Online Learning with Knapsacks: the Best of Both Worlds

Matteo Castiglioni [*1] **Andrea Celli** [*2] **Christian Kroer** [3]

Abstract

We study online learning problems in which a decision maker wants to maximize their expected reward without violating a finite set of m resource constraints. By casting the learning process over a suitably defined space of *strategy mixtures*, we recover strong duality on a Lagrangian relaxation of the underlying optimization problem, even for general settings with non-convex reward and resource-consumption functions. Then, we provide the first best-of-both-worlds type framework for this setting, with no-regret guarantees both under stochastic and adversarial inputs. Our framework yields the same regret guarantees of prior work in the stochastic case. On the other hand, when budgets grow at least linearly in the time horizon, it allows us to provide a *constant* competitive ratio in the adversarial case, which improves over the $O(m \log T)$ competitive ratio of (Immorlica et al., 2019). Moreover, our framework allows the decision maker to handle non-convex reward and cost functions. We provide two game-theoretic applications of our framework to give further evidence of its flexibility.

1. Introduction

In this paper, we study online learning problems in which the decision maker has to satisfy supply or budget constraints. In particular, the decision maker is endowed with $m \geq 1$ limited resources which are consumed over time. For each round t up to the time horizon T, the decision maker chooses a strategy $\boldsymbol{\xi}_t$ which defines a probability measure over the set of actions. Then, they observe some feedback about the reward and resource consumption incurred by playing $\boldsymbol{\xi}_t$. The process stops at time horizon T, or when the total consumption of some resource exceeds its budget. The goal is to maximize the total reward. Our framework can be instantiated both in the *full-information feedback* as well as in the *bandit feedback* setting. In the full information feedback setting, the decision maker observes the reward function f_t and resource-consumption function c_t at each t. In the bandit feedback setting they only get to observe $(f_t(\boldsymbol{x}_t), c_t(\boldsymbol{x}_t))$, where $\boldsymbol{x}_t$ is the realized action selected according to $\boldsymbol{\xi}_t$.

Our framework subsumes the well-known *Bandits with Knapsacks* problem originally introduced by Badanidiyuru et al. (2013; 2018), which has various motivating applications such as dynamic pricing (Besbes & Zeevi, 2009; Babaioff et al., 2012; Besbes & Zeevi, 2012; Wang et al., 2014), dynamic procurement (Badanidiyuru et al., 2012; Singla & Krause, 2013), and online ad allocation (Slivkins, 2013; Combes et al., 2015). Moreover, our framework also subsumes online packing problems, in which the decision maker observes full feedback *before* choosing a strategy (Mehta et al., 2013; Buchbinder & Naor, 2009).

Original contributions We focus on the regime in which $B = \Omega(T)$, that is the budget grows at least linearly in the time horizon T. This is the case, for example, when the decision maker has a fixed per-iteration budget as in most motivating applications, such as online advertising (see, *e.g.*, Balseiro et al. (2020b)). In this setting, we resolve the following two open questions posed by Immorlica et al. (2019):

- *Is it possible to solve both stochastic and adversarial version of BwK with exactly the same algorithm?*
- *Is it possible to obtain a constant-factor competitive ratio in the adversarial case for the regime $B = \Omega(T)$?*

We answer both questions positively. In doing so, by casting the learning process over a suitably defined space of strategy mixtures, we recover strong duality of the Lagrangian relaxation of the underlying optimization problem even when f_t and c_t are non-convex functions, and the set of available actions $\mathcal{X}$ may be non-convex and non-compact. This strictly generalizes the setting studied by Immorlica et al. (2019). We show that our meta-algorithm (Algorithm 1) guarantees a tight regret bound in the stochastic case, matching known

[*]Equal contribution [1]DEIB, Politecnico di Milano, Milan, Italy [2]Department of Computing Sciences, Bocconi University, Milan, Italy [3]IEOR Department, Columbia University, New York, NY. Correspondence to: Matteo Castiglioni <matteo.castiglioni@polimi.it>, Andrea Celli <andrea.celli2@unibocconi.it>, Christian Kroer <christian.kroer@columbia.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

results by Badanidiyuru et al. (2013); Agrawal & Devanur (2014); Immorlica et al. (2019). Moreover, in the adversarial case, it guarantees a constant approximation ratio which is computed as the maximum single-round resource consumption over the per-iteration budget. This improves over the $O(m \log T)$ ratio by Immorlica et al. (2019), and over the recent $O(\log m \log T)$ ratio by Kesselheim & Singla (2020). In doing so, we provide the first *best-of-both-worlds* algorithm for online learning problems with knapsacks. This result allows our framework to achieve good worst-case performance, while being able to take advantage of *well-behaved* problem instances. This makes progress on the line of work initiated by Bubeck & Slivkins (2012). Finally, we describe two novel motivating applications of particular interest. First, we show that our framework may be employed to extend the work by Balcan et al. (2015) to repeated Stackelberg Security Games in which resources are costly, and the planner must satisfy some resource-consumption constraints. Then, we argue that our framework can be adapted to handle budget-pacing mechanisms in the context of repeated first-price auctions. This is particularly relevant for modern auction markets operated by large Internet advertising companies.

2. Related Work

We highlight the most relevant papers with respect to our work. For a more in depth discussion of related work the reader can refer to Slivkins (2019, Chapter 10).

The *Bandits with Knapsacks* (BwK) framework was introduced and optimally solved by Badanidiyuru et al. (2013; 2018). Other regret-optimal algorithms for Stochastic BwK have been proposed by Agrawal & Devanur (2014; 2019), and by Immorlica et al. (2019). The BwK framework has been subsequently extended to numerous settings such as, for example, more general notions of resources and constraints (Agrawal & Devanur, 2014; 2019), contextual bandits (Dudik et al., 2011; Badanidiyuru et al., 2014; Agarwal et al., 2014a; Agrawal et al., 2016), and combinatorial semi-bandits (Sankararaman & Slivkins, 2018).

The *Adversarial Bandits with Knapsacks* setting was first studied by Immorlica et al. (2019), who proved a $O(m \log T)$ competitive ratio. Recently, Kesselheim & Singla (2020) refined that analysis to obtain an $O(\log m \log T)$ competitive ratio for the adversarial setting. Moreover, Cardoso et al. (2019) study a related problem in which the algorithm can continue up to time T, with no stopping rule as in the standard BwK framework.

Best-of-both-worlds-type algorithms usually guarantee optimal regret rates in both adversarial and stochastic settings, without being aware of which environment they are in. Various work study the case of bandits without budget constraints (see, *e.g.*, (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016)). In the context of online allocation problems with fixed per-iteration budget, Balseiro et al. (2020a;b) propose a class of algorithms which attain asymptotically optimal performance in the stochastic case, and they attain an asymptotically optimal constant-factor competitive ratio when the input is adversarial. In their setting, at each round, the input (f_t, c_t) is observed by the decision maker *before* they make a decision. This makes the problem essentially different from ours. Finally, Rangi et al. (2019) present an algorithm for the stochastic and adversarial setting for the special case when there is only one constrained resource, including time (this particular setting admits much stronger performance guarantees). We mention that other results in the simplified setting with one constrained resource have been obtained by György et al. (2007); Tran-Thanh et al. (2010; 2012).

Another line of related work concerns online convex optimization with constraints (see, *e.g.*, (Mahdavi et al., 2012; 2013; Chen et al., 2017; Neely & Yu, 2017; Chen & Giannakis, 2018)), where it is usually assumed that the action set is a convex subset of $\mathbb{R}^m$, in each round rewards (resp., costs) are concave (resp., convex), and, most importantly, resource constraints only apply at the last round, while in BwK the budget constraints hold for all rounds.

3. Preliminaries

We denote vectors by bold fonts. Given vector $\boldsymbol{x}$, let $\boldsymbol{x}[i]$ be its i-th component. The set $\{1, \dots, n\}$, with $n \in \mathbb{N}_{>0}$, is compactly denoted as $[n]$. Finally, given a discrete set S, we denote by Δ^S the $|S|$-simplex.

