\mathcal{P}}^*$ maps $h \in \mathcal{H}$ to a probability distribution over $\mathcal{V}(i_h)$. Observe that randomizing does not help because (i) the costs are independent across time and arcs and (ii) the ambiguity set is rectangular. Hence, without loss of generality, we may assume that $\pi_{f,\mathcal{P}}^*$ actually maps h to the node in $\mathcal{V}(i_h)$ minimizing the worst-case objective function given h. For $h \in \mathcal{H}$, we denote by $X_{\pi_{f,\mathcal{P}}^*}^h$ the random cost-to-go incurred

by following strategy $\pi^*_{f,\mathcal{P}}$, i.e. not including the total cost spent up to this point of the history $T - t_h$. We define π_s as a policy associated with an arbitrary shortest path from i to d with respect to $(m_{ij})_{(i,j)\in\mathcal{A}}$. Specifically, π_s maps the current location i_h to a node in $\mathcal{T}^r(i_h)$, irrespective of the history of the process. Similarly as for $\pi^*_{f,\mathcal{P}}$, we denote by $X^h_{\pi_s}$ the random cost-to-go incurred by following strategy π_s for $h \in \mathcal{H}$. Using ***Bellman's Principle of Optimality*** for $\pi^*_{f,\mathcal{P}}$, we have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}^*}\left[f(t_h - X^h_{\pi^*_{f,\mathcal{P}}})\right] &\geq \inf_{\forall \tau \geq T-t_h, \forall (i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(t - X^h_{\pi^*_{f,\mathcal{P}}})] \\
&\geq \sup_{\pi\in\Pi} \inf_{\forall \tau \geq T-t_h, \forall (i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(t - X^h_{\pi})] \\
&\geq \inf_{\forall \tau \geq T-t_h, \forall (i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(t - X^h_{\pi_s})] \\
&\geq \mathbb{E}_{\mathbf{q}^\tau}[f(t - X^h_{\pi_s})],
\end{aligned}$$

where $(q^\tau_{ij})_{(i,j)\in\mathcal{A},\tau\geq T-t_h}$ is given by the worst-case scenario in the ambiguity sets, i.e.:

$$(q^\tau_{ij})_{(i,j)\in\mathcal{A},\tau\geq T-t_h} \in \operatorname*{argmin}_{\forall \tau \geq T-t_h, \forall (i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(t - X^h_{\pi_s})],$$

which can be shown to exist because the ambiguity sets are compact. Using the last inequality derived, we can prove, using the exact same sequence of inequalities as in Theorem 2.1, that there exists T^r_f such that, for both cases (a) and (b):

$$\mathbb{E}_{\mathbf{p}^*}[X^h_{\pi^*_{f,\mathcal{P}}}] - \mathbb{E}_{\mathbf{q}^\tau}[X^h_{\pi_s}] < \min_{i\neq d} \min_{j\in\mathcal{V}(i), j\notin\mathcal{T}^r(i)} \{m_{ij} + M_j - M_i\} \quad \forall h \in \mathcal{H} \text{ such that } t_h \leq T^r_f, \tag{A.4}$$

with the convention that the minimum of an empty set is equal to infinity. Starting from (A.4), consider $h \in \mathcal{H}$ such that $t_h \leq T^r_f$ and suppose by contradiction that $\pi^*_{f,\mathcal{P}}(h) = j_h \notin \mathcal{T}^r(i_h)$. As mentioned in Theorem 2.1, even though $\pi^*_{f,\mathcal{P}}$ can be fairly complicated, the first action is deterministic and incurs an expected cost of $m_{i_h j_h}$ because the costs are independent across time and arcs. Moreover, when the objective is to minimize the average cost, the optimal strategy among all history-dependent rules is to follow the shortest path with respect to the mean arc costs (once again because the costs are independent across

time and arcs). As a result:

$$\mathbb{E}_{\mathbf{p}^*}[X^h_{\pi^*_{f,\mathcal{P}}}] \geq m_{i_h j_h} + M_{j_h}.$$

Additionally, by definition of $(p^*_{ij})_{(i,j)\in\mathcal{A}}$:

$$\begin{aligned}\mathbb{E}_{\mathbf{q}^\tau}[X^h_{\pi_s}] &\leq \mathbb{E}_{\mathbf{p}^*}[X^h_{\pi_s}] \\ &\leq M_{i_h}.\end{aligned}$$

This implies:

$$\mathbb{E}_{\mathbf{p}^*}[X^h_{\pi^*_{f,\mathcal{P}}}] - \mathbb{E}_{\mathbf{q}^\tau}[X^h_{\pi_s}] \geq m_{i_h j_h} + M_{j_h} - M_{i_h},$$

a contradiction. We conclude that:

$$\pi^*_{f,\mathcal{P}}(h) \in \mathcal{T}^r(i_h) \quad \forall h \in \mathcal{H} \text{ such that } t_h \leq T^r_f.$$

□

A.2.5 Proof of Proposition 2.3.

Proof. Proof of Proposition 2.3. The proof uses a reduction to distributionally robust finite-horizon MDPs in a similar fashion as in Proposition 2.1. Using Theorem 2.2, the optimization problem (2.3) can be equivalently formulated as a discrete-time finite-horizon distributionally robust MDP in the extended space state $(i,t) \in \mathcal{V} \times [T - \delta^{\sup} \cdot \left\lceil \frac{T-T^r_f}{\delta^{\inf}} \right\rceil, T]$ where i is the current location and t is the, possibly negative, remaining budget. Specifically:

- The time horizon is $\left\lceil \frac{T-T^r_f}{\delta^{\inf}} \right\rceil$.
- The initial state is (s,T).
- The set of available actions at state (i,t), for $i \neq d$, is taken as $\mathcal{V}(i)$. Picking $j \in \mathcal{V}(i)$ corresponds to crossing link (i,j) and results in a transition to state $(j, t-\omega)$ with probability $p_{ij}(\omega)\mathrm{d}\omega$.

- The probability of transitions are only known to lie in the rectangular ambiguity set:

$$\prod_{\substack{(i,j)\in\mathcal{A} \\ t\in[T-\delta^{\sup}\cdot\left\lceil\frac{T-T_f^r}{\delta^{\inf}}\right\rceil,T]}} \mathcal{P}_{ij}.$$

- The only available action at a state (d,t) is to remain in this state.

- The transition rewards are all equal to 0.

- The final reward at the epoch $\left\lceil\frac{T-T_f^r}{\delta^{\inf}}\right\rceil$ for any state (i,t) is equal to $f_i(t)$, which is the optimal worst-case expected objective-to-go when following the shortest path tree $\mathcal{T}^r$ starting at node i with remaining budget t. Specifically, the collection of functions $(f_i(\cdot))_{i\in\mathcal{V}}$ is a solution to the following program:

$$\begin{aligned} f_d(t) &= f(t), && t \leq T_f^r \\ f_i(t) &= \max_{j\in\mathcal{T}^r(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega)\cdot f_j(t-\omega)\mathrm{d}\omega && i\neq d, t\leq T_f^r. \end{aligned}$$

As a consequence, we can conclude the proof with Theorem 2.2 of [51] (or equivalently Theorem 1 of [69]).

□

A.2.6 Proof of Proposition 2.4

The proofs are along the same lines as for Proposition 2.2.

Proof. Proof of Proposition 2.4. We deal with each case separately.

Case 1. We make use the following facts:

- The functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ are non-decreasing,

- The functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are non-decreasing,

- The functions $(u_i^{\Delta t}(\cdot))_{i\in\mathcal{V}}$ lower bound the functions $(u_i(\cdot))_{i\in\mathcal{V}}$.

We follow the same recipe as in Proposition 2.2. We start by proving convergence for the discretizattion sequence $(\Delta t_p = \frac{1}{2^p})_{p\in\mathbb{N}}$. Then, we conclude the general study with the exact same argument as in Lemma A.3.

Lemma A.4. *For the regular mesh $(\Delta t_p = \frac{1}{2^p})_{p\in\mathbb{N}}$, the sequence $(u_i^{\Delta t_p}(t))_{p\in\mathbb{N}}$ converges to $u_i(t)$ for almost every point t in $[k_i^{r,min}\cdot\Delta t, T]$.*

Proof. Just like in Lemma A.2 we can prove that the sequence $(u_i^{\Delta t_p}(t))_{p\in\mathbb{N}}$ is non-decreasing for any t and $i\in\mathcal{V}$. Hence, the functions $(u_i^{\Delta t_p}(\cdot))_{i\in\mathcal{V}}$ converge pointwise to some limits $(f_i(\cdot))_{i\in\mathcal{V}}$. Using the preliminary remarks, we get:

$$f_i(t)\leq u_i(t)\quad \forall t\in[k_i^{r,\min}\cdot\Delta t, T], \forall i\in\mathcal{V}.$$

Next, we establish that for any $t\in[k_i^{r,\min}\cdot\Delta t, T]$ and for any $\epsilon>0$, $f_i(t)\geq u_i(t-\epsilon)$. This will enable us to squeeze $f_i(t)$ to finally derive $f_i(t)=u_i(t)$. We start with node d. Observe that, by construction of the approximation, $u_d^{\Delta t}(\cdot)$ converges pointwise to $f(\cdot)$ at every point of continuity of $f(\cdot)$. Furthermore, since $f_d(\cdot)$ and $u_d(\cdot)$ are non-decreasing, we have $f_d(t)\geq u_d(t-\epsilon)$ for all $t\in[k_d^{r,\min}\cdot\Delta t, T]$ and for all $\epsilon>0$. Consider $\epsilon>0$ and a large enough p such that $\epsilon>\frac{1}{2^p}$ which implies $\Delta t_p\cdot\lfloor\frac{t}{\Delta t_p}\rfloor\geq t-\epsilon$. We first show by induction on the level of the nodes in $\mathcal{T}^r$ that:

$$f_i(t)\geq u_i(t-\text{level}(i,\mathcal{T}^r)\cdot\epsilon)\quad \forall t\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor\frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t], \forall i\in\mathcal{V}.$$

The base case follows from the discussion above. Assume that the induction property holds for all nodes of level less than l and consider a node $i\in\mathcal{V}$ of level $l+1$. We have, for $t\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor\frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t]$:

$$\begin{aligned}
u_i^{\Delta t_p}(t) &\geq u_i^{\Delta t_p}\Big(\left\lfloor\frac{t}{\Delta t_p}\right\rfloor\cdot\Delta t_p\Big)\\
&\geq \max_{j\in\mathcal{T}^r(i)}\inf_{p_{ij}\in\mathcal{P}_{ij}}\mathbb{E}_{X\sim p_{ij}}\Big[u_j^{\Delta t_p}\Big(\left\lfloor\frac{t}{\Delta t_p}\right\rfloor\cdot\Delta t_p - X\Big)\Big]\\
&\geq \max_{j\in\mathcal{T}^r(i)}\inf_{p_{ij}\in\mathcal{P}_{ij}}\mathbb{E}_{X\sim p_{ij}}\big[u_j^{\Delta t_p}(t-\epsilon-X)\big].
\end{aligned}$$

Take $j \in \mathcal{T}^r(i)$. Since $u_j^{\Delta t_p}(\cdot)$ is continuous and $\mathcal{P}_{ij}$ is compact, the infimum in the previous inequality is attained for some $p_{ij}^p \in \mathcal{P}_{ij}$ which gives:

$$u_i^{\Delta t_p}(t) \geq \mathbb{E}_{X \sim p_{ij}^p}[u_j^{\Delta t_p}(t - \epsilon - X)].$$

As the sequence $(u_j^{\Delta t_p}(t - \epsilon - \omega))_p$ is non-decreasing for any ω, we have, for any $m \leq p$:

$$u_i^{\Delta t_p}(t) \geq \mathbb{E}_{X \sim p_{ij}^p}[u_j^{\Delta t_m}(t - \epsilon - X)].$$

Because $\mathcal{P}_{ij}$ is a compact set for the weak topology, there exists a subsequence of $(p_{ij}^p)_p$ converging weakly in $\mathcal{P}_{ij}$ to some probability measure p_{ij}. Without loss of generality, we continue to refer to this subsequence as $(p_{ij}^p)_p$. We can now take the limit $p \to \infty$ in the previous inequality which yields:

$$f_i(t) \geq \mathbb{E}_{X \sim p_{ij}}[u_j^{\Delta t_m}(t - \epsilon - X)],$$

since $u_j^{\Delta t_m}(\cdot)$ is continuous. To take the limit $m \to \infty$, note that:

$$u_j^{\Delta t_m}(t - \epsilon - X) \geq u_j^{\Delta t_1}(t - \epsilon - \delta^{\sup}),$$

while $(u_j^{\Delta t_m}(t - \epsilon - X))_{m \in \mathbb{N}}$ is non-decreasing and converges almost surely to $f_j(t - \epsilon - X)$ as $m \to \infty$. Therefore, we can apply the monotone convergence theorem and derive:

$$f_i(t) \geq \mathbb{E}_{X \sim p_{ij}}[f_j(t - \epsilon - X)],$$

which further implies

$$f_i(t) \geq \inf_{p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{X \sim p_{ij}}[f_j(t - \epsilon - X)].$$

As the last inequality holds for any $j \in \mathcal{T}^r(i)$, we finally obtain:

$$f_i(t) \geq \max_{j \in \mathcal{T}^r(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{X \sim p_{ij}}[f_j(t - \epsilon - X)] \quad \forall t \in [k_i^{r,\min} \cdot \Delta t, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor \cdot \Delta t], \forall i \in \mathcal{V}.$$

Using the induction property along with Theorem 2.2, we get:

$$\begin{aligned} f_i(t) &\geq \max_{j \in \mathcal{T}^r(i)} \inf_{p_{ij} \in \mathcal{P}_{ij}} \mathbb{E}_{X \sim p_{ij}}[u_j(t - \text{level}(i, \mathcal{T}^r) \cdot \epsilon - X)] \\ &\geq u_i(t - \text{level}(i, \mathcal{T}^r) \cdot \epsilon), \end{aligned}$$

for all $t \in [k_i^{r,\min} \cdot \Delta t, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor \cdot \Delta t]$, which concludes the induction. We can now prove by induction on m, along the same lines as above, that:

$$f_i(t) \geq u_i(t - (|\mathcal{V}| + m) \cdot \epsilon) \quad \forall t \in [k_i^{r,\min} \cdot \Delta t, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor \cdot \Delta t + m \cdot \delta^{\inf}], \forall i \in \mathcal{V},$$

for all $m \in \mathbb{N}$. This last result can be reformulated as:

$$f_i(t) \geq u_i(t - \epsilon) \quad \forall \epsilon > 0, \forall t \in [k_i^{r,\min} \cdot \Delta t, T], \forall i \in \mathcal{V}.$$

Combining this lower bound with the upper bound previously derived, we get:

$$u_i(t) \geq f_i(t) \geq u_i(t^-) \quad \forall t \in [k_i^{r,\min} \cdot \Delta t, T], \forall i \in \mathcal{V},$$

where $u_i(t^-)$ refers to the left one-sided limit of $u_i(\cdot)$ at t. Since, $u_i(\cdot)$ is non-decreasing, it has countably many discontinuity points and the last inequality shows that $f_i(\cdot) = u_i(\cdot)$ almost everywhere on $[k_i^{r,\min} \cdot \Delta t, T]$.

□

Case 2. The first step consists in proving that the functions $(u_i(\cdot))_{i \in \mathcal{V}}$ are continuous on $(-\infty, T]$. By induction on l, we start by proving that $u_i(\cdot)$ is continuous on $(-\infty, T_f^r]$ for all nodes i of level l in $\mathcal{T}^r$. The base case follows from the continuity of $f(\cdot)$. Assuming the property holds for some $l \geq 1$, we consider a node i of level $l+1$ in $\mathcal{T}^r$, $t \leq T_f^r$ and a sequence $t_n \to_{n \to \infty} t$. Using Theorem 2.2, we have:

$$|u_i(t) - u_i(t_n)| \leq \max_{j \in \mathcal{T}^r(i)} \sup_{p_{ij} \in \mathcal{P}_{ij}} \int_0^\infty p_{ij}(\omega) \cdot |u_j(t - \omega) - u_j(t_n - \omega)| \mathrm{d}\omega.$$

For any $j \in \mathcal{T}^r(i)$, we can use the uniform continuity of $u_j(\cdot)$ on $[t - 2 \cdot \delta^{\sup}, t]$ to prove that this last term converges to 0 as $n \to \infty$. We conclude that all the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $(-\infty, T_f^r]$. By induction on m, we can then show that the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are continuous on $(-\infty, T_f^r + m\cdot\delta^{\inf}]$, to finally conclude that they are continuous on $(-\infty, T]$. We are now able to prove uniform convergence. Since $[T_f^r - |\mathcal{V}| \cdot \delta^{\sup}, T]$ is a compact set, the functions $(u_i(\cdot))_{i\in\mathcal{V}}$ are also uniformly continuous on this set. Take $\epsilon > 0$, there exists $\alpha > 0$ such that:

$$\forall i \in \mathcal{V}, |u_i(\omega) - u_i(\omega')| \leq \epsilon,\ \forall(\omega, \omega') \in [T_f^r - |\mathcal{V}| \cdot \delta^{\sup}, T]^2 \text{ with } |\omega - \omega'| \leq \alpha.$$

Building on this, we can show, by induction on the level of the nodes in $\mathcal{T}^r$, that:

$$\sup_{\omega\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t]} |u_i(\omega) - u_i^{\Delta t}(\omega)| \leq 2 \cdot \text{level}(i, \mathcal{T}^r) \cdot \epsilon, \quad \forall i \in \mathcal{V}.$$

This follows from the sequence of inequalities:

$$\begin{aligned}
\sup_{\omega\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq & \sup_{k\in\{k_i^{r,\min},\cdots,\left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\}} |u_i^{\Delta t}(k\cdot\Delta t) - u_i(k\cdot\Delta t)| \\
& + \sup_{\omega\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t]} |u_i(\omega) - u_i(\left\lfloor \frac{\omega}{\Delta t}\right\rfloor\cdot\Delta t)| \\
& + \sup_{\omega\in[k_i^{r,\min}\cdot\Delta t, \left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\cdot\Delta t]} |u_i(\omega) - u_i(\left\lceil \frac{\omega}{\Delta t}\right\rceil\cdot\Delta t)| \\
& \leq \sup_{k\in\{k_i^{r,\min},\cdots,\left\lfloor \frac{T_f^r}{\Delta t}\right\rfloor\}} \max_{j\in\mathcal{T}(i)} \sup_{p_{ij}\in\mathcal{P}_{ij}} \\
& \{\int_0^\infty p_{ij}(\omega)\cdot|u_j^{\Delta t}(k\cdot\Delta t - \omega) - u_j(k\cdot\Delta t - \omega)|\mathrm{d}\omega\} \\
& + 2\cdot\epsilon \\
& \leq 2\cdot(\text{level}(i,\mathcal{T}^r) - 1)\cdot\epsilon + 2\cdot\epsilon \\
& = 2\cdot\text{level}(i,\mathcal{T}^r)\cdot\epsilon.
\end{aligned}$$

We conclude that:

$$\sup_{\omega \in [k_i^{r,\min} \cdot \Delta t, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor \cdot \Delta t]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq 2 \cdot |\mathcal{V}| \cdot \epsilon, \quad \forall i \in \mathcal{V}.$$

Along the same lines, we can show by induction on m that:

$$\sup_{\omega \in [k_i^{r,\min} \cdot \Delta t, \left\lfloor \frac{T_f^r}{\Delta t} \right\rfloor \cdot \Delta t + m \cdot \delta^{\inf}]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq 2 \cdot (|\mathcal{V}| + m) \cdot \epsilon, \quad \forall i \in \mathcal{V}.$$

This implies:

$$\sup_{\omega \in [k_i^{r,\min} \cdot \Delta t, T]} |u_i^{\Delta t}(\omega) - u_i(\omega)| \leq 2 \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f^r}{\delta^{\inf}} \right\rceil) \cdot \epsilon, \quad \forall i \in \mathcal{V},$$

assuming $\Delta t \leq \delta^{\inf}$. In particular, this shows uniform convergence. To conclude the proof of Case 2, we show that $\pi^{\Delta t}$ is a $o(1)$-approximate optimal solution to (2.3) as $\Delta t \to 0$. We denote by $u_i^{\pi^{\Delta t}}(t)$ the worst-case expected risk function when following policy $\pi^{\Delta t}$ starting at i with remaining budget t. We can show, by induction on the level of the nodes in $\mathcal{T}^r$, that :

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 12 \cdot \text{level}(i, \mathcal{T}^r) \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f^r}{\delta^{\inf}} \right\rceil) \cdot \epsilon, \quad \forall t \in [k_i^{r,\min} \cdot \Delta t, T_f^r], \forall i \in \mathcal{V}.$$

To do so, we can use the same sequence of inequalities as in Case 2 of Proposition 2.2, except that we also take the infimum over $p_{i\pi^{\Delta t}(i,t)} \in \mathcal{P}_{i\pi^{\Delta t}(i,t)}$. We derive:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 12 \cdot |\mathcal{V}| \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f^r}{\delta^{\inf}} \right\rceil) \cdot \epsilon, \ \forall t \in [k_i^{r,\min} \cdot \Delta t, T_f^r], \forall i \in \mathcal{V}.$$

Along the same lines, we can show by induction on m that:

$$u_i^{\pi^{\Delta t}}(t) \geq u_i(t) - 12 \cdot (|\mathcal{V}| + m) \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f^r}{\delta^{\inf}} \right\rceil) \cdot \epsilon, \ \forall t \in [k_i^{r,\min} \cdot \Delta t, T_f^r + m \cdot \delta^{\inf}], \forall i \in \mathcal{V}.$$

We conclude that $u_s^{\pi^{\Delta t}}(T) \geq u_s(T) - 12 \cdot (|\mathcal{V}| + \left\lceil \frac{T - T_f^r}{\delta^{\inf}} \right\rceil)^2 \cdot \epsilon$, which establishes the claim.

Case 3. This case is essentially identical to Case 2 substituting uniform continuity for Lipschitz continuity and the proof mirrors the proof of Case 3 of Proposition 2.2.

□

A.2.7 Proof of Lemma 2.1

Proof. Proof of Lemma 2.1. For a real value x, δ_x refers to the Dirac distribution at x. We denote by $(m_{ij})_{(i,j)\in\mathcal{A}}$ the worst-case expected costs, i.e.:

$$m_{ij} = \sup_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[X] \quad \forall (i,j) \in \mathcal{A}.$$

We first make the crucial observation (which we use repeatedly in what follows) that, for any arc $(i,j) \in \mathcal{A}$, there exists a distribution $p^*_{ij} \in \mathcal{P}_{ij}$ such that:

$$\inf_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[g(X)] = \mathbb{E}_{X\sim p^*_{ij}}[g(X)],$$

for any convex function $g(\cdot)$. Indeed this follows from Jensen's inequality in case (a) since $\delta_{m_{ij}} \in \mathcal{P}_{ij}$ by assumption and this is proved in [21] for case (b). We now move on to prove the result. Observe that $f(\cdot)$ is increasing since $f(\cdot)$ is convex and $f' \to_{-\infty} a > 0$. We use Proposition 2.3 and consider a solution $(\pi^*_{f,\mathcal{P}}, (u_i(\cdot))_{i\in\mathcal{V}})$ to the dynamic programming equation (2.7). We first prove by induction on the level of the nodes in $\mathcal{T}^r$ that:

$$u_i(t) = \max_{j\in\mathcal{T}^r(i)} \mathbb{E}_{X\sim p^*_{ij}}[u_j(t-X)], \quad t \le T^r_f,$$

and that $u_i(\cdot)$ is convex on $(-\infty, T^r_f]$, for all nodes $i \in \mathcal{V}$. The base case is trivial. Assume that the property holds for all nodes of level less than l and consider a node $i \in \mathcal{V}$ of level $l+1$. Since $u_j(\cdot)$ is convex on $(-\infty, T^r_f]$ for $j \in \mathcal{T}^r(i)$ by assumption, we have:

$$\begin{aligned} u_i(t) &= \max_{j\in\mathcal{T}^r(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[u_j(t-X)] \\ &= \max_{j\in\mathcal{T}^r(i)} \mathbb{E}_{X\sim p^*_{ij}}[u_j(t-X)], \end{aligned}$$

for $t \leq T_f^r$ using Theorem 2.2 and the preliminary remark, which concludes the induction given that $\mathbb{E}_{X\sim p}[g(\cdot - X)]$ is convex for any convex function $g(\cdot)$ and for any distribution p. We move on to prove by induction on m that:

$$u_i(t) = \max_{j\in\mathcal{V}(i)} \mathbb{E}_{X\sim p_{ij}^*}[u_j(t-X)], \quad t \leq T_f^r + m\cdot\delta^{\inf}$$

and that $u_i(\cdot)$ is convex on $(-\infty, T_f^r + m\cdot\delta^{\inf}]$ for all nodes $i \in \mathcal{V}$. The base case follows from the previous induction. Assume that the inductive property holds for some $m \in \mathbb{N}$. Consider $i \neq d$. We have, for $t \in (-\infty, T_f^r + (m+1)\cdot\delta^{\inf}]$:

$$\begin{aligned} u_i(t) &= \max_{j\in\mathcal{V}(i)} \inf_{p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p_{ij}}[u_j(t-X)] \\ &= \max_{j\in\mathcal{V}(i)} \mathbb{E}_{X\sim p_{ij}^*}[u_j(t-X)], \end{aligned}$$

using Theorem 2.2, the preliminary remark, and the inductive property. This once again concludes the induction. Hence:

$$u_i(t) = \max_{j\in\mathcal{V}(i)} \mathbb{E}_{X\sim p_{ij}^*}[u_j(t-X)] \quad \forall t \leq T,\ \forall i \in \mathcal{V}.$$

Using Theorem 2.2 and plugging this last expression back into (2.7), this shows that:

$$\begin{aligned} \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p_{ij}^\tau\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T-X_\pi)] &\geq \sup_{\pi\in\Pi} \mathbb{E}_{\mathbf{p}^*}[f(T-X_\pi)] \\ &\geq \sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T-X_\pi)], \end{aligned}$$

where the notation $\mathbf{p}^*$ refers to the fact the costs $(c_{ij})_{(i,j)\in\mathcal{A}}$ are independent and distributed according to $(p_{ij}^*)_{(i,j)\in\mathcal{A}}$. Since (2.3) is a conservative approximation of (2.2), we get:

$$\sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T-X_\pi)] = \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p_{ij}^\tau\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T-X_\pi)],$$

and an optimal strategy for both problems is given by the optimal strategy to the nominal problem (2.1) when the costs $(c_{ij})_{(i,j)\in\mathcal{A}}$ are independent and distributed according to $(p_{ij}^*)_{(i,j)\in\mathcal{A}}$.

□

A.2.8 Proof of Lemma 2.2

Proof. Proof of Lemma 2.2. Without loss of generality, we assume that $f^{(K+1)} > 0$. The proof is almost identical in the converse situation. We use Proposition 2.3 and consider a solution $(\pi^*_{f,\mathcal{P}}, (u_i(\cdot))_{i\in\mathcal{V}})$ to the dynamic programming equation (2.7). We first prove by induction on the level of the nodes i in $\mathcal{G}$ that $u_i^{(K+1)} > 0$ and that, for j the immediate successor of i in $\mathcal{G}$, there exists $p_{ij} \in \mathcal{P}_{ij}$ such that:

$$u_i(t) = \mathbb{E}_{X\sim p_{ij}}[u_j(t-X)] \quad \forall t \leq T \text{ if } i \neq d. \tag{A.5}$$

Assume that the property holds for all nodes of level less than l and consider a node $i \in \mathcal{V}$ of level $l+1$. Let j be the immediate successor of i in $\mathcal{G}$. As $\mathcal{P}_{ij}$ is not empty, Lemma 3.1 from [87] shows that $\mathcal{P}_{ij}$ contains a discrete distribution whose support is a subset of $\{\delta_0, \cdots, \delta_{K+2}\}$ with $\delta_0 = \delta_{ij}^{\text{inf}} < \delta_1 < \cdots < \delta_{K+2} = \delta_{ij}^{\text{sup}}$. For any $n \in \mathbb{N}$, we define the ambiguity set:

$$\mathcal{P}_{ij}^n = \{p \in \mathcal{P}_{ij} | \\ \text{supp}(p) \subset \{\delta_0, \delta_0 + \frac{\delta_1 - \delta_0}{n}, \delta_0 + 2\cdot\frac{\delta_1-\delta_0}{n}, ..., \delta_1, \delta_1 + \frac{\delta_2-\delta_1}{n}, .., \delta_{K+1}\}\},$$

which can be interpreted as a discretization of $\mathcal{P}_{ij}$. Observe that, by design, $\mathcal{P}_{ij}^n$ is not empty. Additionally, we define the sequence of functions $(f_i^n(\cdot))_{n\in\mathbb{N}}$ by:

$$f_i^n(t) = \inf_{p\in\mathcal{P}_{ij}^n} \mathbb{E}_{X\sim p}[u_j(t-X)] \quad \forall t \leq T. \tag{A.6}$$

Since $u_j^{(K+1)} > 0$, the authors of [77] show that there exists $p_{ij}^n \in \mathcal{P}_{ij}^n$ such that:

$$f_i^n(t) = \mathbb{E}_{X\sim p_{ij}^n}[u_j(t-X)] \quad \forall t \leq T.$$

Because $\mathcal{P}_{ij}$ is compact with respect to the weak topology and since $\mathcal{P}_{ij}^n \subset \mathcal{P}_{ij}$, we can take a subsequence of $(p_{ij}^n)_{n\in\mathbb{N}}$ such that $p_{ij}^n \to p_{ij} \in \mathcal{P}_{ij}$ as $n \to \infty$ for the weak topology. Without loss of generality, we continue to denote this sequence $(p_{ij}^n)_{n\in\mathbb{N}}$. Since $u_j(\cdot)$ is continuous, we derive that the sequence of functions $(f_i^n(\cdot))_{n\in\mathbb{N}}$ converges simply to a function $f_i(\cdot)$ which satisfies:

$$f_i(t) = \mathbb{E}_{X\sim p_{ij}}[u_j(t-X)] \quad \forall t \leq T. \tag{A.7}$$

We now move on to show that $f_i(t) = u_i(t)$ for all $t \leq T$. This will conclude the induction because we can take the $(K+1)$th derivative in (A.7) since p_{ij} has compact support. Take $t \leq T$ and $\epsilon > 0$. The function $u_j(\cdot)$ is continuous on $[t-\delta^{\sup}, t-\delta^{\inf}]$ hence, by uniform continuity, there exists $\alpha > 0$ such that:

$$|u_j(t-\omega) - u_j(t-\omega')| \leq \epsilon$$

as soon as $|\omega - \omega'| \leq \alpha$ and $(\omega, \omega') \in [\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]^2$. Consider $n > \frac{\delta_{ij}^{\sup} - \delta_{ij}^{\inf}}{\alpha}$. Using conic duality, Corollary 3.1 of [87] shows that $u_i(t)$ is the optimal value of the infinite linear program:

$$\begin{aligned} \sup_{(a_1,\cdots,a_K,b)\in\mathbb{R}^{K+1}} \quad & \sum_{k=1}^K a_k \cdot m_{ij}^k + b \\ \text{subject to} \quad & \sum_{k=1}^K a_k \cdot \omega^k + b \leq u_j(t-\omega) \quad \forall \omega \in [\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]. \end{aligned} \tag{A.8}$$

Using strong linear programming duality, we also have that $f_i^n(t)$ is the optimal value of the finite linear program:

$$\begin{aligned} \sup_{(a_1,\cdots,a_K,b)\in\mathbb{R}^{K+1}} \quad & \sum_{k=1}^K a_k \cdot m_{ij}^k + b \\ \text{subject to} \quad & \sum_{k=1}^K a_k \cdot \omega^k + b \leq u_j(t-\omega) \\ & \forall \omega \in \{\delta_0, \delta_0 + \frac{\delta_1 - \delta_0}{n}, \delta_0 + 2\cdot\frac{\delta_1-\delta_0}{n}, \cdots, \delta_{K+1}\}. \end{aligned} \tag{A.9}$$

Take $(a_1^n, \cdots, a_K^n, b^n)$ an optimal basic feasible solution to (A.9). By a standard linear programming argument:

$$\max(\max_{k=1,\cdots,K} |a_k^n|, |b^n|) \leq U,$$

where $U = ((K+1) \cdot \max(1, u_j(t - \delta^{\inf}), (\delta_{ij}^{\sup})^K))^{(K+1)}$ does not depend on n. Let us use the shorthand:

$$V = U \cdot (\delta_{ij}^{\sup} - \delta_{ij}^{\inf}) \cdot \sum_{k=1}^{K} k \cdot (\delta_{ij}^{\sup})^{(k-1)},$$

and define $b = b^n - \frac{V}{n} - \epsilon$. We show that $(a_1^n, \cdots, a_K^n, b)$ is feasible for (A.8). For any $w \in [\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$, take $w' \in \{\delta_0, \delta_0 + \frac{\delta_1 - \delta_0}{n}, \delta_0 + 2 \cdot \frac{\delta_1 - \delta_0}{n}, \cdots, \delta_{K+1}\}$ such that $|w - w'| \leq \frac{\delta_{ij}^{\sup} - \delta_{ij}^{\inf}}{n}$. We have:

$$\begin{aligned}
\sum_{k=1}^{K} a_k^n \cdot \omega^k + b &= \sum_{k=1}^{K} a_k^n \cdot (\omega')^k + b^n + \sum_{k=1}^{K} a_k^n \cdot (\omega^k - (\omega')^k) - \frac{V}{n} - \epsilon \\
&\leq u_j(t - \omega') + \sum_{k=1}^{K} |a_k^n| \cdot |\omega^k - (\omega')^k| - \frac{V}{n} - \epsilon \\
&\leq u_j(t - \omega) + \sum_{k=1}^{K} U \cdot k \cdot (\delta_{ij}^{\sup})^{(k-1)} \cdot |\omega - \omega'| - \frac{V}{n} \\
&\leq u_j(t - \omega),
\end{aligned}$$

where we use the fact that $(a_1^n, \cdots, a_K^n, b^n)$ is feasible for (A.9) in the first inequality, the uniform continuity of $u_j(\cdot)$ in the second and the definition of V in the last one. We derive:

$$f_i(t) - \frac{V}{n} - \epsilon \leq u_i(t) \leq f_i(t).$$

Taking $n \to \infty$ and $\epsilon \to 0$, we obtain $f_i(t) = u_i(t)$. This concludes the induction.

As a consequence of (A.5), the infimum in (2.7) is always attained for p_{ij}, irrespective of the remaining budget t, so we can conclude that (2.2) and (2.3) are equivalent.

□

A.2.9 Proof of Lemma 2.3

This result is a direct consequence of the following observations:

- when the risk function is $f(t) = t$, following the shortest path with respect to $(\max_{p\in\mathcal{P}_{ij}} \mathbb{E}_{X\sim p}[X])_{(i,j)\in\mathcal{A}}$ is an optimal strategy for (2.3),
- when the risk function is $f(t) = \exp(t)$, following the shortest path with respect to $(\max_{p\in\mathcal{P}_{ij}} -\log(\mathbb{E}_{X\sim p}[\exp(-X)]))_{(i,j)\in\mathcal{A}}$ is an optimal strategy for (2.3),
- when the risk function is $f(t) = -\exp(-t)$, following the shortest path with respect to $(\max_{p\in\mathcal{P}_{ij}} \log(\mathbb{E}_{X\sim p}[\exp(X)]))_{(i,j)\in\mathcal{A}}$ is an optimal strategy for (2.3).

As a consequence, for any of these risk functions, (2.2) and (2.3) are equivalent. Define $g(\cdot)$ as any of these risk functions. Assuming that $\gamma\cdot g(t)+\beta \geq f(t) \geq a\cdot g(t)+b, \forall t \leq T$, we get:

$$\begin{aligned}
\sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T-X_\pi)] &\leq \sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[\gamma\cdot g(T-X_\pi)+\beta] \\
&\leq \beta + \gamma\cdot \sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[g(T-X_\pi)] \\
&\leq \beta + \gamma\cdot \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[g(T-X_\pi)] \\
&\leq \beta - \frac{\gamma}{a}\cdot b \\
&+ \frac{\gamma}{a}\cdot \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[a\cdot g(T-X_\pi)+b] \\
&\leq \beta - \frac{\gamma}{a}\cdot b \\
&+ \frac{\gamma}{a}\cdot \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T-X_\pi)].
\end{aligned}$$

This last inequality along with:

$$\sup_{\pi\in\Pi} \inf_{\forall(i,j)\in\mathcal{A},\, p_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}}[f(T-X_\pi)] \geq \sup_{\pi\in\Pi} \inf_{\forall\tau,\forall(i,j)\in\mathcal{A},\, p^\tau_{ij}\in\mathcal{P}_{ij}} \mathbb{E}_{\mathbf{p}^\tau}[f(T-X_\pi)]$$

yields the claim with some basic algebra.

A.2.10 Proof of Lemma 2.4

Proof. Proof of Lemma 2.4. For any $k \in \mathbb{N}$, we define $(\pi^k, (u_i^k(\cdot))_{i\in\mathcal{V}})$ as a solution to the dynamic program (2.7) when the ambiguity sets are taken as $(\mathcal{P}^k_{ij})_{(i,j)\in\mathcal{A}}$. Similarly, we

define $(\pi^\infty, (u_i^\infty(\cdot))_{i\in\mathcal{V}})$ as a solution to the dynamic program (2.7) when the ambiguity sets are taken as $(\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k)_{(i,j)\in\mathcal{A}}$. Along the sames lines as what is done in the proof of Proposition 2.4, we can show that the functions $(u_i^k(\cdot))_{k\in\mathbb{N}}$ and $u_i^\infty(\cdot)$ are continuous for any $i \in \mathcal{V}$. Because the ambiguity sets are nested, observe that the sequence $(u_i^k(t))_{k\in\mathbb{N}}$ is non-decreasing for any $t \leq T$, hence it converges to a limit $f_i(t) \leq u_i^\infty(t)$. Moreover, $f_d(t) = f(t)$ for all $t \leq T$. Take $i \neq d$ and $t \leq T$. We have, for any $k \in \mathbb{N}$ and $m \leq k$:

$$\begin{aligned}
f_i(t) &\geq u_i^k(t) \\
&\geq \max_{j\in\mathcal{V}(i)} \inf_{p\in\mathcal{P}_{ij}^k} \int_0^\infty p(\omega)\cdot u_j^k(t-\omega)\mathrm{d}\omega \\
&\geq \max_{j\in\mathcal{V}(i)} \int_0^\infty p_{ij}^k(\omega)\cdot u_j^k(t-\omega)\mathrm{d}\omega \\
&\geq \max_{j\in\mathcal{V}(i)} \int_0^\infty p_{ij}^k(\omega)\cdot u_j^m(t-\omega)\mathrm{d}\omega,
\end{aligned}$$

where $p_{ij}^k \in \mathcal{P}_{ij}^k$ achieves the minimum for any $j \in \mathcal{V}(i)$, which can be shown to exist since $\mathcal{P}_{ij}^k$ is compact and $u_j^k(\cdot)$ is continuous. Because $\mathcal{P}_{ij}^k$ is compact for the weak topology, we can take a subsequence of $(p_{ij}^k)_{k\in\mathbb{N}}$ that converges to a distribution p_{ij}^∞ in $\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k$. Without loss of generality we continue to refer to this sequence as $(p_{ij}^k)_{k\in\mathbb{N}}$. Taking the limit $k \to \infty$ in the last inequality derived yields:

$$f_i(t) \geq \max_{j\in\mathcal{V}(i)} \int_0^\infty p_{ij}^\infty(\omega)\cdot u_j^m(t-\omega)\mathrm{d}\omega.$$

Observing that $u_j^m(t-\omega) \geq u_j^1(t-\omega)$, we can use the monotone convergence theorem for $m \to \infty$ and conclude that:

$$\begin{aligned}
f_i(t) &\geq \max_{j\in\mathcal{V}(i)} \int_0^\infty p_{ij}^\infty(\omega)\cdot f_j(t-\omega)\mathrm{d}\omega \\
&\geq \max_{j\in\mathcal{V}(i)} \inf_{p\in\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k} \int_0^\infty p(\omega)\cdot f_j(t-\omega)\mathrm{d}\omega.
\end{aligned}$$

We use Theorem 2.2 for the ambiguity sets $(\cap_{k\in\mathbb{N}}\mathcal{P}_{ij}^k)_{(i,j)\in\mathcal{A}}$ and denote by T_f^r (resp. $\mathcal{T}^r$) the time budget (resp. the tree) put forth in the statement of the theorem. Using the last sequence of inequalities derived, we can prove, by induction on the levels of the nodes in

$\mathcal{T}^r$ that:

$$f_i(t) \geq u_i^{\infty}(t) \quad \forall t \in [T_f^r - (|\mathcal{V}| - \text{level}(i, \mathcal{T}^r) + 1) \cdot \delta^{\sup}, T_f^r], \forall i \in \mathcal{V},$$

and then by induction on $m \in \mathbb{N}$ that:

$$f_i(t) \geq u_i^{\infty}(t) \quad \forall t \in [T_f^r - (|\mathcal{V}| - \text{level}(i, \mathcal{T}^r) + 1) \cdot \delta^{\sup}, T_f^r + m \cdot \delta^{\inf}], \forall i \in \mathcal{V}.$$

We finally obtain $f_s(T) \geq u_s^{\infty}(T)$ which concludes the proof.

□

A.2.11 Proof of Lemma 2.5

Proof. Proof of Lemma 2.5. Since the optimization problem (2.13) is a conic linear problem over the set of measures on $[\delta_{ij}^{\inf}, \delta_{ij}^{\sup}]$, we can take its Lagrangian dual, in the same sense as defined in [87] using the notion of a polar cone, and use Proposition 3.1 in [87] (established through a conjugate duality approach) to prove strong duality. The assumptions of this proposition are satisfied here since: (i) the value of (2.13) is finite as $\mathcal{P}_{ij}$ is compact and not empty; and (ii) the functions $(g_q^{ij}(\cdot))_{q=1,\cdots,Q_{ij}}$ and $u_j^{\Delta t}(\cdot)$ are continuous, which implies that the set:

$$\left\{ (y_1, \cdots, y_{Q_{ij}}, x_1, \cdots, x_{Q_{ij}}, \kappa) \in \mathbb{R}^{2 \cdot Q_{ij}+1} \mid \begin{array}{l} \exists p \in \mathcal{P}_{ij} \text{ such that :} \\ x_q \leq \mathbb{E}_{X \sim p}[g_q^{ij}(X)], \quad q = 1, \cdots, Q_{ij} \\ y_q \geq \mathbb{E}_{X \sim p}[g_q^{ij}(X)], \quad q = 1, \cdots, Q_{ij} \\ \kappa = \mathbb{E}_{X \sim p}[u_j^{\Delta t}(k \cdot \Delta t - X)] \end{array} \right\}$$

is closed for the standard topology of $\mathbb{R}^{2 \cdot Q_{ij}+1}$.