Basic Setup There are T rounds and m resources. A decision maker has a non-empty set of available strategies $\mathcal{X} \subseteq \mathbb{R}^n$ (this set may be non-convex, integral, and even non-compact). In each round $t \in [T]$, the decision maker chooses $\boldsymbol{x}_t \in \mathcal{X}$, and subsequently observes a reward function $f_t : \mathcal{X} \to [0,1]$, and a function $c_t : \mathcal{X} \to [0,1]^m$ specifying resources consumption (both f_t and c_t need not be convex). Each resource $i \in [m]$ is endowed with a budget B_i to be spent over the T steps. Since $c_t(\boldsymbol{x})[i] \geq 0$, for all t, i, and $\boldsymbol{x}$, budgets cannot be replenished. We denote by $\boldsymbol{\rho} \coloneqq (\rho_1, \dots, \rho_m) \in \mathbb{R}^m_{>0}$ the vector of per-iteration budgets, where for each $i \in [m]$ we have $B_i = T\rho_i$. Without loss of generality we let $\rho_1 = \dots = \rho_m = \rho$, and $B_1 = B_2 = \dots = B_m = B$. A problem with arbitrary budgets can be reduced to this setting by dividing, for each resource $i \in [m]$, all per-round resource consumption $c_t(\cdot)[i]$ by $B_i / \min_j B_j$. We focus on the regime $B = \Omega(T)$, and we study two feedback models: *bandit feedback* (no auxiliary feedback other than $f_t(\boldsymbol{x}_t), c_t(\boldsymbol{x}_t)$ is observed by the decision maker), and *full feedback* (f_t, c_t are observed). Let

$\gamma_t \coloneqq (f_t, c_t)$, and $\boldsymbol{\gamma}_T \coloneqq (\gamma_t)_{t=1}^T$ be the sequence of inputs up to time T. At each step t, the decision maker can condition their decision on $\boldsymbol{\gamma}_{t-1}$, and on the sequence of prior decisions $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}$, but no information about future rewards or resource consumption is available. The repeated decision making process stops at any round $\tau \leq T$ in which the total consumption of any resource i exceeds its budget B_i. The goal of the decision maker is to maximize its total reward. Following previous work (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014; Immorlica et al., 2019), we assume there exists a *void action* $\varnothing \in \mathcal{X}$ with reward 0, and such that $c_t(\varnothing)[i] = 0$ for all resources i. This guarantees the existence of a feasible solution (i.e., a sequence of decisions which do not violate resource constraints).

Regret Minimization A *regret minimizer* for an arbitrary set $\mathcal{W}$ is an abstract model for a decision maker that repeatedly interacts with a black-box environment. At each time t, the regret minimizer can perform two operations: (i) NEXTELEMENT(): this procedure outputs an element $\boldsymbol{w}_t \in \mathcal{W}$; (ii) OBSERVEUTILITY($\ell_t$): this procedure updates the internal state of the regret minimizer using the environment's feedback, in the form of a utility function $\ell_t : \mathcal{W} \to \mathbb{R}$. The utility function can depend adversarially on the sequence of outputs $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{t-1}$. The decision making process encoded by the regret minimizer is *online*: at each time t, the output of the regret minimizer can depend on the sequence $(\boldsymbol{w}_{t'}, \ell_{t'})_{t'=1}^{t-1}$, but no information about future utilities is available. The objective of the regret minimizer is to output a sequence of points in $\mathcal{W}$ so that the *cumulative regret*

$$R^T \coloneqq \sup_{\boldsymbol{w}^* \in \mathcal{W}} \sum_{t=1}^{T} (\ell_t(\boldsymbol{w}^*) - \ell_t(\boldsymbol{w}_t))$$

grows asymptotically sublinearly in the time T. For a review of the various regret minimizers available for the full and bandit feedback setting see Cesa-Bianchi & Lugosi (2006).

4. Strategy Mixtures

We will need to work with the set of probability measures on the Borel sets of $\mathcal{X}$. We refer to this set as the set of *strategy mixtures* and denote it as Ξ. We endow $\mathcal{X}$ with the Lebesgue σ-algebra. We assume that all possible functions f_t, c_t are measurable with respect to every probability measure $\boldsymbol{\xi} \in \Xi$. This ensures that the various expectations taken are well-defined, since the functions are assumed to be bounded above, and are therefore integrable.

It is well-known that the Dirac measures δ_x for $x \in \mathcal{X}$ which assign 1 to a set $\mathcal{A} \subseteq \mathcal{X}$ if and only if $x \in \mathcal{A}$ form the extreme points of the convex set of strategy mixtures Ξ. The Dirac mass $\delta_\varnothing$ is the strategy that deterministically plays the void action. We define $\boldsymbol{\xi}_\varnothing \coloneqq \delta_\varnothing$.

4.1. On Strong Duality

Given two arbitrary measurable functions $f : \mathcal{X} \to [0,1]$, $c : \mathcal{X} \to [0,1]^m$, we define the following linear program, which chooses the strategy $\boldsymbol{\xi}$ that maximizes the reward f, while keeping the expected consumption of every resource $i \in [m]$ given c below the target ρ:

$$\texttt{OPT}_{f,c}^{\texttt{LP}} \coloneqq \begin{cases} \sup_{\boldsymbol{\xi} \in \Xi} \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] \\ \text{s.t. } \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] \leq \boldsymbol{\rho} \end{cases}, \tag{1}$$

where $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c_t(\boldsymbol{x})] = (\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c_t(\boldsymbol{x})[i]])_{i=1}^m \in [0,1]^m$.

By letting $\mathcal{I}$ be an arbitrary set of possible input pairs (f, c), the Lagrangian relaxation of LP (1) is defined as follows.

Definition 4.1 (Lagrangian Function)**.** The *Lagrangian function* $L : \Xi \times \mathbb{R}_{\geq 0}^m \times \mathcal{I} \to \mathbb{R}$ is such that, for any $\boldsymbol{\xi} \in \Xi$, $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^m$, $(f, c) \in \mathcal{I}$ it holds

$$L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) \coloneqq \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] + \langle \boldsymbol{\lambda}, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] \rangle.$$

Next, we show that, when the decision maker is allowed to choose a strategy mixture, we can recover strong duality even if f and c are arbitrary non-convex functions (omitted proofs can be found in Appendix A).

Theorem 4.2. *Let $f : \mathcal{X} \to [0,1]$, $c : \mathcal{X} \to [0,1]^m$, and $(f, c) \in \mathcal{I}$. It holds:*

$$\sup_{\boldsymbol{\xi} \in \Xi} \inf_{\boldsymbol{\lambda} \geq 0} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \inf_{\boldsymbol{\lambda} \geq 0} \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \texttt{OPT}_{f,c}^{LP}.$$

This theorem can be derived from Luenberger (1997, Theorem 1, §8.6). For completeness and ease of readability, we give a proof specific to our setting. The proof is based on standard convex-optimization arguments. In general, a semi-infinite linear optimization problem does not admit strong duality (see the example in Appendix B). However, the existence of a strategy mixture $\boldsymbol{\xi}_\varnothing$ corresponding to playing deterministically the void action $\varnothing$ allows us to follow closely the standard proof of strong duality via Slater's condition, since it yields the existence of a strictly feasible solution.

Then, we show that we can restrict the set of admissible dual vectors to

$$\mathcal{D} \coloneqq \{\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0} : \|\boldsymbol{\lambda}\|_1 \leq 1/\rho\}, \tag{2}$$

while continuing to satisfy strong duality.