□

A.2.12 Proof of Lemma 2.7

Proof. Proof of Lemma 2.7. First observe that, along the same lines as in the general case, the constraint

$$z + (x - y) \cdot l \cdot \Delta t \leq u_j^{\Delta t}((k - l) \cdot \Delta t)$$

does not limit the feasible region if $(l \cdot \Delta t, u_j^{\Delta t}((k - l) \cdot \Delta t))$ is not an extreme point of the upper convex hull of $\{(l \cdot \Delta t, u_j^{\Delta t}(l \cdot \Delta t)),\ l = k - \left\lceil \frac{\delta_{ij}^{\sup}}{\Delta t} \right\rceil, \cdots, k - \left\lfloor \frac{\delta_{ij}^{\inf}}{\Delta t} \right\rfloor\} \cup \{(\delta_{ij}^{\sup}, u_j^{\Delta t}(k \cdot \Delta t - \delta_{ij}^{\sup})), (\delta_{ij}^{\inf}, u_j^{\Delta t}(k \cdot \Delta t - \delta_{ij}^{\inf}))\}$. Hence, we can discard the constraints that do no satisfy this property from (2.17). We denote by S the sorted projection of the set of extreme points onto the first coordinate. Observe that the feasible region is pointed as the polyhedron described by the inequality constraints does not contain any line, therefore there exists a basic optimal feasible solution for which at least three inequality constraints are binding. By definition of S, only two of the constraints

$$z + (x - y) \cdot \omega \leq u_j^{\Delta t}(\omega) \quad \omega \in S$$

can be binding which further implies that at least one of the constraints $x \geq 0$ and $y \geq 0$ must be binding. There are three types of feasible bases depending on whether these last two constraints are binding or if only one of them is. We show that, for each type, we can identify an optimal basis among the bases of the same type by binary search on the first coordinate of the extreme points. This will conclude the proof as it takes constant time to compare the objective function achieved by each of the three potentially optimal bases. Since, by definition of S, $u_j^{\Delta t}(\cdot)$ is convex on S, we can partition S into S_1 and S_2 such that $u_j^{\Delta t}(\cdot)$ is non-increasing on S_1 and non-decreasing on S_2 with $\max(S_1) = \min(S_2)$.

If $x \geq 0$ and $y \geq 0$ are binding then z is the only non-zero variable and the objective is to maximize z. Hence, the optimal basis of this type is given by $x = 0$, $y = 0$ and $z = \min_{\omega \in S} u_j^{\Delta t}(\omega)$ which can be computed by binary search since $u_j^{\Delta t}(\cdot)$ is convex on S.

If only $x \geq 0$ is binding, then the line $\omega \to z - y \cdot \omega$ must be joining two consecutive points in S_1. Since the objective function is precisely the value taken by the line $\omega \to z - y \cdot \omega$ at β^{ij}, the optimal straight line joins two consecutive points in S_1, ω_1 and ω_2, that

satisfy $\omega_1 \leq \beta^{ij} \leq \omega_2$ assuming $\max(S_1) \geq \beta^{ij}$. If $\max(S_1) < \beta^{ij}$, the feasible bases of this type are dominated by the optimal basis of the first type. Computing ω_1 and ω_2 or showing that they do not exist can be done with a single binary search on S.

The discussion is analogous if only $y \geq 0$ is binding instead. The line $\omega \to z + x \cdot \omega$ must be joining two consecutive points in S_2. Since the objective function is precisely the value taken by this line at α^{ij}, the optimal straight line joins two consecutive points in S_2, ω_1 and ω_2, that satisfy $\omega_1 \leq \alpha^{ij} \leq \omega_2$ assuming $\alpha^{ij} \geq \min(S_2)$. If $\min(S_2) > \alpha^{ij}$, the feasible bases of this type are dominated by the optimal basis of the first type. Computing ω_1 and ω_2 or showing that they do not exist can be done with a single binary search on S.

□

Appendix B

Appendix for Chapter 3

B.1 Proof of Theorem 3.1

The assumptions of Theorem 1 in [2] are satisfied for the game $(\ell, \mathcal{Z}, \mathcal{F})$ using Assumption 3.1 and the fact that any loss function ℓ of the form (3.1) is such that $\ell(z, \cdot)$ is continuous for any $z \in \mathcal{Z}$.

B.2 Proof of Lemma 3.1

The proof follows from the repeated use of the von Neumann's minimax theorem developed in [2] (see Theorem 3.1). To simplify the presentation, we prove the result when $T = 2$ but the general proof proceeds along the same lines. Using Theorem 1 in [2], we have:

$$R_2(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathcal{Z}} \inf_{f_2 \in \mathcal{F}} \sup_{z_2 \in \mathcal{Z}} [\sum_{t=1}^{2} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f)].$$

Consider $f_1, f_2 \in \mathcal{F}$ and $z_1 \in \mathcal{Z}$ and define the function $M(z_2) = \sum_{t=1}^{2} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f)$. Observe that M is convex. Indeed, $z_2 \to \ell(z_1, f_1) + \ell(z_2, f_2)$ is affine and $z_2 \to \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f)$ is concave as the infimum of affine functions. Therefore, we have:

$$\sup_{z_2 \in \mathcal{Z}} M(z_2) = \sup_{z_2 \in \text{conv}(\mathcal{Z})} M(z_2).$$

We obtain:

$$R_2(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathcal{Z}} \inf_{f_2 \in \mathcal{F}} \sup_{z_2 \in \mathrm{conv}(\mathcal{Z})} [\, \sum_{t=1}^{2} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f) \,].$$

By randomizing the choice of z_2, we can use the von Neumann's minimax theorem to derive:

$$R_2(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathcal{Z}} \{\, \ell(z_1, f_1) + \sup_{p_2 \in \mathcal{P}(\mathrm{conv}(\mathcal{Z}))} \{\, \inf_{f_2 \in \mathcal{F}} \mathbb{E}_{z \sim p_2} \ell(z, f_2) - \mathbb{E}_{z_2 \sim p_2} \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f) \,\} \,\}.$$

For a fixed $f_1 \in \mathcal{F}$, define:

$$A(z_1) = \ \ell(z_1, f_1) + \sup_{p_2 \in \mathcal{P}(\mathrm{conv}(\mathcal{Z}))} \{\, \inf_{f_2 \in \mathcal{F}} \mathbb{E}_{z \sim p_2} \ell(z, f_2) - \mathbb{E}_{z_2 \sim p_2} \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f) \,\},$$

for any $z_1 \in \mathcal{Z}$. Observe that, for a fixed $p_2 \in \mathcal{P}(\mathrm{conv}(\mathcal{Z}))$, the function:

$$z_1 \to \inf_{f_2 \in \mathcal{F}} \mathbb{E}_{z \sim p_2} \ell(z, f_2) - \mathbb{E}_{z_2 \sim p_2} \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f)$$

is convex as the difference between a constant and the expected value of the infimum of affine functions. Since the supremum of convex functions is convex, A is convex and $\sup_{z_1 \in \mathcal{Z}} A(z_1) = \sup_{z_1 \in \mathrm{conv}(\mathcal{Z})} A(z_1)$. We derive:

$$R_2(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathrm{conv}(\mathcal{Z})} [\, \ell(z_1, f_1) + \sup_{p_2 \in \mathcal{P}(\mathrm{conv}(\mathcal{Z}))} \{\, \inf_{f_2 \in \mathcal{F}} \mathbb{E}_{z \sim p_2} \ell(z, f_2) - \mathbb{E}_{z_2 \sim p_2} \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f) \}].$$

To conclude, we unwind the first step, i.e. we use the minimax theorem in reverse order. This yields:

$$R_2(\ell, \mathcal{Z}, \mathcal{F}) = \inf_{f_1 \in \mathcal{F}} \sup_{z_1 \in \mathrm{conv}(\mathcal{Z})} \inf_{f_2 \in \mathcal{F}} \sup_{z_2 \in \mathrm{conv}(\mathcal{Z})} [\sum_{t=1}^{2} \ell(z_t, f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{2} \ell(z_t, f) \,],$$

i.e. $R_2(\ell, \mathcal{Z}, \mathcal{F}) = R_2(\ell, \text{conv}(\mathcal{Z}), \mathcal{F})$. Moreover, $\mathcal{Z}$ is a compact set which implies that $\text{conv}(\mathcal{Z})$ is also a compact set by a standard topological argument. As a result, the game $(\ell, \text{conv}(\mathcal{Z}), \mathcal{F})$ also satisfies Assumption 3.1.

B.3 Proof of Lemma 3.2

We follow the analysis carried out in the proof of Theorem 19 in [2]. Using Theorem 3.1 with p taken as the distribution defined by i.i.d. copies of Z, we get the lower bound:

$$\begin{aligned}
R_T &\geq T \inf_{f\in\mathcal{F}} \mathbb{E}[\ell(Z_t, f)] - \mathbb{E}[\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)] \\
&\geq T \sup_{f\in\{f_1,f_2\}} \mathbb{E}[\ell(Z_t, f)] - \mathbb{E}[\inf_{f\in\{f_1,f_2\}} \sum_{t=1}^{T} \ell(Z_t, f)] \\
&\geq \mathbb{E}[\max\{\sum_{t=1}^{T} \mathbb{E}[\ell(Z_t, f_1)] - \ell(Z_t, f_1), \sum_{t=1}^{T} \mathbb{E}[\ell(Z_t, f_2)] - \ell(Z_t, f_2)\}] \\
&\geq \mathbb{E}[\max\{0, \sum_{t=1}^{T} \ell(Z_t, f_1) - \ell(Z_t, f_2)\}],
\end{aligned}$$

where we use the fact that $\inf_{f\in\mathcal{F}} \mathbb{E}[\ell(Z_t, f)] = \mathbb{E}[\ell(Z_t, f_1)] = \mathbb{E}[\ell(Z_t, f_2)]$. Since $\ell(Z, f_2) \neq \ell(Z, f_1)$ with positive probability, the random variables $(\ell(Z_t, f_1) - \ell(Z_t, f_2))_{t=1,\cdots,T}$ are i.i.d. with zero mean and positive variance and we can conclude with the central limit theorem since ℓ is bounded.

B.4 Proof of Lemma 3.3

The fact that $R_T \geq 0$ follows from Theorem 3.1 by taking the Z_t's to be deterministic and all equal to any $z \in \mathcal{Z}$. Clearly, if the game is trivial then $R_T = 0$ because this value is attained for $f_1, \cdots, f_T = f^*$ irrespective of the decisions made by the opponent. Conversely, suppose that $R_T = 0$. Take p to be the product of T uniform distributions on $\mathcal{Z}$. Then, using again Theorem 3.1, we have:

$$0 \geq \mathbb{E}[\sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathbb{E}[\ell(Z_t, f_t)] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)\ ],$$

as $Z_1, \cdots, Z_T$ are independent random variables. Since they are also identically distributed, we obtain:

$$0 \geq T \cdot \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)] - \mathbb{E}[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)].$$

Yet $\mathbb{E}[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\sum_{t=1}^{T} \ell(Z_t, f)] = T \cdot \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)]$ and we derive:

$$T \cdot \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)] - \mathbb{E}[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)] = 0.$$

Since ℓ is bounded, $\mathcal{Z}$ is compact, and $\ell(z, \cdot)$ is continuous for any $z \in \mathcal{Z}$, $f \to \mathbb{E}[\ell(Z, f)]$ is continuous by dominated convergence so, since $\mathcal{F}$ is compact, we can take

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)].$$

We obtain:

$$\mathbb{E}[\sum_{t=1}^{T} \ell(Z_t, f^*) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f)] = 0.$$

As $\sum_{t=1}^{T} \ell(Z_t, f^*) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(Z_t, f) \geq 0$, we derive that:

$$(z_1, \cdots, z_T) \to \sum_{t=1}^{T} \ell(z_t, f^*) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(z_t, f) = 0$$

holds almost everywhere on $\mathcal{Z}^T$. If $\mathcal{Z}$ is discrete, this implies equality on $\mathcal{Z}^T$, which in particular implies that $\ell(z, f^*) = \inf_{f \in \mathcal{F}} \ell(z, f)$ for all $z \in \mathcal{Z}$ and we are done. If, on the other hand, $\ell(\cdot, f)$ is continuous for all $f \in \mathcal{F}$, we have:

$$\sum_{t=1}^{T} \ell(z_t, f^*) \leq \sum_{t=1}^{T} \ell(z_t, f), \ \forall f \in \mathcal{F}, \forall (z_1, \cdots, z_T) \in \tilde{\mathcal{Z}},$$

for $\tilde{\mathcal{Z}}$ a subset of $\mathcal{Z}$ with Lebesgue measure equal to that of $\mathcal{Z}$. Since a non-empty open set cannot have Lebesgue measure 0, $\tilde{\mathcal{Z}}$ is dense in $\mathcal{Z}$ and by taking limits in the above

inequality for each $f \in \mathcal{F}$ separately, we conclude that:

$$\sum_{t=1}^{T} \ell(z_t, f^*) \leq \sum_{t=1}^{T} \ell(z_t, f), \ \forall f \in \mathcal{F}, \forall (z_1, \cdots, z_T) \in \mathcal{Z},$$

which in particular implies that $\ell(z, f^*) = \inf_{f \in \mathcal{F}} \ell(z, f)$ for all $z \in \mathcal{Z}$ and the game is trivial.

B.5 Proof of Lemma 3.4

Suppose by contradiction that we cannot find such a finite subset. Since $\mathcal{Z}$ is compact, it is also separable thus it contains a countable dense subset $\{z_n \mid n \in \mathbb{N}\}$. By assumption, the game $(\ell, \{z_k \mid k \leq n\}, \mathcal{F})$ must be trivial for any n, i.e. there exists $f_n \in \mathcal{F}$ such that:

$$\ell(z_k, f_n) \leq \min_{f \in \mathcal{F}} \ell(z, f), \ \forall k \leq n.$$

Since $\mathcal{F}$ is compact, we can find a subsequence of $(f_n)_{n \in \mathbb{N}}$ such that $f_n \to f^* \in \mathcal{F}$. Without loss of generality, we continue to refer to this sequence as $(f_n)_{n \in \mathbb{N}}$. Taking the limit $n \to \infty$ in the above inequality for any fixed $k \in \mathbb{N}$ yields:

$$\ell(z_k, f^*) \leq \ell(z_k, f), \ \forall f \in \mathcal{F}, \forall k \in \mathbb{N}.$$

Consider a fixed $f \in \mathcal{F}$. Since $\{z_n \mid n \in \mathbb{N}\}$ is dense in $\mathcal{Z}$ and since $\ell(\cdot, f^*)$ and $\ell(\cdot, f)$ are continuous, we get:

$$\ell(z, f^*) \leq \ell(z, f), \ \forall f \in \mathcal{F}, \forall z \in \mathcal{Z},$$

which shows that $(\ell, \mathcal{Z}, \mathcal{F})$ is trivial, a contradiction.

B.6 Proof of Theorem 3.2

Without loss of generality we can assume that the game is not trivial and that $\mathcal{X}(z)$ is finite for any $z \in \mathcal{Z}$ since otherwise, if $\mathcal{X}(z)$ is a polyhedron, the maximum in (3.1) must be attained at an extreme point of $\mathcal{X}(z)$ (ℓ is bounded by Assumption 3.1) and there are finitely many such points for any z. Moreover, we can also assume that $\mathcal{Z}$ is discrete by

Lemma 3.4 since, borrowing the notations of Lemma 3.4, we have:

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) \geq R_T(\ell, \tilde{\mathcal{Z}}, \mathcal{F}).$$

Write $\mathcal{Z} = \{z_n \mid 1 \leq n \leq N\}$ and denote by p_0 the uniform distribution on $\mathcal{Z}$, i.e. $p_0(n) = \frac{1}{N}$, for any $n \leq N$. We may assume that there is a single equivalence class in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_0}[\ell(Z, f)]$, otherwise we are done by Lemma 3.2. Take

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_0}[\ell(Z, f)].$$

Since the game $(\ell, \mathcal{Z}, \mathcal{F})$ is not trivial, there exists z_k in $\mathcal{Z}$ and f^{**} in $\mathcal{F}$ such that

$$\ell(z_k, f^{**}) < \ell(z_k, f^*).$$

Therefore, we can find $\epsilon > 0$ small enough such that $(N-1)\epsilon < 1$ and:

$$(1-(N-1)\epsilon)\ell(z_k, f^{**}) + \sum_{n \neq k} \epsilon \ell(z_n, f^{**}) < (1-(N-1)\epsilon)\ell(z_k, f^*) + \sum_{n \neq k} \epsilon \ell(z_n, f^*).$$

Define p_1 as the corresponding distribution, i.e. $p_1(n) = \epsilon$ for $n \neq k$ and $p_1(k) = 1-(N-1)\epsilon$. By construction, the equivalence class of f^* is not in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_1}[\ell(Z, f)]$. Once again, without loss of generality, we may assume that there is a single equivalence class in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_1}[\ell(Z, f)]$, otherwise we are done by Lemma 3.2. Moreover, we can now redefine f^{**} as a representative of the only equivalence class contained in $\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_1}[\ell(Z, f)]$. We now move on to show that there must exist $\alpha \in (0, 1)$ such that there are at least two equivalence classes in $\phi(\alpha) = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{p_\alpha}[\ell(Z, f)]$, where the distribution p_α is defined as $p_\alpha = (1-\alpha)p_0 + \alpha p_1$. This will conclude the proof by Lemma 3.2. For any $f \in \mathcal{F}$, define $I(f) = \{\alpha \in [0, 1] \mid f \in \phi(\alpha)\}$. Since $\alpha \to \mathbb{E}_{p_\alpha}[\ell(Z, f)]$ is linear in α, $I(f)$ is a closed interval. Moreover, note that $\min_{f \in \mathcal{F}} \mathbb{E}_{p_\alpha}[\ell(Z, f)]$ is equal to

the optimal value of the optimization problem:

$$\begin{array}{ll} \min\limits_{q_1,\cdots,q_N,f} & q^\intercal \cdot ((1-\alpha)p_0 + \alpha p_1) \\ \text{subject to} & q = (q_1, \cdots, q_N) \\ & q_n \geq (C(z_n)f + c(z_n))^\intercal x \quad \forall x \in \mathcal{X}(z_n), \forall n = 1, \cdots, N \\ & f \in \mathcal{F}, q_1, \cdots, q_N \in \mathbb{R}. \end{array} \tag{B.1}$$

This is because: (i) for any optimal solution to (B.1) $(q_1, \cdots, q_N, f)$,

$$(\max_{x \in \mathcal{X}(z_1)} (C(z_1)f + c(z_1))^\intercal x, \cdots, \max_{x \in \mathcal{X}(z_N)} (C(z_N)f + c(z_N))^\intercal x, f)$$

is also an optimal solution with objective function $\mathbb{E}_{p_\alpha}[\ell(Z, f)]$, and (ii) any $f \in \mathcal{F}$ can be mapped to a feasible solution of (B.1) with objective function $\mathbb{E}_{p_\alpha}[\ell(Z, f)]$ through the mapping $f \to (\max_{x \in \mathcal{X}(z_1)} (C(z_1)f + c(z_1))^\intercal x, \cdots, \max_{x \in \mathcal{X}(z_N)} (C(z_N)f + c(z_N))^\intercal x, f)$. Since $\mathcal{F}$ is a polyhedron and $\mathcal{X}(z_n)$ is finite for any n, this optimization problem is a linear program. Denoting by $\{f_1, \cdots, f_L\}$ the projections onto the f coordinate of the extreme points of the feasible set of (B.1), there exists, for any $\alpha \in [0, 1]$, $l \in \{1, \cdots, L\}$ such that $f_l \in \phi(\alpha)$. Hence, we can write $[0, 1] = \cup_{l=1}^{L} I(f_l)$. We can further simplify this description by assuming that the f_l's belong to different equivalence classes (because $I(f) = I(f')$ if $f \sim_\ell f'$). Now observe that if $I(f_l) \cap I(f_j) \neq \emptyset$ for some $l \neq j \leq L$, then there are two equivalent classes in $\phi(\alpha)$ for any $\alpha \in I(f_l) \cap I(f_j)$ and we are done by Lemma 3.2. Suppose by contradiction that we cannot find such a pair of indices. Because the only way to partition $[0, 1]$ into $L < \infty$ non-overlapping closed intervals is to have $L = 1$, we get $[0, 1] = I(f_1)$. This implies that $f^* \sim_\ell f_1$ by optimality of f^* for $\alpha = 0$ and $f^{**} \sim_\ell f_1$ by optimality of f^{**} for $\alpha = 1$. We conclude that $f^* \sim_\ell f^{**}$, a contradiction.

B.7 Alternative Proof of Theorem 3.2

Using Lemma 3.1, we can assume without loss of generality that $\mathcal{Z}$ is convex. When ℓ is linear, the procedure developed in the proof of Theorem 3.2 boils down to finding a

point $z \in \text{int}(\mathcal{Z})$ such that $|\operatorname{argmin}_{f\in\mathcal{F}} z^\intercal f| > 1$ and, with further examination, we can also guarantee that there exists $\epsilon > 0, e \in \mathbb{R}^n$ and $f_1, f_2 \in \operatorname{argmin}_{f\in\mathcal{F}} z^\intercal f$ such that $f_1 \in \operatorname{argmin}_{f\in\mathcal{F}}(z - xe)^\intercal f$ while $f_2 \notin \operatorname{argmin}_{f\in\mathcal{F}}(z - xe)^\intercal f$ for all $x \in (0, \epsilon]$ and symmetrically for $x \in [-\epsilon, 0)$. Consider a randomized opponent $Z_t = z + (\epsilon_t \epsilon)e$ for $(\epsilon_t)_{t=1,\cdots,T}$ i.i.d. Rademacher random variables. Then for any player's strategy:

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] = \sum_{t=1}^{T} \mathbb{E}[Z_t]^\intercal f_t - \mathbb{E}[\inf_{f\in\mathcal{F}} f^\intercal \sum_{t=1}^{T} Z_t].$$

This yields:

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] = \sum_{t=1}^{T} z^\intercal f_t - \mathbb{E}[\inf_{f\in\mathcal{F}} f^\intercal \sum_{t=1}^{T} Z_t].$$

We can lower bound the last quantity by:

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] \geq T(z^\intercal f_1) - T\mathbb{E}[\inf_{f\in\mathcal{F}} f^\intercal (z + (\epsilon \cdot \frac{\sum_{t=1}^{T} \epsilon_t}{T})e)],$$

as $f_1 \in \operatorname{argmin}_{f\in\mathcal{F}} z^\intercal f$, but we could have equivalently picked f_2 as $f_1^\intercal z = f_2^\intercal z$. Furthermore, as $|\frac{\sum_{t=1}^{T} \epsilon_t}{T}| \leq 1$, f_1 is optimal in the inner optimization problem when $\sum_{t=1}^{T} \epsilon_t \leq 0$ while f_2 is optimal when $\sum_{t=1}^{T} \epsilon_t \geq 0$. Hence:

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] \geq T(z^\intercal f_1) - T\mathbb{E}[\, f_1^\intercal (z + (\epsilon \cdot \frac{\sum_{t=1}^{T} \epsilon_t}{T})e) \cdot 1_{\sum_{t=1}^{T} \epsilon_t \leq 0} + f_2^\intercal (z + (\epsilon \cdot \frac{\sum_{t=1}^{T} \epsilon_t}{T})e) \cdot 1_{\sum_{t=1}^{T} \epsilon_t \geq 0}\,].$$

Observe that the term $T(z^\intercal f_1)$ cancels out and we get:

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] \geq \frac{\mathbb{E}[|\sum_{t=1} \epsilon_t|]}{T} \cdot \epsilon \cdot (f_1^\intercal e - f_2^\intercal e).$$

By Khintchine's inequality $\mathbb{E}[|\sum_{t=1}^{T} \epsilon_t|] \geq \frac{1}{\sqrt{2}}\sqrt{T}$. Moreover $f_1^\intercal e - f_2^\intercal e > 0$ because $f_2 \in \operatorname{argmin}_{f\in\mathcal{F}}(z + \epsilon e)^\intercal f$ while f_1 does not and $f_1^\intercal z = f_2^\intercal z$. We finally derive

$$\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})] \geq \frac{(f_1^\intercal e - f_2^\intercal e)}{\sqrt{2}}\sqrt{T}.$$

This enables us to conclude $R_T = \Omega(\sqrt{T})$ as this shows that for any player's strategy, there exists a sequence $z_1, \cdots, z_T$ such that

$$r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) \geq \mathbb{E}[r_T((Z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T})].$$

B.8 Proof of Theorem 3.3

Straightforward from Theorem 3.2 since ℓ is jointly continuous.

B.9 Proof of Lemma 3.5

Using Lemma 3.1, we can assume that $\mathcal{Z}$ is convex. Consider $f_1 \neq f_2 \in \mathcal{F}$ and define $e = \frac{f_1 - f_2}{\|f_1 - f_2\|}$. Since $0 \in \operatorname{int}(\mathcal{Z})$, there exists $\epsilon > 0$ such that ϵe and $-\epsilon e$ are in $\mathcal{Z}$. We restrict the opponent's decision set by imposing that, at any round t, the opponent's move be $y_t \epsilon e$ for $y_t \in \tilde{\mathcal{Z}} = \{-1, 1\}$. Since $\ell(y_t \epsilon e, f)$ only depends on f through the scalar product between f and e, the player's decision set can equivalently be described by $\tilde{\mathcal{F}} = \{f^\intercal e \mid f \in \mathcal{F}\}$ which is a closed interval (since $\mathcal{F}$ is convex and compact) and thus a polyhedron. Defining a new loss function as $\tilde{\ell}(y, f) = y\epsilon f$, we have:

$$R_T(\ell, \mathcal{Z}, \mathcal{F}) \geq R_T(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}}).$$

Observe that the game $(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}})$ is linear and not trivial, otherwise there would exist f^* such that $e^\intercal f^* \leq e^\intercal f_2$ and $-e^\intercal f^* \leq -e^\intercal f_1$ which would imply $\|e\| = 0$. With Theorem 3.3, we conclude $R_T(\tilde{\ell}, \tilde{\mathcal{Z}}, \tilde{\mathcal{F}}) = \Omega(\sqrt{T})$ and thus $R_T(\ell, \mathcal{Z}, \mathcal{F}) = \Omega(\sqrt{T})$.

B.10 Proof of Lemma 3.6

Using Lemma 3.1, we can assume that $\mathcal{Z}$ is convex. Since $\mathcal{Z}$ is compact and convex and since $0 \notin \mathcal{Z}$, we can strictly separate 0 from $\mathcal{Z}$ and find $z^* \neq 0$ such that $\mathcal{Z} \subseteq B_2(z^*, \alpha \|z^*\|)$ with $\alpha < 1$. By rescaling $\mathcal{Z}$, we can assume that $\alpha \|z^*\| = 1$ and $\|z^*\| > 1$. In the sequel, σ_{t-1} serves as a shorthand for $\sigma(Z_1, \cdots, Z_{t-1})$. We prove more generally

that, for any choice of random variables $(Z_1, \cdots, Z_T)$ such that $\mathbb{E}[Z_t|\sigma_{t-1}]$ is constant almost surely, the lower bound on regret derived from Theorem 3.1 is $O(1)$. Write $Z_t = z^* + W_t$ and $\mathbb{E}[W_t|\sigma_{t-1}] = c_t$ with $\|W_t\| \leq 1$ and $\|c_t\| \leq 1$. Define $w^* = T \cdot z^* + \sum_{t=1}^T c_t$. Observe that $\|w^*\| \geq T \cdot \|z^*\| - \|\sum_{t=1}^T c_t\| \geq T \cdot (\|z^*\| - 1) > 0$. Write $W_t = X_t \frac{w^*}{\|w^*\|} + \tilde{W}_t + c_t$ with $\tilde{W}_t^\intercal w^* = 0$. Projecting down the equality $\mathbb{E}[W_t - c_t|\sigma_{t-1}] = 0$ onto w^*, we get $\mathbb{E}[X_t|\sigma_{t-1}] = 0$ and $\mathbb{E}[\tilde{W}_t|\sigma_{t-1}] = 0$. The bound that results from an application of Theorem 3.1 is:

$$R_T \geq \mathbb{E}[\|w^* + \sum_{t=1}^T W_t - c_t\|] - \sum_{t=1}^T \|z^* + c_t\|].$$

We now focus on finding an upper bound on the right-hand side. Expanding the first term yields:

$$\|w^* + \sum_{t=1}^T W_t - c_t\| = \sqrt{(1 + \sum_{t=1}^T \frac{X_t}{\|w^*\|})^2 \cdot \|w^*\|^2 + \|\sum_{t=1}^T \tilde{W}_t\|^2}.$$

By concavity of the squared root function:

$$\mathbb{E}[\|w^* + \sum_{t=1}^T W_t - c_t\|] \leq \sqrt{\|w^*\|^2 \cdot \mathbb{E}[(1 + \sum_{t=1}^T \frac{X_t}{\|w^*\|})^2] + \mathbb{E}[\|\sum_{t=1}^T \tilde{W}_t\|^2]}.$$

We expand the two inner terms:

$$\mathbb{E}[(1 + \sum_{t=1}^T \frac{X_t}{\|w^*\|})^2] = 1 + 2\sum_{t=1}^T \frac{\mathbb{E}[X_t]}{\|w^*\|} + \frac{1}{\|w^*\|^2} \cdot \mathbb{E}[(\sum_{t=1}^T X_t)^2].$$

Looking at each term individually, we have $\mathbb{E}[X_t] = \mathbb{E}[\mathbb{E}[X_t|\sigma_{t-1}] = 0$ and:

$$\mathbb{E}[(\sum_{t=1}^T X_t)^2] = \mathbb{E}[(\sum_{t=1}^{T-1} X_t)^2] + 2\mathbb{E}[X_T \cdot (\sum_{t=1}^{T-1} X_t)] + \mathbb{E}[X_T^2],$$

yet $\mathbb{E}[X_T \cdot (\sum_{t=1}^{T-1} X_t)] = \mathbb{E}[\mathbb{E}[X_T|\sigma_{T-1}] \cdot (\sum_{t=1}^{T-1} X_t)] = 0$. Hence, $\mathbb{E}[(1 + \sum_{t=1}^T \frac{X_t}{\|w^*\|})^2] = 1 + \frac{\mathbb{E}[\sum_{t=1}^T X_t^2]}{\|w^*\|^2}$. Similarly $\mathbb{E}[\|\sum_{t=1}^T \tilde{W}_t\|^2] = \sum_{t=1}^T \mathbb{E}[\|\tilde{W}_t\|^2]$. We obtain:

$$\mathbb{E}[\|w^* + \sum_{t=1}^T W_t - c_t\|] \leq \sqrt{\|w^*\|^2 + \sum_{t=1}^T \mathbb{E}[X_t^2 + \|\tilde{W}_t\|^2]}.$$

Remark that $\|W_t - c_t\| \leq \|W_t\| + \|c_t\| \leq 2$. Hence, $X_t^2 + \|\tilde{W}_t\|^2 \leq 2$. We obtain:

$$\mathbb{E}[\|w^* + \sum_{t=1}^{T} W_t - c_t\|] \leq \sqrt{\|w^*\|^2 + 2T}.$$

We have $\sqrt{\|w^*\|^2 + 2T} = \|w^*\| \cdot \sqrt{1 + \frac{2T}{\|w^*\|^2}} \leq \|w^*\| + \frac{T}{\|w^*\|}$ for T big enough as $\|w^*\| \geq T \cdot (\|z^*\| - 1)$. Yet $\|w^*\| = \|\sum_{t=1}^{T} z^* + c_t\| \leq \sum_{t=1}^{T} \|z^* + c_t\|$. Hence, the lower bound derived is:

$$\mathbb{E}[\|w^* + \sum_{t=1}^{T} W_t - c_t\|] - \sum_{t=1}^{T} \|z^* + c_t\| \leq \frac{T}{\|w^*\|} \leq \frac{1}{\|z^*\| - 1} = O(1).$$

B.11 Proof of Theorem 3.5

Without loss of generality, we may assume that $\mathcal{Z}$ is a convex set by Lemma 3.1. Remark that the game $(\ell, \mathcal{Z}, \mathcal{F})$ is trivial if and only if $\mathcal{Z}$ is a segment $[z_1, z_2]$ such that z_1 and z_2 are collinear. Suppose that the game is not trivial. To simplify the presentation, we further suppose that $\mathcal{Z}$ is not a segment, but the proof can easily be extended to deal with this case. Consider three non-collinear points in $\mathcal{Z}$. The projection of $\mathcal{Z}$ onto the two-dimensional space spanned by these three points has non-empty interior. Hence, we can find $z^* \neq 0$, $\alpha \in (0, \frac{1}{32}]$, and e a unit vector orthogonal to z^* such that $\tilde{\mathcal{Z}} = \{z \mid z = z^* + (w\alpha \|z^*\|)e, |w| \leq 1\} \subseteq \mathcal{Z}$, which implies that $R_T(\ell, \mathcal{Z}, \mathcal{F}) \geq R_T(\ell, \tilde{\mathcal{Z}}, \mathcal{F})$, and we can focus on developing a $\Omega(\log(T))$ lower bound on regret for the game $(\ell, \tilde{\mathcal{Z}}, \mathcal{F})$. Using the minimax reformulation of Theorem 3.1, we have:

$$\begin{aligned}
&R_T(\ell, \tilde{\mathcal{Z}}, \mathcal{F}) \\
&= \sup_p \mathbb{E}\left[-\sum_{t=1}^{T} \|z^* + (\alpha \|z^*\| \mathbb{E}[W_t | W_1, \cdots, W_{t-1}])e\| + \left\| Tz^* + (\alpha \|z^*\| \sum_{t=1}^{T} W_t)e \right\| \right] \\
&= \sup_p \mathbb{E}[-\sum_{t=1}^{T} \sqrt{\|z^*\|^2 + (\alpha \|z^*\| \mathbb{E}[W_t | W_1, \cdots, W_{t-1}])^2} \\
&\qquad + \sqrt{T^2 \|z^*\|^2 + (\alpha \|z^*\| \sum_{t=1}^{T} W_t)^2}]
\end{aligned}$$

where the supremum is taken over the distribution p of the random variables $(W_1, \cdots, W_T)$ in $[-1, 1]^T$. Rearranging this expression yields:

$$
\begin{aligned}
&R_T(\ell, \tilde{\mathcal{Z}}, \mathcal{F}) \\
&\quad = \|z^*\| \sup_p \ \mathbb{E}\left[T\sqrt{1 + (\alpha \frac{\sum_{t=1}^T W_t}{T})^2} - \sum_{t=1}^T \sqrt{1 + (\alpha \mathbb{E}[W_t | W_1, \cdots, W_{t-1}])^2} \right] \\
&\quad = \|z^*\| \sup_p \{\mathbb{E}[\ T(1 + \sum_{n=1}^{\infty} \binom{\frac{1}{2}}{n} \alpha^{2n} (\frac{\sum_{t=1}^T W_t}{T})^{2n}) \\
&\qquad\qquad - \sum_{t=1}^T (1 + \sum_{n=1}^{\infty} \binom{\frac{1}{2}}{n} \alpha^{2n} \mathbb{E}[W_t | W_1, \cdots, W_{t-1}]^{2n})]\} \\
&\quad = \|z^*\| \sup_p \{\frac{\alpha^2}{2} \mathbb{E}\left[\frac{(\sum_{t=1}^T W_t)^2}{T} - \sum_{t=1}^T \mathbb{E}[W_t | W_1, \cdots, W_{t-1}]^2 \right] \\
&\qquad\qquad + \sum_{n=2}^{\infty} \binom{\frac{1}{2}}{n} \alpha^{2n} \mathbb{E}\left[\frac{(\sum_{t=1}^T W_t)^{2n}}{T^{2n-1}} - \sum_{t=1}^T \mathbb{E}[W_t | W_1, \cdots, W_{t-1}]^{2n} \right]\},
\end{aligned}
$$

where the second equality results from a series expansion, which is valid since

$$(\alpha \mathbb{E}[W_t | W_1, \cdots, W_{t-1}])^2, (\alpha \frac{\sum_{t=1}^T W_t}{T})^2 \leq \alpha^2 < 1,$$

and the third inequality is derived from Fubini, observing that:

$$\sum_{n=1}^{\infty} |\binom{\frac{1}{2}}{n}| \alpha^{2n} \mathbb{E}[(\frac{\sum_{t=1}^T W_t}{T})^{2n}] \leq \sum_{n=1}^{\infty} \alpha^{2n} = \frac{1}{1 - \alpha^2} < \infty$$

and similarly:

$$\sum_{n=1}^{\infty} |\binom{\frac{1}{2}}{n}| \alpha^{2n} \mathbb{E}[\mathbb{E}[W_t | W_1, \cdots, W_{t-1}]^{2n}] \leq \sum_{n=1}^{\infty} \alpha^{2n} = \frac{1}{1 - \alpha^2} < \infty.$$

Interestingly, the first-order term of this series expansion, i.e.

$$\mathbb{E}\left[\frac{(\sum_{t=1}^T W_t)^2}{T} - \sum_{t=1}^T \mathbb{E}[W_t | W_1, \cdots, W_{t-1}]^2 \right],$$

is precisely the expression of the minimax regret for the game:

$$(\ell(z, f) = (z - f)^2, [-1, 1], [-1, 1])$$

which is known to have optimal regret $\Theta(\log(T))$, see Section 7.3 of [2]. This motivates the introduction of the probability distribution p used in [2] to establish the $\Omega(\log(T))$ lower bound. Specifically, we use the conditional distributions:

$$p_t(W_t = w|W_1, \cdots, W_{t-1}) = \begin{cases} \frac{1+c_t W_{1:t-1}}{2} & \text{if } w = 1 \\ \frac{1-c_t W_{1:t-1}}{2} & \text{if } w = -1 \end{cases} \qquad t = 2, \cdots, T$$

where $W_{1:t-1} = \sum_{\tau=1}^{t-1} W_\tau$ and the sequence $(c_t)_{t=1,\cdots,T}$ is recursively defined as:

$$\begin{aligned} c_T &= \frac{1}{T} \\ c_{t-1} &= c_t + c_t^2 \quad t = T, \cdots, 2. \end{aligned}$$

Together with W_1 taken as a Rademacher random variable, these conditional distributions define a joint distribution p as it can be shown that $c_t \in [0, \frac{1}{t}]$. The authors in [2] show that:

$$\mathbb{E}\left[\frac{(\sum_{t=1}^T W_t)^2}{T} - \sum_{t=1}^T \mathbb{E}[W_t|W_1, \cdots, W_{t-1}]^2 \right] = \log(T) + O(\log\log(T)). \tag{B.2}$$

Hence, it remains to control the terms of order $n \geq 2$ in the series expansion. First observe that, by definition:

$$\begin{aligned} \mathbb{E}[W_{1:T}^{2n}] &= \mathbb{E}[\mathbb{E}[W_{1:T}^{2n}|W_1, \cdots, W_{T-1}]] \\ &= \mathbb{E}[\frac{1 + c_t W_{1:T-1}}{2}(W_{1:T-1} + 1)^{2n} + \frac{1 - c_t W_{1:T-1}}{2}(W_{1:T-1} + 1)^{2n}] \\ &= 1 + \sum_{k=1}^{n} (\binom{2n}{2k} + \binom{2n}{2k-1} c_T) \mathbb{E}[(W_{1:T-1})^{2k}], \end{aligned}$$

which implies that:

$$|c_T^{2n-1}\mathbb{E}[W_{1:T}^{2n}] - (c_T^{2n-1} + 2nc_T^{2n})\mathbb{E}[W_{1:T-1}^{2n}]| \leq \binom{2n}{2(n-1)} c_T + 2\sum_{k=0}^{n}\binom{2n}{2k} c_T^2 \\ \leq 2n^2 c_T + 24^n c_T^2, \tag{B.3}$$

since $c_T|W_{1:T-1}| \leq 1$. Additionally, we have, using the recursive definition of the sequence $(c_t)_{t=1,\cdots,T}$:

$$\begin{aligned} c_{T-1}^{2n-1} &= (c_T + c_T^2)^{2n-1} \\ &= \sum_{k=0}^{2n-1}\binom{2n-1}{k} c_T^{2n-1+k}, \end{aligned}$$

which implies:

$$|c_{T-1}^{2n-1} - (c_T^{2n-1} + (2n-1)c_T^{2n})| \leq 2n^2 c_T^{2n+1} + 4^n c_T^{2n+2}. \tag{B.4}$$

Using $\mathbb{E}[W_T|W_1, \cdots, W_{T-1}] = c_T W_{1:T-1}$, we get:

$$\begin{aligned} |\mathbb{E}[c_T^{2n-1}W_{1:T}^{2n} - \sum_{t=1}^{T}\mathbb{E}[W_t|W_1, \cdots, W_{t-1}]^{2n}]| \\ &\leq |\mathbb{E}[(c_T^{2n-1} + 2(n-1)c_T^{2n})W_{1:T-1}^{2n}] - \sum_{t=1}^{T-1}\mathbb{E}[W_t|W_1, \cdots, W_{t-1}]^{2n}]| \\ &+ 2n^2 c_T + 24^n c_T^2 \\ &\leq |\mathbb{E}[c_{T-1}^{2n-1}W_{1:T-1}^{2n} - \sum_{t=1}^{T-1}\mathbb{E}[W_t|W_1, \cdots, W_{t-1}]^{2n}]| \\ &+ (2n^2 c_T^{2n+1} + 4^n c_T^{2n+2})\mathbb{E}[W_{1:T-1}^{2n}] + 2n^2 c_T + 24^n c_T^2 \\ &\leq 4n^2 c_T + 34^n c_T^2 \\ &\leq 4n^2 \frac{1}{T} + 34^n \frac{1}{T^2}, \end{aligned}$$

where the first (resp. second) inequality is obtained by applying (B.3) (resp. (B.4)) and the fifth inequality is derived using $c_T \in [0, \frac{1}{T}]$ and $|W_{1:T-1}| \leq T-1$. By induction on t, we

get:

$$|\mathbb{E}\left[c_T^{2n-1}W_{1:T}^{2n} - \sum_{t=1}^{T}\mathbb{E}[W_t|W_1,\cdots,W_{t-1}]^{2n}\right]| \leq 4n^2\sum_{t=1}^{T}\frac{1}{t} + 34^n\sum_{t=1}^{T}\frac{1}{t^2}$$
$$\leq 4n^2\log(T) + 4^n\frac{\pi^2}{2}.$$

Bringing everything together, we derive:

$$|\sum_{n=2}^{\infty}\binom{\frac{1}{2}}{n}\alpha^{2n}\mathbb{E}\left[\frac{(\sum_{t=1}^{T}W_t)^{2n}}{T^{2n-1}} - \sum_{t=1}^{T}\mathbb{E}[W_t|W_1,\cdots,W_{t-1}]^{2n}\right]|$$
$$\leq 4(\sum_{n=2}^{\infty}\binom{\frac{1}{2}}{n}\alpha^{2n}n^2)\log(T)$$
$$+ (\sum_{n=2}^{\infty}\binom{\frac{1}{2}}{n}(2\alpha)^{2n})\frac{\pi^2}{2}$$
$$\leq 8\alpha^4(\sum_{n=2}^{\infty}n(n-1)(\alpha^2)^{n-2})\log(T)$$
$$+ (\sum_{n=0}^{\infty}(2\alpha)^{2n})\frac{\pi^2}{2}$$
$$\leq 8\frac{\alpha^4}{(1-\alpha^2)^3}\log(T) + \frac{\pi^2}{2(1-2\alpha)}$$
$$\leq 8\frac{\alpha^4}{(1-\alpha^2)^3}\log(T) + \pi^2,$$

since $\alpha \leq \frac{1}{4}$. Using (B.2), we conclude that:

$$R_T(\ell,\tilde{\mathcal{Z}},\mathcal{F}) \geq \frac{\|z^*\|\,\alpha^2}{2}(1 - 16\frac{\alpha^2}{(1-\alpha^2)^3})\log(T) + O(\log\log(T)),$$

which implies that $R_T(\ell,\tilde{\mathcal{Z}},\mathcal{F}) = \Omega(\log(T))$ as $\alpha^2(1-16\frac{\alpha^2}{(1-\alpha^2)^3}) > 0$ for $\alpha \in (0,\frac{1}{32}]$.

B.12 Proof of Theorem 3.6

The proof is along the same lines as for Theorem 3.4. We start with the same inequality:

$$r_T((z_t)_{t=1,\cdots,T},(f_t)_{t=1,\cdots,T}) \leq \sum_{t=1}^{T}z_t^{\intercal}(f_t - f_{t+1}),$$

and use sensitivity analysis to control this last quantity. Specifically, we show that the mapping $\phi : z \to \operatorname{argmin}_{f \in \mathcal{F}} z^{\mathsf{T}} f$ is well defined and $\frac{1}{q-1}$-Hölder continuous on $\mathcal{Z}$, i.e. there exists $c > 0$ such that:

$$\|\phi(z_1) - \phi(z_2)\|_2 \leq c \|z_1 - z_2\|_2^{\frac{1}{q-1}} \quad \forall (z_1, z_2) \in \mathcal{Z}^2.$$

Using this property, we get:

$$\begin{aligned}
r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) &\leq \sum_{t=1}^{T} \|z_t\|_2 \|f_t - f_{t+1}\|_2 \\
&= O(\sum_{t=1}^{T} \left\| \frac{1}{t-1} \sum_{\tau=1}^{t-1} z_\tau - \frac{1}{t} \sum_{\tau=1}^{t} z_\tau \right\|_2^{\frac{1}{q-1}}) \\
&= O(\sum_{t=1}^{T} \left\| \frac{1}{t(t-1)} \sum_{\tau=1}^{t-1} z_\tau - \frac{1}{t} z_t \right\|_2^{\frac{1}{q-1}}) \\
&= O(\sum_{t=1}^{T} (\frac{1}{t(t-1)} \left\| \sum_{\tau=1}^{t-1} z_\tau \right\|_2 + \frac{1}{t} \|z_t\|_2)^{\frac{1}{q-1}}) \\
&= O(\sum_{t=1}^{T} \frac{1}{t^{\frac{1}{q-1}}}),
\end{aligned}$$

from which we derive that

$$r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) = O(\log(T))$$

if $q = 2$ and

$$r_T((z_t)_{t=1,\cdots,T}, (f_t)_{t=1,\cdots,T}) = O(T^{\frac{q-2}{q-1}})$$

if $q \in (2, 3]$. We move on to show that ϕ is $\frac{1}{q-1}$-Hölder continuous. Just like in Theorem 3.4, we can find $A > 0$ such that $\|z\|_2 \geq A$ for all $z \in \mathcal{Z}$. Take $(z_1, z_2) \in \mathcal{Z}^2$ and $(f_1, f_2) \in \operatorname{argmin}_{f \in \mathcal{F}} z_1^{\mathsf{T}} f \times \operatorname{argmin}_{f \in \mathcal{F}} z_2^{\mathsf{T}} f$. Since we are optimizing a linear function, we may assume, without loss of generality, that $C = 1$ and f_1 and f_2 lie on the boundary of $\mathcal{F}$, i.e. $\|f_1\|_{\mathcal{F}} = \|f_2\|_{\mathcal{F}} = 1$. By definition, we have:

$$\left\| \frac{f_1 + f_2}{2} \right\|_{\mathcal{F}} \leq 1 - \delta_{\mathcal{F}}(\|f_1 - f_2\|_{\mathcal{F}}).$$

As a consequence, we have:

$$\frac{f_1+f_2}{2} - \delta_{\mathcal{F}}(\|f_1-f_2\|_{\mathcal{F}})\frac{z_2}{\|z_2\|_{\mathcal{F}}} \in \mathcal{F}.$$

We get:

$$z_2^\intercal(\frac{f_1+f_2}{2} - \delta_{\mathcal{F}}(\|f_1-f_2\|_{\mathcal{F}})\frac{z_2}{\|z_2\|_{\mathcal{F}}}) \geq \inf_{f\in\mathcal{F}} z_2^\intercal f = z_2^\intercal f_2.$$

Rearranging this last inequality yields:

$$z_2^\intercal \frac{f_1-f_2}{2} \geq \frac{\|z_2\|_2^2}{\|z_2\|_{\mathcal{F}}}\delta_{\mathcal{F}}(\|f_1-f_2\|_{\mathcal{F}}),$$

which implies that:

$$z_2^\intercal \frac{f_1-f_2}{2} \geq K\|f_1-f_2\|_2^q,$$

for some $K>0$ independent of z_1 and z_2 since $\mathcal{Z}$ is compact, $\|z_2\|_2 \geq A > 0$, $\|\ \|_{\mathcal{F}}$ is q-uniformly convex, and by the equivalence of norms in finite dimensions. By optimality of f_1, we also have $z_1^\intercal \frac{f_2-f_1}{2} \geq 0$. Summing up the last two inequalities, we get:

$$(z_2-z_1)^\intercal \frac{f_1-f_2}{2} \geq K\|f_1-f_2\|_2^q,$$

and (by Cauchy-Schwartz):

$$\|z_2-z_1\|_2 \geq 2K\|f_1-f_2\|_2^{q-1},$$

which concludes the proof.

B.13 Proof of Lemma 3.7

Observe that the game $(\ell(z,f)=z^\intercal f, \mathcal{Z}, \mathcal{F})$ is not trivial because $\operatorname{argmin}_{f\in\mathcal{F}} f^\intercal z_3 = \{f^*\}$ while $\operatorname{argmin}_{f\in\mathcal{F}} f^\intercal z_4 = \{f^{**}\}$. For any zero-mean i.i.d. opponent $Z_1,\cdots,Z_T$, we must have $Z_t \in [z_1,z_2]$. Since $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} f^\intercal z$ for $z\in[z_1,z_2]$, we get, irrespective of the

player's strategy:

$$\begin{aligned}\mathbb{E}[r_T((Z_t)_{t=1,\cdots,T},(f_t)_{t=1,\cdots,T})] &= -\mathbb{E}[\inf_{f\in\mathcal{F}} f^\intercal \sum_{t=1}^{T} Z_t] \\ &= -\mathbb{E}[(f^*)^\intercal \sum_{t=1}^{T} Z_t] = 0.\end{aligned}$$

Appendix C

Appendix For Chapter 4

C.1 Extensions

C.1.1 Improving the Multiplicative Factors in the Regret Bounds

C.1.1.a A Single Limited Resource whose Consumption is Deterministic

If the amounts of resource consumed are deterministic, we can substitute the notation μ_k^c for c_k. Moreover, we can take $\lambda = 1$ and, going through the analysis of Lemma 4.5, we can slightly refine the regret bound. Specifically, we have $\mathbb{E}[n_{k,\tau^*}] \leq \frac{16}{(c_k)^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + \frac{\pi^2}{3}$, for any arm k such that $\Delta_k > 0$. Moreover, $\tau^* \leq B/\epsilon + 1$ in this setting since:

$$B \geq \sum_{t=1}^{\tau^*-1} c_{a_t,t} \geq (\tau^* - 1) \cdot \epsilon,$$

by definition of τ^*. As a result, the regret bound derived in Theorem 4.1 turns into:

$$R_B \leq 16(\sum_{k \mid \Delta_k > 0} \frac{1}{c_k \cdot \Delta_k}) \cdot \ln(\frac{B}{\epsilon} + 1) + O(1),$$

which is identical (up to universal constant factors) to the bound obtained in [90]. Note that this bound is scale-free.

C.1.1.b Arbitrarily Many Limited Resources whose Consumptions are Deterministic

We propose another load balancing algorithm that couples bases together. This is key to get a better dependence on K because, otherwise, we have to study each basis independently from the other ones.

Algorithm: Load balancing algorithm $\mathcal{A}_x$ for a feasible basis $x \in \mathcal{B}$.

If x is selected at time t, pull any arm $a_t \in \operatorname{argmax}_{k \in \mathcal{K}_x} \frac{\xi_k^x}{n_{k,t}}$.

Observe that this load balancing algorithm is computationally efficient with a $O(K)$ runtime (once we have computed an optimal basic feasible solution to (4.8)) and requires $O(K)$ memory space. The shortcoming of this approach is that, if there are multiple optimal bases to (4.3), the optimal load balance for each optimal basis will not be preserved since we take into account the number of times we have pulled each arm when selecting any basis (for which we strive to enforce different ratios). Hence, the following assumption will be required for the analysis.

Assumption C.1. *There is a unique optimal basis to* (4.3).

Regret Analysis. All the proofs are deferred to Section C.7. We start by bounding, for each arm k, the number of times this arm can be pulled when selecting any of the suboptimal bases. This is in stark contrast with the analysis carried out in Section 4.5 where we bound the number of times each suboptimal basis has been selected.

Lemma C.1. *For any arm $k \in \{1, \cdots, K\}$, we have:*

$$\mathbb{E}[\sum_{x \in \mathcal{B} \mid k \in \mathcal{K}_x,\, x \neq x^*} n_{k,\tau^*}^x] \leq 16 \frac{\rho \cdot (\sum_{i=1}^C b(i))^2}{\epsilon^2} \cdot \frac{\mathbb{E}[\ln(\tau^*)]}{(\Delta_k)^2} + K \cdot \frac{\pi^2}{3},$$

where $\Delta_k = \min_{x \in \mathcal{B} \mid k \in \mathcal{K}_x,\, x \neq x^*} \Delta_x$.

In contrast to Section 4.5, we can only guarantee that the ratios $(n_{k,t}^x / n_{l,t}^x)_{k,l \in \mathcal{K}_x}$ remain close to the optimal ones $(\xi_k^x / \xi_l^x)_{k,l \in \mathcal{K}_x}$ at all times for the optimal basis $x = x^*$. This

will not allow us to derive distribution-free regret bounds for this particular class of load balancing algorithms.

Lemma C.2. *At any time t and for any arm $k \in \mathcal{K}_{x^*}$, we have:*

$$n_{k,t} \geq n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} - \rho \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1) \tag{C.1}$$

and

$$n_{k,t} \leq n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1. \tag{C.2}$$

Bringing everything together, we are now ready to establish regret bounds.

Theorem C.1. *We have:*

$$R_{B(1),\cdots,B(C)} \leq 32 \frac{\rho^3 \cdot (\sum_{i=1}^C b(i))^3}{\epsilon^3 \cdot b} \cdot (\sum_{k=1}^K \frac{1}{(\Delta_k)^2}) \cdot \ln(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1) + O(1),$$

where the O notation hides universal constant factors.

We derive a distribution-dependent regret bound of order $O(\rho^3 \cdot K \cdot \frac{\ln(B)}{\Delta^2})$ where $\Delta = \min_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x$ but no non-trivial distribution-free regret bound.

C.1.1.c Arbitrarily Many Limited Resources

A straightforward extension of the load balancing algorithm developed in the case of deterministic resource consumption in Section C.1.1.b guarantees that the total number of times any suboptimal basis is pulled is of order $O(K \cdot \ln(T))$. However, in contrast to Section C.1.1.b, this is not enough to get logarithmic regret bounds as $\xi_{k,t}^x$ fluctuates around the optimal load balance $\xi_{k,t}^x$ with a magnitude of order at least $\sim 1/\sqrt{t}$, and, as a result, the ratios $(\mathbb{E}[n_{k,T}^x]/\mathbb{E}[n_{l,T}^x])_{k,l \in \mathcal{K}_x}$ might be very different from the optimal ones $(\xi_k^x/\xi_l^x)_{k,l \in \mathcal{K}_x}$.

Algorithm: Load balancing algorithm $\mathcal{A}_x$ for a feasible basis $x \in \mathcal{B}$.

If x is selected at time t, pull any arm $a_t \in \underset{k \in \mathcal{K}_x}{\operatorname{argmax}} \frac{\xi_{k,t}^x}{n_{k,t}}$.

Lemma C.3. *For any arm $k \in \{1, \cdots, K\}$, we have:*

$$\mathbb{E}[\sum_{x \in \mathcal{B} \mid k \in \mathcal{K}_x,\, \Delta_x > 0} n_{k,T}^x] \leq 16C \cdot \lambda^2 \cdot \frac{\ln(T)}{(\Delta_k)^2} + 2^{12} \frac{K \cdot (C+3)!^2}{\epsilon^6},$$

where $\Delta_k = \min_{x \in \mathcal{B} \mid k \in \mathcal{K}_x,\, \Delta_x > 0} \Delta_x$.

C.1.2 Relaxing Assumption 4.1

The regret bounds obtained in Sections 4.5, 4.6, and 4.7 can be extended when the ratios converge as opposed to being fixed, as precisely stated below, but this requires slightly more work.

Assumption C.2. *For any resource $i \in \{1, \cdots, C\}$, the ratio $B(i)/B(C)$ converges to a finite value $b(i) \in (0, 1]$. Moreover, $b = \min_{i=1,\cdots,C} b(i)$ is a positive quantity.*

To state the results, we need to redefine some notations and to work with the linear program:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K} \in \mathbb{R}_+^K} \quad & \sum_{k=1}^K \mu_k^r \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^K \mu_k^c(i) \cdot \xi_k \leq \frac{B(i)}{B(C)}, \quad i = 1, \cdots, C \end{aligned} \tag{C.3}$$

We redefine $\mathcal{B}$ as the set of bases that are feasible to (C.3) and, for $x \in \mathcal{B}$, Δ_x is redefined as the optimality gap of x with respect to (C.3). We also redefine $\mathcal{O} = \{x \in \mathcal{B} \mid \Delta_x = 0\}$ as the set of optimal bases to (C.3). Moreover, we define $\mathcal{B}_\infty$ (resp. $\mathcal{O}_\infty$) as the set of feasible (resp. optimal) bases to (4.3) and, for $x \in \mathcal{B}_\infty$, Δ_x^∞ is the optimality gap of x with respect to (4.3). Our algorithm remains the same provided that we substitute $b(i)$ with $B(i)/B(C)$ for any resource $i \in \{1, \cdots, C\}$. Specifically, Step-Simplex consists in solving:

$$\begin{aligned} \sup_{(\xi_k)_{k=1,\cdots,K} \in \mathbb{R}_+^K} \quad & \sum_{k=1}^K (\bar{r}_{k,t} + \lambda \cdot \epsilon_{k,t}) \cdot \xi_k \\ \text{subject to} \quad & \sum_{k=1}^K (\bar{c}_{k,t}(i) - \eta_i \cdot \epsilon_{k,t}) \cdot \xi_k \leq \frac{B(i)}{B(C)}, \quad i = 1, \cdots, C \end{aligned} \tag{C.4}$$

and Step-Load-Balance is identical up to the substitution of $b(i)$ with $B(i)/B(C)$.

Regret Analysis. As it turns out, the logarithmic regret bounds established in Theorems 4.3, 4.5, and 4.8 do not always extend when Assumption 4.1 is relaxed even though these bounds appear to be very similar to the one derived in Theorem 4.1 when there is a single limited resource. The fundamental difference is that the set of optimal bases may not converge while it is always invariant in the case of a single limited resource. Typically, the ratios $(B(i)/B(C))_{i=1,\cdots,C}$ may oscillate around $(b(i))_{i=1,\cdots,C}$ in such a way that there exist two optimal bases for (4.3) while there is a unique optimal basis for this same optimization problem whenever the right-hand side of the inequality constraints is slightly perturbed around this limit. This alternately causes one of these two bases to be slightly suboptimal, a situation difficult to identify and to cope with for the decision maker. Nevertheless, this difficulty does not arise in several situations of interest which generalize Assumption 4.1, as precisely stated below. The proofs are deferred to Section C.7.

Arbitrarily many limited resources whose consumptions are deterministic.

Theorem C.2. *Suppose that Assumption C.2 holds. If there exists a unique optimal basis to* (4.3) *or if* $B(i)/B(C) - b(i) = O(\ln(B(C))/B(C))$ *for all resources* $i \in \{1, \cdots, C-1\}$ *then, we have:*

$$R_{B(1),\cdots,B(C)} = O\Big(\frac{\rho \cdot \sum_{i=1}^{C} b(i)}{\epsilon \cdot b} \cdot \Big(\sum_{x \in \mathcal{B}_\infty \mid \Delta_x^\infty > 0} \frac{1}{\Delta_x^\infty}\Big) \cdot \ln\Big(\frac{\sum_{i=1}^{C} b(i) \cdot B(C)}{\epsilon} + 1\Big)\Big) + O\Big(|\mathcal{O}_\infty| \cdot \frac{\ln(B(C))}{\epsilon \cdot b}\Big),$$

where the O notation hides universal constant factors.

A time horizon and another limited resource.

Theorem C.3. *Suppose that Assumptions 4.4, 4.5, and 4.6 hold and that the ratio B/T converges to $b \in (0, 1]$. If there exists a unique optimal basis to* (4.3) *or if $B/T - b =$*

$O(\ln(T)/T)$*, then, we have:*

$$R_{B,T} = O(\frac{\lambda^2}{\epsilon^3} \cdot (\sum_{x \in \mathcal{B}_\infty \mid \Delta_x^\infty > 0} \frac{1}{\Delta_x^\infty}) \cdot \ln(T) + \frac{K^2 \cdot \sigma}{\epsilon^3} \cdot \ln(T)),$$

where the O notation hides universal constant factors.

Arbitrarily many limited resources with a time horizon.

Theorem C.4. *Suppose that Assumptions 4.7, 4.8, and C.2 hold. If there exists a unique optimal basis to* (4.3) *or if $B(i)/T - b(i) = O(\ln(T)/T)$ for all resources $i \in \{1, \cdots, C-1\}$, then, we have:*

$$\begin{aligned} R_{B(1),\cdots,B(C-1),T} = O(&\frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot (\sum_{x \in \mathcal{B}_\infty \mid \Delta_x^\infty > 0} \frac{1}{\Delta_x^\infty}) \cdot \ln(T)) \\ &+ \frac{\sigma \cdot |\mathcal{B}_\infty| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)), \end{aligned}$$

where the O notation hides universal constant factors.

C.2 Proofs for Section 4.3

C.2.1 Proof of Lemma 4.2

The proof can be found in [16]. For the sake of completeness, we reproduce it here. The optimization problem (4.3) is a linear program whose dual reads:

$$\begin{aligned} \inf_{(\zeta_i)_{i=1,\cdots,C}} \quad & \sum_{i=1}^{C} b(i) \cdot \zeta_i \\ \text{subject to} \quad & \sum_{i=1}^{C} \mu_k^c(i) \cdot \zeta_i \geq \mu_k^r, \quad k = 1, \cdots, K \\ & \zeta_i \geq 0, \quad i = 1, \cdots, C. \end{aligned} \tag{C.5}$$

Observe that (4.3) is feasible therefore (4.3) and (C.5) have the same optimal value. Note that (4.3) is bounded under Assumption 4.2 as $\xi_k \in [0, b(i)/\mu_k^c(i)]$ for any feasible point and any resource $i \in \{1, \cdots, C\}$ such that $\mu_k^c(i) > 0$. Hence, (C.5) has an optimal basic feasible solution $(\zeta_1^*, \cdots, \zeta_C^*)$. Consider any non-anticipating algorithm. Let Z_t be the sum of the total payoff accumulated in rounds 1 to t plus the "cost" of the remaining resources, i.e. $Z_t = \sum_{\tau=1}^{t} r_{a_\tau,\tau} + \sum_{i=1}^{C} \zeta_i^* \cdot (B(i) - \sum_{\tau=1}^{t} c_{a_\tau,\tau}(i))$. Observe that $(Z_t)_t$ is a supermartingale with respect to the filtration $(\mathcal{F}_t)_t$ as $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mu_{a_t}^r - \sum_{i=1}^{C} \zeta_i^* \cdot \mu_{a_t}^c(i) \mid \mathcal{F}_{t-1}] + Z_{t-1} \leq Z_{t-1}$ since $(\zeta_1^*, \cdots, \zeta_C^*)$ is feasible for (C.5). Moreover, note that $(Z_t)_t$ has bounded increments since $|Z_t - Z_{t-1}| = |r_{a_t,t} - \sum_{i=1}^{C} \zeta_i^* \cdot c_{a_t,t}(i)| \leq 1 + \sum_{i=1}^{C} \zeta_i^* < \infty$. We also have $\mathbb{E}[\tau^*] < \infty$ as:

$$
\begin{aligned}
\mathbb{E}[\tau^*] &= \sum_{t=1}^{\infty} \mathbb{P}[\tau^* \geq t] \\
&\leq \sum_{t=1}^{\infty} \mathbb{P}[\sum_{\tau=1}^{t-1} c_{a_\tau,\tau}(i) \leq B(i), i = 1, \cdots, C] \\
&\leq 1 + \sum_{t=1}^{\infty} \mathbb{P}[\sum_{\tau=1}^{t} \sum_{i=1}^{C} c_{a_\tau,\tau}(i) \leq t \cdot \epsilon - (t \cdot \epsilon - \sum_{i=1}^{C} B(i))] \\
&\leq (\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 2) + \sum_{t \geq \frac{\sum_{i=1}^{C} B(i)}{\epsilon}}^{\infty} \exp(-\frac{2(t \cdot \epsilon - \sum_{i=1}^{C} B(i))^2}{t}) \\
&< \infty,
\end{aligned}
$$

where the third inequality results from an application of Lemma 4.1 and

$$
\epsilon = \min_{\substack{k=1,\cdots,K \\ i=1,\cdots,C \\ \text{with } \mu_k^c(i)>0}} \mu_k^c(i).
$$

By Doob's optional stopping theorem, $\mathbb{E}[Z_{\tau^*}] \leq \mathbb{E}[Z_0] = \sum_{i=1}^{C} \zeta_i^* \cdot B(i)$. Observe that:

$$
\begin{aligned}
\mathbb{E}[Z_{\tau^*}] &= \mathbb{E}[r_{a_{\tau^*},\tau^*} - \sum_{i=1}^{C} \zeta_i^* \cdot c_{a_{\tau^*},\tau^*}(i) + Z_{\tau^*-1}] \\
&\geq \mathbb{E}[-\sum_{i=1}^{C} \zeta_i^* + \sum_{t=1}^{\tau^*-1} r_{a_t,t}].
\end{aligned}
$$

Using Assumption 4.2 and since $(\zeta_i^*)_{i=1,\cdots,C}$ is a basic feasible solution, for every $i \in \{1,\cdots,C\}$ such that $\zeta_i^* > 0$ there must exist $k \in \{1,\cdots,K\}$ such that $\zeta_i^* \leq \mu_k^r/\mu_k^c(i)$ with $\mu_k^c(i) > 0$. We get:

$$\mathbb{E}[Z_{\tau^*}] \geq \mathbb{E}[\sum_{t=1}^{\tau^*-1} r_{a_t,t}] - \max_{\substack{k=1,\cdots,K \\ i=1,\cdots,C \\ \text{with } \mu_k^c(i)>0}} \frac{\mu_k^r}{\mu_k^c(i)}$$

and finally:

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{\tau^*-1} r_{a_t,t}] &\leq \sum_{i=1}^{C} \zeta_i^* \cdot B(i) + \max_{\substack{k=1,\cdots,K \\ i=1,\cdots,C \\ \text{with } \mu_k^c(i)>0}} \frac{\mu_k^r}{\mu_k^c(i)} \\ &= B \cdot \sum_{i=1}^{C} \zeta_i^* \cdot b(i) + \max_{\substack{k=1,\cdots,K \\ i=1,\cdots,C \\ \text{with } \mu_k^c(i)>0}} \frac{\mu_k^r}{\mu_k^c(i)}.\end{aligned}$$

By strong duality, $\sum_{i=1}^{C} \zeta_i^* \cdot b(i)$ is also the optimal value of (4.3).

C.3 Proofs for Section 4.4

C.3.1 Proof of Lemma 4.4

By definition of τ^*, we have $\sum_{t=1}^{\tau^*-1} c_{a_t,t} \leq B$. Taking expectations on both sides yields:

$$\begin{aligned}B &\geq \mathbb{E}[\sum_{t=1}^{\tau^*-1} c_{a_t,t}] \\ &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot c_{a_t,t}] - 1 \\ &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]] - 1 \\ &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mu_{a_t}^c] - 1 \\ &\geq \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \epsilon] - 1 = \mathbb{E}[\tau^*] \cdot \epsilon - 1,\end{aligned}$$

where we use the fact that $c_{k,t} \leq 1$ for all arms k to derive the second inequality, the fact that τ^* is a stopping time for the second equality, the fact that a_t is deterministically determined by the past, i.e. $a_t \in \mathcal{F}_{t-1}$, for the third equality and Assumption 4.2 for the third inequality. We conclude that $\mathbb{E}[\tau^*] \leq \frac{B+1}{\epsilon}$.

C.3.2 Proof of Lemma 4.5

We break down the analysis in a series of facts. Consider any arm k such that $\Delta_k > 0$. We use the shorthand notation $\beta_k = 2^5(\frac{\lambda}{\mu_k^c})^2 \cdot (\frac{1}{\Delta_k})^2$.

Fact C.1.

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}]. \tag{C.6}$$

Proof. Define the random variable $T_k = \beta_k \cdot \ln(\tau^*)$. We have:

$$\begin{aligned}
\mathbb{E}[n_{k,\tau^*}] &= \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} < T_k}] + \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}] \\
&\leq \beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}].
\end{aligned}$$

Define T_k^* as the first time t such that $n_{k,t} \geq T_k$ and $T_k^* = \infty$ if no such t exists. We have:

$$\begin{aligned}
\mathbb{E}[n_{k,\tau^*} \cdot I_{n_{k,\tau^*} \geq T_k}] &= \mathbb{E}[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}] \\
&= \mathbb{E}[\sum_{t=1}^{T_k^*-1} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}] + \mathbb{E}[\sum_{t=T_k^*}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,\tau^*} \geq T_k}] \\
&\leq \mathbb{E}[n_{k,T_k^*-1} \cdot I_{n_{k,\tau^*} \geq T_k}] + \mathbb{E}[\sum_{t=T_k^*}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq T_k}] \\
&\leq \beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}],
\end{aligned}$$

since, by definition of T_k^*, $n_{k,T_k^*-1} \leq T_k$ if T_k^* if finite, which is always true if $n_{k,\tau^*} \geq T_k$ (the sequence $(n_{k,t})_t$ is non-decreasing and τ^* is finite almost surely as a byproduct of Lemma 4.4). Conversely, $n_{k,t} \geq T_k \geq \beta_k \ln(t)$ for $t \in \{T_k^*, \cdots, \tau^*\}$. Wrapping up, we

obtain:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}].$$

□

Fact C.1 enables us to assume that arm k has been pulled at least $\beta_k \ln(t)$ times out of the last t time periods. The remainder of this proof is dedicated to show that the second term of the right-hand side of (C.6) can be bounded by a constant. Let us first rewrite this term:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{a_t=k} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}] &\leq \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k,t}+E_{k,t} \geq \mathrm{obj}_{k^*,t}+E_{k^*,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k,t} \geq \mathrm{obj}_k + E_{k,t}}] \qquad \text{(C.7)} \\
&+ \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k^*,t} \leq \mathrm{obj}_{k^*} - E_{k^*,t}}] \qquad \text{(C.8)} \\
&+ \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k^*} < \mathrm{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}]. \qquad \text{(C.9)}
\end{aligned}$$

To derive this last inequality, simply observe that if $\mathrm{obj}_{k,t} < \mathrm{obj}_k + E_{k,t}$ and $\mathrm{obj}_{k^*,t} > \mathrm{obj}_{k^*} - E_{k^*,t}$ while $\mathrm{obj}_{k,t} + E_{k,t} \geq \mathrm{obj}_{k^*,t} + E_{k^*,t}$, it must be that $\mathrm{obj}_{k^*} < \mathrm{obj}_k + 2E_{k,t}$. Let us study (C.7), (C.8), and (C.9) separately.

Fact C.2.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k^*} < \mathrm{obj}_k + 2E_{k,t}} \cdot I_{n_{k,t} \geq \beta_k \ln(t)}] \leq \frac{2\pi^2}{3\epsilon^2}.$$

Proof. Observe that when both $n_{k,t} \geq \beta_k \ln(t)$ and $\mathrm{obj}_{k^*} < \mathrm{obj}_k + 2E_{k,t}$, we have:

$$\begin{aligned}
\frac{\Delta_k}{2} &< E_{k,t} \\
&\leq \frac{\lambda}{\bar{c}_{k,t}} \cdot \sqrt{\frac{2}{\beta_k}} \\
&\leq \frac{\mu_k^c}{2\bar{c}_{k,t}} \cdot \frac{\Delta_k}{2},
\end{aligned}$$

by definition of β_k. This implies that $\bar{c}_{k,t} \leq \mu_k^c/2$. Thus:

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*}<\text{obj}_k+2E_{k,t}} \cdot I_{n_{k,t}\geq\beta_k \ln(t)}] \leq \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\bar{c}_{k,t}<\mu_k^c/2} \cdot I_{n_{k,t}\geq\beta_k \ln(t)}]$$

We upper bound this last term using the concentration inequalities of Lemma 4.1 observing that:

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{\infty} I_{\bar{c}_{k,t}<\mu_k^c/2} \cdot I_{n_{k,t}\geq\beta_k \ln(t)}] &= \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{k,t} < \frac{\mu_k^c}{2} \ ; \ n_{k,t} \geq \beta_k \ln(t)] \\ &\leq \sum_{t=1}^{\infty} \sum_{s=\beta_k \ln(t)}^{t} \mathbb{P}[\bar{c}_{k,t} < \mu_k^c - \frac{\mu_k^c}{2} \ ; \ n_{k,t} = s].\end{aligned}$$

Denote by $t_1, \cdots, t_s$ the first s random times at which arm k is pulled (these random variables are finite almost surely). We have:

$$\mathbb{P}[\bar{c}_{k,t} < \mu_k^c - \frac{\mu_k^c}{2} \ ; \ n_{k,t} = s] \leq \mathbb{P}[\sum_{l=1}^{s} c_{k,t_l} < s \cdot \mu_k^c - s \cdot \frac{\mu_k^c}{2}].$$

Observe that, for any $l \leq s$:

$$\begin{aligned}\mathbb{E}[c_{k,t_l} \mid c_{k,t_1}, \cdots, c_{k,t_{l-1}}] &= \mathbb{E}[\sum_{\tau=1}^{\infty} I_{t_l=\tau} \cdot \mathbb{E}[c_{k,\tau} \mid \mathcal{F}_{\tau-1}] \mid c_{k,t_1}, \cdots, c_{k,t_{l-1}}] \\ &= \mathbb{E}[\sum_{\tau=1}^{\infty} I_{t_l=\tau} \cdot \mu_k^c \mid c_{k,t_1}, \cdots, c_{k,t_{l-1}}] \\ &= \mu_k^c,\end{aligned}$$

since the algorithm is not randomized ($\{t_l = \tau\} \in \mathcal{F}_{\tau-1}$) and using the tower property.

Hence, we can apply Lemma 4.1 to get:

$$\begin{aligned}\sum_{t=1}^{\infty}\mathbb{P}[\bar{c}_{k,t}<\frac{\mu_k^c}{2}\ ;\ n_{k,t}\geq\beta_k\ln(t)] &\leq \sum_{t=1}^{\infty}\sum_{s=\beta_k\ln(t)}^{\infty}\exp(-s\cdot\frac{(\mu_k^c)^2}{2})\\ &\leq \sum_{t=1}^{\infty}\frac{\exp(-\frac{(\mu_k^c)^2}{2}\cdot\beta_k\ln(t))}{1-\exp(-\frac{(\mu_k^c)^2}{2})}\\ &\leq \frac{1}{1-\exp(-\frac{(\mu_k^c)^2}{2})}\sum_{t=1}^{\infty}\frac{1}{t^2}\\ &\leq \frac{2\pi^2}{3(\mu_k^c)^2}\\ &\leq \frac{2\pi^2}{3\epsilon^2},\end{aligned}$$

where we use the fact that $\beta_k \geq 2^5(\frac{1+\kappa}{\mu_k^c})^2\cdot(\frac{\mu_{k^*}^c}{\mu_{k^*}^r})^2 \geq 2^5(\frac{1+\frac{1}{\kappa}}{\mu_k^c})^2 \geq \frac{4}{(\mu_k^c)^2}$ for the third inequality (using Assumption 4.3), the fact that $\exp(-x)\leq 1-\frac{x}{2}$ for $x\in[0,1]$ for the fourth inequality, and Assumption 4.2 for the last inequality. □

Let us now elaborate on (C.7).

Fact C.3.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k,t}\geq\mathrm{obj}_k+E_{k,t}}]\leq\frac{\pi^2}{3}.$$

Proof. Note that if $\bar{r}_{k,t}/\bar{c}_{k,t}=\mathrm{obj}_{k,t}\geq\mathrm{obj}_k+E_{k,t}=\mu_k^r/\mu_k^c+E_{k,t}$, then either $\bar{r}_{k,t}\geq\mu_k^r+\epsilon_{k,t}$ or $\bar{c}_{k,t}\leq\mu_k^c-\epsilon_{k,t}$, otherwise we would have:

$$\begin{aligned}\frac{\bar{r}_{k,t}}{\bar{c}_{k,t}}-\frac{\mu_k^r}{\mu_k^c} &= \frac{(\bar{r}_{k,t}-\mu_k^r)\mu_k^c+(\mu_k^c-\bar{c}_{k,t})\mu_k^r}{\bar{c}_{k,t}\cdot\mu_k^c}\\ &< \frac{\epsilon_{k,t}}{\bar{c}_{k,t}}+\frac{\epsilon_{k,t}}{\bar{c}_{k,t}}\cdot\frac{\mu_k^r}{\mu_k^c}\\ &\leq (1+\kappa)\cdot\frac{\epsilon_{k,t}}{\bar{c}_{k,t}}\\ &= E_{k,t},\end{aligned}$$

a contradiction. Therefore:

$$
\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k,t} \geq \mathrm{obj}_k + E_{k,t}}] &\leq \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] + \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \sqrt{\frac{2\ln(t)}{s}} \ ; \ n_{k,t} = s] \\
&+ \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \sqrt{\frac{2\ln(t)}{s}} \ ; \ n_{k,t} = s] \\
&= \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\sum_{l=1}^{s} r_{k,t_l} \geq s \cdot \mu_k^r + \sqrt{s \cdot 2\ln(t)} \ ; \ n_{k,t} = s] \\
&+ \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\sum_{l=1}^{s} c_{k,t_l} \leq s \cdot \mu_k^c - \sqrt{s \cdot 2\ln(t)} \ ; \ n_{k,t} = s] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^{t} 2\exp(-4\ln(t)) \\
&= \frac{\pi^2}{3},
\end{aligned}
$$

where the random variables $(t_l)_l$ are defined similarly as in the proof of Fact C.2 and the last inequality results from an application of Lemma 4.1. □

What remains to be done is to bound (C.8).

Fact C.4.

$$
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{k^*,t} \leq \mathrm{obj}_{k^*} - E_{k^*,t}}] \leq \frac{\pi^2}{3}.
$$

Proof. We proceed along the same lines as in the proof of Fact C.3. As a matter of fact, the situation is perfectly symmetric because, in the course of proving Fact C.3, we do not rely on the fact that we have pulled arm k more than $\beta_k \ln(t)$ times at any time t. If $\bar{r}_{k^*,t}/\bar{c}_{k^*,t} = \mathrm{obj}_{k^*,t} \leq \mathrm{obj}_{k^*} - E_{k^*,t} = \mu_{k^*}^r/\mu_{k^*}^c - E_{k^*,t}$, then we have either $\bar{r}_{k^*,t} \leq \mu_{k^*}^r - \epsilon_{k^*,t}$ or $\bar{c}_{k^*,t} \geq \mu_{k^*}^c + \epsilon_{k^*,t}$, otherwise we would have:

$$
\begin{aligned}
\frac{\bar{r}_{k^*,t}}{\bar{c}_{k^*,t}} - \frac{\mu_{k^*}^r}{\mu_{k^*}^c} &= \frac{(\bar{r}_{k^*,t} - \mu_{k^*}^r)\mu_{k^*}^c + (\mu_{k^*}^c - \bar{c}_{k^*,t})\mu_{k^*}^r}{\bar{c}_{k^*,t} \cdot \mu_{k^*}^c} \\
&> -\frac{\epsilon_{k^*,t}}{\bar{c}_{k^*,t}} - \frac{\epsilon_{k^*,t}}{\bar{c}_{k^*,t}} \cdot \frac{\mu_{k^*}^r}{\mu_{k^*}^c} \\
&\geq -(1+\kappa) \cdot \frac{\epsilon_{k^*,t}}{\bar{c}_{k^*,t}} = -E_{k^*,t},
\end{aligned}
$$

a contradiction. Therefore:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{k^*,t} \leq \text{obj}_{k^*} - E_{k^*,t}}] &\leq \mathbb{E}[\sum_{t=1}^{\infty} I_{\bar{r}_{k^*,t} \leq \mu^r_{k^*} - \epsilon_{k,t}} + I_{\bar{c}_{k^*,t} \geq \mu^c_{k^*} + \epsilon_{k,t}}] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\bar{r}_{k^*,t} \leq \mu^r_{k^*} - \sqrt{\frac{2\ln(t)}{s}} \ ; \ n_{k^*,t} = s] \\
&+ \sum_{t=1}^{\infty} \sum_{s=1}^{t} \mathbb{P}[\bar{c}_{k^*,t} \geq \mu^c_{k^*} + \sqrt{\frac{2\ln(t)}{s}} \ ; \ n_{k^*,t} = s] \\
&\leq \sum_{t=1}^{\infty} \sum_{s=1}^{t} \frac{2}{t^4} \\
&= \frac{\pi^2}{3},
\end{aligned}$$

where the third inequality is obtained using Lemma 4.1 as in Fact C.3. □

We conclude:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \frac{4\pi^2}{3\epsilon^2}.$$

C.3.3 Proof of Theorem 4.1

First observe that:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] &= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mathbb{E}[r_{a_t,t} \mid \mathcal{F}_{t-1}]] \\
&= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mu^r_{a_t}] \\
&= \sum_{k=1}^{K} \mu^r_k \cdot \mathbb{E}[n_{k,\tau^*}],
\end{aligned}$$

since τ^* is a stopping time. Plugging this equality into (4.9) yields:

$$\begin{aligned}
R_B &\leq B \cdot \frac{\mu^r_{k^*}}{\mu^c_{k^*}} - \sum_{k=1}^{K} \mu^r_k \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\
&= \frac{\mu^r_{k^*}}{\mu^c_{k^*}} \cdot (B - \sum_{k \mid \Delta_k = 0} \mu^c_k \cdot \mathbb{E}[n_{k,\tau^*}]) - \sum_{k \mid \Delta_k > 0} \mu^r_k \cdot \mathbb{E}[n_{k,\tau^*}] + O(1).
\end{aligned}$$

Along the same lines as for the rewards, we can show that $\mathbb{E}[\sum_{t=1}^{\tau^*} c_{a_t,t}] = \sum_{k=1}^{K} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}]$. By definition of τ^*, we have $B \leq \sum_{t=1}^{\tau^*} c_{a_t,t}$ almost surely. Taking expectations on both sides yields:

$$\begin{aligned} B &\leq \sum_{k=1}^{K} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}] \\ &= \sum_{k \mid \Delta_k = 0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}] + \sum_{k \mid \Delta_k > 0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau^*}]. \end{aligned}$$

Plugging this inequality back into the regret bound, we get:

$$\begin{aligned} R_B &\leq \sum_{k \mid \Delta_k > 0} (\frac{\mu_{k^*}^r}{\mu_{k^*}^c} \cdot \mu_k^c - \mu_k^r) \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\ &= \sum_{k \mid \Delta_k > 0} \mu_k^c \cdot \Delta_k \cdot \mathbb{E}[n_{k,\tau^*}] + O(1). \end{aligned} \tag{C.10}$$

Using the upper bound of Lemma 4.4, the concavity of the logarithmic function, and Lemma 4.5, we derive:

$$\begin{aligned} R_B &\leq 2^6 \lambda^2 \cdot (\sum_{k \mid \Delta_k > 0} \frac{1}{\mu_k^c \cdot \Delta_k}) \cdot \ln(\frac{B+1}{\epsilon}) + \frac{4\pi^2}{3\epsilon^2} \cdot (\sum_{k \mid \Delta_k > 0} \mu_k^c \cdot \Delta_k) + O(1) \\ &\leq 2^6 \lambda^2 \cdot (\sum_{k \mid \Delta_k > 0} \frac{1}{\mu_k^c \cdot \Delta_k}) \cdot \ln(\frac{B+1}{\epsilon}) + K \cdot \frac{4\pi^2 \kappa}{3\epsilon^2} + O(1), \end{aligned}$$

since $\Delta_k \leq \mu_{k^*}^r / \mu_{k^*}^c \leq \kappa$ and $\mu_k^c \leq 1$ for any arm k.

C.4 Proofs for Section 4.5

C.4.1 Proof of Lemma 4.8

Consider any suboptimal basis $x \in \mathcal{B}$. The proof is along the same lines as for Lemma 4.5 and follows the exact same steps. We use the shorthand notation $\beta_x = 8\rho \cdot (\frac{\sum_{k=1}^{K} \xi_k^x}{\Delta_x})^2$.