Lemma 4.3. *Let $\mathcal{D}$ be defined as in Equation (2). Given $f : \mathcal{X} \to [0,1]$, $c : \mathcal{X} \to [0,1]^m$, $(f, c) \in \mathcal{I}$, it holds*

$$\sup_{\boldsymbol{\xi} \in \Xi} \inf_{\boldsymbol{\lambda} \in \mathcal{D}} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \inf_{\boldsymbol{\lambda} \in \mathcal{D}} \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \texttt{OPT}_{f,c}^{LP}.$$

From Lemma 4.3 we have that $\boldsymbol{\lambda}$ is chosen from a compact set. By noting that the supremum over a set of lower semi-continuous (LSC) functions is LSC (Aliprantis et al., 2006,

Lemma 2.41), we get that $\sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c)$ is LSC as a function of $\boldsymbol{\lambda}$. It follows by a generalization of Weierstrass' theorem that an optimal $\boldsymbol{\lambda}^*$ exists (Aliprantis et al., 2006, Theorem 2.43). Therefore, going forward, we will replace all infima over $\mathcal{D}$ by minima when needed.

4.2. Baselines

In this section we provide details on the baselines for the case of adversarial and stochastic inputs, respectively.

Baseline adversarial setting Given a sequence of inputs $\boldsymbol{\gamma}_T$, the baseline for the adversarial setting is the total expected reward of the best *fixed* policy in Ξ, such that strategies are drawn from the same fixed mixture until the budget is fully depleted, and the void action is selected afterwards. Following the notation of Immorlica et al. (2019), we denote its value by $\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}}$. Given $\tau \in [T]$, we write $\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma},\tau}$ to denote the expected reward of the best fixed policy for inputs restricted to $(\gamma_1, \dots, \gamma_\tau)$. Moreover, for any sequence of inputs $\boldsymbol{\gamma}_T$, and for any $\tau \in [T]$, let $\tilde{f}_\tau : \mathcal{X} \to [0,1]$ and $\tilde{c}_\tau : \mathcal{X} \to [0,1]^m$ be such that, for each $\boldsymbol{x} \in \mathcal{X}$:

$$\tilde{f}_\tau(\boldsymbol{x}) \coloneqq \frac{1}{\tau}\sum_{t=1}^{\tau} f_t(\boldsymbol{x}) \quad \text{and} \quad \tilde{c}_\tau(\boldsymbol{x}) \coloneqq \frac{1}{\tau}\sum_{t=1}^{\tau} c_t(\boldsymbol{x}). \tag{3}$$

Then, for $\tau \in [T]$, we define $\texttt{OPT}^{\texttt{LP}}_{\tilde{f}_\tau,\tilde{c}_\tau}$ according to Equation (1). The value of these LPs will be essential during the regret analysis.

Baseline stochastic setting In the stochastic version of the problem, each input $\gamma_t = (f_t, c_t)$ is drawn i.i.d. from some unknown distribution $\mathcal{P}$ over a set of possible input pairs $\mathcal{I}$. Let $\bar{f} : \mathcal{X} \to [0,1]$ be the expected reward function, and $\bar{c} : \mathcal{X} \to [0,1]^m$ be the expected resource-consumption function (where both expectations are taken w.r.t. $\mathcal{P}$). Let Ψ be the set of dynamic policies specifying the current strategy mixture $\boldsymbol{\xi}_t$ as a function of the past history. The baseline for the stochastic setting $\texttt{OPT}^{\texttt{DP}}$ is given by $\texttt{OPT}^{\texttt{DP}} \coloneqq \sup_{\psi\in\Psi} \mathbb{E}_{\boldsymbol{\gamma}\sim\mathcal{P}}[\texttt{OPT}^{\texttt{DP}}_{\psi,\boldsymbol{\gamma}}]$, where $\texttt{OPT}^{\texttt{DP}}_{\psi,\boldsymbol{\gamma}}$ is the value of the policy ψ under the sequence of inputs $\boldsymbol{\gamma}$. Intuitively, $\texttt{OPT}^{\texttt{DP}}$ is the value of the best dynamic policy when the decision maker knows $\mathcal{P}$, but gets to observe the realized γ_t only after having made the decision at t. In the following, we use the solution to LP (1) initialized with reward function $\bar{f}$, and cost function $\bar{c}$, as an upper bound to the value of the optimal policy $\texttt{OPT}^{\texttt{DP}}$. In particular, we prove the following.

Lemma 4.4. *Given a distribution over inputs $\mathcal{P}$, let $\bar{f} : \mathcal{X} \to [0,1]$ be the expected reward function, and $\bar{c} : \mathcal{X} \to [0,1]^m$ be the expected resource-consumption function. Then, $T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{c}} \geq \texttt{OPT}^{\texttt{DP}}$.*

5. Meta-Algorithm

Our algorithm is based on the classic primal-dual approach usually employed in online problems with packing constraints (see, *e.g.*, (Balseiro et al., 2020b; Immorlica et al., 2019)). Our framework assumes access to two regret minimizers with the following characteristics. The first one, which we denote by $\mathcal{R}^{\texttt{P}}$, is the *primal regret minimizer* which outputs strategy mixtures in Ξ, and receives as feedback the linear utility $\ell^{\texttt{P}}_t : \Xi \to \mathbb{R}$ such that, for each $\boldsymbol{\xi} \in \Xi$, $\ell^{\texttt{P}}_t(\boldsymbol{\xi}) \coloneqq \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle$. The second regret minimizer, which we denote by $\mathcal{R}^{\texttt{D}}$, is the *dual regret minimizer*, and it outputs points in the space of dual variables $\mathcal{D}$. Moreover, $\mathcal{R}^{\texttt{D}}$ receives as feedback the linear utility $\ell^{\texttt{D}}_t : \mathcal{D} \to \mathbb{R}$ such that, for each $\boldsymbol{\lambda} \in \mathcal{D}$, $\ell^{\texttt{D}}_t(\boldsymbol{\lambda}) \coloneqq -\langle \boldsymbol{\lambda}, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle$. The primal regret minimizer $\mathcal{R}^{\texttt{P}}$ may have either bandit or full feedback, depending on the setting of interest. The dual regret minimizer $\mathcal{R}^{\texttt{D}}$ has full feedback by construction. Finally, we denote by $\mathcal{E}^{\texttt{P}}_T$ (resp., $\mathcal{E}^{\texttt{D}}_T$) the upper bound on the cumulative regret guaranteed by $\mathcal{R}^{\texttt{P}}$ (resp., $\mathcal{R}^{\texttt{D}}$).

Remark 5.1. In order to guarantee convergence in both the stochastic and adversarial setting, it will be enough to set $\mathcal{D} \coloneqq \{\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0} : \|\boldsymbol{\lambda}\|_1 \leq 1/\rho\}$ (see Sections 6 and 7). Therefore, a natural choice for the dual regret minimizing algorithm $\mathcal{R}^{\texttt{D}}$ is, for example, *online mirror descent* (OMD) with negative entropy as reference function, which guarantees a regret upper bound of $\mathcal{E}^{\texttt{D}}_T = O(1/\rho\sqrt{T\log(m+1)})$ (Nemirovskij & Yudin, 1983; Beck & Teboulle, 2003).

Algorithm 1 summarizes the structure of our meta-algorithm. For each t, the meta-algorithm first computes a primal and dual decision through $\mathcal{R}^{\texttt{P}}$ and $\mathcal{R}^{\texttt{D}}$, respectively (see the invocation of NEXTELEMENT()). The action played by the decision maker at t is going to be $\boldsymbol{x}_t \sim \boldsymbol{\xi}_t$. Then, (f_t, c_t) are observed, and the budget consumption is updated according to the realized cost vector $c_t(\boldsymbol{x}_t)$. Finally, the internal state of the two regret minimizer is updated according to the feedback specified by $\ell^{\texttt{P}}_t, \ell^{\texttt{D}}_t$ (see the invocation of OBSERVEUTILITY()). Notice that the primal regret minimizer $\mathcal{R}^{\texttt{P}}$ may as well receive partial feedback (*i.e.*, observe only $(f_t(\boldsymbol{x}_t), c_t(\boldsymbol{x}_t))$ for each t). In the following sections we show that, with an appropriate choice of $\mathcal{R}^{\texttt{P}}$, Algorithm 1 can be adapted also to this setting. The algorithm terminates when the agent has no sufficient budget, or when the time horizon T is reached.

6. Regret Bound for the Adversarial Setting

In this section, we assume that, for each $t \in [T]$, the request (f_t, c_t) is chosen by an oblivious adversary, and we look at the worst-case performance over all possible inputs. We show that, in this setting, Algorithm 1 is α-competitive, with $\alpha \coloneqq 1/\rho$. This is, to the best of our knowledge, the first constant-factor competitive ratio for the adversarial setting

Algorithm 1 Meta-algorithm for strategy mixture Ξ.

Input: parameters B, T, primal regret minimizer $\mathcal{R}^{\mathrm{P}}$, dual regret minimizer $\mathcal{R}^{\mathrm{D}}$
Initialization: $\forall i \in [m], B_{i,1} \leftarrow B$, $\boldsymbol{\rho} \leftarrow \mathbf{1} \cdot B/T$, and initialize $\mathcal{R}^{\mathrm{P}}, \mathcal{R}^{\mathrm{D}}$.
for $t = 1, 2, \dots, T$ **do**
Primal decision: $\Xi \ni \boldsymbol{\xi}_t \leftarrow \mathcal{R}^{\mathrm{P}}.\textsc{NextElement}()$,

$$\boldsymbol{x}_t \leftarrow \begin{cases} \boldsymbol{x} \sim \boldsymbol{\xi}_t & \text{if } B_{i,t} \geq 1,\ \forall i \in [m] \\ \varnothing & \text{otherwise} \end{cases}.$$

Dual decision: $\mathcal{D} \ni \boldsymbol{\lambda}_t \leftarrow \mathcal{R}^{\mathrm{D}}.\textsc{NextElement}()$
Observe request: observe (f_t, c_t) and update available resources: $B_{i,t+1} \leftarrow B_{i,t} - c_t(\boldsymbol{x}_t)[i]$, $\forall i \in [m]$.
Primal update:
- $\ell_t^{\mathrm{P}} \leftarrow$ linear utility defined as

$$\ell_t^{\mathrm{P}} : \Xi \ni \boldsymbol{\xi} \mapsto \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle$$

- $\mathcal{R}^{\mathrm{P}}.\textsc{ObserveUtility}(\ell_t^{\mathrm{P}})$

Dual update:
- $\ell_t^{\mathrm{D}} \leftarrow$ linear utility defined as

$$\ell_t^{\mathrm{D}} : \mathcal{D} \ni \boldsymbol{\lambda} \mapsto -\langle \boldsymbol{\lambda}, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle$$

- $\mathcal{R}^{\mathrm{D}}.\textsc{ObserveUtility}(\ell_t^{\mathrm{D}})$

end for

in the $B = \Omega(T)$ regime. The competitive ratio α, being defined as the maximum cost (*i.e.*, 1 in our setting) over the per-round budget, captures the *relative wealthiness* of the decision maker. As one may expect, bidding strategies may perform poorly when budgets are small compared to the costs, but better performance can be guaranteed with bigger budgets. In particular, we provide the following convergence guarantees of Meta-Algorithm 1 in the adversarial setting.

Theorem 6.1. *Consider Meta-Algorithm 1 equipped with two arbitrary regret minimizers $\mathcal{R}^P$ and $\mathcal{R}^D$ for the sets Ξ and $\mathcal{D}$, respectively. Suppose they guarantee cumulative regret up to time T which is upper bounded by $\mathcal{E}_T^P$ and $\mathcal{E}_T^D$, respectively. Suppose requests are chosen by an oblivious adversary. Letting $\alpha := 1/\rho$, for each $\delta > 0$ we have*

$$\mathtt{OPT}_{\boldsymbol{\gamma}}^{FP} - \alpha\, \mathtt{REW}_{\boldsymbol{\gamma}} \leq O(\alpha^2 \sqrt{T \ln(T/\delta)}) + \mathcal{E}_T^P + \mathcal{E}_T^D$$

with probability at least $1 - \delta$, where $\mathtt{REW}_{\boldsymbol{\gamma}} := \sum_t f_t(\boldsymbol{x}_t)$ is the reward of the algorithm for the sequence of inputs $\boldsymbol{\gamma}$.

Proof. Let τ be the stopping time of Algorithm 1, i.e. when $B_{i,t} < 1$. We proceed in three steps.

Step 1: lower bound on the reward up to τ. First, we provide a lower bound on the reward guaranteed by Algorithm 1 up to the stopping time τ. The cumulative external regret of $\mathcal{R}^{\mathrm{P}}$ up to the stopping time τ is

$$R_\tau^{\mathrm{P}} = \sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} \big(\ell_t^{\mathrm{P}}(\boldsymbol{\xi}) - \ell_t^{\mathrm{P}}(\boldsymbol{\xi}_t)\big) \leq \mathcal{E}_\tau^{\mathrm{P}}.$$

By definition of ℓ_t^{P}, we have

$$\sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} (\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] + \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle) \leq \mathcal{E}_\tau^{\mathrm{P}}.$$

Then, by rearranging,

$$\sum_{t=1}^{\tau} \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] \geq \sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} (\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle + \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x}_t)]\rangle) - \mathcal{E}_\tau^{\mathrm{P}}. \quad (4)$$

By definition of dual regret minimizer $\mathcal{R}^{\mathrm{D}}$, for any $\boldsymbol{\lambda} \in \mathcal{D}$ (we will specify a precise value for $\boldsymbol{\lambda}$ in the final step of the proof), we have $\sum_{t=1}^{\tau} (\ell_t^{\mathrm{D}}(\boldsymbol{\lambda}) - \ell_t^{\mathrm{D}}(\boldsymbol{\lambda}_t)) \leq \mathcal{E}_\tau^{\mathrm{D}}$. Then, by definition of ℓ_t^{D},

$$\sum_{t=1}^{\tau} \langle \boldsymbol{\lambda}_t, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle \geq \sum_{t=1}^{\tau} (\langle \boldsymbol{\lambda}_t, \boldsymbol{\rho}\rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\rho}\rangle + \langle \boldsymbol{\lambda}, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle) - \mathcal{E}_\tau^{\mathrm{D}}.$$

By substituting in Equation (4),

$$\sum_{t=1}^{\tau} \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] \geq -\mathcal{E}_\tau^{\mathrm{P}} - \mathcal{E}_\tau^{\mathrm{D}} + \sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} (\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle). \quad (5)$$

Then, we bound the term

$$\text{Ⓐ} := \sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} (\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle).$$

Let $\mathtt{OPT}_{\boldsymbol{\gamma},\tau}^* := \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{t=1}^{\tau} f_t(\boldsymbol{x})$, that is, $\mathtt{OPT}_{\boldsymbol{\gamma},\tau}^*$ is the supremum of the unconstrained problem. Then, for each $\epsilon > 0$, there exists an $\boldsymbol{x}^* \in \mathcal{X}$ such that $\sum_t f_t(\boldsymbol{x}^*) \geq \mathtt{OPT}_{\boldsymbol{\gamma},\tau}^* - \epsilon$. Then, we show that

$$\begin{aligned} \text{Ⓐ} &\geq \max_{\boldsymbol{x}\in\{\boldsymbol{x}^*,\varnothing\}} \sum_{t=1}^{\tau} (\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[f_t(\boldsymbol{x})] + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}}[c_t(\boldsymbol{x})]\rangle) \\ &\geq \rho\, \mathtt{OPT}_{\boldsymbol{\gamma},\tau}^* - \epsilon. \end{aligned} \quad (6)$$