Fact C.5.

$$\mathbb{E}[n_{x,\tau^*}] \leq 2\beta_x \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{x_t = x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}]. \tag{C.11}$$

We omit the proof as it is analogous to the proof of Fact C.1. As in Lemma 4.5, we break down the second term in three terms and bound each of them by a constant:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{x_t=x} \cdot I_{n_{x,t}\geq\beta_x \ln(t)}] &= \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x,t}+E_{x,t}\geq\mathrm{obj}_{x^*,t}+E_{x^*,t}} \cdot I_{n_{x,t}\geq\beta_x \ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}] && \text{(C.12)} \\
&+ \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*,t}\leq\mathrm{obj}_{x^*}-E_{x^*,t}}] && \text{(C.13)} \\
&+ \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}} \cdot I_{n_{x,t}\geq\beta_x \ln(t)}]. && \text{(C.14)}
\end{aligned}$$

Fact C.6.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}} \cdot I_{n_{x,t}\geq\beta_x \ln(t)}] = 0.$$

Proof. If $\mathrm{obj}_{x^*} < \mathrm{obj}_x + 2E_{x,t}$, we get:

$$\begin{aligned}
\frac{\Delta_x}{2} &< \sum_{k\in\mathcal{K}_x} \xi_k^x \cdot \sqrt{\frac{2\ln(t)}{n_{k,t}}} \\
&\leq \sum_{k\in\mathcal{K}_x} \xi_k^x \cdot \sqrt{\frac{2\ln(t)}{\rho + n_{k,t}^x}} \\
&\leq \sqrt{\sum_{k\in\mathcal{K}_x} \xi_k^x} \cdot \sum_{k\in\mathcal{K}_x} \sqrt{\xi_k^x} \cdot \sqrt{\frac{2\ln(t)}{n_{x,t}}},
\end{aligned}$$

where we use (4.10) and Lemma 4.6 for each $k \in \mathcal{K}_x$ such that $\xi_k^x \neq 0$ (otherwise, if $\xi_k^x = 0$, the inequality is trivial). This implies:

$$\begin{aligned}
n_{x,t} &< 8\rho \cdot (\frac{\sum_{k=1}^K \xi_k^x}{\Delta_x})^2 \cdot \ln(t) \\
&= \beta_x \cdot \ln(t),
\end{aligned}$$

using the Cauchy−Schwarz inequality and the fact that a basis involves at most ρ arms.

□

Fact C.7.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}}] \leq \rho \cdot \frac{\pi^2}{6}.$$

Proof. If $\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}$, there must exist $k \in \mathcal{K}_x$ such that $\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}$, otherwise:

$$\begin{aligned}
\mathrm{obj}_{x,t} - \mathrm{obj}_x &= \sum_{k \in \mathcal{K}_x} (\bar{r}_{k,t} - \mu_k^r) \cdot \xi_k^x \\
&< \sum_{k \in \mathcal{K}_x} \epsilon_{k,t} \cdot \xi_k^x \\
&= E_{x,t},
\end{aligned}$$

where the inequality is strict because there must exist $l \in \mathcal{K}_x$ such that $\xi_l^x > 0$ as a result of Assumption 4.2 (at least one resource constraint is binding for a feasible basis to (4.3) aside from the basis $\tilde{x}$ associated with $\mathcal{K}_{\tilde{x}} = \emptyset$). We obtain:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}}] &\leq \sum_{k \in \mathcal{K}_x} \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] \\
&\leq \rho \cdot \frac{\pi^2}{6},
\end{aligned}$$

where the last inequality is derived along the same lines as in the proof of Fact C.3. □

Fact C.8.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*,t} \leq \mathrm{obj}_{x^*} - E_{x^*,t}}] \leq \rho \cdot \frac{\pi^2}{6}.$$

Proof. Similar to Fact C.7. □

C.4.2 Proof of Theorem 4.3

The proof proceeds along the same lines as for Theorem 4.1. We build upon (4.4):

$$\begin{aligned}
R_{B(1),\cdots,B(C)} &\leq B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\
&= B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \sum_{k=1}^{K} \sum_{x \in \mathcal{B}} r_{k,t} \cdot I_{x_t=x, a_t=k}] + O(1) \\
&= B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \sum_{k=1}^{K} \sum_{x \in \mathcal{B}} I_{x_t=x, a_t=k} \cdot \mathbb{E}[r_{k,t} \mid \mathcal{F}_{t-1}]] + O(1) \\
&= B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}^x] + O(1),
\end{aligned}$$

where we use the fact that x_t and a_t are determined by the events of the first $t-1$ rounds and that τ^* is a stopping time. Using Lemma 4.6, we derive:

$$\begin{aligned}
R_{B(1),\cdots,B(C)} &\leq B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k \in \mathcal{K}_x} \{\mu_k^r \cdot \frac{\xi_k^x}{\sum_{l=1}^{K} \xi_l^x} \cdot \mathbb{E}[n_{x,\tau^*}] - \rho\} + O(1) \\
&= B \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \{\frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^{K} \xi_k^x} \cdot (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) - (\rho)^2\} + O(1) \\
&= (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*}) \cdot (B - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^{K} \xi_k^x}) \\
&- \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \{(\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \frac{\mathbb{E}[n_{x,\tau^*}]}{\sum_{k=1}^{K} \xi_k^x}\} + O(1).
\end{aligned}$$

Now observe that, by definition, at least one resource is exhausted at time τ^*. Hence, there exists $i \in \{1, \cdots, C\}$ such that the following holds almost surely:

$$\begin{aligned}
B(i) &\leq \sum_{x \in \mathcal{B}} \sum_{k \in \mathcal{K}_x} c_k(i) \cdot n_{k,\tau^*}^x \\
&\leq \sum_{x \in \mathcal{B}} \sum_{k \in \mathcal{K}_x} [c_k(i) \cdot (\frac{\xi_k^x}{\sum_{l=1}^{K} \xi_l^x} \cdot n_{x,\tau^*} + 1)] \\
&= |\mathcal{B}| \cdot \rho + \sum_{x \in \mathcal{B}} \frac{n_{x,\tau^*}}{\sum_{k=1}^{K} \xi_k^x} \cdot \sum_{k \in \mathcal{K}_x} c_k(i) \cdot \xi_k^x \\
&\leq |\mathcal{B}| \cdot \rho + b(i) \cdot \sum_{x \in \mathcal{B}} \frac{n_{x,\tau^*}}{\sum_{k=1}^{K} \xi_k^x},
\end{aligned}$$

where we use Lemma 4.6 again and the fact that any basis $x \in \mathcal{B}$ satisfies all the constraints of (4.3). We conclude that the inequality:

$$\sum_{x \in \mathcal{B} \mid \Delta_x = 0} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi_k^x} \geq B - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{n_{x,\tau^*}}{\sum_{k=1}^K \xi_k^x} - \frac{|\mathcal{B}| \cdot \rho}{b}$$

holds almost surely. Taking expectations on both sides and plugging the result back into the regret bound yields:

$$\begin{aligned} R_{B(1),\cdots,B(C)} &\leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{(\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k=1}^K \mu_k^r \cdot \xi_k^x)}{\sum_{k=1}^K \xi_k^x} \cdot \mathbb{E}[n_{x,\tau^*}] \qquad \text{(C.15)} \\ &\quad + (\sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*}) \cdot \frac{|\mathcal{B}| \cdot \rho}{b} + O(1) \\ &\leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\Delta_x}{\sum_{k=1}^K \xi_k^x} \cdot \mathbb{E}[n_{x,\tau^*}] + O(1), \qquad \text{(C.16)} \end{aligned}$$

where we use the fact that:

$$\begin{aligned} \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} &\leq \sum_{k=1}^K \frac{\sum_{i=1}^C c_k(i)}{\epsilon} \cdot \xi_k^{x^*} \\ &= \frac{1}{\epsilon} \cdot \sum_{i=1}^C \sum_{k=1}^K c_k(i) \cdot \xi_k^{x^*} \qquad \text{(C.17)} \\ &\leq \frac{\sum_{i=1}^C b(i)}{\epsilon}, \end{aligned}$$

using Assumption 4.2 and the fact that x^* is a feasible basis to (4.3). Using Lemma 4.7, Lemma 4.8, and the concavity of the logarithmic function, we obtain:

$$\begin{aligned} R_{B(1),\cdots,B(C)} &\leq 16\rho \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\sum_{k=1}^K \xi_k^x}{\Delta_x}) \cdot \ln(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1) \\ &\quad + \frac{\pi^2}{3}\rho \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{\Delta_x}{\sum_{k=1}^K \xi_k^x}) + O(1) \\ &\leq 16\frac{\rho \cdot \sum_{i=1}^C b(i)}{\epsilon} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(\frac{\sum_{i=1}^C b(i) \cdot B}{\epsilon} + 1) + O(1). \end{aligned}$$

To derive this last inequality, we use: (i) $\Delta_x \leq \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{i=1}^C b(i)/\epsilon$ (see (C.17)), (ii) the fact that, for any basis $x \in \mathcal{B}$, at least one of the first C inequalities is binding in

(4.3), which implies that there exists $i \in \{1, \cdots, C\}$ such that:

$$\begin{aligned}\sum_{k=1}^{K} \xi_k^x &\geq \sum_{k=1}^{K} c_k(i) \cdot \xi_k^x \\ &= b(i) \\ &\geq b,\end{aligned}$$

and (iii) the inequality:

$$\begin{aligned}\sum_{k=1}^{K} \xi_k^x &\leq \sum_{k=1}^{K} \frac{\sum_{i=1}^{C} c_k(i)}{\epsilon} \cdot \xi_k^x \\ &= \frac{1}{\epsilon} \cdot \sum_{i=1}^{C} \sum_{k=1}^{K} c_k(i) \cdot \xi_k^x \\ &\leq \frac{\sum_{i=1}^{C} b(i)}{\epsilon},\end{aligned}$$

for any basis $x \in \mathcal{B}$.

As a side note, observe a possibly better regret bound is given by:

$$R_{B(1),\cdots,B(C)} \leq 16\rho \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(T) + O(1),$$

if time is a limited resource since, in this case, $\tau^* \leq T$ and the constraint $\sum_{k=1}^{K} \xi_k^x \leq 1$ is part of (4.3).

C.4.3 Proof of Theorem 4.4

Along the same lines as for the case of a single limited resource, we start from inequality (C.16) derived in the proof of Theorem 4.3 and apply Lemma 4.8 only if Δ_x is big enough, taking into account the fact that:

$$\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x,\tau^*}] \leq \mathbb{E}[\tau^*] \leq \frac{\sum_{i=1}^{C} b(i) \cdot B}{\epsilon} + 1.$$

Specifically, we have:

$$
\begin{aligned}
& R_{B(1),\cdots,B(C)} \\
\leq & \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq \frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x>0} \min\Big(\frac{\Delta_x}{\sum_{k=1}^{K}\xi_k^x}\cdot n_x, \\
& \qquad 16\rho\cdot\frac{\sum_{k=1}^{K}\xi_k^x}{\Delta_x}\cdot\ln\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)+\frac{\pi^2}{3}\rho\cdot\frac{\Delta_x}{\sum_{k=1}^{K}\xi_k^x}\Big)\ \} + O(1) \\
\leq & \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq \frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x>0} \min\Big(\frac{\Delta_x}{\sum_{k=1}^{K}\xi_k^x}\cdot n_x, \\
& \qquad 16\rho\cdot\frac{\sum_{k=1}^{K}\xi_k^x}{\Delta_x}\cdot\ln\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)\Big)\ \} + \frac{\pi^2}{3}\frac{|\mathcal{B}|\cdot\rho}{\epsilon} + O(1) \\
\leq & \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq \frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x>0} \sqrt{16\rho\cdot\ln\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)\cdot n_x}\ \} + O(1) \\
\leq & 4\sqrt{\rho\cdot\ln\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)}\cdot \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq \frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x>0} \sqrt{n_x}\ \} + O(1) \\
\leq & 4\sqrt{\rho\cdot|\mathcal{B}|\cdot\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)\cdot\ln\Big(\frac{\sum_{i=1}^{C} b(i)\cdot B}{\epsilon}+1\Big)} + O(1),
\end{aligned}
$$

where we use $\Delta_x \leq \sum_{i=1}^{C} b(i)/\epsilon$ and $\sum_{k=1}^{K}\xi_k^x \geq b$ for the second inequality (see the end of the proof of Theorem 4.3), we maximize over $\Delta_x/\sum_{k=1}^{K}\xi_k^x \geq 0$ for each $x\in\mathcal{B}$ to derive the third inequality, and we use Cauchy-Schwartz for the last inequality.

C.5 Proofs for Section 4.6

C.5.1 Proof of Lemma 4.9

For any $T\in\mathbb{N}$ and any arm $k\in\{1,\cdots,K\}$, we denote by $n_{k,T}^{\mathrm{opt}}$ the expected number of times that arm k is pulled by the optimal non-anticipating algorithm (which is characterized by a high-dimensional dynamic program) when the time horizon is T and the budget is $B=b\cdot T$. We prove the claim in two steps. First, we show that if $T-n_{k^*,T}^{\mathrm{opt}}=\Omega(\sqrt{T})$

(Case A) or $T - n^{\mathrm{opt}}_{k^*,T} = o(\sqrt{T})$ (Case B) then $\mathrm{ER}_{\mathrm{OPT}}(B,T) = T \cdot \mathrm{obj}_{x^*} - \Omega(\sqrt{T})$. This is enough to establish the result because if $T - n^{\mathrm{opt}}_{k^*,T} \neq \Omega(\sqrt{T})$ then we can extract a subsequence of $(T - n^{\mathrm{opt}}_{k^*,T})/\sqrt{T}$ that converges to 0 and we can conclude with Case B.

Case A: $T - n^{\mathrm{opt}}_{k^*,T} = \Omega(\sqrt{T})$**.** Consider the linear program:

$$\begin{array}{lll} \sup\limits_{(\xi_k)_{k=1,\cdots,K} \in \mathbb{R}^K_+} & \sum\limits_{k=1}^K \mu^r_k \cdot \xi_k & \\ \text{subject to} & \sum\limits_{k=1}^K \mu^c_k \cdot \xi_k \leq b & \\ & \sum\limits_{k=1}^K \xi_k \leq 1 & \\ & \xi_{k^*} \leq \Gamma & \end{array} \tag{C.18}$$

parametrized by Γ and its dual:

$$\begin{array}{lll} \inf\limits_{(\zeta_1,\zeta_2,\zeta_3) \in \mathbb{R}^3_+} & b \cdot \zeta_1 + \zeta_2 + \Gamma \cdot \zeta_3 & \\ \text{subject to} & \mu^c_k \cdot \zeta_1 + \zeta_2 \geq \mu^r_k, \quad k \neq k^* & \\ & \mu^c_{k^*} \cdot \zeta_1 + \zeta_2 + \zeta_3 \geq \mu^r_{k^*} & \end{array} \tag{C.19}$$

Since the vector $(\xi_k)_{k=1,\cdots,K}$ determined by $\xi_{k^*} = 1$ and $\xi_k = 0$ for $k \neq k^*$ is the only optimal solution to (C.18) when $\Gamma = 1$ (by assumption), we can find a strictly complementary optimal solution to the dual (C.19) $\zeta^*_1, \zeta^*_2, \zeta^*_3 > 0$. Moreover, by definition of $n^{\mathrm{opt}}_{k^*,T}$, we can show, along the same lines as in the proof of Lemma 4.2, that $\mathrm{ER}_{\mathrm{OPT}}(B,T)$ is no larger than T times the value of (C.18) when $\Gamma = n^{\mathrm{opt}}_{k^*,T}/T$ (up to a constant additive term of order $O(1)$). By weak duality, and since $(\zeta^*_1, \zeta^*_2, \zeta^*_3)$ is feasible for (C.19) for any Γ, this implies:

$$\begin{aligned} \mathrm{ER}_{\mathrm{OPT}}(B,T) &\leq T \cdot (b \cdot \zeta^*_1 + \zeta^*_2 + \frac{n^{\mathrm{opt}}_{k^*,T}}{T} \cdot \zeta^*_3) + O(1) \\ &\leq T \cdot (b \cdot \zeta^*_1 + \zeta^*_2 + \zeta^*_3 - \frac{T - n^{\mathrm{opt}}_{k^*,T}}{T} \cdot \zeta^*_3) + O(1) \\ &\leq T \cdot \mathrm{obj}_{x^*} - \Omega(\sqrt{T}), \end{aligned}$$

where we use the fact that $b \cdot \zeta_1^* + \zeta_2^* + \zeta_3^*$ is the optimal value of (C.18) when $\Gamma = 1$ by strong duality (both (C.18) and (C.19) are feasible) and $\zeta_3^* > 0$.

Case B: $T - n_{k^*,T}^{\mathrm{opt}} = o(\sqrt{T})$**.** First observe that since the vector $(\xi_k)_{k=1,\cdots,K}$ determined by $\xi_{k^*} = 1$ and $\xi_k = 0$ for $k \neq k^*$ is the only optimal solution to (4.3), it must be that $\mu_{k^*}^r > 0$ (since 0 is a feasible solution to (4.3) with objective value 0). For any $t \in \mathbb{N}$, denote by a_t the arm pulled by the optimal non-anticipating algorithm at time t and define τ_T^* as the corresponding stopping time when the time horizon is T. We have:

$$\begin{aligned}
\mathrm{ER}_{\mathrm{OPT}}(B,T) &= \mathbb{E}[\sum_{t=1}^{\tau_T^*-1} r_{a_t,t}] \\
&= \sum_{k=1}^{K} \mu_k^r \cdot n_{k,T}^{\mathrm{opt}} \\
&\leq \sum_{k\neq k^*} n_{k,T}^{\mathrm{opt}} + \mu_{k^*}^r \cdot n_{k^*,T}^{\mathrm{opt}} \\
&\leq (T - n_{k^*,T}^{\mathrm{opt}}) + \mu_{k^*}^r \cdot (\mathbb{E}[\tau_T^*] - 1) \\
&= T \cdot \mu_{k^*}^r - \mu_{k^*}^r \cdot (T - \mathbb{E}[\tau_T^*] + 1) + o(\sqrt{T}) \\
&= T \cdot \mathrm{obj}_{x^*} - \mu_{k^*}^r \cdot \mathbb{E}[\sum_{t=\tau_T^*}^{T} 1] + o(\sqrt{T}) \\
&\leq T \cdot \mathrm{obj}_{x^*} - \mu_{k^*}^r \cdot \mathbb{E}[(\sum_{t=\tau_T^*}^{T} c_{k^*,t} + \sum_{t=1}^{\tau_T^*-1} c_{a_t,t} - B)_+] \\
&\leq T \cdot \mathrm{obj}_{x^*} - \mu_{k^*}^r \cdot \left(\mathbb{E}[(\sum_{t=1}^{T}\{c_{k^*,t} - b\})_+] - \sum_{k\neq k^*} n_{k,T}^{\mathrm{opt}}\right) + o(\sqrt{T}) \\
&= T \cdot \mathrm{obj}_{x^*} - \mu_{k^*}^r \cdot \mathbb{E}[(\sum_{t=1}^{T}\{c_{k^*,t} - b\})_+] + o(\sqrt{T}).
\end{aligned}$$

The first inequality is obtained using the fact that the rewards are bounded by 1. The second inequality is obtained using the fact that $\sum_{k=1}^{K} n_{k,T}^{\mathrm{opt}} = \mathbb{E}[\tau_T^*] - 1 \leq T$. The third inequality is obtained along the same lines as in the proof of Lemma 4.3, using $\sum_{t=1}^{\tau_T^*-1} c_{a_t,t} \leq B$ by definition of τ_T^*. We use the inequality $(y+z)_+ \geq y_+ - |z|$ (true for any $(y,z) \in \mathbb{R}^2$) and the fact the amount of resource consumed at any step is no larger than 1 for the fourth inequality. Since $(c_{k^*,t} - b)_{t\in\mathbb{N}}$ is an i.i.d. zero-mean bounded stochastic process with

positive variance, $\frac{1}{\sqrt{T}} \cdot \mathbb{E}[(\sum_{t=1}^{T}\{c_{k,t} - b\})_+]$ converges to a positive value and we conclude:

$$\mathrm{ER}_{\mathrm{OPT}}(B,T) \leq T \cdot \mathrm{obj}_{x^*} - \Omega(\sqrt{T}),$$

since $\mu^r_{k^*} > 0$.

C.5.2 Proof of Lemma 4.10

Consider x either an infeasible basis to (4.3) or a pseudo-basis for (4.3). Without loss of generality, we can assume that x involves two arms (one of which may be a dummy arm introduced in the specification of the algorithm given in Section 4.6) and that $\mathcal{K}_x = \{k, l\}$ with $\mu^c_k, \mu^c_l > b$ (the situation is symmetric if the reverse inequality holds). Defining $\beta_x = 32/\epsilon^3$, we have:

$$\begin{aligned}
\mathbb{E}[n_{x,T}] &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{b_{x,t} \geq n_{x,t} \cdot (b+\epsilon/2)}] \\
&+ \sum_{t=1}^{T} \sum_{s=\beta_x \ln(T)}^{t} \mathbb{P}[b_{x,t} < s \cdot (b+\epsilon) - s \cdot \epsilon/2, n_{x,t} = s] \\
&\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{b_{x,t} \geq n_{x,t} \cdot (b+\epsilon/2)}] \\
&+ \sum_{t=1}^{T} \sum_{s=\beta_x \ln(T)}^{\infty} \exp(-s\frac{\epsilon^2}{2}) \\
&\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{b_{x,t} \geq n_{x,t} \cdot b + \epsilon\beta_x/4 \cdot \ln(t)}] + \frac{2\pi^2}{3\epsilon^2}.
\end{aligned}$$

The first inequality is derived along the same lines as in Fact C.1. The third inequality is obtained by observing that, as a result of Assumption 4.6, the average amount of resource consumed any time basis x is selected at Step-Simplex is at least $b + \epsilon$ no matter which of arm k or l is pulled. Finally, we use the same bounds as in Fact C.2 for the last two inequalities. Observe that if x is selected at time t, either $\bar{c}_{k,t} - \epsilon_{k,t} \leq b$ or $\bar{c}_{l,t} - \epsilon_{l,t} \leq b$, otherwise x would have been infeasible for (4.8). Moreover, if $n_{x,t} \geq \beta_x \ln(T)$, then we

have either $n^x_{k,t} \geq \beta_x/2\ln(T)$ or $n^x_{l,t} \geq \beta_x/2\ln(T)$ since there are only two arms in $\mathcal{K}_x$. By symmetry, we study the first situation and look at:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n^x_{k,t}\geq\beta_x/2\cdot\ln(t)} \cdot I_{b_{x,t}\geq n_{x,t}\cdot b+\epsilon\beta_x/4\cdot\ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n^x_{k,t}\geq\beta_x/2\cdot\ln(t)} \cdot I_{b_{x,t}\geq n_{x,t}\cdot b+\epsilon\beta_x/4\cdot\ln(t)} \cdot I_{\bar{c}_{k,\tau_q}-\epsilon_{k,\tau_q}\geq b,\ q=q_t-\epsilon\beta_x/4\ln(t),\cdots,q_t}] \\
&+ \sum_{t=1}^{T}\sum_{\tau=1}^{t}\sum_{s=\beta_x/4\cdot\ln(t)}^{t} \mathbb{P}[\bar{c}_{k,\tau} < b + \frac{\epsilon}{2}, n_{k,\tau} = s]
\end{aligned}$$

and

$$\sum_{t=1}^{T}\sum_{\tau=1}^{t}\sum_{s=\beta_x/4\cdot\ln(t)}^{t} \mathbb{P}[\bar{c}_{k,\tau} < b + \frac{\epsilon}{2}, n_{k,\tau} = s] \leq \sum_{t=1}^{T}\sum_{\tau=1}^{t}\sum_{s=\beta_x/4\cdot\ln(t)}^{\infty} \exp(-s\cdot\frac{\epsilon^2}{2}) \leq \frac{2\pi^2}{3\epsilon^2},$$

where $(\tau_q)_{q\in\mathbb{N}}$ denote the random times at which basis x is selected and, for a time t at which basis x is selected, q_t denotes the index $q \in \mathbb{N}$ such that $\tau_q = t$. The first inequality is a consequence of $n^x_{k,\tau_q} = n^x_{k,t} - (q_t - q) \geq n^x_{k,t} - \epsilon\beta_x/4\ln(t) \geq \beta_x/4\ln(t)$ for $q = q_t - \epsilon\beta_x/4\ln(t), \cdots, q_t$ and $n_{k,\tau_q} \geq n^x_{k,\tau_q}$, which implies $\epsilon_{k,\tau_q} \leq \epsilon/2$. We use the same bounds as in Fact C.2 for the last two inequalities. Now observe that, for any $q \in \{q_t - \epsilon\beta_x/4\ln(t), \cdots, q_t\}$, we have $\bar{c}_{l,\tau_q} - \epsilon_{l,\tau_q} \leq b$ since $\bar{c}_{k,\tau_q} - \epsilon_{k,\tau_q} \geq b$ and since x is feasible basis to (4.8) at time τ_q (by definition). This implies that, for any $q \in \{q_t - \epsilon\beta_x/4\ln(t), \cdots, q_t\}$, arm l was pulled at time τ_q by definition of the load balancing algorithm since the amount of resource consumed at any round cannot be larger than 1 and $b_{x,\tau_q} \geq b_{x,t} - (q_t - q) \geq n_{x,t}\cdot b + \epsilon\beta_x/4\cdot\ln(t) - (q_t - q) \geq n_{x,t}\cdot b \geq n_{x,\tau_q} b$. We get:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n^x_{k,t}\geq\beta_x/2\cdot\ln(t)} \cdot I_{b_{x,t}\geq n_{x,t}\cdot b+\epsilon\beta_x/4\cdot\ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} I_{n^x_{l,t}\geq\epsilon\beta_x/4\cdot\ln(t)} \cdot I_{\bar{c}_{l,t}-\epsilon_{l,t}\leq b}] + \frac{2\pi^2}{3\epsilon^2} \\
&\leq \sum_{t=1}^{T} \mathbb{P}[\bar{c}_{l,t} \leq b + \frac{\epsilon}{2}, n^x_{l,t} \geq \epsilon\beta_x/4\cdot\ln(t)] + \frac{2\pi^2}{3\epsilon^2} \\
&\leq \sum_{t=1}^{T}\sum_{s=\epsilon\beta_x/4\cdot\ln(t)}^{\infty} \exp(-s\cdot\frac{\epsilon^2}{2}) + \frac{2\pi^2}{3\epsilon^2} \leq \frac{4\pi^2}{3\epsilon^2}.
\end{aligned}$$

Bringing everything together, we derive:

$$\mathbb{E}[n_{x,T}] \leq \frac{2^6}{\epsilon^3} \cdot \ln(T) + \frac{10\pi^2}{3\epsilon^2}.$$

C.5.3 Proof of Lemma 4.11

Without loss of generality, we can assume that x involves two arms (one of which may be a dummy arm introduced in the specification of the algorithm given in Section 4.6) and that $\mathcal{K}_x = \{k, l\}$ with $\mu_k^c > b > \mu_l^c$. We say that a "swap" occurred at time τ if basis x was selected at time τ and $\bar{c}_{k,\tau} - \epsilon_{k,\tau} \leq b \leq \bar{c}_{l,\tau} - \epsilon_{l,\tau}$. We define $n_{x,t}^{\text{swap}}$ as the total number of swaps that have occurred before time t, i.e. $n_{x,t}^{\text{swap}} = \sum_{\tau=1}^{t-1} I_{x_\tau = x} \cdot I_{\bar{c}_{k,\tau} - \epsilon_{k,\tau} \leq b \leq \bar{c}_{l,\tau} - \epsilon_{l,\tau}}$. Consider $u \geq 1$ and define $\gamma_x = (4/\epsilon)^2$. First note that:

$$\begin{aligned}
\mathbb{P}[n_{x,t}^{\text{swap}} \geq \gamma_x \ln(t)] &\leq \sum_{q=\gamma_x \ln(t)}^{t} \mathbb{P}[\bar{c}_{k,\tau_q} - \epsilon_{k,\tau_q} \leq b \leq \bar{c}_{l,\tau_q} - \epsilon_{l,\tau_q}] \\
&\leq \sum_{q=\gamma_x \ln(t)}^{t} \mathbb{P}[\bar{c}_{k,\tau_q} \leq b + \frac{\epsilon}{2}, n_{k,\tau_q} \geq \frac{\gamma_x}{2} \ln(t)] \\
&+ \sum_{q=\gamma_x \ln(t)}^{t} \mathbb{P}[b - \frac{\epsilon}{2} \leq \bar{c}_{l,\tau_q}, n_{l,\tau_q} \geq \frac{\gamma_x}{2} \ln(t)] \\
&\leq 2 \sum_{q=1}^{t} \sum_{s=\gamma_x/2 \cdot \ln(t)}^{\infty} \exp(-s \cdot \frac{\epsilon^2}{2}) \\
&\leq \frac{8}{\epsilon^2 \cdot t^2},
\end{aligned}$$

where $(\tau_q)_{q \in \mathbb{N}}$ are defined as the times at which basis x is selected. The first inequality is derived observing that if $n_{x,t}^{\text{swap}} \geq \gamma_x \ln(t)$ then it must be that basis x was selected for the qth time, for some $q \geq \gamma_x \ln(t)$, and that we had $\bar{c}_{k,\tau_q} - \epsilon_{k,\tau_q} \leq b \leq \bar{c}_{l,\tau_q} - \epsilon_{l,\tau_q}$. To obtain the second inequality, we observe that, at any time τ, at least one of arm k and l must have been pulled $n_{x,\tau}/2$ times and that $\epsilon_{k,\tau} \leq \epsilon/2$ when $n_{k,\tau} \geq \gamma_x/2 \ln(t)$ (a similar inequality holds for arm l). The last two inequalities are derived in the same fashion as in Lemma

4.10. This yields:

$$\begin{aligned}
&\mathbb{P}[|b_{x,t} - n_{x,t} \cdot b| \geq u + \gamma_x \ln(t)] \\
&\quad \leq \mathbb{P}[|b_{x,t} - n_{x,t} \cdot b| \geq u + \gamma_x \ln(t) \ ; \ n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)] + \mathbb{P}[n_{x,t}^{\text{swap}} \geq \gamma_x \ln(t)] \\
&\quad \leq \mathbb{P}[|b_{x,t} - n_{x,t} \cdot b| \geq u + \gamma_x \ln(t) \ ; \ n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)] + \frac{8}{\epsilon^2 \cdot t^2}.
\end{aligned}$$

Note that, by definition of the load balancing algorithm, we are led to pull arm k (resp. arm l) at time τ_q if the budget spent so far when selecting basis x, denoted by b_{x,τ_q}, is below (resp. above) the "target" of $n_{x,\tau_q} \cdot b$ assuming there is no "swap" at time τ_q (i.e. $\bar{c}_{k,\tau_q} - \epsilon_{k,\tau_q} \geq \bar{c}_{l,\tau_q} - \epsilon_{l,\tau_q}$). Hence, if $b_{x,t} - n_{x,t} \cdot b \geq u + \gamma_x \ln(t)$ and $n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)$, we must have been pulling arm l for at least $s \geq \lfloor u \rfloor$ rounds $t_1 \leq \cdots \leq t_s \leq t-1$ where basis x was selected since the last time, denoted by t_0, where basis x was selected and the budget was below the target, i.e. $b_{x,t_0} \leq n_{x,t_0} \cdot b$ (because the amounts of resource consumed at each round are almost surely bounded by 1). Moreover, we have:

$$\begin{aligned}
\sum_{\tau=t_0+1}^{t-1} I_{x_\tau = x} \cdot (c_{k,\tau} \cdot I_{\bar{c}_{k,\tau} - \epsilon_{k,\tau} \geq \bar{c}_{l,\tau} - \epsilon_{l,\tau}} + c_{l,\tau} \cdot I_{\bar{c}_{k,\tau} - \epsilon_{k,\tau} < \bar{c}_{l,\tau} - \epsilon_{l,\tau}}) \\
&= b_{x,t} - b_{x,t_0+1} \\
&\geq (n_{x,t} - n_{x,t_0}) \cdot b + u - 1 + \gamma_x \ln(t) \\
&\geq s \cdot b + u - 1 + \gamma_x \ln(t).
\end{aligned}$$

This implies:

$$\sum_{q=1}^{s} c_{l,t_q} \geq s \cdot b + u - 1$$

since $\sum_{\tau=t_0+1}^{t-1} I_{x_\tau=x} \cdot I_{\bar{c}_{k,\tau}-\epsilon_{k,\tau}<b} \cdot c_{k,\tau} \leq n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)$. Hence, if $u \geq 1$:

$$\begin{aligned}
&\mathbb{P}[b_{x,t} - n_{x,t} \cdot b \geq u + \gamma_x \ln(t) \ ; \ n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)] \\
&\quad \leq \sum_{s=\lfloor u \rfloor}^{t} \mathbb{P}[\sum_{q=1}^{s} c_{l,t_q} \geq s \cdot b + u - 1] \\
&\quad = \sum_{s=\lfloor u \rfloor}^{t} \mathbb{P}[\sum_{q=1}^{s} c_{l,t_q} \geq s \cdot \mu_l^c + s \cdot (b - \mu_l^c)] \\
&\quad \leq \sum_{s=\lfloor u \rfloor}^{t} \mathbb{P}[\sum_{q=1}^{s} c_{l,t_q} \geq s \cdot \mu_l^c + s \cdot \epsilon] \\
&\quad \leq \sum_{s=\lfloor u \rfloor}^{t} \exp(-2\epsilon^2 \cdot s) \\
&\quad \leq \frac{\exp(-2\epsilon^2 \cdot \lfloor u \rfloor)}{1 - \exp(-2\epsilon^2)} \\
&\quad \leq \frac{2}{\epsilon^2} \cdot \exp(-\epsilon^2 \cdot u),
\end{aligned}$$

where we use Lemma 4.1 for the third inequality and the fact that $\exp(-2v) \leq 1 - v/2$ for $v \in [0, 1]$ for the last inequality. With a similar argument, we conclude:

$$\mathbb{P}[|b_{x,t} - n_{x,t}| \cdot b \geq u + \gamma_x \ln(t) \ ; \ n_{x,t}^{\text{swap}} \leq \gamma_x \ln(t)] \leq \frac{4}{\epsilon^2} \cdot \exp(-\epsilon^2 \cdot u).$$

This last result enables us to show that:

$$\begin{aligned}
&|\mathbb{E}[b_{x,T}] - \mathbb{E}[n_{x,T}] \cdot b| \\
&\quad \leq \mathbb{E}[|b_{x,T} - n_{x,T} \cdot b|] \\
&\quad = \int_0^T \mathbb{P}[|b_{x,T} - n_{x,T} \cdot b| \geq u] \mathrm{d}u \\
&\quad \leq \int_0^T \mathbb{P}[|b_{x,T} - n_{x,T} \cdot b| \geq u \ ; \ n_{x,T}^{\text{swap}} \leq \gamma_x \ln(T)] \mathrm{d}u + T \cdot \mathbb{P}[n_{x,T}^{\text{swap}} \geq \gamma_x \ln(T)] \\
&\quad \leq \int_0^T \mathbb{P}[|b_{x,T} - n_{x,T} \cdot b| \geq u + 1 + \gamma_x \ln(T) \ ; \ n_{x,T}^{\text{swap}} \leq \gamma_x \ln(T)] \mathrm{d}u + 1 + \gamma_x \ln(T) + \frac{8}{\epsilon^2} \\
&\quad \leq \frac{4}{\epsilon^2} \cdot \int_0^T \exp(-\epsilon^2 \cdot u) \mathrm{d}u + 1 + \gamma_x \ln(T) + \frac{8}{\epsilon^2} \\
&\quad = \frac{13}{\epsilon^4} + (\frac{4}{\epsilon})^2 \ln(T).
\end{aligned}$$

We get $\mathbb{E}[n^x_{k,T}] \cdot \mu^c_k + \mathbb{E}[n^x_{l,T}] \cdot \mu^c_l = \mathbb{E}[b_{x,T}] \geq \mathbb{E}[n_{x,T}] \cdot b - \frac{13}{\epsilon^4} - (\frac{4}{\epsilon})^2 \ln(T)$, which, in combination with $\mathbb{E}[n^x_{k,T}] + \mathbb{E}[n^x_{l,T}] = \mathbb{E}[n_{x,T}]$, shows that:

$$\begin{aligned}\mathbb{E}[n^x_{k,T}] &\geq (\frac{b-\mu^c_l}{\mu^c_k - \mu^c_l}) \cdot \mathbb{E}[n_{x,T}] - \frac{13}{\epsilon^4 \cdot (\mu^c_k - \mu^c_l)} - \frac{4^2}{\epsilon^2 \cdot (\mu^c_k - \mu^c_l)} \ln(T) \\ &\geq \xi^x_k \cdot \mathbb{E}[n_{x,T}] - \frac{13}{\epsilon^5} - \frac{16}{\epsilon^3} \ln(T).\end{aligned}$$

C.5.4 Proof of Lemma 4.12

Consider any suboptimal basis $x \in \mathcal{B}$ and define $\beta_x = \frac{2^8}{\epsilon^3} \cdot (\frac{\lambda}{\Delta_x})^2$. Without loss of generality, we can assume that both x and x^* involve two arms (one of which may be a dummy arm introduced in the specification of the algorithm given in Section 4.6) and that $\mathcal{K}_{x^*} = \{k^*, l^*\}$ with $\mu^c_{k^*} > b > \mu^c_{l^*}$ and $\mathcal{K}_x = \{k, l\}$ with $\mu^c_k > b > \mu^c_l$. The proof is along the same lines as for Lemmas 4.5 and 4.8. We break down the analysis in a series of facts where we stress the main differences. We start with an inequality analogous to Fact C.1.

$$\begin{aligned}&\mathbb{E}[n_{x,T}] \leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\ &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{\bar{c}_{k^*,t} - \epsilon_{k^*,t} \leq \mu^c_{k^*}} \cdot I_{\bar{c}_{l^*,t} - \epsilon_{l^*,t} \leq \mu^c_{l^*}}] \\ &+ \sum_{t=1}^{T} \mathbb{P}[\bar{c}_{l^*,t} > \mu^c_{l^*} + \epsilon_{l^*,t}] + \mathbb{P}[\bar{c}_{k^*,t} > \mu^c_{k^*} + \epsilon_{k^*,t}] \\ &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{\bar{c}_{k^*,t} - \epsilon_{k^*,t} \leq \mu^c_{k^*}} \cdot I_{\bar{c}_{l^*,t} - \epsilon_{l^*,t} \leq \mu^c_{l^*}}] + \frac{\pi^2}{3} \\ &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{\text{obj}_{x,t} + E_{x,t} \geq \sum_{k \in \{k^*, l^*\}} (\bar{r}_{k,t} + \lambda\epsilon_{k,t}) \cdot \xi^{x^*}_k} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] + \frac{\pi^2}{3} \\ &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{\text{obj}_{x,t} + E_{x,t} \geq \text{obj}_{x^*}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\ &+ \sum_{t=1}^{T} \mathbb{P}[\bar{r}_{l^*,t} < \mu^r_{l^*} - \epsilon_{l^*,t}] + \mathbb{P}[\bar{r}_{k^*,t} < \mu^r_{k^*} - \epsilon_{k^*,t}] + \frac{\pi^2}{3} \\ &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{\text{obj}_{x,t} + E_{x,t} \geq \text{obj}_{x^*}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] + \frac{2\pi^2}{3}.\end{aligned}$$

The first inequality is derived in the same fashion as in Fact C.1 substituting k with x. The third and last inequalities are obtained using Lemma 4.1 in the same fashion as in Fact C.3.

The fourth inequality is obtained by observing that (i) if $x_t = x$ then x_t is optimal for (4.8) and (ii) $(\xi_k^*)_{k=1,\cdots,K}$ is feasible for (4.8) if $\bar{c}_{l^*,t} - \epsilon_{l^*,t} \leq \mu_{l^*}^c$ and $\bar{c}_{k^*,t} - \epsilon_{k^*,t} \leq \mu_{k^*}^c$. The fifth inequality results from $\lambda \geq 1$ and $\mathrm{obj}_{x^*} = \sum_{k \in \{k^*,l^*\}} \mu_k^r \cdot \xi_k^{x^*}$. The second term in the last upper bound can be broken down in two terms similarly as in Lemmas 4.5 and 4.8:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x_t=x} \cdot I_{\mathrm{obj}_{x,t}+E_{x,t} \geq \mathrm{obj}_{x^*}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\quad \leq \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \qquad \text{(C.20)} \\
&\quad + \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\mathrm{obj}_{x^*} \leq \mathrm{obj}_x + 2E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}]. \qquad \text{(C.21)}
\end{aligned}$$

We carefully study each term separately.