To do so, we consider two cases. First, if $\sum_{t=1}^{\tau} f_t(\boldsymbol{x}^*) \geq \sum_{t=1}^{\tau} \langle \boldsymbol{\lambda}_t, c_t(\boldsymbol{x}^*)\rangle$, then the value of the function for $\boldsymbol{x}^*$ is

at least

$$\begin{aligned}
\text{Ⓐ} &\geq \sum_{t=1}^{\tau}(f_t(\boldsymbol{x}^*) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - c_t(\boldsymbol{x}^*)\rangle) \\
&\geq \sum_{t=1}^{\tau}(f_t(\boldsymbol{x}^*) + \langle \boldsymbol{\lambda}_t, \rho \cdot c_t(\boldsymbol{x}^*) - c_t(\boldsymbol{x}^*)\rangle) \\
&\geq \sum_{t=1}^{\tau} f_t(\boldsymbol{x}^*) - (1-\rho)\sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}_t, c_t(\boldsymbol{x}^*)\rangle \\
&\geq \rho \sum_{t=1}^{\tau} f_t(\boldsymbol{x}^*) = \rho\, \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} - \epsilon,
\end{aligned}$$

where the second inequality holds since $c_t(\cdot) \in [0,1]^m$, for each $t \in [T]$. Otherwise, if $\sum_t f_t(\boldsymbol{x}^*) < \sum_t \langle \boldsymbol{\lambda}_t, c_t(\boldsymbol{x}^*)\rangle$, we have that the null action $\varnothing$ has value at least

$$\begin{aligned}
\text{Ⓐ} &\geq \sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}_t, \boldsymbol{\rho}\rangle \geq \sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}_t, \rho \cdot c_t(\boldsymbol{x}^*)\rangle \\
&\geq \rho \sum_{t=1}^{\tau} f_t(\boldsymbol{x}^*) \geq \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} - \epsilon.
\end{aligned}$$

This shows that Equation (6) holds. Hence, $\text{Ⓐ} \geq \rho\, \texttt{OPT}^*_{\boldsymbol{\gamma},\tau}$. Then, by substituting this in Equation (5), we have

$$\begin{aligned}
&\sum_{t=1}^{\tau}(\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle) \\
&\qquad\qquad \geq -\mathcal{E}^{\texttt{P}}_{\tau} - \mathcal{E}^{\texttt{D}}_{\tau} + \rho \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} - \tau\,\langle \boldsymbol{\lambda}, \boldsymbol{\rho}\rangle. \quad (7)
\end{aligned}$$

Step 2: relating expectations and their realizations. Now, we have to relate the lefthand side of the above inequality to its realized value $\sum_{t=1}^{\tau}(f_t(\boldsymbol{x}_t) - \langle \boldsymbol{\lambda}, c_t(\boldsymbol{x}_t)\rangle)$. In order to do this, let

$$W_t \coloneqq -f_t(\boldsymbol{x}_t) + \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] + \langle \boldsymbol{\lambda}, c_t(\boldsymbol{x}_t) - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle,$$

and observe that $W_1, \dots, W_\tau$ is a martingale difference sequence, with $|W_t| \leq 1 + \|\boldsymbol{\lambda}\|_\infty$. Then, by the Azuma-Hoeffding inequality, we have that, for any $\tau \in [T]$,

$$\Pr\left[\sum_{t=1}^{\tau} W_t > (1 + \|\boldsymbol{\lambda}\|_\infty)\sqrt{2\tau \ln(1/\delta)}\right] \leq \delta.$$

Let, $q(\tau, \boldsymbol{\lambda}, \delta) \coloneqq (1 + \|\boldsymbol{\lambda}\|_\infty)\sqrt{2\tau \ln(1/\delta)}$. Then, by taking a union bound, we have

$$\Pr\left[\forall \tau \in [T],\ \sum_{t=1}^{\tau} W_t \leq q(\tau, \boldsymbol{\lambda}, \delta)\right] \geq 1 - T\,\delta.$$

Therefore, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned}
\sum_{t=1}^{\tau}(f_t(\boldsymbol{x}_t) - \langle \boldsymbol{\lambda}, c_t(\boldsymbol{x}_t)\rangle) &\geq -q(\tau, \boldsymbol{\lambda}, \delta/T) \\
&\quad + \sum_{t=1}^{\tau}(\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[f_t(\boldsymbol{x})] - \langle \boldsymbol{\lambda}, \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}[c_t(\boldsymbol{x})]\rangle) \\
&\geq -\underbrace{\left(q(\tau, \boldsymbol{\lambda}, \delta/T) + \mathcal{E}^{\texttt{P}}_{\tau} + \mathcal{E}^{\texttt{D}}_{\tau}\right)}_{\text{Ⓑ}} + \rho\, \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} - \tau\,\langle \boldsymbol{\lambda}, \boldsymbol{\rho}\rangle,
\end{aligned}$$

where the last inequality is by Equation (7).

Step 3: putting everything together. First, we rewrite $\texttt{OPT}^*_{\boldsymbol{\gamma},\tau}$ as a function of the baseline $\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}}$. In particular, we have

$$\rho\, \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} \geq \rho\, \texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma},\tau} \geq \rho\left(\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}} - T + \tau\right).$$

By definition $\texttt{REW}_{\boldsymbol{\gamma}} \coloneqq \sum_{t=1}^{\tau}(f_t(\boldsymbol{x}_t))$. Then,

$$\begin{aligned}
\texttt{REW}_{\boldsymbol{\gamma}} &\geq \rho\, \texttt{OPT}^*_{\boldsymbol{\gamma},\tau} - \sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}, \boldsymbol{\rho} - c_t(\boldsymbol{x}_t)\rangle - \text{Ⓑ} \\
&\geq \frac{\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}} - T + \tau}{\alpha} - \sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}, \boldsymbol{\rho} - c_t(\boldsymbol{x}_t)\rangle - \text{Ⓑ}. \quad (8)
\end{aligned}$$

If $\tau = T$ (*i.e.*, the stopping time coincides with the time horizon T), in order to get the result it is enough to set $\boldsymbol{\lambda} = \mathbf{0}$, and to substitute the above expression in the definition of regret. Otherwise, if $\tau < T$, it means that there exists a resource $i^{\texttt{S}} \in [m]$ for which

$$\sum_{t=1}^{\tau} c_t(\boldsymbol{x}_t)[i^{\texttt{S}}] + 1 \geq \rho T, \quad (9)$$

where, in our setting, 1 is the maximum observable cost. Then, we set $\boldsymbol{\lambda}$ as follows: $\boldsymbol{\lambda}[i^{\texttt{S}}] = 1/\rho$, and $\boldsymbol{\lambda}[i] = 0$ for all $i \neq i^{\texttt{S}}$. For this choice of $\boldsymbol{\lambda}$, and by exploiting Equation (9), we have

$$\begin{aligned}
\sum_{t=1}^{\tau}\langle \boldsymbol{\lambda}, \boldsymbol{\rho} - c_t(\boldsymbol{x}_t)\rangle &= \alpha \sum_{t=1}^{\tau}(\rho - c_t(\boldsymbol{x}_t)[i^{\texttt{S}}]) \\
&\leq \tau - T + \alpha.
\end{aligned}$$

Then, by substituting the above expression in Equation (8),

$$\texttt{REW}_{\boldsymbol{\gamma}} \geq \frac{\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}} - T + \tau}{\alpha} - (\tau - T) - \alpha - \text{Ⓑ}.$$

Finally, we have

$$\begin{aligned}
\texttt{OPT}^{\texttt{FP}}_{\boldsymbol{\gamma}} - \alpha\, \texttt{REW}_{\boldsymbol{\gamma}} &\leq (T - \tau) - \alpha(T - \tau) + \alpha^2 + \alpha \cdot \text{Ⓑ} \\
&\leq \alpha^2 + \alpha\left(q(\tau, \boldsymbol{\lambda}, \delta/T) + \mathcal{E}^{\texttt{P}}_{\tau} + \mathcal{E}^{\texttt{D}}_{\tau}\right) \\
&\leq \alpha^2 + \alpha(1+\alpha)\sqrt{2T \ln(T/\delta)} + \mathcal{E}^{\texttt{P}}_{T} + \mathcal{E}^{\texttt{D}}_{T},
\end{aligned}$$

where the last inequality holds because the error terms are increasing in t. This concludes the proof. □