Fact C.9.

$$\mathbb{E}[\sum_{t=1}^{T} I_{x \in \mathcal{B}_t} \cdot I_{\mathrm{obj}_{x^*} \leq \mathrm{obj}_x + 2E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \leq \frac{6\pi^2}{\epsilon^2}.$$

Proof. Using the shorthand notations $\alpha_x = 8(\frac{\lambda}{\Delta_x})^2$ and $\gamma_x = (\frac{4}{\epsilon})^2$, we have:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x \in \mathcal{B}_t} \cdot I_{\mathrm{obj}_{x^*} \leq \mathrm{obj}_x + 2E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\quad \leq \mathbb{E}[\sum_{t=1}^{T} I_{\Delta_x \leq 2\lambda \cdot \max(\epsilon_{k,t}, \epsilon_{l,t})} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\quad = \mathbb{E}[\sum_{t=1}^{T} I_{\min(n_{k,t}, n_{l,t}) \leq \alpha_x \ln(t)} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\quad \leq \sum_{t=1}^{T} \mathbb{P}[n_{l,t} \leq \alpha_x \ln(t) \ ; \ n_{x,t} \geq \beta_x \ln(t)] + \sum_{t=1}^{T} \mathbb{P}[n_{k,t} \leq \alpha_x \ln(t) \ ; \ n_{x,t} \geq \beta_x \ln(t)].
\end{aligned}$$

The first inequality is derived using $E_{x,t} = \lambda \cdot (\xi_{k,t}^x \cdot \epsilon_{k,t} + \xi_{l,t}^x \cdot \epsilon_{l,t})$ and $\xi_{k,t}^x + \xi_{l,t}^x \leq 1$ (this is imposed as a constraint in (4.8)). Observe now that α_x / β_x is a constant factor independent of Δ_x. Thus, we just have to show that if x has been selected at least $\beta_x \ln(t)$ times, then both k and l have been pulled at least a constant fraction of the time with high probability. This is the only time the load balancing algorithm comes into play in the proof of Lemma

4.12. We study the first term and we conclude the study by symmetry. We have:

$$\begin{aligned}
&\mathbb{P}[n_{l,t} \leq \alpha_x \ln(t) \ ; \ n_{x,t} \geq \beta_x \ln(t)] \\
&\quad \leq \mathbb{P}[n_{l,t} \leq \alpha_x \ln(t) \ ; \ n_{x,t} \geq \beta_x \ln(t) \ ; \ \sum_{\tau=1}^{t} I_{x_\tau = x} \cdot I_{a_\tau = k} \cdot c_{k,\tau} \geq (b + \epsilon/2) \cdot n_{k,t}^x] \\
&\quad + \sum_{s=4/\epsilon^2 \ln(t)}^{t} \mathbb{P}[\sum_{\tau=1}^{t} I_{x_\tau = x} \cdot I_{a_\tau = k} \cdot c_{k,\tau} \leq (b + \epsilon/2) \cdot s \ ; \ n_{k,t}^x = s] \\
&\quad \leq \mathbb{P}[n_{l,t} \leq \alpha_x \ln(t) \ ; \ n_{x,t} \geq \beta_x \ln(t) \ ; \ b_{x,t} - n_{x,t} \cdot b \geq \epsilon/2 \cdot n_{k,t}^x - n_{l,t}^x] \\
&\quad + \sum_{s=4/\epsilon^2 \ln(t)}^{t} \exp(-s \cdot \frac{\epsilon^2}{2}) \\
&\quad \leq \mathbb{P}[b_{x,t} - n_{x,t} \cdot b \geq 2\gamma_x \ln(t)] + (\frac{2}{\epsilon \cdot t})^2 \\
&\quad \leq \frac{16}{\epsilon^2 \cdot t^2}.
\end{aligned}$$

The first inequality is obtained observing that if $n_{l,t} \leq \alpha_x \ln(t)$ and $n_{x,t} \geq \beta_x \ln(t)$, we have:

$$n_{k,t}^x = n_{x,t} - n_{l,t}^x \geq (\frac{2^8}{\epsilon^3} - 8) \cdot (\frac{\lambda}{\Delta_x})^2 \cdot \ln(t) \geq \frac{4}{\epsilon^2} \cdot \ln(t)$$

because $\lambda \geq 1$ and $\Delta_x \leq \text{obj}_{x^*} = \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^{K} \xi_k^{x^*} \leq 1$ since x^* is a feasible basis to (4.3). To derive the second inequality, we use Lemma 4.1 for the second term and remark that:

$$b_{x,t} - n_{x,t} \cdot b \geq (\sum_{\tau=1}^{t} I_{x_\tau = x} \cdot I_{a_\tau = k} \cdot c_{k,\tau} - n_{k,t}^x \cdot b) - n_{l,t}^x$$

since $b \leq 1$. The third inequality is derived using:

$$\epsilon/2 \cdot n_{k,t}^x - n_{l,t}^x \geq \epsilon/2 \cdot n_{x,t} - 2 \cdot n_{l,t}^x \geq (\frac{2^7}{\epsilon^2} - 16) \cdot \ln(t) \geq 2\gamma_x \ln(t)$$

and the last inequality is obtained with Lemma 4.11.

□

Fact C.10.

$$\mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \leq \frac{3\pi^2}{\epsilon^2}.$$

Proof. First observe that:

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{n_{k,t} \geq \beta_x/2 \cdot \ln(t)}] \\
&+ \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{n_{l,t} \geq \beta_x/2 \cdot \ln(t)}] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} > b}] \\
&+ \mathbb{E}[\sum_{t=1}^{T} I_{x_t \in \mathcal{B}_t} \cdot I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{l,t} - \epsilon_{l,t} < b}] \\
&+ \sum_{t=1}^{T} \sum_{s=\beta_x/2 \cdot \ln(t)}^{t} \mathbb{P}[\bar{c}_{k,t} - \epsilon_{k,t} \leq b, n_{k,t} = s] + \mathbb{P}[\bar{c}_{l,t} - \epsilon_{l,t} \geq b, n_{l,t} = s] \\
&\leq 2 \cdot \mathbb{E}[\sum_{t=1}^{T} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} \geq b \geq \bar{c}_{l,t} - \epsilon_{l,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} > \bar{c}_{l,t} - \epsilon_{l,t}}] \\
&+ \sum_{t=1}^{T} \sum_{s=\beta_x/2 \cdot \ln(t)}^{t} \mathbb{P}[\bar{c}_{k,t} \leq b + \epsilon/2, n_{k,t} = s] + \mathbb{P}[\bar{c}_{l,t} \geq b + \epsilon/2, n_{l,t} = s] \\
&\leq 2 \cdot \mathbb{E}[\sum_{t=1}^{T} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} \geq b \geq \bar{c}_{l,t} - \epsilon_{l,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} > \bar{c}_{l,t} - \epsilon_{l,t}}] \\
&+ 2 \cdot \sum_{t=1}^{T} \sum_{s=\beta_x/2 \cdot \ln(t)}^{t} \exp(-s\frac{\epsilon^2}{2}) \\
&\leq 2 \cdot \mathbb{E}[\sum_{t=1}^{T} I_{\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} \geq b \geq \bar{c}_{l,t} - \epsilon_{l,t}} \cdot I_{\bar{c}_{k,t} - \epsilon_{k,t} > \bar{c}_{l,t} - \epsilon_{l,t}}] + \frac{4\pi^2}{3\epsilon^2}.
\end{aligned}
$$

The third inequality is obtained by observing that $\epsilon_{k,t}, \epsilon_{l,t} \leq \frac{\epsilon}{2}$ for $n_{k,t}, n_{l,t} \geq \frac{\beta_x}{2} \ln(t)$ (because $\lambda \geq 1$ and $\Delta_x \leq \text{obj}_{x^*} = \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^{K} \xi_k^{x^*} \leq 1$) and that, if $x_t \in \mathcal{B}_t$ and (for example) $\bar{c}_{l,t} - \epsilon_{l,t} < b$, it must be that $\bar{c}_{k,t} - \epsilon_{k,t} \geq b$. The last two inequalities are obtained in the same fashion as in Lemma 4.10 observing that $\beta_x/2 \geq 4/\epsilon^2$. At this point, the key observation is that if $\text{obj}_{x,t} \geq \text{obj}_x + E_{x,t}$, $\bar{c}_{k,t} - \epsilon_{k,t} \geq b \geq \bar{c}_{l,t} - \epsilon_{l,t}$, and $\bar{c}_{k,t} - \epsilon_{k,t} > \bar{c}_{l,t} - \epsilon_{l,t}$, at least one of the following six events occurs: $\{\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}\}$, $\{\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}\}$, $\{\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}\}$, $\{\bar{c}_{k,t} \geq \mu_k^c + \epsilon_{k,t}\}$, $\{\bar{c}_{l,t} \leq \mu_l^c - \epsilon_{l,t}\}$ or $\{\bar{c}_{l,t} \geq \mu_l^c + \epsilon_{l,t}\}$.

Otherwise, using the shorthand notations $\tilde{c}_k = \bar{c}_{k,t} - \epsilon_{k,t}$ and $\tilde{c}_l = \bar{c}_{l,t} - \epsilon_{l,t}$, we have:

$$
\begin{aligned}
&\mathrm{obj}_{x,t} - \mathrm{obj}_x \\
&\quad = [\frac{\tilde{c}_k - b}{\tilde{c}_k - \tilde{c}_l} \cdot \bar{r}_{l,t} + \frac{b - \tilde{c}_l}{\tilde{c}_k - \tilde{c}_l} \cdot \bar{r}_{k,t}] - [\frac{\mu_k^c - b}{\mu_k^c - \mu_l^c} \cdot \mu_l^r + \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \cdot \mu_k^r] \\
&\quad < [\frac{\tilde{c}_k - b}{\tilde{c}_k - \tilde{c}_l} \cdot (\mu_l^r + \epsilon_{l,t}) + \frac{b - \tilde{c}_l}{\tilde{c}_k - \tilde{c}_l} \cdot (\mu_k^r + \epsilon_{k,t})] - [\frac{\mu_k^c - b}{\mu_k^c - \mu_l^c} \cdot \mu_l^r + \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c} \cdot \mu_k^r] \\
&\quad = \frac{1}{\lambda} \cdot E_{x,t} + (\mu_k^r - \mu_l^r) \cdot [\frac{b - \tilde{c}_l}{\tilde{c}_k - \tilde{c}_l} - \frac{b - \mu_l^c}{\mu_k^c - \mu_l^c}] \\
&\quad = \frac{1}{\lambda} \cdot E_{x,t} + \frac{(\mu_k^r - \mu_l^r)}{(\tilde{c}_k - \tilde{c}_l) \cdot (\mu_k^c - \mu_l^c)} \cdot [(\mu_k^c - b)(\mu_l^c - \tilde{c}_l) + (b - \mu_l^c)(\mu_k^c - \tilde{c}_k)] \\
&\quad \leq \frac{1}{\lambda} \cdot E_{x,t} + \frac{\kappa}{\tilde{c}_k - \tilde{c}_l} \cdot |(\mu_k^c - b)(\mu_l^c - \tilde{c}_l) + (b - \mu_l^c)(\mu_k^c - \tilde{c}_k)| \\
&\quad \leq \frac{1}{\lambda} \cdot E_{x,t} + 2\frac{\kappa}{\tilde{c}_k - \tilde{c}_l} \cdot [(\tilde{c}_k - b) \cdot \epsilon_{l,t} + (b - \tilde{c}_l) \cdot \epsilon_{k,t}] \\
&\quad = \frac{1}{\lambda} \cdot E_{x,t} + \frac{2\kappa}{\lambda} \cdot E_{x,t} \\
&\quad = E_{x,t},
\end{aligned}
$$

a contradiction. The first inequality is strict because either $\tilde{c}_k > b$ or $\tilde{c}_l < b$. The second inequality is derived using Assumption 4.5. The third inequality is derived from the observation that the expression $(\mu_k^c - b)(\mu_l^c - \tilde{c}_l) + (b - \mu_l^c)(\mu_k^c - \tilde{c}_k)$ is a linear function of (μ_k^c, μ_l^c) (since the cross term $\mu_k^c \cdot \mu_l^c$ cancels out) so that $|(\mu_k^c - b)(\mu_l^c - \tilde{c}_l) + (b - \mu_l^c)(\mu_k^c - \tilde{c}_k)|$ is convex in (μ_k^c, μ_l^c) and the maximum of this expression over the polyhedron $[\bar{c}_{k,t} - \epsilon_{k,t}, \bar{c}_{k,t} + \epsilon_{k,t}] \times [\bar{c}_{l,t} - \epsilon_{l,t}, \bar{c}_{l,t} + \epsilon_{l,t}]$ is attained at an extreme point. We obtain:

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} I_{x \in \mathcal{B}_t} I_{\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}] \\
&\quad \leq \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{k,t} \geq \mu_k^r + \epsilon_{k,t}] + \mathbb{P}[\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}] \\
&\quad + \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{l,t} \geq \mu_l^c + \epsilon_{l,t}] + \mathbb{P}[\bar{c}_{k,t} \geq \mu_k^c + \epsilon_{k,t}] \\
&\quad + \sum_{t=1}^{\infty} \mathbb{P}[\bar{c}_{k,t} \leq \mu_k^c - \epsilon_{k,t}] + \mathbb{P}[\bar{c}_{l,t} \leq \mu_l^c - \epsilon_{l,t}] + \frac{4\pi^2}{3\epsilon^2} \leq \frac{3\pi^2}{\epsilon^2},
\end{aligned}
$$

using the same argument as in Fact C.3. □

C.5.5 Proof of Theorem 4.5

We build upon (4.4):

$$\begin{aligned}
R_{B,T} &\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\
&= T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \mathbb{E}[\sum_{t=\tau^*+1}^{T} r_{a_t,t}] + O(1) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \sigma \cdot \mathbb{E}[\sum_{t=\tau^*+1}^{T} c_{a_t,t}] + O(1) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \sigma \cdot \mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t} - B)_+] + O(1).
\end{aligned}$$

The second inequality is a consequence of Assumption 4.4:

$$\begin{aligned}
\mathbb{E}[\sum_{t=\tau^*+1}^{T} c_{a_t,t}] &= \mathbb{E}[\sum_{t=1}^{T} c_{a_t,t}] - \mathbb{E}[\sum_{t=1}^{\tau^*} c_{a_t,t}] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]] - \mathbb{E}[\sum_{t=1}^{\infty} I_{\tau^* \geq t} \cdot \mathbb{E}[c_{a_t,t} \mid \mathcal{F}_{t-1}]] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mu_{a_t}^c] - \mathbb{E}[\sum_{t=1}^{\infty} I_{\tau^* \geq t} \cdot \mu_{a_t}^c] \\
&= \mathbb{E}[\sum_{t=\tau^*+1}^{T} \mu_{a_t}^c] \\
&\geq \frac{1}{\sigma} \cdot \mathbb{E}[\sum_{t=\tau^*+1}^{T} \mu_{a_t}^r] = \frac{1}{\sigma} \cdot \mathbb{E}[\sum_{t=\tau^*+1}^{T} r_{a_t,t}],
\end{aligned}$$

since τ^* is a stopping time. To derive the third inequality, observe that $\tau^* = T + 1$ implies:

$$\sum_{t=\tau^*+1}^{T} c_{a_t,t} = 0 \leq (\sum_{t=1}^{T} c_{a_t,t} - B)_+,$$

while if $\tau^* < T + 1$ we have run out of resources before round T, i.e. $\sum_{t=1}^{\tau^*} c_{a_t,t} \geq B$, which implies:

$$\sum_{t=\tau^*+1}^{T} c_{a_t,t} \leq \sum_{t=\tau^*+1}^{T} c_{a_t,t} + \sum_{t=1}^{\tau^*} c_{a_t,t} - B \leq (\sum_{t=1}^{T} c_{a_t,t} - B)_+.$$

Now observe that:

$$\begin{aligned}
\mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t} - B)_+] &= \mathbb{E}[(\sum_{x \text{ pseudo-basis for (4.3)}} \{b_{x,T} - n_{x,T} \cdot b\})_+] \\
&\leq \sum_{x \in \mathcal{B}} \mathbb{E}[|b_{x,T} - n_{x,T} \cdot b|] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}] \\
&= O(\frac{K^2}{\epsilon^3} \ln(T)),
\end{aligned}$$

where we use the fact that $c_{k,t} \leq 1$ at any time t and for all arms k for the first inequality and Lemma 4.10 along with the proof of Lemma 4.11 for the last equality. Plugging this last inequality back into the regret bound yields:

$$\begin{aligned}
R_{B,T} &\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,T}^x] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&= \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (T - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \mathbb{E}[n_{x,T}]) - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] \quad \text{(C.22)} \\
&+ O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&= \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \mathbb{E}[n_{x,T}] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}]) \\
&- \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x \cdot \mathbb{E}[n_{x,T}] + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \quad \text{(C.23)} \\
&\leq 2^9 \frac{\lambda^2}{\epsilon^3} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(T) + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)),
\end{aligned}$$

where we use Lemma 4.11 for the third inequality, Lemma 4.10 along with $\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^{K} \xi_k^{x^*} \leq 1$ for the fourth inequality, and Lemma 4.12 for the last inequality.

C.5.6 Proof of Theorem 4.6

Along the same lines as for the case of a single limited resource, we start from inequality (C.23) derived in the proof of Theorem 4.5 and apply Lemma 4.12 only if Δ_x is big enough, taking into account the fact that $\sum_{x \in \mathcal{B}} \mathbb{E}[n_{x,T}] \leq T$. Specifically, we have:

$$\begin{aligned}
R_{B,T} &\leq \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, 2^9 \frac{\lambda^2}{\epsilon^3} \cdot \frac{\ln(T)}{\Delta_x} + \frac{10\pi^2}{3\epsilon^2} \cdot \Delta_x) \} \\
&\quad + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \{ \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, 2^9 \frac{\lambda^2}{\epsilon^3} \cdot \frac{\ln(T)}{\Delta_x}) \} + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \{ \sum_{x \in \mathcal{B}} \sqrt{2^9 \frac{\lambda^2}{\epsilon^3} \cdot \ln(T) \cdot n_x} \} + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq 2^5 \frac{\lambda}{\epsilon^{3/2}} \cdot \sqrt{\ln(T)} \cdot \sup_{\substack{(n_x)_{x \in \mathcal{B}} \geq 0 \\ \sum_{x \in \mathcal{B}} n_x \leq T}} \{ \sum_{x \in \mathcal{B}} \sqrt{n_x} \} + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)) \\
&\leq 2^5 \frac{\lambda}{\epsilon^{3/2}} \cdot \sqrt{|\mathcal{B}| \cdot T \cdot \ln(T)} + O(\frac{K^2 \cdot \sigma}{\epsilon^3} \ln(T)),
\end{aligned}$$

where we use the fact that $\Delta_x \leq \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^K \xi_k^{x^*} \leq 1$ for the second inequality, we maximize over each $\Delta_x \geq 0$ to derive the third inequality, and we use Cauchy-Schwartz for the last inequality.

C.5.7 Proof of Theorem 4.7

When $b \leq \epsilon/2$, the analysis almost falls back to the case of a single limited resource. Indeed, we have $\tau^* = \tau(B)$ with high probability given that:

$$\begin{aligned}
\mathbb{P}[\tau(B) > T + t] &= \mathbb{P}[\sum_{\tau=1}^{T+t} c_{a_\tau,\tau} \leq B] \\
&\leq \mathbb{P}[\frac{1}{T+t} \cdot \sum_{t=1}^{T+t} c_{a_t,t} \leq \epsilon - (\epsilon - b)] \\
&\leq \exp(-(T+t) \cdot \frac{\epsilon^2}{2}),
\end{aligned}$$

for any $t \in \mathbb{N}$ using Lemma 4.1. Now observe that, since $b \leq \epsilon/2$, the feasible bases for (4.3) are exactly the bases x such that $\mathcal{K}_x = \{k\}$ and $\mathcal{C}_x = \{1\}$ for some $k \in \{1, \cdots, K\}$, which we denote by $(x_k)_{k=1,\cdots,K}$. This shows that $\mathrm{ER}_{\mathrm{OPT}}(B,T) = B \cdot \max_{k=1,\cdots,K} \mu_k^r/\mu_k^c$. Moreover note that Assumption 4.6 is automatically satisfied when $b \leq \epsilon/2$. Hence, with a minor modification of the proof of Lemma 4.10, we get:

$$\mathbb{E}[n_{x,\tau(B)}] \leq \frac{2^{12}}{\epsilon^3} \cdot \mathbb{E}[\ln(\tau(B))] + \frac{40\pi^2}{3\epsilon^2},$$

for any pseudo-basis x involving two arms or any basis x such that $\mathcal{K}_x = \{k\}$ and $\mathcal{C}_x = \{2\}$ for some arm $k \in \{1, \cdots, K\}$. Similarly, a minor modification of Lemma 4.12 yields:

$$\mathbb{E}[n_{x_k,\tau(B)}] \leq 2^{12}\frac{\lambda^2}{\epsilon^3} \cdot \frac{\mathbb{E}[\ln(\tau(B))]}{\Delta_{x_k}^2} + \frac{40\pi^2}{\epsilon^2},$$

for any $k \in \{1, \cdots, K\}$. What is left is to refine the analysis of Theorem 4.5 as follows:

$$\begin{aligned}
R_{B,T} &\leq B \cdot \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\
&\leq B \cdot \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c} - \mathbb{E}[\sum_{t=1}^{\tau(B)} r_{a_t,t}] + T \cdot \mathbb{P}[\tau(B) > T] + \mathbb{E}[(\tau(B) - T)_+] + O(1) \\
&\leq B \cdot \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c} - \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + O(1) \\
&\leq \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c} \cdot (B - \sum_{k \mid \Delta_{x_k}=0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}]) - \sum_{k \mid \Delta_{x_k}>0} \mu_k^r \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + O(1).
\end{aligned}$$

By definition of $\tau(B)$, we have $B \leq \sum_{t=1}^{\tau(B)} c_{a_t,t}$. Taking expectations on both sides yields:

$$\begin{aligned}
B &\leq \sum_{k=1}^{K} \mu_k^c \cdot \mathbb{E}[n_{k,\tau(B)}] \\
&= \sum_{k \mid \Delta_{x_k}=0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + \sum_{k \mid \Delta_{x_k}>0} \mu_k^c \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,\tau(B)}].
\end{aligned}$$

Plugging this inequality back into the regret bound, we get:

$$\begin{aligned}
R_{B,T} &= \sum_{k \mid \Delta_{x_k}>0} \mu_k^c \cdot \Delta_{x_k} \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + \max_{k=1,\cdots,K} \frac{\mu_k^r}{\mu_k^c} \cdot \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,\tau(B)}] + O(1) \\
&\leq \sum_{k \mid \Delta_{x_k}>0} \mu_k^c \cdot \Delta_{x_k} \cdot \mathbb{E}[n_{k,\tau(B)}^{x_k}] + \frac{2^{12}K^2 \cdot \kappa}{\epsilon^3} \cdot \mathbb{E}[\ln(\tau(B))] + O(1) \\
&\leq 2^{12}\frac{\lambda^2}{\epsilon^3} \cdot (\sum_{k \mid \Delta_{x_k}>0} \frac{1}{\Delta_{x_k}}) \cdot \ln(\frac{B+1}{\epsilon}) + \frac{2^{12}K^2 \cdot \kappa}{\epsilon^3} \cdot \ln(\frac{B+1}{\epsilon}) + O(1),
\end{aligned}$$

where we use Assumption 4.5 for the second inequality and Lemma 4.4 for the third inequality. A distribution-independent regret bound of order $O(\sqrt{K \cdot \frac{B+1}{\epsilon} \cdot \ln(\frac{B+1}{\epsilon})} + \frac{K^2 \cdot \kappa}{\epsilon^3} \cdot \ln(\frac{B+1}{\epsilon}))$ can be derived from the penultimate inequality along the same lines as in Theorem 4.6.

C.6 Proofs for Section 4.7

C.6.1 Preliminary work for the proofs of Section 4.7

Concentration inequality. We will use the following inequality repeatedly. For a given round $\tau \geq \mathrm{t_{ini}}$ and a basis x:

$$\begin{aligned}
&\mathbb{P}[\exists (k,i) \in \mathcal{K}_x \times \{1,\cdots,C\}, |\bar{c}_{k,\tau}(i) - \mu_k^c(i)| > \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}] \\
&\quad \leq \sum_{s=\mathrm{t_{ini}}/K}^{T} \sum_{\substack{k \in \mathcal{K}_x \\ i \in \{1,\cdots,C\}}} \mathbb{P}[|\bar{c}_{k,\tau}(i) - \mu_k^c(i)| > \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}, n_{k,\tau} = s] \\
&\quad \leq 2 \cdot C^2 \cdot \sum_{s=\mathrm{t_{ini}}/K}^{\infty} \exp(-s \cdot \frac{\epsilon^6}{2^7 \cdot (C+2)!^4}) \\
&\quad \leq 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6 \cdot \mathrm{T}^2},
\end{aligned} \tag{C.24}$$

using Lemma 4.1, the inequality $\exp(-x) \leq 1 - x/2$ for $x \in [0,1]$, and the fact that we pull each arm $\mathrm{t_{ini}}/K \geq 2^8 \frac{(\mathrm{C}+2)!^4}{\epsilon^6} \cdot \ln(T)$ times during the initialization phase.

Useful matrix inequalities. For any basis x, assume that $\{|\bar{c}_{k,t}(i) - \mu_k^c(i)| \leq \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}\}$, for any arm $k \in \mathcal{K}_x$ and resource $i \in \mathcal{C}_x$. We have:

$$
\begin{aligned}
&|\det(\bar{A}_{x,t}) - \det(A_x)| \\
&\quad = |\sum_{\sigma \in S(\mathcal{K}_x, \mathcal{C}_x)} [\prod_{k \in \mathcal{K}_x} \bar{c}_{k,t}(\sigma(k)) - \prod_{k \in \mathcal{K}_x} \mu_k^c(\sigma(k))]| \\
&\quad = |\sum_{\sigma \in S(\mathcal{K}_x, \mathcal{C}_x)} \sum_{l \in \mathcal{K}_x} \prod_{k<l} \bar{c}_{k,t}(\sigma(k)) \cdot [\bar{c}_{l,t}(\sigma(l)) - \mu_l^c(\sigma(l))] \cdot \prod_{k>l} \mu_k^c(\sigma(k))| \\
&\quad \leq \sum_{\sigma \in S(\mathcal{K}_x, \mathcal{C}_x)} \sum_{l \in \mathcal{K}_x} |\bar{c}_{l,t}(\sigma(l)) - \mu_l^c(\sigma(l))| \\
&\quad \leq \frac{\epsilon^3}{16 \cdot (C+2)!} \\
&\quad \leq \frac{\epsilon}{2},
\end{aligned}
$$

since the amounts of resources consumed at any round are no larger than 1. This yields:

$$
|\det(\bar{A}_{x,t})| \geq |\det(A_x)| - |\det(\bar{A}_{x,t}) - \det(A_x)| \geq \frac{\epsilon}{2}, \tag{C.25}
$$

using Assumption 4.8. Now consider any vector c such that $\|c\|_\infty \leq 1$, we have:

$$
\begin{aligned}
&|c^\intercal \bar{A}_{x,t}^{-1} b_{\mathcal{K}_x} - c^\intercal A_x^{-1} b_{\mathcal{K}_x}| \\
&\quad = \frac{1}{|\det(A_x)| \cdot |\det(\bar{A}_{x,t})|} | \cdot |\det(A_x) \cdot c^\intercal \operatorname{adj}(\bar{A}_{x,t}) b_{\mathcal{K}_x} - \det(\bar{A}_{x,t}) \cdot c^\intercal \operatorname{adj}(A_x) b_{\mathcal{K}_x}| \\
&\quad \leq \frac{1}{|\det(\bar{A}_{x,t})|} | \cdot |c^\intercal (\operatorname{adj}(\bar{A}_{x,t}) - \operatorname{adj}(A_x)) b_{\mathcal{K}_x}| \\
&\quad + \frac{1}{|\det(A_x)| \cdot |\det(\bar{A}_{x,t})|} | \cdot |\det(\bar{A}_{x,t}) - \det(A_x)| \cdot |c^\intercal \operatorname{adj}(A_x) b_{\mathcal{K}_x}| \\
&\quad \leq \frac{\epsilon^2}{8} + \frac{\epsilon}{8 \cdot (C+2)!} \cdot \|c\|_2 \cdot \|\operatorname{adj}(A_x) b_{\mathcal{K}_x}\|_2 \\
&\quad \leq \frac{\epsilon}{4}.
\end{aligned}
$$

The second inequality is obtained using Assumption 4.8 and (C.25) by proceeding along the same lines as above to bound the difference between two determinants for each component of $\operatorname{adj}(\bar{A}_{x,t}) - \operatorname{adj}(A_x)$. The last inequality is obtained using $\|c\|_2 \leq \sqrt{C}$ and the fact that

each component of A_x is smaller than 1. If we take $c = e_k$ for $k \in \mathcal{K}_x$, this yields:

$$|\xi^x_{k,t} - \xi^x_k| \leq \frac{\epsilon}{4}. \tag{C.26}$$

If we take $c = (\mu^c_k(i))_{k\in\mathcal{K}_x}$, for any $i \in \{1, \cdots, C\}$, we get:

$$|\sum_{k\in\mathcal{K}_x} \mu^c_k(i) \cdot \xi^x_{k,t} - \sum_{k\in\mathcal{K}_x} \mu^c_k(i) \cdot \xi^x_k| \leq \frac{\epsilon}{4} \tag{C.27}$$

and

$$\begin{aligned}
|\sum_{k\in\mathcal{K}_x} \bar{c}_{k,t}(i) \cdot \xi^x_{k,t} - \sum_{k\in\mathcal{K}_x} \mu^c_k(i) \cdot \xi^x_k| &= |\bar{c}^\intercal \bar{A}^{-1}_{x,t} b_{\mathcal{K}_x} - c^\intercal A^{-1}_x b_{\mathcal{K}_x}| \\
&\leq |(\bar{c} - c)^\intercal \bar{A}^{-1}_{x,t} b_{\mathcal{K}_x}| + |c^\intercal \bar{A}^{-1}_{x,t} b_{\mathcal{K}_x} - c^\intercal A^{-1}_x b_{\mathcal{K}_x}| \\
&\leq \|\bar{c} - c\|_2 \cdot \frac{1}{|\det(\bar{A}_{x,t})|} \cdot \left\|\mathrm{adj}(\bar{A}_{x,t}) b_{\mathcal{K}_x}\right\|_2 + \frac{\epsilon}{4} \\
&\leq \sqrt{C} \cdot \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2} \cdot \frac{2}{\epsilon} \cdot C! + \frac{\epsilon}{4} \\
&\leq \frac{\epsilon}{2},
\end{aligned} \tag{C.28}$$

where $\bar{c} = (\bar{c}_{k,t}(i))_{k\in\mathcal{K}_x}$.

C.6.2 Proof of Lemma 4.13

First, consider a basis $x \notin \mathcal{B}$. Since x is ϵ-non-degenerate by Assumption 4.8, there must exist $k \in \mathcal{K}_x$ such that $\xi^x_k \leq -\epsilon$ or $i \in \{1, \cdots, C\}$ such that $\sum_{k=1}^K \mu^c_k(i) \cdot \xi^x_k \geq b(i) + \epsilon$. Let us assume that we are in the first situation (the proof is symmetric in the other scenario). Using (C.26) in the preliminary work of Section C.6.1, we have:

$$\xi^x_{k,t} = \xi^x_k + (\xi^x_{k,t} - \xi^x_k) \leq -\frac{\epsilon}{2}, \tag{C.29}$$

if $\{|\bar{c}_{k,t}(i)-\mu_k^c(i)| \leq \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}\}$ for all arms $k \in \mathcal{K}_x$ and resources $i \in \{1,\cdots,C\}$. Hence:

$$\begin{aligned}
\mathbb{E}[n_{x,T}] &= \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x}] \\
&\leq \sum_{t=\mathrm{t_{ini}}}^{T} \mathbb{P}[\xi_{k,t}^x \geq 0] \\
&\leq \sum_{t=\mathrm{t_{ini}}}^{T} \sum_{\substack{k\in\mathcal{K}_x \\ i\in\{1,\cdots,C\}}} \mathbb{P}[|\bar{c}_{k,t}(i) - \mu_k^c(i)| > \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}] \\
&\leq 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6},
\end{aligned}$$

where the third inequality is derived with (C.24).

Second, consider a pseudo-basis x for (4.3) that is not a basis. Since $\det(A_x) = 0$, either every component of A_x is 0, in which case $\det(\bar{A}_{x,t}) = 0$ at every round t and x can never be selected, or there exists a basis $\tilde{x}$ for (4.3) with $\mathcal{K}_{\tilde{x}} \subset \mathcal{K}_x$ and $\mathcal{C}_{\tilde{x}} \subset \mathcal{C}_x$ along with coefficients $(\alpha_{kl})_{k\in\mathcal{K}_x-\mathcal{K}_{\tilde{x}}, l\in\mathcal{K}_{\tilde{x}}}$ such that $\mu_k^c(i) = \sum_{l\in\mathcal{K}_{\tilde{x}}} \alpha_{kl}\cdot\mu_l^c(i)$ for any resource $i \in \mathcal{C}_x$. Assuming we are in the second scenario and since $\tilde{x}$ is ϵ-non-degenerate by Assumption 4.8, we have $\sum_{k=1}^{K} \mu_k^c(j)\cdot\xi_k^{\tilde{x}} \leq b(j) - \epsilon$ for any $j \notin \mathcal{C}_{\tilde{x}}$. Take $i \in \mathcal{C}_x - \mathcal{C}_{\tilde{x}}$. Suppose that x is feasible for (4.8) at round t and assume by contradiction that $\{|\bar{c}_{k,t}(j) - \mu_k^c(j)| \leq \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}\}$ for all arms $k \in \mathcal{K}_x$ and resources $j \in \{1,\cdots,C\}$. Using the notations $\tilde{\xi}_{k,t}^x = \xi_{k,t}^x + \sum_{l\in\mathcal{K}_x-\mathcal{K}_{\tilde{x}}} \alpha_{kl}\cdot\xi_{l,t}^x$, we have, for any resource $j \in \mathcal{C}_{\tilde{x}}$:

$$\begin{aligned}
b(j) &= \sum_{l\in\mathcal{K}_x} \bar{c}_{l,t}(j)\cdot\xi_{l,t}^x \\
&= \alpha(j) + \sum_{l\in\mathcal{K}_x} \mu_l^c(j)\cdot\xi_{l,t}^x \\
&= \alpha(j) + \sum_{l\in\mathcal{K}_{\tilde{x}}} \mu_l^c(j)\cdot\tilde{\xi}_{l,t}^x,
\end{aligned}$$

where $|\alpha(j)| \leq \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}$ since $\xi_k \in [0,1]\ \forall k \in \{1,\cdots,K\}$ for any feasible solution to (4.8). We get $A_{\tilde{x}}\tilde{\xi}^x_{\mathcal{K}_{\tilde{x}},t} = b_{\mathcal{C}_{\tilde{x}}} - \alpha_{\mathcal{C}_{\tilde{x}}}$ while $A_{\tilde{x}}\xi^{\tilde{x}}_{\mathcal{K}_{\tilde{x}}} = b_{\mathcal{C}_{\tilde{x}}}$. We derive:

$$
\begin{aligned}
&|\sum_{k=1}^{K}\bar{c}_{k,t}(i)\cdot\xi^x_{k,t} - \sum_{k=1}^{K}\mu^c_k(i)\cdot\xi^{\tilde{x}}_k| \\
&\quad\leq |\sum_{k=1}^{K}\mu^c_k(i)\cdot\xi^x_{k,t} - \sum_{k=1}^{K}\mu^c_k(i)\cdot\xi^{\tilde{x}}_k| + \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2} \\
&\quad\leq |\sum_{k\in\mathcal{K}_{\tilde{x}}}\mu^c_k(i)\cdot(\tilde{\xi}^x_{k,t} - \xi^{\tilde{x}}_k)| + \frac{\epsilon}{4} \\
&\quad= \frac{\epsilon}{4} + |\mu^c_{\mathcal{K}_{\tilde{x}}}(i)^\intercal A^{-1}_{\tilde{x}}\alpha_{\mathcal{C}_{\tilde{x}}}| \\
&\quad\leq \frac{\epsilon}{4} + \sqrt{C}\cdot\frac{1}{|\det(A_{\tilde{x}})|}\cdot\|\mathrm{adj}(A_{\tilde{x}})\alpha_{\mathcal{C}_{\tilde{x}}}\|_2 \\
&\quad\leq \frac{\epsilon}{4} + \frac{(C+1)!}{\epsilon}\cdot\frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2} \\
&\quad\leq \frac{\epsilon}{2}.
\end{aligned}
$$

Thus we obtain:

$$
\sum_{k=1}^{K}\bar{c}_{k,t}(i)\cdot\xi^x_{k,t} \leq \sum_{k=1}^{K}\mu^c_k(i)\cdot\xi^{\tilde{x}}_k + \frac{\epsilon}{2} \leq b(i) - \frac{\epsilon}{2} < b(i),
$$

a contradiction since this inequality must be binding by definition if x is selected at round t. We finally conclude:

$$
\begin{aligned}
\mathbb{E}[n_{x,T}] &= \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x}] \\
&\leq \sum_{t=\mathrm{t_{ini}}}^{T}\sum_{\substack{k\in\mathcal{K}_x \\ j\in\{1,\cdots,C\}}}\mathbb{P}[|\bar{c}_{k,t}(j) - \mu^c_k(j)| > \frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}] \\
&\leq 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6}.
\end{aligned}
$$

C.6.3 Proof of Lemma 4.14

Proof of (4.12)**.** Consider a resource $i \in \mathcal{C}_x$ and $u \geq 1$. We study $\mathbb{P}[b_{x,t}(i) - n_{x,t}\cdot b(i) \geq u]$ but the same technique can be used to bound $\mathbb{P}[b_{x,t}(i) - n_{x,t}\cdot b(i) \leq -u]$. If $b_{x,t}(i) - n_{x,t}\cdot$

$b(i) \geq u$, it must be that $e^x_{i,\tau} = -1$ for at least $s \geq \lfloor u \rfloor$ rounds $\tau = t_1 \leq \cdots \leq t_s \leq t-1$ where x was selected at Step-Simplex since the last time, denoted by $t_0 < t_1$, where x was selected at Step-Simplex and the budget was below the target, i.e. $b_{x,t_0}(i) \leq n_{x,t_0} \cdot b(i)$ (because the amounts of resources consumed at each round are bounded by 1). Moreover, we have:

$$\begin{aligned}
\sum_{q=1}^{s} c_{a_{t_q},t_q}(i) &= \sum_{\tau=t_0+1}^{t-1} I_{x_\tau=x} \cdot c_{a_\tau,\tau}(i) \\
&= b_{x,t}(i) - b_{x,t_0+1}(i) \\
&\geq (n_{x,t} - n_{x,t_0}) \cdot b(i) + u - 1 \\
&\geq s \cdot b(i) + u - 1.
\end{aligned}$$

Hence:

$$\begin{aligned}
&\mathbb{P}[b_{x,t}(i) - n_{x,t} \cdot b(i) \geq u] \\
&\leq \sum_{s=\lfloor u \rfloor}^{t} \mathbb{P}[\sum_{q=1}^{s} c_{a_{t_q},t_q}(i) \geq s \cdot b(i) + u - 1 \ ; \ e^x_{i,t_q} = -1 \ \forall q \in \{1, \cdots, s\}] \\
&\leq \sum_{s=\lfloor u \rfloor}^{t} \mathbb{P}[\sum_{q=1}^{s} c_{a_{t_q},t_q}(i) \geq s \cdot b(i) \ ; \ \sum_{k=1}^{K} \mu^c_k(i) \cdot p^x_{k,t_q} \leq b(i) - \frac{\epsilon^2}{4 \cdot (\mathrm{C}+1)!} \ \forall q \in \{1, \cdots, s\}] \\
&+ \sum_{\tau=\mathrm{t_{ini}}}^{T} \sum_{\substack{k \in \mathcal{K}_x \\ j \in \{1,\cdots,C\}}} \mathbb{P}[|\bar{c}_{k,\tau}(j) - \mu^c_k(j)| > \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}] \\
&\leq \sum_{s=\lfloor u \rfloor}^{\infty} \exp(-2s \cdot (\frac{\epsilon^2}{4 \cdot (\mathrm{C}+1)!})^2) + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6} \cdot \frac{1}{T} \\
&\leq 16 \frac{(C+1)!^2}{\epsilon^4} \cdot \exp(-u \cdot (\frac{\epsilon^2}{4 \cdot (\mathrm{C}+1)!})^2) + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6} \cdot \frac{1}{T}.
\end{aligned}$$

The last inequality is obtained using $\exp(-x) \leq 1 - x/2$ for $x \in [0,1]$. The third inequality is derived using Lemma 4.1 for the first term and (C.24) for the second term. The second inequality is obtained by observing that if x was selected at time τ and $|\bar{c}_{k,\tau}(j) - \mu^c_k(j)| \leq \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}$ for any arm $k \in \mathcal{K}_x$ and resource $j \in \{1, \cdots, C\}$, then we must have $\delta^*_{x,\tau} \geq$

$\frac{\epsilon^2}{4\cdot(C+1)!}$. Indeed, using (C.25) and (C.26), we have, for $\delta \leq \frac{\epsilon^2}{4\cdot(C+1)!}$ and any arm $k \in \mathcal{K}_x$:

$$\begin{aligned}
c^\intercal \bar{A}_{x,\tau}^{-1}(b_{\mathcal{C}_x} + \delta \cdot e_\tau^x) &= \xi_{k,\tau}^x + \delta \cdot c^\intercal \bar{A}_{x,\tau}^{-1} e_\tau^x \\
&\geq \xi_k^x - |\xi_k^x - \xi_{k,\tau}^x| - \delta \cdot |c^\intercal \bar{A}_{x,\tau}^{-1} e_\tau^x| \\
&\geq \frac{\epsilon}{2} - \delta \cdot \left\| \bar{A}_{x,\tau}^{-1} e_\tau^x \right\|_2 \\
&\geq \frac{\epsilon}{2} - \delta \cdot \frac{1}{|\det(\bar{A}_{x,\tau})|} \cdot \left\| \mathrm{adj}(\bar{A}_{x,\tau}) e_\tau^x \right\|_2 \\
&\geq \frac{\epsilon}{2} - \frac{2\delta}{\epsilon} \cdot \sqrt{C} \cdot C! \\
&\geq 0,
\end{aligned}$$

where $c = e_k$ and since x is ϵ-non-degenerate by Assumption 4.8. Similarly, using (C.25) and (C.28), we have, for $\delta \leq \frac{\epsilon^2}{4\cdot(C+1)!}$ and any resource $j \notin \mathcal{C}_x$:

$$\begin{aligned}
c^\intercal \bar{A}_{x,\tau}^{-1}(b_{\mathcal{C}_x} + \delta \cdot e_\tau^x) &= \sum_{k \in \mathcal{K}_x} \bar{c}_{k,\tau}(j) \cdot \xi_{k,\tau}^x + \delta \cdot c^\intercal \bar{A}_{x,\tau}^{-1} e_\tau^x \\
&\leq \sum_{k \in \mathcal{K}_x} \mu_k^c(j) \cdot \xi_{k,\tau}^x + |\sum_{k \in \mathcal{K}_x} \bar{c}_{k,\tau}(j) \cdot \xi_{k,\tau}^x - \sum_{k \in \mathcal{K}_x} \mu_k^c(j) \cdot \xi_{k,\tau}^x| \\
&+ \delta \cdot |c^\intercal \bar{A}_{x,\tau}^{-1} e_\tau^x| \\
&\leq b(j) - \frac{\epsilon}{2} + \delta \cdot \sqrt{C} \cdot \left\| \bar{A}_{x,\tau}^{-1} e_\tau^x \right\|_2 \\
&\leq b(j) - \frac{\epsilon}{2} + \frac{2\delta}{\epsilon} \cdot (C+1)! \\
&\leq b(j),
\end{aligned}$$

where $c = (\bar{c}_{k,\tau}(j))_{k \in \mathcal{K}_x}$ and since x is ϵ-non-degenerate by Assumption 4.8.