Remark 6.2. Let $\hat{\ell}_t^{\mathrm{P}} : \mathcal{X} \ni \boldsymbol{x} \mapsto f_t(\boldsymbol{x}) - \langle \boldsymbol{\lambda}_t, c_t(\boldsymbol{x}) \rangle$. Then, by definition of the set of strategy mixtures Ξ (see Section 4), for each $\tau \in [T]$ it holds

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{\tau} \hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}) = \sup_{\boldsymbol{\xi} \in \Xi} \sum_{t=1}^{\tau} \ell_t^{\mathrm{P}}(\boldsymbol{\xi}).$$

Remark 6.3. The guarantees of Theorem 6.1 can be extended, with minor modifications, to the bandit feedback setting. In particular, let $\hat{\ell}_t^{\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ be defined as in Remark 6.2, and $\mathcal{R}^{\mathrm{P}}$ be a primal regret minimizer guaranteeing, with probability at least $1-\delta$, that $\sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{\tau} \left(\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}) - \hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}_t) \right) \leq \mathcal{E}_{\tau,\delta}^{\mathrm{P}}$. By Remark 6.2, the above high-probability regret bound implies that, with probability at least $1-2\delta$, $R_\tau^{\mathrm{P}} \leq \mathcal{E}_{\tau,\delta}^{\mathrm{P}} + O(\sqrt{T \ln(T/\delta)}/\rho)$ (see Appendix A.2). Then, it is possible to follow the proof of Theorem 6.1, and to recover, via the application of an additional union bound, the same guarantees with probability at least $1-3\delta$.

7. Regret Bound for the Stochastic Setting

In this section, we prove an optimal regret upper bound matching that of Badanidiyuru et al. (2013); Immorlica et al. (2019). We employ the notion of *expected Lagrangian game* employed by Immorlica et al. (2019). However, by working in the space of strategy mixtures Ξ, we provide a simplified analysis. In particular, working on the mixtures Ξ, we can provide deterministic bounds on the regret, while Immorlica et al. (2019) works with regret minimizers in high probability. Moreover, our approach allows to generalize the result of Immorlica et al. (2019) to general, *e.g.*, non-convex, problems.

Theorem 7.1. *Consider Meta-Algorithm 1 equipped with two arbitrary regret minimizers $\mathcal{R}^P$ and $\mathcal{R}^D$ for the sets Ξ and $\mathcal{D}$, respectively. In particular, assume that they guarantee a cumulative regret up to time T which is upper bounded by $\mathcal{E}_T^P$ and $\mathcal{E}_T^D$, respectively. For each $t \in [T]$, let the inputs(f_t, c_t) be i.i.d. samples from a fixed but unknown distribution $\mathcal{P}$ over the set of possible requests $\mathcal{I}$. For $\delta > 0$, with probability at least $1-\delta$ we have*

$$\mathtt{OPT}^{DP} - \mathtt{REW}_{\boldsymbol{\gamma}} \leq O\left(\frac{1}{\rho} \sqrt{2T \log(mT/\delta)} \right) + \mathcal{E}_T^P + \mathcal{E}_T^D,$$

where $\mathtt{REW}_{\boldsymbol{\gamma}} \coloneqq \sum_t f_t(\boldsymbol{x}_t)$ is the reward of the algorithm for the sequence of inputs $\boldsymbol{\gamma}$.

The proof is based on two steps. First, by applying the Azuma-Hoeffding inequality, we show that, in the first $\tau \in [T]$ rounds, the average reward and cost for each resource i up to τ is *close*, with high probability, to $\mathbb{E}_{\boldsymbol{x} \sim \bar{\boldsymbol{\xi}}}[\bar{f}(\boldsymbol{x})]$ and $\mathbb{E}_{\boldsymbol{x} \sim \bar{\boldsymbol{\xi}}}[\bar{c}(\boldsymbol{x})[i]]$, where $\bar{\boldsymbol{\xi}}$ is the average of the strategy mixtures selected by the primal regret minimizer. Then, we define a two-player, zero-sum, *expected Lagrangian game* such that, by Lemma 4.3, the value at the Nash equilibrium of game is equal to $\mathtt{OPT}_{\bar{f},\bar{c}}^{\mathrm{LP}}$. Finally, we show that $\bar{\boldsymbol{\xi}}$ is an approximation of the equilibrium strategy of one of the two players.

Analogously to the adversarial case (Remark 6.3), Theorem 7.1 extends to the bandit feedback setting with minor modifications, whenever a primal regret minimizer for set $\mathcal{X}$ with high-probability regret guarantees with respect to $\hat{\ell}_t^{\mathrm{P}}$ is available.

8. Applications

In this section, we first provide an explicit instantiation of our algorithm in the classical multi-armed bandit with knapsack setting of Badanidiyuru et al. (2013). Then, we describe two applications of our framework to well known game-theoretic problems. This provide further evidence of the flexibility of our framework.

8.1. Multi-Armed Bandits with Knapsacks

Consider a multi-armed bandit problem with K arms (*i.e.*, $\mathcal{X} = [K]$), and per-arm utility and cost defined as in Section 3.

Let the primal regret minimizer for bandit feedback $\mathcal{R}^{\mathrm{P}}$ be EXP3.P (Auer et al., 2002). By its definition (see Algorithm 1) and by Equation (2), the loss ℓ_t^{P} is such that $\ell_t^{\mathrm{P}} : \Xi \to [-1/\rho, 1]$. Then, at time T, EXP3.P guarantees that, with probability at least $1-\delta$, $\mathcal{E}_{T,\delta}^{\mathrm{P}} \leq O(\sqrt{KT \log(T/\delta)}/\rho)$.

Corollary 8.1. *Consider Meta-Algorithm 1, and let the primal regret minimizer for bandit feedback $\mathcal{R}^P$ be EXP3.P, and let the dual regret minimizer with full feedback $\mathcal{R}^D$ be online mirror descent with negative entropy reference function (see Remark 5.1). We have the following two cases:*

- *If requests are chosen by an oblivious adversary. Letting $\alpha \coloneqq 1/\rho$, for each $\delta > 0$ we have*

$$\mathtt{OPT}_{\boldsymbol{\gamma}}^{FP} - \alpha\, \mathtt{REW}_{\boldsymbol{\gamma}} \leq O(\alpha^2 \sqrt{KT \log(Tm/\delta)});$$

- *if, for each $t \in [T]$, the inputs (f_t, c_t) are i.i.d. samples from a fixed but unknown distribution $\mathcal{P}$ over the set of possible requests $\mathcal{I}$. For $\delta > 0$, we have*

$$\mathtt{OPT}^{DP} - \mathtt{REW}_{\boldsymbol{\gamma}} \leq O\left(\sqrt{2KT \log(Tm/\delta)}/\rho \right);$$

with probability at least $1-\delta$, where $\mathtt{REW}_{\boldsymbol{\gamma}} \coloneqq \sum_t f_t(\boldsymbol{x}_t)$ is the reward of the algorithm for the sequence of inputs $\boldsymbol{\gamma}$.

8.2. Repeated Stackelberg Games with Knapsacks

In Stackelberg games a *leader* commits to a (possibly mixed) strategy, and then a *follower* best responds to that strategy

(von Stackelberg, 1934). Stackelberg games have recently received significant attention for their applications in security domains (Tambe, 2011). In such settings, Online *Stackelberg Security Games* (SSG) have been introduced to circumvent the assumption that the leader must know the attacker's utility function (Balcan et al., 2015). We show that our framework could be extended to model repeated SSGs in which there are hard budget constraints with respect to deployed defensive resources.