Proof of (4.14)**.** First observe that, using (4.12), we have:

$$\begin{aligned}
&\max_{i\in\mathcal{C}_x} |\mathbb{E}[b_{x,T}(i)] - \mathbb{E}[n_{x,T}]\cdot b(i)| \\
&\quad\leq \mathbb{E}[\max_{i\in\mathcal{C}_x} |b_{x,T}(i) - n_{x,T}\cdot b(i)|] \\
&\quad= \int_0^T \mathbb{P}[\max_{i\in\mathcal{C}_x} |b_{x,T}(i) - n_{x,T}\cdot b(i)| \geq u]\mathrm{d}u \\
&\quad= \sum_{i\in\mathcal{C}_x} \int_0^T \mathbb{P}[|b_{x,T}(i) - n_{x,T}\cdot b(i)| \geq u]\mathrm{d}u \\
&\quad\leq 32C\cdot\frac{(C+1)!^2}{\epsilon^4}\cdot\int_0^T \exp(-u\cdot(\frac{\epsilon^2}{4\cdot(\mathrm{C}+1)!})^2)\mathrm{d}u + C + C\cdot 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\
&\quad= 2^9 C\cdot\frac{(C+1)!^4}{\epsilon^8} + C + C\cdot 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\
&\quad\leq 2^{10}C\cdot\frac{(C+3)!^4}{\epsilon^8}.
\end{aligned}$$

Now observe that, for any resource $i\in\mathcal{C}_x$, we have $\mathbb{E}[b_{x,T}(i)] = \sum_{k\in\mathcal{K}_x}\mu_k^c(i)\cdot\mathbb{E}[n_{k,T}^x]$. Hence, defining the vector $p = (\frac{\mathbb{E}[n_{k,T}^x]}{\mathbb{E}[n_{x,T}]})_{k\in\mathcal{K}_x}$, we get:

$$\begin{aligned}
\mathbb{E}[n_{x,T}]\cdot\|p-\xi^x\|_2 &= \mathbb{E}[n_{x,T}]\cdot\left\|A_x^{-1}A_x(p-\xi^x)\right\|_2 \\
&\leq \left\|\left|A_x^{-1}\right|\right\|_2\cdot\|\mathbb{E}[n_{x,T}]\cdot A_x(p-\xi^x)\|_2 \\
&= \frac{1}{|\det(A_x)|}\cdot\|\mathrm{adj}(A_x)\|_2\cdot\|(\mathbb{E}[b_{x,T}(i)])_{i\in\mathcal{C}_x} - (\mathbb{E}[n_{x,T}]\cdot b(i))_{i\in\mathcal{C}_x}\|_2 \\
&\leq \frac{1}{\epsilon}\cdot(C+1)!\cdot\sqrt{C}\cdot 2^{10}C\cdot\frac{(C+3)!^4}{\epsilon^8} \\
&\leq 2^{10}\frac{(C+3)!^5}{\epsilon^9},
\end{aligned}$$

using Assumption 4.8. Finally we obtain:

$$\begin{aligned}
\mathbb{E}[n_{x,T}]\cdot\xi_k^x - \mathbb{E}[n_{k,T}^x] &\leq \mathbb{E}[n_{x,T}]\cdot\|p-\xi^x\|_2 \\
&\leq 2^{10}\frac{(C+3)!^5}{\epsilon^9},
\end{aligned}$$

for any arm $k\in\mathcal{K}_x$.

Proof of (4.13)**.** Consider a resource $i \notin \mathcal{C}_x$ and assume that $b_{x,t}(i) - n_{x,t} \cdot b(i) \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)$. By contradiction, suppose that:

- $|b_{x,t}(j) - n_{x,t} \cdot b(j)| \leq 16 \frac{(C+1)!^2}{\epsilon^4} \cdot \ln(T)$ for all resources $j \in \mathcal{C}_x$,

- $|\mu_k^c(j) - \tilde{c}_k(j)| \leq \frac{\epsilon^2}{8 \cdot (C+2)!}$ for all resources $j \in \{1, \cdots, C\}$ and for all arms $k \in \mathcal{K}_x$ such that $n_{k,t}^x \geq 2^6 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T)$, where $\tilde{c}_k(j)$ denotes the empirical average amount of resource j consumed when selecting basis x and pulling arm k, i.e. $\tilde{c}_k(j) = \frac{1}{n_{k,t}^x} \cdot \sum_{\tau=\mathrm{t_{ini}}}^{t-1} I_{x_\tau = x} \cdot I_{a_\tau = k} \cdot c_{k,\tau}(j)$.

Observe that if $b_{x,t}(i) - n_{x,t} \cdot b(i) \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)$, it must be that x has been selected at least $2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)$ times at Step-Simplex since $\mathrm{t_{ini}}$, i.e. $n_{x,t} \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)$. We can partition $\mathcal{K}_x$ into two sets $\mathcal{K}_x^{\mathbb{1}}$ and $\mathcal{K}_x^2$ such that $n_{k,t}^x \geq 2^6 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T)$ for all $k \in \mathcal{K}_x^{\mathbb{1}}$ and $n_{k,t}^x < 2^6 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T)$ for all $k \in \mathcal{K}_x^2$. We get, for any $j \in \mathcal{C}_x$:

$$\begin{aligned}
16 \frac{(C+1)!^2}{\epsilon^4} \cdot \ln(T) &\geq |b_{x,t}(j) - n_{x,t} \cdot b(j)| \\
&\geq n_{x,t} \cdot |\sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(j) \cdot p_k - b(j)| - \sum_{k \in \mathcal{K}_x^2} n_{k,t}^x \\
&\geq n_{x,t} \cdot |\sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(j) \cdot p_k - b(j)| - 2^6 C \cdot \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T),
\end{aligned}$$

where $p_k = \frac{n_{k,t}^x}{n_{x,t}}$ for $k \in \mathcal{K}_1^x$ and $p_k = 0$ otherwise. Hence:

$$\begin{aligned}
|\sum_{k \in \mathcal{K}_x} \mu_k^c(j) \cdot p_k - b(j)| &\leq \max_{k \in \mathcal{K}_x^{\mathbb{1}}} |\mu_k^c(j) - \tilde{c}_k(j)| + |\sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(j) \cdot p_k - b(j)| \\
&\leq \frac{\epsilon^2}{8 \cdot (C+2)!} + \frac{(16 \frac{(C+1)!^2}{\epsilon^4} + 2^6 C \cdot \frac{(C+2)!^2}{\epsilon^4}) \cdot \ln(T)}{n_{x,t}} \\
&\leq \frac{\epsilon^2}{4 \cdot (C+2)!},
\end{aligned}$$

where we use the fact that $\sum_{k=1}^{K} p_k \leq 1$ and $p_k \geq 0$ for any arm k for the first inequality and $n_{x,t} \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)$ for the last one. We get:

$$\begin{aligned}
\|p - \xi^x\|_2 &= \left\|A_x^{-1} A_x (p - \xi^x)\right\|_2 \\
&\leq \left\|\left\|A_x^{-1}\right\|\right\|_2 \cdot \|A_x(p - \xi^x)\|_2 \\
&\leq \frac{1}{|\det(A_x)|} \cdot \left\|\left\|\mathrm{adj}(A_x)\right\|\right\|_2 \cdot \sqrt{C} \cdot \frac{\epsilon^2}{4 \cdot (C+2)!} \\
&\leq \frac{1}{\epsilon} \cdot (C+1)! \cdot \sqrt{C} \cdot \frac{\epsilon^2}{4 \cdot (C+2)!} \\
&\leq \frac{\epsilon}{4 \cdot \sqrt{C}},
\end{aligned}$$

using Assumption 4.8. Hence:

$$\begin{aligned}
2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T) &\leq b_{x,t}(i) - n_{x,t} \cdot b(i) \\
&\leq n_{x,t} \cdot \Big(\sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(i) \cdot p_k - b(i) \Big) + \sum_{k \in \mathcal{K}_x^2} n_{k,t}^x \\
&\leq n_{x,t} \cdot \Big(\sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(i) \cdot p_k - b(i) \Big) + 2^6 C \cdot \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T).
\end{aligned}$$

Using the shorthand notation $c = (\mu_k^c(i))_{k \in \mathcal{K}_x}$, this implies:

$$\begin{aligned}
0 &\leq \sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_k(i) \cdot p_k - b(i) \\
&\leq \sum_{k \in \mathcal{K}_x^{\mathbb{1}}} \mu_k^c(i) \cdot p_k + \frac{\epsilon^2}{8 \cdot (C+2)!} - \Big(\sum_{k \in \mathcal{K}_x} \mu_k^c(i) \cdot \xi_k^x + \epsilon \Big) \\
&\leq c^{\intercal}(p - \xi^x) - \frac{\epsilon}{2} \\
&\leq \sqrt{C} \cdot \|p - \xi^x\|_2 - \frac{\epsilon}{2} \\
&< 0,
\end{aligned}$$

a contradiction. Note that we use the fact that $\sum_{k=1}^{K} p_k \leq 1$ and $p_k \geq 0$ for any arm k and Assumption 4.8 for the second inequality. We conclude that:

$$\begin{aligned}
&\mathbb{P}[b_{x,t}(i) - n_{x,t} \cdot b(i) \geq 2^8 \frac{(C+3)!^3}{\epsilon^6} \cdot \ln(T)] \\
&\leq \sum_{j \in \mathcal{C}_x} \mathbb{P}[|b_{x,t}(j) - n_{x,t} \cdot b(j)| \geq 16 \frac{(C+1)!^2}{\epsilon^4} \cdot \ln(T)] \\
&+ \sum_{\substack{k \in \mathcal{K}_x \\ j \in \{1,\cdots,C\}}} \mathbb{P}[|\mu_k^c(j) - \tilde{c}_k(j)| \geq \frac{\epsilon^2}{8 \cdot (C+2)!} \; ; \; n_{k,t}^x \geq 2^6 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(T)] \\
&\leq 2^5 C \cdot \frac{(C+1)!^2}{\epsilon^4 \cdot T} + C \cdot T \cdot 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6 \cdot \mathrm{T}^2} \\
&+ \sum_{\substack{k \in \mathcal{K}_x \\ j \in \{1,\cdots,C\}}} \sum_{s = 2^6 \cdot (C+2)!^2 / \epsilon^4 \cdot \ln(T)} \mathbb{P}[|\mu_k^c(j) - \tilde{c}_k(j)| \geq \frac{\epsilon^2}{8 \cdot (C+2)!} \; ; \; n_{k,t}^x = s] \\
&\leq 2C \cdot T \cdot 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6 \cdot \mathrm{T}^2} + 2^8 C^2 \cdot \frac{(C+2)!^2}{\epsilon^4 \cdot T^2} \\
&\leq 2^{10} \frac{(C+4)!^4}{\epsilon^6 \cdot T},
\end{aligned}$$

where we use (4.12) for the second inequality and Lemma 4.1 for the third inequality.

C.6.4 Proof of Lemma 4.15

Consider any suboptimal basis $x \in \mathcal{B}$. The proof is along the same lines as for Lemmas 4.5, 4.8, and 4.12. We break down the analysis in a series of facts where we emphasize the main differences. We start off with an inequality similar to Fact C.1. We use the shorthand notation $\beta_x = 2^{10} \frac{(C+3)!^3}{\epsilon^6} \cdot (\frac{\lambda}{\Delta_x})^2$.

Fact C.11.

$$\begin{aligned}
\mathbb{E}[n_{x,T}] \leq & 2\beta_x \cdot \ln(T) + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\
& + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t = x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}]. \qquad \text{(C.30)}
\end{aligned}$$

Proof. Similarly as in Fact C.1, we have:

$$\mathbb{E}[n_{x,T}] \leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)}].$$

This yields:

$$\mathbb{E}[n_{x,T}] \leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}] + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x^* \notin \mathcal{B}_t}].$$

Using (C.26), (C.28), and Assumption 4.8, we have:

$$\xi_{k,t}^{x^*} = \xi_k^{x^*} - (\xi_k^{x^*} - \xi_{k,t}^{x^*}) \geq \frac{\epsilon}{2} \geq 0,$$

for any $k \in \mathcal{K}_{x^*}$ and

$$\begin{aligned}
\sum_{k \in \mathcal{K}_{x^*}} \bar{c}_{k,t}(i) \cdot \xi_{k,t}^{x^*} &= \sum_{k \in \mathcal{K}_{x^*}} \mu_k^c(i) \cdot \xi_k^{x^*} + (\sum_{k \in \mathcal{K}_{x^*}} \bar{c}_{k,t}(i) \cdot \xi_{k,t}^{x^*} - \sum_{k \in \mathcal{K}_{x^*}} \mu_k^c(i) \cdot \xi_k^{x^*}) \\
&\leq b(i) - \epsilon + \frac{\epsilon}{2} \\
&\leq b(i) - \frac{\epsilon}{2} \\
&\leq b(i),
\end{aligned}$$

for any resource $i \notin \mathcal{C}_{x^*}$ if $\{|\bar{c}_{l,t}(j) - \mu_l^c(j)| \leq \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}\}$ for any arm $l \in \mathcal{K}_{x^*}$ and resource $j \in \{1, \cdots, C\}$. Hence:

$$\begin{aligned}
\mathbb{E}[n_{x,T}] &\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}] \\
&\quad + \sum_{t=\mathrm{t_{ini}}}^{T} \sum_{\substack{l \in \mathcal{K}_{x^*} \\ j \in \{1, \cdots, C\}}} \mathbb{P}[|\bar{c}_{l,t}(j) - \mu_l^c(j)| > \frac{\epsilon^3}{16 \cdot (\mathrm{C}+2)!^2}] \\
&\leq 2\beta_x \cdot \ln(T) + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x} \cdot I_{n_{x,t} \geq \beta_x \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}] + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6},
\end{aligned}$$

where we bound the third term appearing in the right-hand side using (C.24). □

The remainder of this proof is dedicated to show that the last term in (C.30) can be bounded

by a constant. This term can be broken down in three terms similarly as in Lemmas 4.5 and 4.8.

$$\begin{aligned}
&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t=x}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}\cdot I_{x^*\in\mathcal{B}_t}]\\
&\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}+E_{x,t}\geq\mathrm{obj}_{x^*,t}+E_{x^*,t}}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}\cdot I_{x\in\mathcal{B}_t,x^*\in\mathcal{B}_t}]\\
&\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}] &\text{(C.31)}\\
&+\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*,t}\leq\mathrm{obj}_{x^*}-E_{x^*,t}}\cdot I_{x^*\in\mathcal{B}_t}] &\text{(C.32)}\\
&+\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}]. &\text{(C.33)}
\end{aligned}$$

Fact C.12.

$$\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}]\leq 2^{11}\frac{(C+4)!^4}{\epsilon^6}.$$

Proof. Using the shorthand notation $\alpha_x = 8(\frac{\lambda}{\Delta_x})^2$, we have:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}]\\
&\qquad\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\Delta_x<2\lambda\cdot\max_{k\in\mathcal{K}_x}\epsilon_{k,t}}\cdot I_{n_{x,t}\geq\beta_x\ln(t)}]\\
&\qquad\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\min_{k\in\mathcal{K}_x} n_{k,t}\leq\alpha_x\ln(t)}\cdot I_{n_{x,t}\geq\beta_x\cdot\ln(t)}]\\
&\qquad\leq \sum_{t=\mathrm{t_{ini}}}^{T}\sum_{k\in\mathcal{K}_x}\mathbb{P}[n_{k,t}\leq\alpha_x\ln(t)\ ;\ n_{x,t}\geq\beta_x\ln(t)],
\end{aligned}$$

since $\sum_{l=1}^{K}\xi_{l,t}^x\leq 1$ and $\xi_{l,t}^x\geq 0$ for any arm l when x is feasible for (4.8) at time t. Consider $k\in\mathcal{K}_x$ and assume that $n_{k,t}\leq\alpha_x\cdot\ln(t)$ and $n_{x,t}\geq\beta_x\cdot\ln(t)$. Suppose, by contradiction, that $|b_{x,t}(i)-n_{x,t}\cdot b(i)|\leq 32\frac{(C+1)!^2}{\epsilon^4}\cdot\ln(t)$ for any resource $i\in\mathcal{C}_x$ and that $|\mu_l^c(i)-\tilde{c}_l(i)|\leq\frac{\epsilon^2}{8\cdot(C+2)!}$ for any resource $i\in\{1,\cdots,C\}$ for all arms $l\in\mathcal{K}_x$ such that $n_{l,t}^x\geq 2^7\frac{(C+2)!^2}{\epsilon^4}\cdot\ln(t)$, where $\tilde{c}_l(i)$ is the empirical average amount of resource i consumed

when selecting basis x and pulling arm l, i.e. $\tilde{c}_l(i) = \frac{1}{n^x_{l,t}} \cdot \sum_{\tau=\mathrm{t_{ini}}}^{t-1} I_{x_\tau = x} \cdot I_{a_\tau = l} \cdot c_{l,\tau}(i)$. We can partition $\mathcal{K}_x - \{k\}$ into two sets $\mathcal{K}_x^{\mathbb{1}}$ and $\mathcal{K}_x^2$ such that $n^x_{l,t} \geq 2^7 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(t)$ for all $l \in \mathcal{K}_x^{\mathbb{1}}$ and $n^x_{l,t} < 2^7 \frac{(C+2)!^2}{\epsilon^4} \cdot \ln(t)$ for all $l \in \mathcal{K}_x^2$. Similarly as in the proof of Lemma 4.14, we have, for any resource $i \in \mathcal{C}_x$:

$$\begin{aligned}
32\frac{(C+1)!^2}{\epsilon^4} \cdot \ln(t) &\geq |b_{x,t}(i) - n_{x,t} \cdot b(i)| \\
&\geq n_{x,t} \cdot |\sum_{l \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_l(i) \cdot p_l - b(i)| - n_{k,t} - \sum_{l \in \mathcal{K}_x^2} n^x_{l,t} \\
&\geq n_{x,t} \cdot |\sum_{l \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_l(i) \cdot p_l - b(i)| - \alpha_x \cdot \ln(t) - C \cdot \frac{2^7 \cdot (C+2)!^2}{\epsilon^4} \cdot \ln(t),
\end{aligned}$$

where $p_l = \frac{n^x_{l,t}}{n_{x,t}}$ for $l \in \mathcal{K}_1^x$ and $p_l = 0$ otherwise. Hence:

$$\begin{aligned}
|\sum_{l \in \mathcal{K}_x} \mu_l^c(i) \cdot p_l - b(i)| &\leq \max_{l \in \mathcal{K}_x^{\mathbb{1}}} |\mu_l^c(i) - \tilde{c}_l(i)| + |\sum_{l \in \mathcal{K}_x^{\mathbb{1}}} \tilde{c}_l(i) \cdot p_l - b(i)| \\
&\leq \frac{\epsilon^2}{8 \cdot (C+2)!} + \frac{(\alpha_x + C \cdot \frac{2^7 \cdot (C+2)!^2}{\epsilon^4}) \cdot \ln(t)}{n_{x,t}} \\
&\leq \frac{\epsilon^2}{4 \cdot (C+2)!}.
\end{aligned}$$

To derive the first inequality, we use the fact that $\sum_{l=1}^K p_l \leq 1$ and $p_l \geq 0$ for any arm l. For the last inequality, we use $n_{x,t} \geq 2^{10} \frac{(C+3)!^3}{\epsilon^6} \cdot (\frac{\lambda}{\Delta_x})^2 \cdot \ln(t)$ along with $\lambda \geq 1$ and $\Delta_x \leq \mathrm{obj}_{x^*} = \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^K \xi_k^{x^*} \leq 1$ because of the time constraint imposed in (4.3). We get:

$$\begin{aligned}
\xi_k^x &\leq \|p - \xi^x\|_2 \\
&= \left\|A_x^{-1} A_x (p - \xi^x)\right\|_2 \\
&\leq \left\|\left|A_x^{-1}\right|\right\|_2 \cdot \|A_x(p - \xi^x)\|_2 \\
&\leq \frac{1}{|\det(A_x)|} \cdot \left\|\left|\mathrm{adj}(A_x)\right|\right\|_2 \cdot \sqrt{C} \cdot \frac{\epsilon^2}{4 \cdot (C+2)!} \\
&\leq \frac{1}{\epsilon} \cdot (C+1)! \cdot \sqrt{C} \cdot \frac{\epsilon^2}{4 \cdot (C+2)!} \\
&\leq \frac{\epsilon}{2},
\end{aligned}$$

a contradiction since x is ϵ-non-degenerate by Assumption 4.8. We conclude that:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_x+2E_{x,t}} \cdot I_{x\in\mathcal{B}_t} \cdot I_{n_{x,t}\geq\beta_x\cdot\ln(t)}] \\
&\quad\leq C\cdot\sum_{t=\mathrm{t_{ini}}}^{T}\sum_{i\in\mathcal{C}_x}\mathbb{P}[|b_{x,t}(i)-n_{x,t}\cdot b(i)|\geq 32\frac{(C+1)!^2}{\epsilon^4}\cdot\ln(t)] \\
&\quad+C\cdot\sum_{t=\mathrm{t_{ini}}}^{T}\sum_{\substack{l\in\mathcal{K}_x\\ i\in\mathcal{C}_x}}\mathbb{P}[|\mu_l^c(i)-\tilde{c}_l(i)|\geq\frac{\epsilon^2}{8\cdot(C+2)!}\ ;\ n_{l,t}^x\geq 2^7\frac{(C+2)!^2}{\epsilon^4}\cdot\ln(t)] \\
&\quad\leq 32C^2\cdot\frac{(C+1)!^2}{\epsilon^4}\cdot\frac{\pi^2}{6}+C\cdot 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\
&\quad+\sum_{\substack{l\in\mathcal{K}_x\\ j\in\mathcal{C}_x}}\sum_{s=2^7\cdot(C+2)!^2/\epsilon^4\cdot\ln(T)}\mathbb{P}[|\mu_l^c(j)-\tilde{c}_l(j)|\geq\frac{\epsilon^2}{8\cdot(C+2)!}\ ;\ n_{l,T}^x=s] \\
&\quad\leq 4C\cdot 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6}+2^7C^2\cdot\frac{(C+2)!^2}{\epsilon^4}\cdot\frac{\pi^2}{6} \\
&\quad\leq 2^{11}\frac{(C+4)!^4}{\epsilon^6},
\end{aligned}$$

where we use (4.12) for the second inequality and Lemma 4.1 for the third inequality.

□

Fact C.13.

$$\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}]\leq 2^{10}\frac{(C+3)!^2}{\epsilon^6}.$$

Proof. First observe that:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}] \\
&\quad\leq\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{|\det(\bar{A}_{x,t})|\geq\epsilon/2}]+\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{|\det(\bar{A}_{x,t})|<\epsilon/2}] \\
&\quad\leq\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{|\det(\bar{A}_{x,t})|\geq\epsilon/2}] \\
&\quad+\sum_{t=\mathrm{t_{ini}}}^{T}\sum_{\substack{k\in\mathcal{K}_x\\ i\in\mathcal{C}_x}}\mathbb{P}[|\bar{c}_{k,t}(i)-\mu_k^c(i)|>\frac{\epsilon^3}{16\cdot(\mathrm{C}+2)!^2}] \\
&\quad\leq\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq\mathrm{obj}_x+E_{x,t}}\cdot I_{x\in\mathcal{B}_t}\cdot I_{|\det(\bar{A}_{x,t})|\geq\epsilon/2}]+2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6},
\end{aligned}$$

where we use the preliminary work of Section C.6.1 and in particular (C.25). The key observation now is that if $\mathrm{obj}_{x,t} \geq \mathrm{obj}_x + E_{x,t}$, $x \in \mathcal{B}_t$, and $|\det(\bar{A}_{x,t})| \geq \epsilon/2$, at least one of the following events $\{\bar{r}_{k,t} \geq \mu^r_k + \epsilon_{k,t}\}$, for $k \in \mathcal{K}_x$, or $\{|\bar{c}_{k,t}(i) - \mu^c_k(i)| \geq \epsilon_{k,t}\}$, for $k \in \mathcal{K}_x$ and $i \in \{1, \cdots, C\}$, occurs. Otherwise we have:

$$\begin{aligned}
\mathrm{obj}_{x,t} - \mathrm{obj}_x &= (\bar{A}^{-1}_{x,t} b_{\mathcal{C}_x})^\intercal \bar{r}_{\mathcal{K}_x,t} - (A^{-1}_x b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x} \\
&< (\bar{A}^{-1}_{x,t} b_{\mathcal{C}_x})^\intercal (\mu^r_{\mathcal{K}_x} + \epsilon_{\mathcal{K}_x,t}) - (A^{-1}_x b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x} \\
&= \frac{1}{\lambda} \cdot E_{x,t} + ((\bar{A}^{-1}_{x,t} - A^{-1}_x) b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x},
\end{aligned}$$

where the first inequality is a consequence of the fact that x is feasible for (4.8), i.e. $\bar{A}^{-1}_{x,t} b_{\mathcal{C}_x} \geq 0$ with at least one non-zero coordinate since $\bar{c}_{k,t}(i) \geq \epsilon$ for all arms k and resource i by Assumption 4.8. Writing $\mu^c_k(i) = \bar{c}_{k,t}(i) + u_{i,k} \cdot \epsilon_{k,t}$, with $u_{i,k} \in [-1, 1]$ for all $(i,k) \in \mathcal{C}_x \times \mathcal{K}_x$, and defining the matrix $U = (u_{i,k})_{(i,k) \in \mathcal{C}_x \times \mathcal{K}_x}$, we get:

$$\begin{aligned}
&\mathrm{obj}_{x,t} - \mathrm{obj}_x \\
&< \frac{E_{x,t}}{\lambda} + |(\bar{A}^{-1}_{x,t} - (\bar{A}_{x,t} + U \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}))^{-1}) b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x}| \\
&= \frac{E_{x,t}}{\lambda} + |(\bar{A}^{-1}_{x,t} U (I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U)^{-1} \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x}| \\
&\leq \frac{E_{x,t}}{\lambda} \\
&+ \frac{|\det(\bar{A}_{x,t} + U \operatorname{diag}(\epsilon_{\mathcal{K}_x,t})) \cdot (\bar{A}^{-1}_{x,t} U (I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U)^{-1} \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x}|}{\epsilon} \\
&= \frac{E_{x,t}}{\lambda} + \frac{1}{\epsilon} \cdot |(\operatorname{adj}(\bar{A}_{x,t}) U \operatorname{adj}(I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U) \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} b_{\mathcal{C}_x})^\intercal \mu^r_{\mathcal{K}_x}| \\
&\leq \frac{E_{x,t}}{\lambda} + \frac{1}{\epsilon} \cdot \left\| \operatorname{adj}(\bar{A}_{x,t}) U \operatorname{adj}(I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U) \right\|_2 \cdot \left\| \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} b_{\mathcal{C}_x} \right\|_2 \cdot \left\| \mu^r_{\mathcal{K}_x} \right\|_2 \\
&\leq \frac{E_{x,t}}{\lambda} + \frac{1}{\epsilon} \cdot \left\| \operatorname{adj}(\bar{A}_{x,t}) U \right\|_2 \cdot \left\| \operatorname{adj}(I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U) \right\|_2 \cdot \frac{E_{x,t}}{\lambda} \cdot \sqrt{C} \\
&\leq \frac{E_{x,t}}{\lambda} + \frac{1}{\epsilon} \cdot (C+1)! \cdot \left\| \operatorname{adj}(I + \operatorname{diag}(\epsilon_{\mathcal{K}_x,t}) \bar{A}^{-1}_{x,t} U) \right\|_2 \cdot \frac{E_{x,t}}{\lambda} \\
&\leq \frac{E_{x,t}}{\lambda} + \frac{2}{\epsilon} \cdot (C+1)!^2 \cdot \frac{E_{x,t}}{\lambda} \\
&= E_{x,t}.
\end{aligned}$$

We use the Woodbury matrix identity to derive the first equality and the matrix determinant lemma for the second equality. The second inequality is derived from Assumption 4.8 since $\epsilon \leq \det(A_x) = \det(\bar{A}_{x,t} + U\,\mathrm{diag}(\epsilon_{\mathcal{K}_x,t}))$ by definition of U. The fourth inequality is derived from the observation that $\mathrm{diag}(\epsilon_{\mathcal{K}_x,t})\bar{A}_{x,t}^{-1}b_{\mathcal{C}_x}$ is the vector $(\epsilon_{k,t}\cdot\xi_{k,t}^x)_{k\in\mathcal{K}_x}$. The fifth inequality is obtained by observing that the components of $\bar{A}_{x,t}$ and U are all smaller than 1 in absolute value. The sixth inequality is obtained by observing that the elements of $\bar{A}_{x,t}^{-1}$ are smaller than $\frac{2}{\epsilon}\cdot(C-1)!$ since $\det(\bar{A}_{x,t}) \geq \frac{\epsilon}{2}$ and that $\epsilon_{k,t} \leq \frac{\epsilon}{2\cdot C!}$ for all arms k as a result of the initialization phase. We get:

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x,t}\geq \mathrm{obj}_x + E_{x,t}}\cdot I_{x\in\mathcal{B}_t}] \\
&\quad\leq \sum_{t=\mathrm{t_{ini}}}^{T}\sum_{\substack{k\in\mathcal{K}_x\\ i\in\mathcal{C}_x}} \mathbb{P}[|c_{k,t}(i)-\mu_k^c(i)|\geq\epsilon_{k,t}] + \sum_{t=\mathrm{t_{ini}}}^{T}\sum_{k\in\mathcal{K}_x}\mathbb{P}[r_{k,t}\geq\mu_k^r+\epsilon_{k,t}] + 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\
&\quad\leq 2\cdot 2^9\frac{(\mathrm{C}+3)!^4}{\epsilon^6}\,.
\end{aligned}
$$

□

Fact C.14.

$$
\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*,t}\leq \mathrm{obj}_{x^*}-E_{x,t}}\cdot I_{x^*\in\mathcal{B}_t}] \leq 2^{10}\frac{(C+3)!^2}{\epsilon^6}.
$$

We omit the proof since it is almost identical to the proof of Fact C.13.

C.6.5 Proof of Theorem 4.8

Along the same lines as for Theorem 4.5, we build upon (4.4):

$$
\begin{aligned}
R_{B(1),\cdots,B(C-1),T} &\leq T\cdot\sum_{k=1}^{K}\mu_k^r\cdot\xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] + O(1) \\
&= T\cdot\sum_{k=1}^{K}\mu_k^r\cdot\xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \mathbb{E}[\sum_{t=\tau^*+1}^{T} r_{a_t,t}] + O(1) \\
&\leq T\cdot\sum_{k=1}^{K}\mu_k^r\cdot\xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \sigma\cdot\mathbb{E}[\min_{i=1,\cdots,C}\sum_{t=\tau^*+1}^{T} c_{a_t,t}(i)] + O(1) \\
&\leq T\cdot\sum_{k=1}^{K}\mu_k^r\cdot\xi_k^{x^*} - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}] + \sigma\cdot\sum_{i=1}^{C}\mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t}(i) - B(i))_+] + O(1).
\end{aligned}
$$

The second inequality is a direct consequence of Assumption 4.7. To derive the last inequality, observe that if $\tau^* = T+1$, we have:

$$\sum_{t=\tau^*+1}^{T} c_{a_t,t}(j) = 0 \leq \sum_{i=1}^{C} \mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t}(i) - B(i))_+],$$

for any $j \in \{1, \cdots, C\}$ while if $\tau^* < T+1$ we have run out of resources before the end of the game, i.e. there exists $j \in \{1, \cdots, C\}$ such that $\sum_{t=1}^{\tau^*} c_{a_t,t}(j) \geq B(j)$, which implies that:

$$\begin{aligned}
\min_{i=1,\cdots,C} \sum_{t=\tau^*+1}^{T} c_{a_t,t}(i) &\leq \sum_{t=\tau^*+1}^{T} c_{a_t,t}(j) \\
&\leq \sum_{t=\tau^*+1}^{T} c_{a_t,t}(j) + \sum_{t=1}^{\tau^*} c_{a_t,t}(j) - B(j) \\
&= (\sum_{t=1}^{T} c_{a_t,t}(j) - B(j))_+ \leq \sum_{i=1}^{C} (\sum_{t=1}^{T} c_{a_t,t}(i) - B(i))_+.
\end{aligned}$$

Now observe that, for any resource $i \in \{1, \cdots, C\}$:

$$\begin{aligned}
&\mathbb{E}[(\sum_{t=1}^{T} c_{a_t,t}(i) - B)_+] \\
&\leq \mathbb{E}[(\sum_{t=\mathrm{t_{ini}}}^{T} c_{a_t,t}(i) - b(i))_+] + K \cdot 2^8 \frac{(\mathrm{C}+2)!^4}{\epsilon^6} \cdot \ln(T) \\
&= \mathbb{E}[(\sum_{x \text{ basis for (4.3)}} \{b_{x,T}(i) - n_{x,T} \cdot b(i)\})_+] + O(K \cdot \frac{(C+2)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq \sum_{x \in \mathcal{B}} \mathbb{E}[(b_{x,T}(i) - n_{x,T} \cdot b(i))_+] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] \\
&+ \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}] + O(K \cdot \frac{(C+2)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq \sum_{x \in \mathcal{B}} \int_0^T \mathbb{P}[b_{x,T}(i) - n_{x,T} \cdot b(i) \geq u] \mathrm{d}u + O(K \cdot \frac{(C+3)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq \sum_{x \in \mathcal{B}} T \cdot \mathbb{P}[b_{x,T}(i) - n_{x,T} \cdot b(i) \geq 2^8 \frac{(C+3)!^4}{\epsilon^6} \cdot \ln(T)] \\
&+ 2^8 |\mathcal{B}| \cdot \frac{(C+3)!^4}{\epsilon^6} \cdot \ln(T) + O(K \cdot \frac{(C+3)!^4}{\epsilon^6} \cdot \ln(T)) = O(\frac{|\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)),
\end{aligned}$$

where we use the fact that the amounts of resources consumed at any time period are no larger than 1 for the first and second inequalities, Lemma 4.13 for the third inequality and inequalities (4.12) and (4.13) from Lemma 4.14 along with the fact that there are at least K feasible basis for (4.3) (corresponding to single-armed strategies) for the last equality. Plugging this back into the regret bound yields:

$$
\begin{aligned}
& R_{B(1),\cdots,B(C-1),T} && \text{(C.34)}\\
& \leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} r_{a_t,t}] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T))\\
& = T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} \sum_{k=1}^{K} \mu_k^r \cdot \mathbb{E}[n_{k,T}^x] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T))\\
& \leq T \cdot \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} - \sum_{x \in \mathcal{B}} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T))\\
& = \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (T - \sum_{x \in \mathcal{B} \mid \Delta_x = 0} \mathbb{E}[n_{x,T}]) && \text{(C.35)}\\
& - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T))\\
& = \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \cdot (\mathrm{t_{ini}} + \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \mathbb{E}[n_{x,T}] + \sum_{x \notin \mathcal{B}} \mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} \mathbb{E}[n_{x,T}])\\
& - \sum_{x \in \mathcal{B} \mid \Delta_x > 0} (\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^x) \cdot \mathbb{E}[n_{x,T}] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T))\\
& \leq \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \Delta_x \cdot \mathbb{E}[n_{x,T}] + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)) && \text{(C.36)}\\
& \leq 2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot (\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}) \cdot \ln(T) + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)),
\end{aligned}
$$

where we use (4.14) from Lemma 4.14 for the second inequality, Lemma 4.13 for the third inequality. To derive the last inequality, we use Lemma 4.15 and the fact that $\Delta_x \leq \mathrm{obj}_{x^*} = \sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*} \leq \sum_{k=1}^{K} \xi_k^{x^*} \leq 1$.

C.6.6 Proof of Theorem 4.9

Along the same lines as for the proof of Theorem 4.6, we start from inequality (C.36) derived in the proof of Theorem 4.8 and apply Lemma 4.15 only if Δ_x is big enough, taking into account the fact that $\sum_{x\in\mathcal{B}} \mathbb{E}[n_{x,T}] \leq T$. Specifically, we have:

$$
\begin{aligned}
&R_{B(1),\cdots,B(C-1),T} \\
&= \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq T}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, 2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot \frac{\ln(T)}{\Delta_x} + 2^{11} \frac{(C+4)!^2}{\epsilon^6} \cdot \Delta_x) \}) \\
&+ O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)) \\
&= \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq T}} \{ \sum_{x\in\mathcal{B} \mid \Delta_x > 0} \min(\Delta_x \cdot n_x, 2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot \frac{\ln(T)}{\Delta_x}) \} \\
&+ O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq T}} \{ \sum_{x\in\mathcal{B}} \sqrt{2^{10} \frac{(C+3)!^3 \cdot \lambda^2}{\epsilon^6} \cdot \ln(T) \cdot n_x} \} + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq 2^5 \frac{(C+3)!^2 \cdot \lambda}{\epsilon^3} \cdot \sqrt{\ln(T)} \cdot \sup_{\substack{(n_x)_{x\in\mathcal{B}}\geq 0 \\ \sum_{x\in\mathcal{B}} n_x \leq T}} \{ \sum_{x\in\mathcal{B}} \sqrt{n_x} \} + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)) \\
&\leq 2^5 \frac{(C+3)!^2 \cdot \lambda}{\epsilon^3} \cdot \sqrt{\sigma \cdot |\mathcal{B}| \cdot T \cdot \ln(T)} + O(\frac{\sigma \cdot |\mathcal{B}| \cdot (C+3)!^4}{\epsilon^6} \cdot \ln(T)),
\end{aligned}
$$

where we use the fact that $\Delta_x \leq 1$ (see the proof of Theorem 4.8) for the second equality, we maximize over each $\Delta_x \geq 0$ to derive the first inequality, and we use Cauchy-Schwartz for the last inequality.

C.6.7 Proof of Theorem 4.10

Define $\tilde{b}(i) = B(i)/\tilde{T}$ for any $i \in \{1, \cdots, C\}$. If the decision maker stops pulling arms at round $\tilde{T}$ at the latest, all the results derived in Section 4.7 hold as long as we substitute T

with $\tilde{T}$ and we get:

$$\tilde{T} \cdot \tilde{\text{opt}} - \mathbb{E}[\sum_{t=1}^{\min(\tau^*, \tilde{T})} r_{a_t,t}] \leq X,$$

where X denotes the right-hand side of the regret bound derived in either Theorem 4.8 or Theorem 4.9 and $\tilde{\text{opt}}$ denotes the optimal value of (4.3) when $b(i)$ is substituted with $\tilde{b}(i)$ for any $i \in \{1, \cdots, C\}$. The key observation is that $\tilde{T} \cdot \tilde{\text{opt}} = T \cdot \text{opt}$, where opt denotes the optimal value of (4.3), because the time constraint is redundant in (4.3) even when $b(i)$ is substituted with $\tilde{b}(i)$ for any $i \in \{1, \cdots, C\}$. This is enough to show the claim as we get:

$$X \geq T \cdot \text{opt} - \mathbb{E}[\sum_{t=1}^{\tau^*} r_{a_t,t}] \geq R_{B(1),\cdots,B(C-1),T},$$

where we use Lemma 4.2 for the last inequality.

C.6.8 Proof of Theorem 4.11

The only difference with the proofs of Theorems 4.8 and 4.9 lies in the following bounds:

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=\tau^*+1}^{T} r_{a_t,t}] \\
&\leq \mathbb{E}[(T-\tau^*)_+] \\
&= \sum_{t=0}^{T} \mathbb{P}[\tau^* \leq T-t] \\
&\leq \sum_{i=1}^{C}\sum_{t=0}^{T} \mathbb{P}[\sum_{\tau=1}^{T-t} c_{a_\tau,\tau}(i) \geq B(i)] \\
&= \sum_{i=1}^{C}\sum_{t=0}^{T} \mathbb{P}[\sum_{\tau=1}^{T-t} (c_{a_\tau,\tau}(i) - b(i)) \geq t\cdot b(i)] \\
&= \sum_{i=1}^{C}\sum_{t=0}^{T} \mathbb{P}[\mathrm{t_{ini}} + \sum_{x\notin\mathcal{B}} n_{x,T} + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} n_{x,T} + \sum_{x\in\mathcal{B}} b_{x,T-t}(i) - n_{x,T-t}\cdot b(i) \geq t\cdot b(i)] \\
&= \sum_{i=1}^{C}\sum_{t=0}^{T} \mathbb{P}[\mathrm{t_{ini}} + \sum_{x\notin\mathcal{B}} n_{x,T} + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}} n_{x,T} \geq t\cdot\frac{b(i)}{2}] \\
&+ \sum_{i=1}^{C}\sum_{t=0}^{T}\sum_{x\in\mathcal{B}} \mathbb{P}[b_{x,T-t}(i) - n_{x,T-t}\cdot b(i) \geq t\cdot\frac{b(i)}{2\cdot|\mathcal{B}|}] \\
&\leq 2\sum_{i=1}^{C}\sum_{t=1}^{T} \frac{\mathrm{t_{ini}} + \sum_{x\notin\mathcal{B}}\mathbb{E}[n_{x,T}] + \sum_{\substack{x \text{ pseudo-basis for (4.3)} \\ \text{with } \det(A_x)=0}}\mathbb{E}[n_{x,T}]}{t\cdot b(i)} \\
&+ 2|\mathcal{B}|\cdot\sum_{i=1}^{C}\sum_{t=1}^{T}\sum_{x\in\mathcal{B}} \frac{\mathbb{E}[|b_{x,T-t}(i) - n_{x,T-t}\cdot b(i)|]}{t\cdot b(i)} + O(1) = O(\frac{(C+4)!^4\cdot|\mathcal{B}|^2}{b\cdot\epsilon^6}\cdot\ln^2(T)),
\end{aligned}
$$

where we use Lemma 4.13 and we bound $\mathbb{E}[|b_{x,T-t}(i) - n_{x,T-t}\cdot b(i)|]$ in the same fashion as in the proof of Theorem 4.8 using Lemma 4.14.

C.7 Proofs for Section 2.6

C.7.1 Proof of Lemma C.1

The proof follows the same steps as for Lemma 4.8. We use the shorthand notations $\beta_k = 8\frac{\rho\cdot(\sum_{i=1}^{C} b(i))^2}{\epsilon^2}\cdot(\frac{1}{\Delta_k})^2$ and $n_{k,t}^{\neq x^*} = \sum_{x\in\mathcal{B}\,|\,k\in\mathcal{K}_x, x\neq x^*} n_{k,t}^x$. Along the same lines as in Fact C.1,

we have:

$$\mathbb{E}[n_{k,\tau^*}] \leq 2\beta_k \cdot \mathbb{E}[\ln(\tau^*)] + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}],$$

and we can focus on bounding the second term, which can be broken down as follows:

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{\tau^*} I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}] \\
&\quad = \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x_t,t}+E_{x_t,t} \geq \mathrm{obj}_{x^*,t}+E_{x^*,t}} \cdot I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}] \\
&\quad \leq \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x_t,t} \geq \mathrm{obj}_{x_t}+E_{x_t,t}}] \\
&\quad + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*,t} \leq \mathrm{obj}_{x^*}-E_{x^*,t}}] \\
&\quad + \mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x^*} < \mathrm{obj}_{x_t}+2E_{x_t,t}} \cdot I_{x_t \neq x^*} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}].
\end{aligned}$$

We study each term separately, just like in Lemma 4.8.

Fact C.15.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\mathrm{obj}_{x_t,t} \geq \mathrm{obj}_{x_t}+E_{x_t,t}}] \leq K \cdot \frac{\pi^2}{6}.$$

Proof. If $\mathrm{obj}_{x_t,t} \geq \mathrm{obj}_{x_t} + E_{x_t,t}$, there must exist $l \in \mathcal{K}_{x_t}$ such that $\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}$, otherwise:

$$\begin{aligned}
\mathrm{obj}_{x_t,t} - \mathrm{obj}_{x_t} &= \sum_{l \in \mathcal{K}_{x_t}} (\bar{r}_{l,t} - \mu_l^r) \cdot \xi_l^{x_t} \\
&< \sum_{l \in \mathcal{K}_{x_t}} \epsilon_{l,t} \cdot \xi_l^{x_t} \\
&= E_{x_t,t},
\end{aligned}$$

where the inequality is strict because there must exist $l \in \mathcal{K}_{x_t}$ such that $\xi_l^{x_t} > 0$ (at least one resource constraint is binding for a feasible basis to (4.3) aside from the basis $\tilde{x}$ associated

with $\mathcal{K}_{\tilde{x}} = \emptyset$). We obtain:

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x_t,t} \geq \text{obj}_{x_t} + E_{x_t,t}}] &\leq \mathbb{E}[\sum_{t=1}^{\tau^*} \sum_{l=1}^{K} I_{\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}}] \\ &\leq \sum_{l=1}^{K} \sum_{t=1}^{\infty} \mathbb{P}[\bar{r}_{l,t} \geq \mu_l^r + \epsilon_{l,t}] \\ &\leq K \cdot \frac{\pi^2}{6},\end{aligned}$$

where the last inequality is derived along the same lines as in the proof of Fact C.3. □

Similarly, we can show that:

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*,t} \leq \text{obj}_{x^*} - E_{x^*,t}}] \leq K \cdot \frac{\pi^2}{6}.$$

We move on to study the last term.

Fact C.16.

$$\mathbb{E}[\sum_{t=1}^{\tau^*} I_{\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}} \cdot I_{x_t \neq x^*} \cdot I_{a_t = k} \cdot I_{n_{k,t}^{\neq x^*} \geq \beta_k \ln(t)}] = 0.$$

Proof. If $\text{obj}_{x^*} < \text{obj}_{x_t} + 2E_{x_t,t}$, $x_t \neq x^*$, and $a_t = k$, we have:

$$\begin{aligned}\frac{\Delta_k}{2} &\leq \frac{\Delta_{x_t}}{2} \\ &< \sum_{l \in \mathcal{K}_{x_t}} \xi_l^{x_t} \cdot \sqrt{\frac{2\ln(t)}{n_{l,t}}} \\ &\leq \sum_{l \in \mathcal{K}_{x_t}} \sqrt{\frac{2\xi_l^{x_t} \cdot \xi_k^{x_t} \ln(t)}{n_{k,t}}},\end{aligned}$$

where we use the fact that, by definition of the load balancing algorithm and since $a_t = k$, $\xi_k^{x_t} \neq 0$ (otherwise arm k would not have been selected) and:

$$n_{l,t} \geq \frac{\xi_l^{x_t}}{\xi_k^{x_t}} n_{k,t}, \tag{C.37}$$

for all arms $l \in \mathcal{K}_{x_t}$. We get:

$$\begin{aligned} n_{k,t} &< \frac{8}{(\Delta_k)^2} \cdot \xi_k^{x_t} \cdot (\sum_{l \in \mathcal{K}_{x_t}} \sqrt{\xi_l^{x_t}})^2 \cdot \ln(t) \\ &\leq \frac{8}{(\Delta_k)^2} \cdot \xi_k^{x_t} \cdot \rho \cdot \sum_{l \in \mathcal{K}_{x_t}} \xi_l^{x_t} \cdot \ln(t) \leq \frac{8}{(\Delta_k)^2} \cdot \rho \cdot (\sum_{l \in \mathcal{K}_{x_t}} \xi_l^{x_t})^2 \cdot \ln(t), \end{aligned}$$

using the Cauchy−Schwarz inequality and the fact that a basis involves at most ρ arms. Now observe that:

$$\sum_{l \in \mathcal{K}_{x_t}} \xi_l^{x_t} \leq \sum_{l \in \mathcal{K}_{x_t}} \frac{\sum_{i=1}^{C} c_l(i)}{\epsilon} \cdot \xi_l^{x_t} \leq \frac{\sum_{i=1}^{C} b(i)}{\epsilon}$$

as x_t is a feasible basis to (4.8) and using Assumption 4.2. We obtain:

$$n_{k,t}^{\neq x^*} \leq n_{k,t} < 8 \cdot \frac{\rho \cdot (\sum_{i=1}^{C} b(i))^2}{\epsilon^2 \cdot (\Delta_k)^2} \cdot \ln(t) = \beta_k \cdot \ln(t).$$

□

C.7.2 Proof of Lemma C.2

We first show (C.2) by induction on t. The base case is straightforward. Suppose that the inequality holds at time $t-1$. There are three cases:

- arm k is not pulled at time $t-1$, in which case the left-hand side of the inequality remains unchanged while the right-hand side can only increase, hence the inequality still holds at time t,
- arm k is pulled at time $t-1$ after selecting $x_{t-1} \neq x^*$, in which case both sides of the inequality increase by one and the inequality still holds at time t,
- arm k is pulled at time $t-1$ after selecting $x_{t-1} = x^*$. First observe that there must exist $l \in \mathcal{K}_{x^*}$ such that $n_{l,t-1} \leq (t-1) \cdot \frac{\xi_l^{x^*}}{\sum_{r=1}^{K} \xi_r^{x^*}}$. Suppose otherwise, we have:

$$t-1 = \sum_{l=1}^{K} n_{l,t} \geq \sum_{l \in \mathcal{K}_{x^*}} n_{l,t} > \sum_{l \in \mathcal{K}_{x^*}} (t-1) \cdot \frac{\xi_l^{x^*}}{\sum_{r=1}^{K} \xi_r^{x^*}} = t-1,$$

a contradiction. Suppose now by contradiction that inequality (C.2) no longer holds at time t, we have:

$$\begin{aligned} n_{k,t-1} &= n_{k,t} - 1 \\ &> n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} \\ &\geq (n_{x^*,t} + \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t}) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} = (t-1) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}}, \end{aligned}$$

which implies, using the preliminary remark above, that $\frac{\xi_k^{x^*}}{n_{k,t-1}} < \max_{l \in \mathcal{K}_{x^*}} \frac{\xi_l^{x^*}}{n_{l,t-1}}$, a contradiction given the definition of the load balancing algorithm.

We conclude that inequality (C.2) holds for all times t and arms $k \in \mathcal{K}_{x^*}$. We also derive inequality (C.1) as a byproduct, since, at any time t and for any arm $k \in \mathcal{K}_{x^*}$:

$$\begin{aligned} n_{k,t} &\geq n_{x^*,t} - \sum_{l \in \mathcal{K}_{x^*}, l \neq k} n_{l,t} \\ &\geq n_{x^*,t} \cdot (1 - \frac{\sum_{l \in \mathcal{K}_{x^*}, l \neq k} \xi_l^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}}) - \rho \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1) \\ &= n_{x^*,t} \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^K \xi_l^{x^*}} - \rho \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,t} + 1), \end{aligned}$$

as a basis involves at most ρ arms.

C.7.3 Proof of Theorem C.1

The proof proceeds along the same lines as for Theorem 4.3. We build upon (4.4):

$$\begin{aligned} R_{B(1),\cdots,B(C)} &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k=1}^K \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\ &\leq B \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} - \sum_{k \in \mathcal{K}_{x^*}} \mu_k^r \cdot \mathbb{E}[n_{k,\tau^*}] + O(1) \\ &\leq (B - \frac{\mathbb{E}[n_{x^*,\tau^*}]}{\sum_{k=1}^K \xi_k^{x^*}}) \cdot \sum_{k=1}^K \mu_k^r \cdot \xi_k^{x^*} + \rho^2 \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,\tau^*}] + O(1), \end{aligned}$$

where we use (C.1) to derive the third inequality. Now observe that, by definition, at least one resource is exhausted at time τ^*. Hence, there exists $i \in \{1, \cdots, C\}$ such that the following holds almost surely:

$$\begin{aligned}
B(i) &\leq \sum_{k=1}^{K} c_k(i) \cdot n_{k,\tau^*} \\
&\leq \sum_{k \notin \mathcal{K}_{x^*}} n_{k,\tau^*} + \sum_{k \in \mathcal{K}_{x^*}} c_k(i) \cdot n_{k,\tau^*} \\
&\leq \sum_{x \in \mathcal{B}, x \neq x^*} n_{x,\tau^*} + \sum_{k \in \mathcal{K}_{x^*}} c_k(i) \cdot n_{k,\tau^*} \\
&\leq \rho \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,\tau^*} + 2) + n_{x^*,\tau^*} \cdot \sum_{k \in \mathcal{K}_{x^*}} c_k(i) \cdot \frac{\xi_k^{x^*}}{\sum_{l=1}^{K} \xi_l^{x^*}} \\
&\leq \rho \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,\tau^*} + 2) + b(i) \cdot \frac{n_{x^*,\tau^*}}{\sum_{k=1}^{K} \xi_k^{x^*}},
\end{aligned}$$

where we use (C.2) and the fact that x^* is a feasible basis to (4.3). Rearranging yields:

$$\frac{n_{x^*,\tau^*}}{\sum_{k=1}^{K} \xi_k^{x^*}} \geq B - \frac{\rho}{b} \cdot (\sum_{x \in \mathcal{B}, x \neq x^*} n_{x,\tau^*} + 2),$$

almost surely. Plugging this last inequality back into the regret bound, we get:

$$\begin{aligned}
R_{B(1),\cdots,B(C)} &\leq \rho \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] \cdot (\frac{\sum_{k=1}^{K} \mu_k^r \cdot \xi_k^{x^*}}{b} + \rho) + O(1) \\
&\leq \rho \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] \cdot (\frac{\sum_{k=1}^{K} \sum_{i=1}^{C} c_k(i) \cdot \xi_k^{x^*}}{\epsilon \cdot b} + \rho) + O(1) \\
&\leq (\frac{\rho \cdot \sum_{i=1}^{C} b(i)}{\epsilon \cdot b} + (\rho)^2) \cdot \sum_{x \in \mathcal{B}, x \neq x^*} \mathbb{E}[n_{x,t}] + O(1) \\
&= (\frac{\rho \cdot \sum_{i=1}^{C} b(i)}{\epsilon \cdot b} + (\rho)^2) \cdot \sum_{k=1}^{K} \mathbb{E}[\sum_{x \in \mathcal{B} \mid k \in \mathcal{K}_x, x \neq x^*} n_{k,\tau^*}^x] + O(1) \\
&\leq 32 \frac{\rho^3 \cdot (\sum_{i=1}^{C} b(i))^3}{\epsilon^3 \cdot b} \cdot (\sum_{k=1}^{K} \frac{1}{(\Delta_k)^2}) \cdot \mathbb{E}[\ln(\tau^*)] + O(1) \\
&\leq 32 \frac{\rho^3 \cdot (\sum_{i=1}^{C} b(i))^3}{\epsilon^3 \cdot b} \cdot (\sum_{k=1}^{K} \frac{1}{(\Delta_k)^2}) \cdot \ln(\frac{\sum_{i=1}^{C} b(i) \cdot B}{\epsilon} + 1) + O(1),
\end{aligned}$$

where we use the fact that x^* is a feasible basis to (4.3) for the third inequality, Lemma C.1 for the fourth inequality, the concavity of the logarithmic function along with Lemma 4.7

for the last inequality.

C.7.4 Proof of Lemma C.3

We use the shorthand notations $\beta_k = 8C \cdot (\frac{\lambda}{\Delta_k})^2$ and, for any round t:

$$n_{k,t}^{\notin \mathcal{O}} = \sum_{x \in \mathcal{B} \mid k \in \mathcal{K}_x,\, x \notin \mathcal{O}} n_{k,t}^{x}.$$

Similarly as in Fact C.11, we have:

$$\begin{aligned}\mathbb{E}[n_{k,T}^{\notin \mathcal{O}}] \leq\ & 2\beta_k \cdot \ln(T) + 2^9 \frac{(\mathrm{C}+3)!^4}{\epsilon^6} \\ & + \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t \notin \mathcal{O}} \cdot I_{a_t = k} \cdot I_{n_{k,t}^{\notin \mathcal{O}} \geq \beta_k \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}],\end{aligned}$$

and what remains to be done is to bound the second term, which we can break down as follows:

$$\begin{aligned}&\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{x_t \notin \mathcal{O}} \cdot I_{a_t = k} \cdot I_{n_{k,t}^{\notin \mathcal{O}} \geq \beta_k \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}] \\ &\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x_t,t} + E_{x_t,t} \geq \mathrm{obj}_{x^*,t} + E_{x^*,t}} \cdot I_{x_t \notin \mathcal{O}} \cdot I_{a_t = k} \cdot I_{n_{k,t}^{\notin \mathcal{O}} \geq \beta_k \ln(t)} \cdot I_{x^* \in \mathcal{B}_t}] \\ &\leq \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x_t,t} \geq \mathrm{obj}_{x_t} + E_{x_t,t}} \cdot I_{x_t \in \mathcal{B}_t}] \\ &+ \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*,t} \leq \mathrm{obj}_{x^*} - E_{x^*,t}} \cdot I_{x^* \in \mathcal{B}_t}] \\ &+ \mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*} < \mathrm{obj}_{x_t} + 2E_{x_t,t}} \cdot I_{x_t \notin \mathcal{O}} \cdot I_{a_t = k} \cdot I_{n_{k,t}^{\notin \mathcal{O}} \geq \beta_k \ln(t)}].\end{aligned}$$

The study of the second term is the same as in the proof of Lemma 4.15. We can also bound the first term in the same fashion as in the proof of Lemma 4.15 since there is no reference to the load balancing algorithm in the proof of Fact C.13. The major difference with the proof of Lemma 4.15 lies in the study of the last term.

Fact C.17.

$$\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x_t,t}\geq \mathrm{obj}_{x_t}+E_{x_t,t}} \cdot I_{x\in\mathcal{B}_t}] \leq 2^{10}\frac{K\cdot(C+3)!^2}{\epsilon^6}.$$

Proof. The only difference with the proof of Fact C.13 is that the number of arms that belong to $\mathcal{K}_x$ for x ranging in $\{\tilde{x}\in\mathcal{B} \mid k\in\mathcal{K}_{\tilde{x}},\ \tilde{x}\notin\mathcal{O}\}$ can be as big as K, as opposed to C when we are considering one basis at a time. This increases the bound by a multiplicative factor K. □

Fact C.18.

$$\mathbb{E}[\sum_{t=\mathrm{t_{ini}}}^{T} I_{\mathrm{obj}_{x^*}<\mathrm{obj}_{x_t}+2E_{x_t,t}} \cdot I_{x_t\notin\mathcal{O}} \cdot I_{a_t=k} \cdot I_{n_{k,t}^{\notin\mathcal{O}}\geq\beta_k\ln(t)}] = 0.$$

Proof. Assume that $\mathrm{obj}_{x^*} < \mathrm{obj}_{x_t} + 2E_{x_t,t}$, $x_t\notin\mathcal{O}$, and $a_t = k$. We have:

$$\begin{aligned}\frac{\Delta_k}{2} &\leq \frac{\Delta_{x_t}}{2}\\ &< \lambda\cdot\sum_{l\in\mathcal{K}_{x_t}}\xi_{l,t}^{x_t}\cdot\sqrt{\frac{2\ln(t)}{n_{l,t}}}\\ &\leq \lambda\cdot\sum_{l\in\mathcal{K}_{x_t}}\sqrt{\frac{2\xi_{l,t}^{x_t}\cdot\xi_{k,t}^{x_t}\ln(t)}{n_{k,t}}},\end{aligned}$$

where we use the fact that, by definition of the load balancing algorithm and since $a_t = k$, $\xi_{k,t}^{x_t}\neq 0$ (otherwise arm k would not have been selected) and:

$$n_{l,t}\geq\frac{\xi_{l,t}^{x_t}}{\xi_{k,t}^{x_t}}\cdot n_{k,t}, \tag{C.38}$$

for any arm $l\in\mathcal{K}_{x_t}$. We get:

$$\begin{aligned}n_{k,t}^{\notin\mathcal{O}} &\leq n_{k,t}\\ &< 8(\frac{\lambda}{\Delta_k})^2\cdot\xi_{k,t}^{x_t}\cdot(\sum_{l\in\mathcal{K}_{x_t}}\sqrt{\xi_{l,t}^{x_t}})^2\cdot\ln(t)\\ &\leq 8(\frac{\lambda}{\Delta_k})^2\cdot\xi_{k,t}^{x_t}\cdot C\cdot\sum_{l\in\mathcal{K}_{x_t}}\xi_{l,t}^{x_t}\cdot\ln(t)\\ &\leq 8(\frac{\lambda}{\Delta_k})^2\cdot C\cdot(\sum_{l\in\mathcal{K}_{x_t}}\xi_{l,t}^{x_t})^2\cdot\ln(t)\leq 8C\cdot(\frac{\lambda}{\Delta_k})^2\cdot\ln(t),\end{aligned}$$

using the Cauchy−Schwarz inequality, the fact that a basis involves at most C arms, and the fact that x_t is feasible for (4.8) whose linear constraints include $\sum_{l=1}^{K} \xi_l \leq 1$ and $\xi_l \geq 0, \forall l \in \{1, \cdots, K\}$. We get $n_{k,t}^{\notin \mathcal{O}} < \beta_k \ln(t)$ by defintion of β_k. □

C.7.5 Proof of Theorem C.2

Substituting $b(i)$ with $B(i)/B(C)$ for every resource $i \in \{1, \cdots, C\}$, the regret bound obtained in Theorem 4.3 turns into:

$$R_{B(1),\cdots,B(C)} \leq 16\frac{\rho}{\epsilon} \cdot \frac{\sum_{i=1}^{C} B(i)}{B(C)} \cdot \Big(\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x} \Big) \cdot \ln\Big(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1\Big) + O(1). \quad \text{(C.39)}$$

Observe that $\mathcal{B}^{\complement}$ and $\mathcal{O}^{\complement}$ are defined by strict inequalities that are linear in the vector $(B(1)/B(C), \cdots, B(C-1)/B(C))$. Hence, for $B(C)$ large enough, $\mathcal{B}_{\infty}^{\complement} \subset \mathcal{B}^{\complement}$ and $\mathcal{O}_{\infty}^{\complement} \subset \mathcal{O}^{\complement}$ and thus $\mathcal{B} \subset \mathcal{B}_{\infty}$ and $\mathcal{O} \subset \mathcal{O}_{\infty}$. We now move on to prove each claim separately.

First claim. Suppose that there exists a unique optimal basis to (4.3), which we denote by x^*. Then, we must have $\mathcal{O} = \{x^*\} = \mathcal{O}_{\infty}$ for $B(C)$ large enough. Indeed, using the set inclusion relations shown above, we have $\mathcal{O} \subset \mathcal{O}_{\infty} = \{x^*\}$ and $\mathcal{O}$ can never be empty as there exists at least one optimal basis to (4.3) (this linear program is feasible and bounded). We get $\mathcal{O}^{\complement} \cap \mathcal{B} \subset \mathcal{O}_{\infty}^{\complement} \cap \mathcal{B}_{\infty}$ for $B(C)$ large enough. Note moreover that, for any $x \in \mathcal{B}$, Δ_x converges to Δ_x^{∞} (because both the objective value of a feasible basis and the optimal value of a linear program are Lipschitz in the right-hand side of the inequality constraint), which implies that $\Delta_x > \frac{\Delta_x^{\infty}}{2} > 0$ when $x \in \mathcal{B} \cap \mathcal{O}^{\complement}$ for $B(C)$ large enough. We conclude with (C.39) that:

$$R_{B(1),\cdots,B(C)} \leq 32\frac{\rho}{\epsilon} \cdot \frac{\sum_{i=1}^{C} B(i)}{B(C)} \cdot \Big(\sum_{x \in \mathcal{B}_{\infty} \mid \Delta_x^{\infty} > 0} \frac{1}{\Delta_x} \Big) \cdot \ln\Big(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1\Big) + O(1),$$

for $B(C)$ large enough. This yields the result since $B(i)/B(C) \to b(i) > 0$ for any resource $i = 1, \cdots, C-1$.

Second claim. Suppose that $\frac{B(i)}{B(C)} - b(i) = O(\frac{\ln(B(C))}{B(C)})$ for any resource $i \in \{1, \cdots, C-1\}$.

Starting from (C.16) derived in the proof of Theorem 4.3 and applying Lemma 4.8 only if Δ_x is big enough, we have:

$$
\begin{aligned}
&R_{B(1),\cdots,B(C)} \\
&\leq O(1) + \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min\{ \quad \frac{\Delta_x}{\sum_{k=1}^{K} \xi_k^x} \cdot \mathbb{E}[n_{x,\tau^*}], \\
&\qquad\qquad\qquad\qquad 16\rho \cdot \frac{\sum_{k=1}^{K} \xi_k^x}{\Delta_x} \cdot \ln(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1) + \frac{\pi^2}{3}\rho \cdot \frac{\Delta_x}{\sum_{k=1}^{K} \xi_k^x}\} \\
&\leq O(1) + \sum_{x \in \mathcal{B} \mid \Delta_x > 0} \min\{ \quad \Delta_x \cdot \frac{B(C)}{\min_{i=1,\cdots,C} B(i)} \cdot \frac{\sum_{i=1}^{C} B(i)}{\epsilon}, \\
&\qquad\qquad\qquad\qquad 16\frac{\rho \cdot \sum_{i=1}^{C} B(i)/B(C)}{\epsilon} \cdot \frac{1}{\Delta_x} \cdot \ln(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1)\}.
\end{aligned}
$$

Thus, we get:

$$
\begin{aligned}
&R_{B(1),\cdots,B(C)} \\
&\leq 16\frac{\rho \cdot \sum_{i=1}^{C} B(i)/B(C)}{\epsilon} \cdot (\sum_{x \in \mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_{\infty}^{\complement}} \frac{1}{\Delta_x}) \cdot \ln(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1) \\
&+ (\sum_{x \in \mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_{\infty}} \Delta_x) \cdot \frac{B(C)}{\min_{i=1,\cdots,C} B(i)} \cdot \frac{\sum_{i=1}^{C} B(i)}{\epsilon} + O(1), \text{`}
\end{aligned}
$$

where we use:

$$
\sum_{k=1}^{K} \xi_k^x \in [\min_{i=1,\cdots,C} B(i)/B(C), \frac{\sum_{i=1}^{C} B(i)/B(C)}{\epsilon}]
$$

and

$$
\Delta_x \leq \frac{\sum_{i=1}^{C} B(i)/B(C)}{\epsilon},
$$

as shown in the proof of Theorem 4.3 (substituting b with $\min_{i=1,\cdots,C} B(i)/B(C)$). For $x \in \mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_{\infty}^{\complement}$, we have $x \in \mathcal{B}_{\infty}$ and $\Delta_x > \frac{\Delta_x^{\infty}}{2} > 0$ for $B(C)$ large enough, as shown for the first claim. For $x \in \mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_{\infty}$, we have $\Delta_x = O(\ln(B(C))/B(C))$ as both the objective value of a feasible basis and the optimal value of a linear program are Lipschitz in

the right-hand side of the inequality constraints. We conclude that, for $B(C)$ large enough:

$$\begin{aligned}
&R_{B(1),\cdots,B(C)} \\
&\leq 32\frac{\rho \cdot \sum_{i=1}^{C} B(i)/B(C)}{\epsilon} \cdot (\sum_{x \in \mathcal{B}_\infty \cap \mathcal{O}_\infty^{\complement}} \frac{1}{\Delta_x^\infty}) \cdot \ln(\frac{\sum_{i=1}^{C} B(i)}{\epsilon} + 1) \\
&+ \frac{1}{\epsilon} \cdot \frac{B(C)}{\min_{i=1,\cdots,C} B(i)} \cdot \sum_{x \in \mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_\infty} O(\ln(B(C))) + O(1).
\end{aligned}$$

This yields the result since $|\mathcal{B} \cap \mathcal{O}^{\complement} \cap \mathcal{O}_\infty| \leq |\mathcal{O}_\infty|$ and $B(i)/B(C) \to b(i) > 0$ for any resource $i = 1, \cdots, C-1$.

C.7.6 Proof of Theorem C.3

The proof is along the same lines as for Theorem C.2. Specifically, in a first step, we observe that all the proofs of Section 4.6 remain valid (up to universal constant factors) for T large enough as long as we substitute b with B/T. Indeed, for T large enough, we have $\frac{B}{T} \leq 2$ and $|\mu_k^c - \frac{B}{T}| > \frac{\epsilon}{2}$ for all arms $k \in \{1, \cdots, K\}$ under Assumption 4.6. In a second step, just like in the proof of Theorem C.2, we show that we can substitute $\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}$ with $\sum_{x \in \mathcal{B}_\infty \mid \Delta_x^\infty > 0} \frac{1}{\Delta_x^\infty}$ in the regret bound up to universal constant factors.

C.7.7 Proof of Theorem C.4

The proof is along the same lines as for Theorem C.2. Specifically, in a first step, we observe that all the proofs of Section 4.7 remain valid (up to universal constant factors) for T large enough as long as we substitute b with $\min_{i=1,\cdots,C-1} B(i)/T$. Indeed, for T large enough, we have $\min_{i=1,\cdots,C-1} B(i)/T \leq 2$ and, under Assumption 4.8, any basis to (C.3) has determinannt larger than $\epsilon/2$ in absolute value and is $\epsilon/2-$non-degenerate by continuity of linear functions. In a second step, just like in the proof of Theorem C.2, we show that we can substitute $\sum_{x \in \mathcal{B} \mid \Delta_x > 0} \frac{1}{\Delta_x}$ with $\sum_{x \in \mathcal{B}_\infty \mid \Delta_x^\infty > 0} \frac{1}{\Delta_x^\infty}$ in the regret bound up to universal constant factors.

Appendix D

Appendix For Chapter 5

D.1 Proof of Lemma 5.1

Consider any non-anticipating algorithm. The expected reward obtained at period $t \in \mathbb{N}$ is:

$$\begin{aligned}
\mathbb{E}[(v_t - p_t) \cdot \mathbb{1}_{b_t \geq p_t}] &= \mathbb{E}[\mathbb{E}[(v_t - p_t) \cdot \mathbb{1}_{b_t \geq p_t} \mid \tilde{\mathcal{F}}_{t-1}]] \\
&= \mathbb{E}[(\mathbb{E}[v_t \mid \tilde{\mathcal{F}}_{t-1}, b_t] - p_t) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \mathbb{E}[(x_t^\intercal \theta_* - p_t) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\leq \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+].
\end{aligned}$$

To derive the first equality, we use the fact that $((x_\tau, v_\tau, p_\tau))_{\tau \in \mathbb{N}}$ is an i.i.d. sequence, that (v_t, p_t) is independent of b_t conditioned on x_t since the algorithm is non-anticipating, and that v_t is independent of p_t conditioned on x_t. This shows that:

$$\mathrm{ER}_{\mathrm{OPT}}(T) \leq \sum_{t=1}^{T} \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+].$$

Moreover, this last inequality is in fact an equality since bidding $b_t = x_t^\intercal \theta_*$ at any time period $t \in \mathbb{N}$ yields the expected reward:

$$\begin{aligned}
\mathbb{E}[(v_t - p_t) \cdot \mathbb{1}_{x_t^\intercal \theta_* \geq p_t}] &= \mathbb{E}[\mathbb{E}[(v_t - p_t) \cdot \mathbb{1}_{x_t^\intercal \theta_* \geq p_t} \mid \tilde{\mathcal{F}}_{t-1}]] \\
&= \mathbb{E}[(\mathbb{E}[v_t \mid \tilde{\mathcal{F}}_{t-1}] - p_t) \cdot \mathbb{1}_{x_t^\intercal \theta_* \geq p_t}] \\
&= \mathbb{E}[(x_t^\intercal \theta_* - p_t) \cdot \mathbb{1}_{x_t^\intercal \theta_* \geq p_t}] \\
&= \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+].
\end{aligned}$$

D.2 Proof of Lemma 5.2

This is almost a direct consequence of Theorems 1 and 2 of [1] with the minor change (in their notations): $\eta_t = (v_t - x_t^\intercal \theta_*) \cdot \mathbb{1}_{b_t \geq p_t}$, $X_t = \mathbb{1}_{b_t \geq p_t} \cdot x_t$, and $Y_t = v_t \cdot \mathbb{1}_{b_t \geq p_t}$. Defining the σ-algebra $F_t = \sigma(x_1, \cdots, x_{t+1}, p_1, \cdots, p_{t+1}, v_1, \cdots, v_t)$, observe that X_t is F_{t-1}-measurable since (5.5) defines a non-anticipating algorithm, that η_t is F_t-measurable, and that $\eta_t \in [-1, 1]$ and has mean 0 conditioned on F_{t-1} since v_t is independent of p_t conditioned on x_t with mean $x_t^\intercal \theta_*$ and since $((x_\tau, v_\tau, p_\tau))_{\tau \in \mathbb{N}}$ is an i.i.d. sequence. This implies that the assumptions of Theorems 1 and 2 of [1] are satisfied with $R = 1$, $V = I_d$, $S = \sqrt{d}$ (since $\|\theta_*\|_\infty \leq 1$), $\delta = 1/T$, and $L = \sqrt{d}$ (since $\|x_t\|_\infty \leq 1$).

D.3 Proof of Theorem 5.1

At any time period $t \in \mathbb{N}$, we denote by $\tilde{\theta}_t$ an arbitrary element of $\operatorname{argmax}_{\theta \in \mathcal{C}_t} x_t^\intercal \theta$ so that $b_t = \max(0, \min(1, x_t^\intercal \tilde{\theta}_t))$. Using Lemma 5.1, we have:

$$\begin{aligned}
R_T &= \sum_{t=1}^T \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+] - \sum_{t=1}^T \mathbb{E}[(v_t - p_t) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \sum_{t=1}^T \mathbb{E}[(x_t^\intercal \theta_* - p_t)_+] - \sum_{t=1}^T \mathbb{E}[(x_t^\intercal \theta_* - p_t) \cdot \mathbb{1}_{b_t \geq p_t}].
\end{aligned}$$

The second equality is derived by conditioning on $\tilde{\mathcal{F}}_{t-1}$ in the same fashion as done in the proof of Lemma 5.1 since b_t is entirely determined by $\tilde{\mathcal{F}}_{t-1}$, see (5.5). Observe that:

$$\begin{aligned}(x_t^\intercal\theta_* - p_t)_+ &= (x_t^\intercal\theta_* - p_t)_+ \cdot \mathbb{1}_{x_t^\intercal\theta_* \geq p_t > b_t} + (x_t^\intercal\theta_* - p_t)_+ \cdot \mathbb{1}_{b_t \geq p_t} \\ &\leq (x_t^\intercal\theta_* - p_t)_+ \cdot \mathbb{1}_{x_t^\intercal\theta_* > b_t} + (x_t^\intercal\theta_* - p_t)_+ \cdot \mathbb{1}_{b_t \geq p_t} \\ &\leq \mathbb{1}_{x_t^\intercal\theta_* > b_t} + (x_t^\intercal\theta_* - p_t)_+ \cdot \mathbb{1}_{b_t \geq p_t},\end{aligned}$$

since $v_t \in [0,1]$ (which implies that $x_t^\intercal\theta_* = \mathbb{E}[v_t|x_t] \in [0,1]$) and $p_t \geq 0$. Plugging this inequality back into the regret bound yields:

$$\begin{aligned}R_T &\leq \sum_{t=1}^T \mathbb{P}[x_t^\intercal\theta_* > b_t] + \mathbb{E}[((x_t^\intercal\theta_* - p_t)_+ - (x_t^\intercal\theta_* - p_t)) \cdot \mathbb{1}_{b_t \geq p_t}] \\ &= \sum_{t=1}^T \mathbb{P}[x_t^\intercal\theta_* > b_t] + \mathbb{E}[(p_t - x_t^\intercal\theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}].\end{aligned} \tag{D.1}$$

Since $x_t^\intercal\theta_* \in [0,1]$, $x_t^\intercal\theta_* > b_t$ implies that $x_t^\intercal\theta_* > \max_{\theta \in \mathcal{C}_t} x_t^\intercal\theta$ and we conclude that $\theta^* \notin \mathcal{C}_t$. Using Lemma 5.2, we get:

$$\sum_{t=1}^T \mathbb{P}[x_t^\intercal\theta_* > b_t] \leq \sum_{t=1}^T \mathbb{P}[\theta_* \notin \mathcal{C}_t] \leq 1.$$

What remains to be done is to upper bound the second term in the right-hand side of (D.1). Using Fubini's theorem, we have:

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}[(p_t - x_t^\intercal\theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}] &= \int_0^\infty \mathbb{E}[\sum_{t=1}^T \mathbb{1}_{p_t - x_t^\intercal\theta_* \geq u} \cdot \mathbb{1}_{b_t \geq p_t}] \mathrm{d}u \\ &\leq \int_0^\infty \mathbb{E}[\sum_{t=1}^T \mathbb{1}_{b_t - x_t^\intercal\theta_* \geq u} \cdot \mathbb{1}_{b_t \geq p_t}] \mathrm{d}u \\ &= \mathbb{E}[\sum_{t=1}^T (b_t - x_t^\intercal\theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}].\end{aligned} \tag{D.2}$$

Using Lemma 5.2, we have that $\theta^* \in \mathcal{C}_t$ (which implies $\left\|\tilde{\theta}_t - \theta^*\right\|_{M_t} \leq 2\delta_T$) for all $t \in \{t, \cdots, T\}$ with probability at least $1 - 1/T$. Using the shorthand $E = \{\theta_* \in \cap_{t=1}^T \mathcal{C}_t\}$, we

get:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[(b_t - x_t^\intercal\theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}] &\leq T \cdot \mathbb{P}[E^\complement] + \mathbb{E}[\mathbb{1}_E \cdot \sum_{t=1}^{T}(b_t - x_t^\intercal\theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}] \\ &\leq 1 + \mathbb{E}[\mathbb{1}_E \cdot \sum_{t=1}^{T}|x_t^\intercal\tilde{\theta}_t - x_t^\intercal\theta_*| \cdot \mathbb{1}_{b_t \geq p_t}] \\ &\leq 1 + \mathbb{E}[\mathbb{1}_E \cdot \sum_{t=1}^{T}\|\mathbb{1}_{b_t \geq p_t} \cdot x_t\|_{M_t^{-1}} \cdot \left\|\tilde{\theta}_t - \theta_*\right\|_{M_t}] \\ &\leq 1 + 2\delta_T \cdot \mathbb{E}[\sum_{t=1}^{T}\|\mathbb{1}_{b_t \geq p_t} \cdot x_t\|_{M_t^{-1}}] \\ &\leq 1 + 2\delta_T \cdot \sqrt{d \cdot T \cdot \ln(T)},\end{aligned}$$

where we use $b_t \in [0,1]$ and $x_t^\intercal\theta_* \in [0,1]$ for the first inequality and where the last inequality is derived in Lemma 11 of [11].

D.4 Proof of Theorem 5.2

At any time period $t \in \mathbb{N}$, we denote by $\tilde{\theta}_t$ an arbitrary element of $\operatorname{argmax}_{\theta \in \mathcal{C}_{\tau_t}} x_t^\intercal\theta$. The proof is along the same lines as for Theorem 5.1 except for two inequalities. First, we now bound the first term in the right-hand side of D.1 as follows:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{P}[x_t^\intercal\theta_* > b_t] &\leq \sum_{t=1}^{T}\mathbb{P}[\theta_* \notin \mathcal{C}_{\tau_t}] \\ &\leq \sum_{t=1}^{T}\mathbb{P}[\theta_* \notin \cap_{\tau=1}^{T}\mathcal{C}_\tau] \\ &\leq 1,\end{aligned}$$

which leads to the same conclusion. Second, using the shorthand $E = \{\theta_* \in \cap_{t=1}^{T} \mathcal{C}_t\}$, we bound the right-hand side of (D.2) as follows:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}(b_t - x_t^\intercal \theta_*)_+ \cdot \mathbb{1}_{b_t \geq p_t}] &\leq T \cdot \mathbb{P}[E^\complement] + \mathbb{E}[\mathbb{1}_E \cdot \sum_{t=1}^{T} |x_t^\intercal \tilde{\theta}_t - x_t^\intercal \theta_*| \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\leq 1 + 2\sqrt{1+A} \cdot \delta_T \cdot \mathbb{E}[\sum_{t=1}^{T} \|\mathbb{1}_{b_t \geq p_t} \cdot x_t\|_{M_t^{-1}}] \\
&\leq 1 + 2\sqrt{1+A} \cdot \delta_T \cdot \sqrt{d \cdot T \cdot \ln(T)},
\end{aligned}$$

where the second inequality is a direct consequence of the proof of Theorem 4 in [1] and the last inequality is derived in Lemma 11 of [11] just like for Theorem 5.1.

D.5 Proof of Lemma 5.3

There are two cases depending on whether $\mathbb{E}[P] \leq \beta$ or not.

Case 1: $\mathbb{E}[P] \leq \beta$.

In this case $\lambda_* = 0$ and the total expected reward obtained by any non-anticipating algorithm is:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*-1} v_t \cdot \mathbb{1}_{b_t \geq p_t}] &\leq \mathbb{E}[\sum_{t=1}^{T} v_t] \\
&= T \cdot \mathbb{E}[V] \\
&= T \cdot \mathbb{E}[\mathbb{E}[V \mid X]] \\
&= T \cdot \mathbb{E}[g(X)],
\end{aligned}$$

which shows that $\mathrm{ER}_{\mathrm{OPT}}(B, T) \leq T \cdot R(\lambda_*, \mathcal{C})$.

Case 2: $\mathbb{E}[P] > \beta$.

The total expected reward obtained by any non-anticipating algorithm can be bounded as

follows:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{\tau^*-1} v_t \cdot \mathbb{1}_{b_t \geq p_t}] &\leq \mathbb{E}[\sum_{t=1}^{\tau^*} v_t \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot v_t \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \mathbb{E}[v_t \mid \tilde{\mathcal{F}}_{t-1}, b_t] \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \sum_{t=1}^{\infty} \mathbb{E}[I_{\tau^* \geq t} \cdot \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \mathbb{E}[\sum_{t=1}^{\tau^*} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\leq \mathbb{E}[\sum_{t=1}^{\tau^*-1} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] + 1,
\end{aligned}$$

where we use the the fact that $((x_t, v_t, p_t))_{t \in \mathbb{N}}$ is an i.i.d. sequence, that (v_t, p_t) is independent of b_t conditioned on x_t since the algorithm is non-anticipating, that v_t is independent of p_t conditioned on x_t, and that τ^* is a stopping time with respect to $((x_t, v_t, p_t))_{t \in \mathbb{N}}$. As a result, up to an additive term of order $O(1)$ in the final bound, we just need to bound the performance of any non-anticipating algorithm when the reward obtained at round t is $\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}$ as opposed to $v_t \cdot \mathbb{1}_{b_t \geq p_t}$. Observe that, in this setting, the total reward (resp. cost) obtained (resp. incurred) by any non-anticipating algorithm can be written as $\sum_{t=1}^{T} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot y_t$ (resp. $\sum_{t=1}^{T} p_t \cdot y_t$) where $y_t = \mathbb{1}_{b_t \geq p_t}$ for $t < \tau^*$ and $y_t = 0$ for $t \geq \tau^*$. Remark that $y_t \in [0, 1]$ for all $t \in \{1, \cdots, T\}$ and that, by definition of τ^*, $\sum_{t=1}^{T} p_t \cdot y_t \leq B$. Thus $(y_t)_{t=1,\cdots,T}$ is always a feasible solution to the knapsack problem:

$$\begin{aligned}
\sup_{(\xi_t)_{t=1,\cdots,T}} \quad & \sum_{t=1}^{T} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \xi_t \\
\text{subject to} \quad & \sum_{t=1}^{T} p_t \cdot \xi_t \leq B \\
& \xi_t \in [0, 1], \quad t = 1, \cdots, T.
\end{aligned} \tag{D.3}$$

As a consequence, we conclude that the expected total reward obtained by any non-anticipating algorithm is always no larger than the expected optimal value of (D.3). This reduces the problem of bounding $\mathrm{ER}_{\mathrm{OPT}}(B, T)$ to a stochastic i.i.d. knapsack problem with T items

and a knapsack capacity of B when the t-th item has value $\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta)$ and weights p_t. [63] study the expected optimal value of the knapsack problem:

$$\begin{aligned} \sup_{(\xi_t)_{t=1,\cdots,T}} \quad & \sum_{t=1}^{T} Y_t \cdot \xi_t \\ \text{subject to} \quad & \sum_{t=1}^{T} W_t \cdot \xi_t \leq B \\ & \xi_t \in [0,1], \quad t = 1, \cdots, T \end{aligned} \tag{D.4}$$

when $((Y_t, W_t))_{t=1,\cdots,T}$ is an i.i.d. stochastic process with distribution $\tilde{\nu}$ and when $B = \beta \cdot T$. Denoting by (Y, W) a 2-dimensional vector of random variables with distribution $\tilde{\nu}$, the author of [63] shows that, under the five conditions listed below, the expected optimal value of (D.4) is $\mathbb{E}[Y \cdot \mathbb{1}_{Y \geq \gamma_* \cdot W}] + O(1)$ as $T \to \infty$ where $\gamma_* \geq 0$ satisfies $\mathbb{E}[W \cdot \mathbb{1}_{Y \geq \gamma_* \cdot W}] = \beta$. Defining the mapping $\Phi : \gamma \in \mathbb{R}_+ \to \mathbb{E}[W \cdot \mathbb{1}_{Y \geq \gamma \cdot W}]$, the five conditions mentioned above are:

1. Y and W are non-negative bounded random variables,

2. there exists $\gamma_* \geq 0$ such that $\Phi(\gamma_*) = \beta$,

3. $\Phi(\cdot)$ is differentiable in a neighborhood of γ_*,

4. $\Phi'(\cdot)$ is negative in a neighborhood of γ_*,

5. $\Phi'(\cdot)$ is Lipschitz in a neighborhood of γ_*.