We take the perspective of the leader that, at each time $t \in [T]$, plays a game against a follower of an unknown type. The leader has a finite set of available actions $\mathcal{A}_{\text{L}}$ with $n_{\text{L}} := |\mathcal{A}_{\text{L}}|$, and strategies $\mathcal{X} := \Delta(\mathcal{A}_{\text{L}})$, while the follower has a set of available actions $\mathcal{A}_{\text{F}}$ with $n_{\text{F}} := |\mathcal{A}_{\text{F}}|$, and strategies $\mathcal{Y} := \Delta(\mathcal{A}_{\text{F}})$. The utility function of the follower at time t is denoted by $u_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We assume that, for each $t \in [T]$, $\boldsymbol{x} \in \mathcal{X}$, and $\boldsymbol{y} \in \mathcal{Y}$, $u_t(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top U_t \boldsymbol{y}$, for $U_t \in \mathbb{R}^{n_{\text{L}} \times n_{\text{F}}}$. Moreover, we assume that the follower has a finite set of possible types $\mathcal{K}$ and hence, for each t, $U_t \in \{U_k\}_{k \in \mathcal{K}}$. At each t, the leader commits to a strategy $\boldsymbol{x}_t \in \mathcal{X}$. Then, the follower will play their *best-response* given $\boldsymbol{x}_t$. Formally, for type $k \in \mathcal{K}$, the follower plays the strategy $\boldsymbol{y}^k_{\boldsymbol{x}_t} := \boldsymbol{e}_a$, where $a \in \arg\max_{a' \in \mathcal{A}_{\text{F}}} \boldsymbol{x}_t^\top U_k \boldsymbol{e}_{a'}$, and $\boldsymbol{e}_a$ denotes the vector where component a is equal to 1 and the others are equal to 0. As it is customary in the literature, we assume that the follower breaks ties in favour of the leader. Then, the leader's utility function $f_t : \mathcal{X} \to [0, 1]$ is such that, for each $\boldsymbol{x} \in \mathcal{X}$, and t, $f_t(\boldsymbol{x}) := \boldsymbol{x}^\top U_{\text{L}} \boldsymbol{y}^{k_t}_{\boldsymbol{x}}$, where k_t is the follower type at round t. This function is upper semicontinuous, and it is therefore Borel measurable.

At each t, the leader pays a cost based on the strategy they commit to. In particular, for each t there's a cost matrix $C_t \in [0, 1]^{n_{\text{L}} \times m}$, that specifies a vector of m costs for leader's actions. The cost incurred by the leader at time t is then $\boldsymbol{x}_t^\top C_t$, where $\boldsymbol{x}_t$ is the strategy played by the leader at time t. The leader has an overall budget $B \in \mathbb{R}_{\geq 0}$ for each resource. Let ρ be the per-iteration budget (defined as in Section 3), Moreover, we assume the leader has a void action which yields no reward and no consumption of any other resource other than time.

In order to apply Algorithm 1, we show that there exists a regret minimizer for the leader. We show that this is possible despite the fact that the leader's utility is non-convex, and not even continuous. By Remark 6.2, we can safely restrict our attention to regret minimizers that provide no-regret with respect to the optimal fixed strategy in $\boldsymbol{x}^* \in \mathcal{X}$. As a first step, we show that for each sequence of follower's types $(k_t)_{t=1}^\tau$, there always exists an optimal mixed strategy belonging to a *finite* set of strategies $\mathcal{X}^* \subset \mathcal{X}$. Moreover, we show that $\mathcal{X}^*$ is independent from the sequence of types. In order to define the restricted set $\mathcal{X}^*$, for each type $k \in \mathcal{K}$ and action $a \in \mathcal{A}_{\text{F}}$, let $\mathcal{X}^{k,a} \subseteq \mathcal{X}$ be the set of leader's strategy in which a is a best response for the follower of type k, *i.e.*, $\mathcal{X}^{k,a} := \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{y}^k_{\boldsymbol{x}} = \boldsymbol{e}_a\}$. Let $\boldsymbol{a} \in \mathcal{A}_{\text{L}}^{|\mathcal{K}|}$ be a tuple with an action per follower's type, and $\mathcal{X}^{\boldsymbol{a}}$ be the polytope such that each action $\boldsymbol{a}[k]$ is optimal for the corresponding type k, *i.e.*, $\mathcal{X}^{\boldsymbol{a}} := \cap_{k \in \mathcal{K}} \mathcal{X}^{k, \boldsymbol{a}[k]}$. Finally, we let $\mathcal{X}^* := \cup_{\boldsymbol{a} \in \mathcal{A}_{\text{L}}^{|\mathcal{K}|}} V(\mathcal{X}^{\boldsymbol{a}})$, where $V(\mathcal{X}^{\boldsymbol{a}})$ denotes the set of the vertexes of the polytope $\mathcal{X}^{\boldsymbol{a}}$.

Lemma 8.2. *Let $\ell_{\text{L},t}(\boldsymbol{x}, \boldsymbol{\lambda}) := f_t(\boldsymbol{x}) - \langle \boldsymbol{\lambda}, \boldsymbol{x}^\top C_t \rangle$ for all pairs $(\boldsymbol{x}, \boldsymbol{\lambda})$. Then, for each $\tau \in [T]$, each sequence of receiver's types $(k_t)_{t=1}^\tau$, and each sequence $(\boldsymbol{\lambda}_t)_{t=1}^\tau$, it holds:*

$$\max_{\boldsymbol{x}^* \in \mathcal{X}^*} \sum_{t=1}^{\tau} \ell_{\text{L},t}(\boldsymbol{x}^*, \boldsymbol{\lambda}_t) = \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{\tau} \ell_{\text{L},t}(\boldsymbol{x}, \boldsymbol{\lambda}_t).$$

Then, we bound the cardinality of the restricted set $\mathcal{X}^*$.

Lemma 8.3. *It holds $|\mathcal{X}^*| \leq (|\mathcal{K}| n_{\text{F}}^2)^{n_{\text{L}}-1}$*

Lemma 8.2 implies that to build a regret minimizer for $\mathcal{X}$, it is sufficient to have small regret with respect to the optimal action in $\mathcal{X}^*$. Thus, we can focus on a regret minimizer for $\mathcal{X}^*$. Since $\mathcal{X}^*$ has finite support, the set of randomized strategy Ξ^* is the simplex over $\mathcal{X}^*$, *i.e.*, $\Xi^* = \Delta^{\mathcal{X}^*}$. As a primal regret minimizer, we can employ OMD with a negative entropy regularizer that provides regret upper bound $O(\sqrt{T \log(|\mathcal{X}^*|)})$. Therefore, we proved the existence of a regret minimizer for the primal decision space.

Theorem 8.4. *There exists a primal regret minimizer $\mathcal{R}^P$ for the Strackelberg problem with regret $\mathcal{E}_T^P = O(\sqrt{T n_{\text{L}} \log(|\mathcal{K}| n_{\text{F}})})$.*

Equipped with the above result, we can directly apply Theorems 6.1 and 7.1 to our setting.

8.3. Budget-Pacing in Repeated First-Price Auctions

Internet advertising platforms typically offer advertisers the possibility to pace the rate at which their budget is depleted, through *budget-pacing mechanisms* (Agarwal et al., 2014b; Conitzer et al., 2021; Balseiro et al., 2021). These mechanisms are essential to ensure that the advertisers' budget is not depleted too early (thereby missing potentially valuable future advertising opportunities), while being fully depleted within the planned duration of the campaign. We focus on budget pacing in the context of first-price auctions, which is particularly relevant for selling display ads (*e.g.*, in 2019 Google announced a shift to first-price auctions for its Ad-Manager exchange).[1] Dual mirror descent schemes which are usually employed in the context of repeated second price auctions (Balseiro & Gur, 2019; Balseiro et al., 2020a; Celli et al., 2022) cannot be applied to our setting, because they rely on particular features of second price auctions.[2]

[1] See https://tinyurl.com/chv5nxys.