As a result, showing that these five conditions hold for the choice $(Y = g(X), W = P)$ will conclude the proof. Observe that the first condition is satisfied since $(g(X), P) \in [0,1]^2$ because $\max_{\theta \in \mathcal{C}} X^\intercal \theta \geq X^\intercal \theta_* \geq 0$ as $\theta_* \in \mathcal{C}$. Adapting notations, we need to show that there exists $\lambda_* \geq 0$ such that $\phi(\lambda_*, \mathcal{C}) = \beta$ to establish that the second condition holds. This is true since: (i) $\phi(0, \mathcal{C}) = \mathbb{E}[P] > \beta$ (by assumption), (ii) $\phi(2/r, \mathcal{C}) = 0$ since

$g(X) \in [0,1]$ and $P \geq r$, and (iii) $\phi(\cdot, \mathcal{C})$ is continuous since:

$$\begin{aligned}
\phi(\lambda, \mathcal{C}) &= \mathbb{E}[\mathbb{E}[P \cdot \mathbb{1}_{g(X) \geq \lambda \cdot P} \mid X]] \\
&= \mathbb{E}[\int_0^{g(X)/\lambda} w \cdot f_X(w) \mathrm{d}w] \\
&= \mathbb{E}[\int_0^{1/\lambda} g(X) \cdot w \cdot f_X(g(X) \cdot w) \mathrm{d}w] \\
&= \int_0^{1/\lambda} \mathbb{E}[g(X) \cdot w \cdot f_X(g(X) \cdot w)] \mathrm{d}w,
\end{aligned} \tag{D.5}$$

for any $\lambda \geq 0$ and since:

$$\int_0^{\infty} \mathbb{E}[g(X) \cdot w \cdot f_X(g(X) \cdot w)] \mathrm{d}w = \mathbb{E}[P] \leq 1 < \infty$$

by Fubini's theorem. We move on to show that the third condition is satisfied. Observe that since $\mathbb{E}[P] > \beta$ we must have $\lambda_* > 0$. Moreover, using (D.5), $\phi(\cdot, \mathcal{C})$ is continuously differentiable on $\mathbb{R}_+^*$ by continuity under the integral sign with:

$$\phi'(\lambda, \mathcal{C}) = -\frac{1}{\lambda^3} \cdot \mathbb{E}[g(X) \cdot f_X(g(X)/\lambda)]$$

for $\lambda > 0$ since $f_x(\cdot)$ is continuous for all $x \in \mathcal{X}$ and $g(X) \cdot f_X(g(X)/\lambda)$ lies in $[0, \bar{L}]$ by Assumption 5.2. The fourth condition is satisfied since $\phi(\cdot, \mathcal{C})$ is non-increasing. Finally, we show that $\phi'(\cdot, \mathcal{C})$ is Lipschitz on $\mathbb{R}_+^*$. Indeed, for any $\lambda_1 \geq \lambda_2 > 0$, we have:

$$\begin{aligned}
|\phi'(\lambda_1, \mathcal{C}) - \phi'(\lambda_2, \mathcal{C})| &= |\mathbb{E}[\frac{g(X)}{\lambda_1^3} \cdot f_X(\frac{g(X)}{\lambda_1})] - \mathbb{E}[\frac{g(X)}{\lambda_2^3} \cdot f_X(\frac{g(X)}{\lambda_2})]| \\
&\leq \mathbb{E}[\frac{1}{\lambda_1^2} \cdot |\frac{g(X)}{\lambda_1} \cdot f_X(\frac{g(X)}{\lambda_1}) - \frac{g(X)}{\lambda_2} \cdot f_X(\frac{g(X)}{\lambda_2})|] \\
&+ \mathbb{E}[|\frac{1}{\lambda_1^2} - \frac{1}{\lambda_2^2}| \cdot \frac{1}{\lambda_2} \cdot |g(X) \cdot f_X(\frac{g(X)}{\lambda_2})|].
\end{aligned}$$

Observe that:

$$\begin{aligned}|\frac{1}{\lambda_1^2}-\frac{1}{\lambda_2^2}|\cdot\frac{1}{\lambda_2}\cdot|g(X)\cdot f_X(\frac{g(X)}{\lambda_2})| &\leq 2(\lambda_1-\lambda_2)\cdot|\frac{g(X)}{\lambda_2^4}\cdot f_X(\frac{g(X)}{\lambda_2})| \\ &\leq 2\frac{\bar{L}}{g(X)^3}\cdot(\lambda_1-\lambda_2) \\ &\leq 2\frac{\bar{L}}{(X^\intercal\theta_*)^3}\cdot(\lambda_1-\lambda_2),\end{aligned}$$

irrespective of whether $g(X)/\lambda_2 \leq 1$ or $g(X)/\lambda_2 > 1$ (in which case $f_X(g(X)/\lambda_2) = 0$). We use $\theta_* \in \mathcal{C}$ to derive the last inequality. Similarly, we have:

$$\begin{aligned}\mathbb{E}[\frac{1}{\lambda_1^2}\cdot|\frac{g(X)}{\lambda_1}\cdot f_X(\frac{g(X)}{\lambda_1})-\frac{g(X)}{\lambda_2}\cdot f_X(\frac{g(X)}{\lambda_2})|] &\leq \frac{K}{g(X)^3}\cdot(\lambda_1-\lambda_2) \\ &\leq \frac{K}{(X^\intercal\theta_*)^3}\cdot(\lambda_1-\lambda_2)\end{aligned}$$

irrespective of whether $g(X)/\lambda_1 \in \operatorname{supp} f_X(\cdot)$ (which implies $g(X)/\lambda_1 \leq 1$), $g(X)/\lambda_1 \notin \operatorname{supp} f_X(\cdot)$, $g(X)/\lambda_2 \in \operatorname{supp} f_X(\cdot)$ (which implies $g(X)/\lambda_2 \leq 1$), or whether $g(X)/\lambda_2 \notin \operatorname{supp} f_X(\cdot)$ (using $\theta_* \in \mathcal{C}$ for the last inequality). Bringing everything together, we get:

$$|\phi'(\lambda_1,\mathcal{C})-\phi'(\lambda_2,\mathcal{C})| \leq (2\bar{L}+K)\cdot\mathbb{E}[\frac{1}{(X^\intercal\theta_*)^3}]\cdot|\lambda_1-\lambda_2|,$$

and $(2\bar{L}+K)\cdot\mathbb{E}[1/(X^\intercal\theta_*)^3] < \infty$ by Assumption 5.2.

D.6 Proof of Lemma 5.4

For any $\lambda_1 \geq \lambda_2 \geq 0$, we have:

$$\begin{aligned}|R(\lambda_1,\mathcal{C})-R(\lambda_2,\mathcal{C})| &= \mathbb{E}[g(X)\cdot\mathbb{1}_{g(X)/\lambda_2 \geq P > g(X)/\lambda_1}] \\ &\leq \frac{1}{r}\cdot\mathbb{E}[P\cdot\mathbb{1}_{g(X)/\lambda_2 \geq P > g(X)/\lambda_1}] \\ &= \frac{1}{r}\cdot|\phi(\lambda_1,\mathcal{C})-\phi(\lambda_2,\mathcal{C})|,\end{aligned}$$

where the first inequality is a consequence of $g(X) \in [0,1]$ and $P \geq r$.

D.7 Proof of Lemma 5.5

For any $\lambda_2 \geq \lambda_1 > 0$, we have:

$$\begin{aligned}
|\phi(\lambda_2, \mathcal{C}) - \phi(\lambda_1, \mathcal{C})| &= \mathbb{E}[P \cdot \mathbb{1}_{\min(1, g(X)/\lambda_1) \geq P > \min(1, g(X)/\lambda_2)}]` \\
&\leq \mathbb{E}[\mathbb{1}_{\min(1, g(X)/\lambda_1) \geq P > \min(1, g(X)/\lambda_2)}] \\
&= \mathbb{E}[\int_{\min(1, g(X)/\lambda_2)}^{\min(1, g(X)/\lambda_1)} f_X(w) \mathrm{d}w] \\
&\leq \bar{L} \cdot \mathbb{E}[\min(1, g(X)/\lambda_1) - \min(1, g(X)/\lambda_2)] \\
&\leq \bar{L} \cdot \mathbb{E}[\frac{1}{g(X)}] \cdot |\lambda_1 - \lambda_2| \\
&\leq \bar{L} \cdot \mathbb{E}[\frac{1}{X^\intercal \theta_*}] \cdot |\lambda_1 - \lambda_2|,
\end{aligned}$$

where the first inequality is obtained using $P \in [0, 1]$, the second inequality is a consequence of Assumption 5.2, the third inequality actually holds almost surely irrespective of whether $g(X)/\lambda_1 \leq 1$, $g(X)/\lambda_1 > 1$, $g(X)/\lambda_2 \leq 1$, or $g(X)/\lambda_2 > 1$, and the last inequality is obtained using the fact that $\theta_* \in \mathcal{C}$. Also, observe that the last inequality holds even when $\lambda_1 = 0$.

D.8 Proof of Lemma 5.6

For any phase $k \in \mathbb{N}$, we denote by t_k the time period at which phase k starts. First note that we can reason conditionally on $\mathcal{F}_{t_{k-1}}$ since $((x_t, v_t, p_t))_{t \in \mathbb{N}}$ is an i.i.d. stochastic process. We use the Rademacher complexity approach to concentration inequalities for empirical processes to derive the result, see, for example, [20] and [29]. Specifically, the class of functions of interest is $\mathcal{F} = \{\ell_\lambda : (x, y) \in [0, 1] \times [r, 1] \to y \cdot \mathbb{1}_{x \geq \lambda \cdot y} \mid \lambda \in [\underline{\lambda}_k, 2/r]\}$. Observe that $\ell_\lambda(x, y) \in [0, 1]$ for any $(x, y) \in [0, 1] \times [r, 1]$ and that $\phi(\lambda, \mathcal{C}) = \mathbb{E}[\ell_\lambda(g(X), P)]$. Moreover, note that N_k samples, denoted by $(\min(1, \max_{\theta \in \mathcal{C}} X_n^\intercal \theta), P_n)_{n=1,\cdots,N_k}$, have been generated according to the same distribution as $(g(X), P)$ in an i.i.d. fashion at the end of

phase k. Using Theorem 3.2 from [29], we get:

$$\mathbb{P}[\exists \lambda \in [\underline{\lambda}_k, 2/r] \mid \hat{\phi}_k(\lambda, \mathcal{C}) - \phi(\lambda, \mathcal{C})| \geq 2\mathcal{R}_{N_k}(\mathcal{F}) + t \mid \mathcal{F}_{t_{k-1}}] \leq \exp(-2N_k \cdot t^2) \quad \forall t \geq 0, \tag{D.6}$$

where $\mathcal{R}_{N_k}(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$ for N_k samples. What remains to be done is to upper bound this last quantity. By definition, we have, for N_k independent Rademacher variables $(\epsilon_n)_{n=1,\cdots,N_k}$ that are independent of $(X_n, P_n)_{n=1,\cdots,N_k}$:

$$\begin{aligned}
\mathcal{R}_{N_k}(\mathcal{F}) &= \frac{1}{N_k} \cdot \mathbb{E}[\sup_{\lambda \in [\underline{\lambda}_k, 2/r]} |\sum_{n=1}^{N_k} \epsilon_n \cdot \ell_\lambda(\min(1, \max_{\theta \in \mathcal{C}} X_n^\intercal \theta), P_n)|] \\
&= \frac{1}{N_k} \cdot \mathbb{E}[\mathbb{E}[\sup_{z \in S((X_n, P_n)_{n=1,\cdots,N_k})} |\sum_{n=1}^{N_k} \epsilon_n \cdot z_n| \mid (X_n, P_n)_{n=1,\cdots,N_k}]] \\
&\leq \frac{1}{N_k} \cdot \mathbb{E}[\sqrt{2N_k \cdot \ln(2|S((X_n, P_n)_{n=1,\cdots,N_k})|)}] \\
&\leq \sqrt{\frac{2\ln(2(N_k+1))}{N_k}} \\
&\leq \sqrt{\frac{2\ln(2T)}{N_k}},
\end{aligned}$$

with:

$$S((X_n, P_n)_{n=1,\cdots,N_k}) = \{(P_{f(n)} \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} X_{f(n)}^\intercal \theta)/P_{f(n)} \geq \lambda})_{n=1,\cdots,N_k} \mid \lambda \geq 0\},$$

where the permutation $f(\cdot)$ of $\{1, \cdots, N_k\}$ is determined by:

$$\min(1, \max_{\theta \in \mathcal{C}} X_{f(n)}^\intercal \theta)/P_{f(n)} \leq \cdots \leq \min(1, \max_{\theta \in \mathcal{C}} X_{f(1)}^\intercal \theta)/P_{f(1)}.$$

The second equality is obtained by reindexing the vector $(z_1, \cdots, z_{N_k})$ according to the mapping $f(\cdot)$ which does not change the inner expectation since $(\epsilon_n)_{n=1,\cdots,N_k}$ is independent of $(X_n, P_n)_{n=1,\cdots,N_k}$. Note that $S((X_n, P_n)_{n=1,\cdots,N_k})$ is always a finite set with cardinality no larger than $N_k + 1$, which yields the second and third inequality using standard bounds on the Rademacher complexity of a finite set, see Theorem 3.3 of [29]. Plugging

$t = \sqrt{\frac{2\ln(2T)}{N_k}}$ in (D.6) and using the definition of Δ_k, we conclude that:

$$\mathbb{P}[\exists \lambda \in [\underline{\lambda}_k, 2/r] \mid \hat{\phi}_k(\lambda, \mathcal{C}) - \phi(\lambda, \mathcal{C})| \geq \Delta_k \mid \mathcal{F}_{t_{k-1}}] \leq \exp(-4\ln(2T)) \leq 1/T,$$

which, in particular, implies that:

$$\mathbb{P}[\sup_{\lambda \in [\underline{\lambda}_k, 2/r]} |\hat{\phi}_k(\lambda, \mathcal{C}) - \phi(\lambda, \mathcal{C})| \leq \Delta_k] \geq 1 - 1/T.$$

D.9 Proof of Lemma 5.7

By definition, we have:

$$\begin{aligned} T &\geq \sum_{k=0}^{\bar{k}_T - 1} N_k \\ &\geq 3 \sum_{k=0}^{\bar{k}_T - 1} 4^k \cdot \ln^2(T) \\ &\geq (4^{\bar{k}_T} - 1) \cdot \ln^2(T), \end{aligned}$$

which implies $4^{\bar{k}_T} \leq T/\ln^2(T) + 1$. Since $\ln^2(T) \geq 1$ for $T \geq 3$, we get $4^{\bar{k}_T} \leq T + 1$. Taking logarithms yields the claim since $\ln(4) \geq 1$.

D.10 Proof of Proposition 5.1

To simplify the discussion, we assume that $\mathbb{E}[P] \leq \beta$ so that $\phi(\lambda_*, \mathcal{C}) = \beta$ but the discussion would be almost identical if $\mathbb{E}[P] > \beta$ (in which case $\lambda_* = 0$). For any $k \in \{0, \cdots, \bar{k}_T\}$, we define the event:

$$A_k = \{\lambda_* \geq \underline{\lambda}_k,\ |\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \beta| \leq 4C \cdot |I_k|,\ |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| \leq 3C \cdot |I_{k+1}|\}.$$

Using the shorthand $E = \cap_{k=0}^{\bar{k}_T} A_k$, we have:

$$\mathbb{P}[E^{\complement}] \leq \sum_{k=0}^{\bar{k}_T} \mathbb{P}[A_k^{\complement}].$$

Note that we exclude the condition $|\phi(\underline{\lambda}_0, \mathcal{C}) - \beta| \leq 3C \cdot |I_0|$ from the definition of E since this condition is automatically satisfied almost surely given Lemma 5.5. By induction, we have:

$$\begin{aligned}\mathbb{P}[A_k^{\complement}] &\leq \mathbb{P}[A_0^{\complement}] + \sum_{j=0}^{k-1} \mathbb{P}[A_{j+1}^{\complement} \cap A_j] \\ &= \sum_{j=0}^{k-1} \mathbb{P}[A_{j+1}^{\complement} \cap A_j]\end{aligned}$$

for any $k > 0$ since, by construction, $\lambda_* \in [\underline{\lambda}_0, \bar{\lambda}_0] = [0, 2/r]$ which implies that $\mathbb{P}[A_0^{\complement}] = 0$. Rearranging yields:

$$\begin{aligned}\mathbb{P}[E^{\complement}] &\leq \sum_{k=0}^{\bar{k}_T} (\bar{k}_T - k) \cdot \mathbb{P}[A_{k+1}^{\complement} \cap A_k] \\ &\leq \sum_{k=0}^{\bar{k}_T} (\bar{k}_T - k) \cdot \mathbb{P}[B_k^{\complement}] + \sum_{k=0}^{\bar{k}_T} (\bar{k}_T - k) \cdot \mathbb{P}[A_{k+1}^{\complement} \cap A_k \cap B_k] \\ &\leq \frac{1}{T} \cdot \bar{k}_T \cdot (\bar{k}_T + 1) + \sum_{k=0}^{\bar{k}_T} (\bar{k}_T - k) \cdot \mathbb{P}[A_{k+1}^{\complement} \cap A_k \cap B_k] \\ &\leq \frac{\ln(T+1)^2}{T} + \sum_{k=0}^{\bar{k}_T} (\bar{k}_T - k) \cdot \mathbb{P}[A_{k+1}^{\complement} \cap A_k \cap B_k],\end{aligned}$$

where $B_k = \{\sup_{\lambda \in [\underline{\lambda}_k, 2/r]} |\hat{\phi}_k(\lambda, \mathcal{C}) - \phi(\lambda, \mathcal{C})| \leq \Delta_k\}$. We use Lemma 5.6 to derive the third inequality and Lemma 5.7 for the last inequality. What remains to be done is to show that the second term in the right-hand side is 0. Consider $k \in \{1, \cdots, \bar{k}_T\}$ and suppose that A_{k-1} and B_{k-1} hold. We show that A_k must hold which will imply that

$\mathbb{P}[A_k^{\complement} \cap A_{k-1} \cap B_{k-1}] = 0$. First observe that we have:

$$\begin{aligned}
|\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \beta| &\leq |\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \phi(\underline{\lambda}_k, \mathcal{C})| + |\phi(\underline{\lambda}_k, \mathcal{C}) - \beta| \\
&\leq \Delta_k + 3C \cdot |I_k| \\
&\leq 4C \cdot |I_k|,
\end{aligned}$$

where we use the fact that A_{k-1} and B_{k-1} hold for the first inequality and the fact that $T \geq \exp(8r^2/C^2)$ for the last inequality. At the end of Algorithm 5 for the k-th phase, we end up with an interval $[\underline{\gamma}_k, \bar{\gamma}_k]$ of length $|I_k|$ such that either (i) $\underline{\gamma}_k > \underline{\lambda}_k$ or (ii) $\underline{\gamma}_k = \underline{\lambda}_k$. In situation (i), by definition of the ending criterion of Algorithm 5, we must have $\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) \leq \beta + \Delta_k$ and $\hat{\phi}_k(\underline{\gamma}_k, \mathcal{C}) > \beta + \Delta_k$. This last inequality, combined with the fact that B_{k-1} holds, implies that $\phi(\underline{\gamma}_k, \mathcal{C}) > \beta$ and thus we have $\underline{\gamma}_k \leq \lambda_*$. In situation (ii), we automatically have $\underline{\gamma}_k \leq \lambda_*$ since A_{k-1} holds. Moreover, by definition of the ending criterion of Algorithm 5, we must have $\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) \leq \beta + \Delta_k$. We conclude that $\underline{\gamma}_k \leq \lambda_*$ and:

$$\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) \leq \beta + \Delta_k \tag{D.7}$$

irrespective of whether (i) or (ii) holds. There are several cases to consider at this point depending on the value of $|\hat{\phi}_k(1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k, \mathcal{C}) - \beta|$. We show that, in any case, we have $\underline{\lambda}_{k+1} \leq \lambda_*$ and $|\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| \leq 3C \cdot |I_{k+1}|$ which will conclude the proof.

Case 1: $\hat{\phi}_k(1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k, \mathcal{C}) < \beta - \Delta_k$.

In this case, we have $\underline{\lambda}_{k+1} = \underline{\gamma}_k \leq \lambda_*$ and $\bar{\lambda}_{k+1} = 1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k$. Using $\hat{\phi}_k(\bar{\lambda}_{k+1}, \mathcal{C}) < \beta - \Delta_k$ along with the fact that B_{k-1} holds, we get $\phi(\bar{\lambda}_{k+1}, \mathcal{C}) < \beta$ which implies that $\lambda_* \in [\underline{\lambda}_{k+1}, \bar{\lambda}_{k+1}]$ and, as a result, $|\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| = |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq C \cdot |I_{k+1}|$ using Lemma 5.5.

Case 2: $|\hat{\phi}_k(1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k, \mathcal{C}) - \beta| \leq \Delta_k$.

In this case, we have $\underline{\lambda}_{k+1} = \underline{\gamma}_k \leq \lambda_*$ and $\bar{\lambda}_{k+1} = 1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k$. We get:

$$
\begin{aligned}
&|\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| \\
&\quad = |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \\
&\quad \leq |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\bar{\lambda}_{k+1}, \mathcal{C})| + |\phi(\bar{\lambda}_{k+1}, \mathcal{C}) - \hat{\phi}_k(\bar{\lambda}_{k+1}, \mathcal{C})| + |\hat{\phi}_k(\bar{\lambda}_{k+1}, \mathcal{C}) - \beta| \\
&\quad \leq C \cdot |I_{k+1}| + \Delta_k + \Delta_k \\
&\quad \leq 3C \cdot |I_{k+1}|,
\end{aligned}
$$

where we use Lemma 5.5, the fact B_{k-1} hold, and $|\hat{\phi}_k(1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k, \mathcal{C}) - \beta| \leq \Delta_k$ for the second inequality while we use $T \geq \exp(8r^2/C^2)$ for the last inequality.

Case 3: $\hat{\phi}_k(1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k, \mathcal{C}) > \beta + \Delta_k$.

In this case, $\underline{\lambda}_{k+1} = 1/2\underline{\gamma}_k + 1/2\bar{\gamma}_k$ and $\bar{\lambda}_{k+1} = \bar{\gamma}_k$. Since B_{k-1} holds, we get $\phi(\underline{\lambda}_{k+1}, \mathcal{C}) > \beta$ and thus $\underline{\lambda}_{k+1} \leq \lambda_*$. Using (D.7), we have either (a) $\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) < \beta - \Delta_k$ or (b) $|\hat{\phi}_k(\bar{\gamma}_k, \mathcal{C}) - \beta| \leq \Delta_k$. If (a) is true then, since B_{k-1} holds, it must be that $\phi(\bar{\lambda}_{k+1}, \mathcal{C}) < \beta$ and thus we get $\lambda_* \in [\underline{\lambda}_{k+1}, \bar{\lambda}_{k+1}]$ which implies that $|\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| = |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq C \cdot |I_{k+1}|$ using Lemma 5.5. If (b) is true then we have:

$$
\begin{aligned}
&|\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \beta| \\
&\quad = |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \\
&\quad \leq |\phi(\underline{\lambda}_{k+1}, \mathcal{C}) - \phi(\bar{\lambda}_{k+1}, \mathcal{C})| + |\phi(\bar{\lambda}_{k+1}, \mathcal{C}) - \hat{\phi}_k(\bar{\lambda}_{k+1}, \mathcal{C})| + |\hat{\phi}_k(\bar{\lambda}_{k+1}, \mathcal{C}) - \beta| \\
&\quad \leq C \cdot |I_{k+1}| + \Delta_k + \Delta_k \\
&\quad \leq 3C \cdot |I_{k+1}|.
\end{aligned}
$$

where we use (D.7), the fact that B_{k-1} holds, and (b) for the second inequality while we use $T \geq \exp(8r^2/C^2)$ for the last inequality.

D.11 Proof of Theorem 5.3

For any phase $k \in \mathbb{N}$, we denote by t_k the time period at which phase k starts. Using Lemma 5.3, we have:

$$\begin{aligned} R_{B,T} &\leq T \cdot R(\lambda_*, \mathcal{C}) - \mathbb{E}[\sum_{t=1}^{\tau^*-1} v_t \cdot \mathbb{1}_{b_t \geq p_t}] + O(1) \\ &= T \cdot R(\lambda_*, \mathcal{C}) - \mathbb{E}[\sum_{t=1}^{\tau^*} v_t \cdot \mathbb{1}_{b_t \geq p_t}] + O(1). \end{aligned}$$

Since τ^* is a stopping time with respect to the sequence $((x_t, v_t, p_t))_{t \in \mathbb{N}}$ and since $b_t = \min(1, \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta)/\lambda_t)$ is $\tilde{\mathcal{F}}_{t-1}$-measurable, we have:

$$\begin{aligned} \mathbb{E}[\sum_{t=1}^{\tau^*} v_t \cdot \mathbb{1}_{b_t \geq p_t}] &= \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{1}_{\tau^* \geq t} \cdot \mathbb{E}[v_t \mid \tilde{\mathcal{F}}_{t-1}] \cdot \mathbb{1}_{b_t \geq p_t}] \\ &= \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{1}_{\tau^* \geq t} \cdot \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\ &= \mathbb{E}[\sum_{t=1}^{\tau^*} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\ &= \sum_{t=1}^{T} \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] - \mathbb{E}[\sum_{t=\tau^*+1}^{T} \min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\ &\geq \sum_{t=1}^{T} \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] - \frac{1}{r} \cdot \mathbb{E}[\sum_{t=\tau^*+1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t}], \end{aligned}$$

where we use $\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \leq 1$, $p_t \geq r$, and the fact that v_t is independent of p_t conditioned on x_t. Observe that:

$$\sum_{t=\tau^*+1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} = 0 \leq (\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+,$$

if $\tau^* = T + 1$ while:

$$\begin{aligned} \sum_{t=\tau^*+1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} &\leq \sum_{t=\tau^*+1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} + \sum_{t=1}^{\tau^*} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B \\ &\leq (\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+ \end{aligned}$$

if $\tau^* < T+1$ since, in this case, we have $\sum_{t=1}^{\tau^*} p_t \cdot \mathbb{1}_{b_t \geq p_t} \geq B$. We derive:

$$R_{B,T} \leq T \cdot R(\lambda_*, \mathcal{C}) - \sum_{t=1}^{T} \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] + \frac{1}{r} \cdot \mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+] + O(1). \tag{D.8}$$

We bound the two terms appearing in the right-hand side of (D.8) separately starting with the first one. Using the shorthand notation:

$$E = \cap_{k=0}^{\bar{k}_T} \{|\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq 4C \cdot |I_k|,\ |\phi(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \leq 3C \cdot |I_k|\},$$

where $C = \bar{L} \cdot \mathbb{E}[\frac{1}{X^\intercal \theta_*}]$, we have:

$$\begin{aligned}
&T \cdot R(\lambda_*, \mathcal{C}) - \sum_{t=1}^{T} \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&= \sum_{t=1}^{T} \{R(\lambda_*, \mathcal{C}) - \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \geq \lambda_t \cdot p_t}]\} \\
&= \sum_{t=1}^{T} \{R(\lambda_*, \mathcal{C}) - \mathbb{E}[\mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \geq \lambda_t \cdot p_t} \mid \mathcal{F}_{t-1}]]\} \\
&= \sum_{t=1}^{T} \{R(\lambda_*, \mathcal{C}) - \mathbb{E}[R(\lambda_t, \mathcal{C})]\} \\
&\leq \sum_{t=1}^{T} \mathbb{E}[|R(\lambda_*, \mathcal{C}) - R(\lambda_t, \mathcal{C})|] \\
&\leq \frac{1}{r} \cdot \sum_{t=1}^{T} \mathbb{E}[|\phi(\lambda_*, \mathcal{C}) - \phi(\lambda_t, \mathcal{C})|] \\
&\leq \frac{1}{r} \cdot \sum_{k=0}^{\bar{k}_T} N_k \cdot \mathbb{E}[|\phi(\lambda_*, \mathcal{C}) - \phi(\underline{\lambda}_k, \mathcal{C})|] \\
&\leq \frac{T}{r} \cdot \mathbb{P}[E^\complement] + \frac{1}{r} \cdot \sum_{k=0}^{\bar{k}_T} N_k \cdot \mathbb{E}[|\phi(\lambda_*, \mathcal{C}) - \phi(\underline{\lambda}_k, \mathcal{C})| \cdot \mathbb{1}_E] \\
&\leq 2\frac{\ln^2(T)}{r} + \frac{3C}{r} \cdot \sum_{k=0}^{\bar{k}_T} N_k \cdot |I_k| \\
&\leq 2\frac{\ln^2(T)}{r} + \frac{18C}{r^2} \cdot \sum_{k=0}^{\bar{k}_T} 2^k \cdot \ln^2(T) \\
&\leq 2\frac{\ln^2(T)}{r} + \frac{36C}{r^2} \cdot \sqrt{T} \cdot \ln(T).
\end{aligned}$$

To derive the third equality we use the fact that λ_t is $\mathcal{F}_{t-1}$-measurable. For the second inequality, we use Lemma 5.4 For the fourth inequality, we use $\phi(\underline{\lambda}_k, \mathcal{C}), \beta \in [0,1]$. We use Proposition 5.1 to derive the fifth inequality while we use Lemma 5.7 for the last one. We can now focus on the second term appearing in the right-hand side of (D.8):

$$
\begin{aligned}
\mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+] &= \mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \geq \lambda_t \cdot p_t} - B)_+] \\
&\leq \sum_{k=0}^{\bar{k}_T - 1} \mathbb{E}[(\sum_{t=t_k}^{t_{k+1}-1} p_t \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \geq \underline{\lambda}_k \cdot p_t} - N_k \cdot \beta)_+] \\
&+ \mathbb{E}[(\sum_{t=t_{\bar{k}_T}}^{T} p_t \cdot \mathbb{1}_{\min(1, \max_{\theta \in \mathcal{C}} x_t^\intercal \theta) \geq \underline{\lambda}_{\bar{k}_T} \cdot p_t} - (T - \sum_{k=0}^{\bar{k}_T - 1} N_k) \cdot \beta)_+] \\
&\leq \sum_{k=0}^{\bar{k}_T - 1} N_k \cdot \mathbb{E}[(\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C}))_+] \\
&+ (T - \sum_{k=0}^{\bar{k}_T - 1} N_k) \cdot \mathbb{E}[(\hat{\phi}_{\bar{k}_T}(\underline{\lambda}_{\bar{k}_T}, \mathcal{C}) - \phi(\lambda_*, \mathcal{C}))_+] \\
&\leq T \cdot \mathbb{P}[E^{\complement}] + \sum_{k=0}^{\bar{k}_T} N_k \cdot \mathbb{E}[|\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}) - \phi(\lambda_*, \mathcal{C})| \cdot \mathbb{1}_E] \\
&\leq 2 \ln^2(T) + 4C \cdot \sum_{k=0}^{\bar{k}_T} N_k \cdot |I_k| \\
&\leq 2 \ln^2(T) + \frac{24C}{r} \cdot \sum_{k=0}^{\bar{k}_T} 2^k \cdot \ln^2(T) \\
&\leq 2 \ln^2(T) + \frac{48C}{r} \cdot \sqrt{T} \cdot \ln(T).
\end{aligned}
$$

To derive the second inequality, we use $\beta \geq \phi(\lambda_*, \mathcal{C})$. To derive the third inequality, we use $\hat{\phi}_k(\underline{\lambda}_k, \mathcal{C}), \phi(\lambda_*, \mathcal{C}) \in [0,1]$ and $N_{\bar{k}_T} \geq T - \sum_{k=0}^{\bar{k}_T - 1} N_k$. We use Proposition 5.1 to derive the fourth inequality while we use Lemma 5.7 for the last one.

D.12 Proof of Lemma 5.8

We have:

$$\begin{aligned}\det(M_0)\cdot(1+A)^Q &\leq \det(M_T)\\ &\leq \det((T\cdot d)I_d)\\ &= (T\cdot d)^d,\end{aligned}$$

by definition of Q. The second inequality is obtained using $\|x_t\|_\infty \leq 1$ (which implies that $dI_d - x_t x_t^\intercal$ is positive semidefinite) and the fact that $\det(B+C) \geq \det(B)$ for positive semidefinite matrices B and C. Taking logarithms yields the claim.

D.13 Proof of Theorem 5.4

For any master phase $q \in \{0,\cdots,\bar{Q}\}$, we denote by $t_q \in \mathbb{N}$ the round at which phase q starts. For any master phase $q \in \{0,\cdots,\bar{Q}\}$, any phase $k \in \{0,\cdots,\bar{k}_T\}$, and any $\lambda \geq 0$, we denote by $\hat{\phi}_{q,k}(\lambda,\mathcal{C}_q)$ the empirical estimate of $\phi(\lambda,\mathcal{C}_q)$ using all N_k samples obtained during the k-th phase of the binary search that runs during the q-th master phase. We also use the shorthand notations $E = \{\theta_* \in \cap_{t=1}^T \mathcal{C}_t\}$ and:

$$E_q = \cap_{k=0}^{\bar{k}_T}\{|\hat{\phi}_{q,k}(\underline{\lambda}_{q,k},\mathcal{C}_q) - \phi(\lambda_{q,*},\mathcal{C}_q)| \leq 4C\cdot|I_k|,\ |\phi(\underline{\lambda}_{q,k},\mathcal{C}_q) - \phi(\lambda_{q,*},\mathcal{C}_q)| \leq 3C\cdot|I_k|\},$$

for any $q \in \{0,\cdots,\bar{Q}\}$. Using the same analysis as in the proof of Theorem 5.3 with $\mathcal{C} = \{\theta_*\}$ (see (D.8)), we derive:

$$R_{B,T} \leq T\cdot R(\lambda_*,\{\theta_*\}) - \sum_{t=1}^T \mathbb{E}[x_t^\intercal\theta_*\cdot\mathbb{1}_{b_t\geq p_t}] + \frac{1}{r}\cdot\mathbb{E}[(\sum_{t=1}^T p_t\cdot\mathbb{1}_{b_t\geq p_t} - B)_+] + O(1). \tag{D.9}$$

We first study the third term in (D.9). Observe that, along the same lines as what is done in the proof of Theorem 5.3, we have:

$$\begin{aligned}
\mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+] &\leq \mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{b_t \geq p_t} - B)_+ \cdot \mathbb{1}_E] + T \cdot \mathbb{P}[E] \\
&\leq \mathbb{E}[\sum_{q=0}^{Q} \sum_{k=0}^{\bar{k}_q} N_k \cdot |\hat{\phi}_{q,k}(\lambda_{q,k}, \mathcal{C}_q) - \phi(\lambda_{q,*}, \mathcal{C}_q)| \cdot \mathbb{1}_E] + 1 \\
&\leq \mathbb{E}[\sum_{q=0}^{Q} \sum_{k=0}^{\bar{k}_q} 4N_k \cdot C \cdot |I_k|] + T \cdot \sum_{q=0}^{\bar{Q}} \mathbb{P}[E_q^{\complement} \cap E] + O(1) \\
&\leq \frac{24C}{r} \cdot \sum_{q=0}^{\bar{Q}} \sum_{k=0}^{\bar{k}_T} 2^k \cdot \ln^2(T) + 2\ln^2(T) \cdot (\bar{Q}+1) + O(1) \\
&\leq \frac{48C}{r} \cdot \sqrt{T} \cdot \ln(T) \cdot (\bar{Q}+1) + 2\ln^2(T) \cdot (\bar{Q}+1) + O(1) \\
&= O(\frac{d \cdot C}{r \cdot \ln(1+A)} \cdot \sqrt{T} \cdot \ln^2(T \cdot d)) \\
&= \tilde{O}(\frac{d \cdot C}{r \cdot \ln(1+A)} \cdot \sqrt{T}).
\end{aligned}$$

We use the same analysis as in the proof of Theorem 5.3 along with Lemma 5.2 to derive the second inequality. We use Proposition 5.1 for the fourth inequality and we use Lemma 5.8 to get the final asymptotic bound. We move on to study the second term in (D.9). Denoting by $\tilde{\theta}_t$ an arbitrary element of $\operatorname{argmax}_{\theta \in \mathcal{C}_{\tau_t}} x_t^\intercal \theta$, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[x_t^\intercal \theta_* \cdot \mathbb{1}_{b_t \geq p_t}] &= \sum_{t=1}^{T} \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_{\tau_t}} x_t^\intercal \theta) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\quad - \mathbb{E}[\sum_{t=1}^{T} (\min(1, \max_{\theta \in \mathcal{C}_{\tau_t}} x_t^\intercal \theta) - x_t^\intercal \theta_*) \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\geq \sum_{t=1}^{T} \mathbb{E}[R(\lambda_t, \mathcal{C}_{\tau_t})] - \mathbb{E}[\sum_{t=1}^{T} |x_t^\intercal \tilde{\theta}_t - x_t^\intercal \theta_*| \cdot \mathbb{1}_{b_t \geq p_t}] \\
&\geq \sum_{t=1}^{T} \mathbb{E}[R(\lambda_t, \mathcal{C}_{\tau_t})] + \tilde{O}(d \cdot \sqrt{A \cdot T}),
\end{aligned}$$

where the last inequality is obtained in the proof of Theorem 5.2. Hence, what remains to be done to get the regret bound is to upper bound:

$$T \cdot R(\lambda_*, \{\theta_*\}) - \sum_{t=1}^{T} \mathbb{E}[R(\lambda_t, \mathcal{C}_{\tau_t})].$$

First note that:

$$\begin{aligned}
&\mathbb{E}[|\sum_{t=1}^{T} R(\lambda_t, \mathcal{C}_{\tau_t}) - \sum_{q=0}^{Q}(t_{q+1} - t_q) \cdot R(\lambda_{q,*}, \mathcal{C}_q)|] \\
&\leq \mathbb{E}[\sum_{q=0}^{Q}\sum_{k=0}^{\bar{k}_q} N_k \cdot |R(\underline{\lambda}_{q,k}, \mathcal{C}_q) - R(\lambda_{q,*}, \mathcal{C}_q)|] \\
&\leq \frac{1}{r} \cdot \mathbb{E}[\sum_{q=0}^{Q}\sum_{k=0}^{\bar{k}_q} N_k \cdot |\phi(\underline{\lambda}_{q,k}, \mathcal{C}_q) - \phi(\lambda_{q,*}, \mathcal{C}_q)|] \\
&\leq \frac{1}{r} \cdot (T \cdot \mathbb{P}[E] + \sum_{q=0}^{\bar{Q}} T \cdot \mathbb{P}[E_q^{\complement} \cap E] + \sum_{q=0}^{\bar{Q}}\sum_{k=0}^{\bar{k}_T} 3N_k \cdot C \cdot |I_k|) \\
&\leq \frac{1}{r} \cdot (1 + 2\ln^2(T) \cdot (\bar{Q} + 1) + \frac{48C}{r} \cdot \sqrt{T} \cdot \ln(T) \cdot (\bar{Q} + 1)) \\
&= \tilde{O}(\frac{d \cdot C}{r^2 \cdot \ln(1 + A)} \cdot \sqrt{T}).
\end{aligned}$$

We derive the second inequality using Lemma 5.4 We derive the fourth inequality using Lemma 5.2 and Proposition 5.1 in the same fashion as done for the third term in (D.9). We conclude that all that is left to be done is to upper bound:

$$T \cdot R(\lambda_*, \{\theta_*\}) - \mathbb{E}[\sum_{q=0}^{Q}(t_{q+1} - t_q) \cdot R(\lambda_{q,*}, \mathcal{C}_q)],$$

which we do next. Using Lemma 5.3, observe that, conditioned on $\mathcal{F}_{t_q - 1}$ and assuming that $\theta_* \in \mathcal{C}_q$, $R(\lambda_{q,*}, \mathcal{C}_q)$ is almost surely larger than $\mathrm{ER}_{\mathrm{OPT}}(B, T)/T + O(1/T)$ by definition of $\lambda_{q,*}$ when $\mathcal{C} = \mathcal{C}_q$. Note that bidding $\tilde{b_t} = \min(x_t^{\intercal}\theta_*/\lambda_*, 1)$ at any time period t is a valid

algorithm for this problem that yields an expected total reward:

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{\tau^*} v_t \cdot \mathbb{1}_{\tilde{b}_t \geq p_t}] \\
&\quad \geq T \cdot \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_q} X^\intercal \theta) \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P}] - \frac{1}{r} \cdot \mathbb{E}[(\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{\tilde{b}_t \geq p_t} - B)_+] \\
&\quad \geq T \cdot \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_q} X^\intercal \theta) \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P}] - \frac{1}{r} \cdot \mathbb{E}[|\sum_{t=1}^{T} p_t \cdot \mathbb{1}_{\tilde{b}_t \geq p_t} - T \cdot \phi(\lambda_*, \{\theta_*\})|] \\
&\quad \geq T \cdot \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_q} X^\intercal \theta) \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P}] - \frac{\sqrt{T}}{r},
\end{aligned}
$$

where the expectations are all conditioned on $\mathcal{F}_{t_q - 1}$ and the inequalities hold almost surely. The first inequality is derived in the same fashion as done in the proof of Theorem 5.3 to derive (D.8). The second inequality is a consequence of $B = \beta \cdot T$ and $\phi(\lambda_*, \{\theta_*\}) \leq \beta$. The third inequality is obtained with Khintchine's inequality (by symmetrization) since $p_t \in [0, 1]$ and $(p_t \cdot \mathbb{1}_{\tilde{b}_t \geq p_t})_{t \in \mathbb{N}}$ is an i.i.d. stochastic process with mean $\phi(\lambda_*, \{\theta_*\})$. We conclude that:

$$
\begin{aligned}
R(\lambda_{q,*}, \mathcal{C}_q) &\geq \mathbb{E}[\min(1, \max_{\theta \in \mathcal{C}_q} X^\intercal \theta) \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P} | \mathcal{F}_{t_q - 1}] - \frac{1}{r \cdot \sqrt{T}} + O(\frac{1}{T}) \\
&\geq \mathbb{E}[X^\intercal \theta_* \cdot \mathbb{1}_{X^\intercal \theta_* \geq \lambda_* \cdot P} | \mathcal{F}_{t_q - 1}] - \frac{1}{r \cdot \sqrt{T}} + O(\frac{1}{T}) \\
&= R(\lambda_*, \{\theta_*\}) - \frac{1}{r \cdot \sqrt{T}} + O(\frac{1}{T})
\end{aligned}
$$

almost surely as long as $\theta_* \in \mathcal{C}_q$. This implies that:

$$
\begin{aligned}
&T \cdot R(\lambda_*, \{\theta_*\}) - \mathbb{E}[\sum_{q=0}^{Q} (t_{q+1} - t_q) \cdot R(\lambda_{q,*}, \mathcal{C}_q)] \\
&\quad = \mathbb{E}[\sum_{q=0}^{Q} (t_{q+1} - t_q) \cdot (R(\lambda_*, \{\theta_*\}) - R(\lambda_{q,*}, \mathcal{C}_q))] \\
&\quad \leq \mathbb{E}[\sum_{q=0}^{Q} (t_{q+1} - t_q) \cdot (\mathbb{1}_E + \frac{1}{r \cdot \sqrt{T}} + O(\frac{1}{T}))] \\
&\quad \leq T \cdot \mathbb{P}[E] + \frac{\sqrt{T}}{r} + O(1) \\
&\quad = O(\frac{\sqrt{T}}{r}),
\end{aligned}
$$

where we use Lemma 5.2 for the last step. This concludes the proof.