[2] In particular, in the primal update step, $\boldsymbol{x}_t$ is chosen by maximizing a function of f_t and c_t. Therefore, in general, this assumes

We consider the problem faced by a bidder that takes part to a sequence of first-price auctions. At each round $t \in [T]$, the bidder observes their valuation v_t extracted by a finite set of possible valuations $\mathcal{V} \subset [0,1]$, with $n_{\mathrm{v}} := |\mathcal{V}|$.[3] Then, the bidder chooses $b_t \in \mathcal{B}$, where $\mathcal{B} \subset [0,1]$ is a finite set of n_{b} possible bids. Then, the utility function of the bidder depends on the maximum among the competing bids, which we denote by $\hat{b}_t$. In particular, if $b_t \geq \hat{b}_t$, the bidder wins the auction, pays to the auctioneer b_t, and has utility $f_t(b_t) = v_t - b_t$. Moreover, the bidder incurs in a cost $c_t(b_t) = b_t$. Otherwise, the bidder does not win the item, $f_t(b_t) = 0$, and $c_t(b_t) = 0$. Finally, the bidder has a budget $B \in \mathbb{R}_+$, which limits the total amount that the agent can spend throughout the T rounds. As a benchmark to evaluate the performance of the algorithm, we consider the best static policy $\pi : \mathcal{V} \to \mathcal{B}$. Next, we show that the problem can be easily addressed via our framework. The set of static policies can be represented by $\mathcal{X} := \mathcal{B}^{n_{\mathrm{v}}}$, where a vector $\boldsymbol{b} \in B^{n_{\mathrm{v}}}$ is such that $\boldsymbol{b}[v]$ is the bid played by the policy with valuation v. Then, the utility function is such that $f_t(\boldsymbol{b}) = (v_t - \boldsymbol{b}[v_t]) I_{\{\boldsymbol{b}[v_t] \geq \hat{b}_t\}}$, where I denotes the indicator function, and the cost is $c_t(\boldsymbol{b}) = \boldsymbol{b}[v_t] I_{\{\boldsymbol{b}[v_t] \geq \hat{b}_t\}}$. The set of strategy mixtures is given by the set $\Xi := \Delta^{\mathcal{X}}$.

Let $\ell_t^{\mathrm{B}} : \Xi \ni \boldsymbol{\xi} \mapsto \mathbb{E}_{\boldsymbol{b} \sim \boldsymbol{\xi}}[f_t(\boldsymbol{b})] - \lambda_t \mathbb{E}_{\boldsymbol{b} \sim \boldsymbol{\xi}}[c_t(\boldsymbol{b})]$ be the primal loss function. We want to show the existence of a regret minimizer for the set Ξ. To do that, by Remark 6.2, we know that it is enough to design a regret minimizer for $\mathcal{X}$. Then, by letting $\Xi^* := (\Delta^{\mathcal{B}})^{n_{\mathrm{v}}}$, and since $\max_{\boldsymbol{b} \in \mathcal{X}} \sum_{t=1}^{\tau} f_t(\boldsymbol{b}) - \lambda_t c_t(\boldsymbol{b}) = \max_{\boldsymbol{\xi} \in \Xi^*} \sum_{t=1}^{\tau} \ell_t^{\mathrm{B}}(\boldsymbol{\xi})$, it is enough to design a regret minimizer for the set Ξ^*. Since the primal loss function ℓ_t^{B} is linear in $\boldsymbol{\xi}$, we can apply OMD with negative entropy regularizer to get a regret upper bound of $O(\sqrt{T n_{\mathrm{v}} \log(n_{\mathrm{b}})})$ (see, *e.g.*, Farina et al. (2021)).

Theorem 8.5. *There exists a primal regret minimizer $\mathcal{R}^P$ for the problem of bidding in first-price auctions with regret upper bound $O(\sqrt{T n_{\mathrm{v}} \log(n_{\mathrm{b}})})$.*

This immediately implies that Theorems 6.1 and 7.1 hold in the full information setting (*i.e.*, when the bidder observes $\hat{b}_t$ for each t). In the bandit setting, one can obtain analogous results by instantiating an appropriate regret minimizer (*e.g.*, EXP3 by Auer et al. (2002)) for each $v \in \mathcal{V}$. We remark that our model is clearly a simplification of real budget-pacing systems. We leave the problem of studying the general setting (with arbitrary sets $\mathcal{V}$ and $\mathcal{B}$) within our framework as an interesting future research direction.

that the decision maker observers (f_t, c_t) before taking the decision at time t. The fact that costs are determined through a second price auction allows the decision maker to implicitly take the max without actually observing the costs, by bidding their *adjusted* valuation. However, this is not possible when allocations are determined through first-price auctions.

[3] In ad auctions this models the fact that the auctioneer shares with the advertisers some targeting information about users.

Acknowledgements

Christian Kroer is supported by the Office of Naval Research Young Investigator Program under grant N00014-22-1-2530.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014a.

Agarwal, D., Ghosh, S., Wei, K., and You, S. Budget pacing for targeted online advertisements at linkedin. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1613–1619, 2014b.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006. ACM, 2014.

Agrawal, S. and Devanur, N. R. Bandits with global convex constraints and objective. *Operations Research*, 67(5): 1486–1502, 2019.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *29th Annual Conference on Learning Theory (COLT)*, 2016.

Aliprantis, C. D., Border, K. C., et al. Infinite dimensional analysis. *Springer Books*, 2006.

Auer, P. and Chiang, C. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Conf. on Learning Theory (COLT)*, 2016.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Babaioff, M., Dughmi, S., Kleinberg, R., and Slivkins, A. Dynamic pricing with limited supply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 74–91, 2012.

Badanidiyuru, A., Kleinberg, R., and Singer, Y. Learning on a budget: posted price mechanisms for online procurement. In *Proceedings of the 13th ACM conference on electronic commerce*, pp. 128–145, 2012.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science, FOCS 2013*, pp. 207–216. IEEE, 2013.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. *J. ACM*, 65(3), 2018.

Balcan, M.-F., Blum, A., Haghtalab, N., and Procaccia, A. D. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pp. 61–78, 2015.

Balseiro, S., Lu, H., and Mirrokni, V. The best of many worlds: Dual mirror descent for online allocation problems. *arXiv preprint arXiv:2011.10124*, 2020a.

Balseiro, S., Lu, H., and Mirrokni, V. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pp. 613–628. PMLR, 2020b.

Balseiro, S., Kim, A., Mahdian, M., and Mirrokni, V. Budget-management strategies in repeated auctions. *Operations Research*, 2021.

Balseiro, S. R. and Gur, Y. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Besbes, O. and Zeevi, A. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Besbes, O. and Zeevi, A. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Buchbinder, N. and Naor, J. S. The design of competitive online algorithms via a primal–dual approach. *Foundations and Trends® in Theoretical Computer Science*, 3 (2–3):93–263, 2009.

Cardoso, A. R., Abernethy, J., Wang, H., and Xu, H. Competing against Nash equilibria in adversarially changing zero-sum games. In *International Conference on Machine Learning*, pp. 921–930. PMLR, 2019.

Celli, A., Colini-Baldeschi, R., Kroer, C., and Sodomka, E. The parity ray regularizer for pacing in auction markets. In *Proceedings of the ACM Web Conference 2022*, pp. 162–172, 2022.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, T. and Giannakis, G. B. Bandit convex optimization for scalable and dynamic iot management. *IEEE Internet of Things Journal*, 2018.

Chen, T., Ling, Q., and Giannakis, G. B. An online convex optimization approach to proactive network resource allocation. *IEEE Transactions on Signal Processing*, 65(24): 6350–6364, 2017.

Combes, R., Jiang, C., and Srikant, R. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1): 245–257, 2015.

Conitzer, V., Kroer, C., Sodomka, E., and Stier-Moses, N. E. Multiplicative pacing equilibria in auction markets. *Operations Research*, 2021.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.

Farina, G., Kroer, C., and Sandholm, T. Better regularization for sequential decision spaces: Fast convergence rates for Nash, correlated, and team equilibria. In *ACM Conference on Economics and Computation*, 2021.

György, A., Kocsis, L., Szabó, I., and Szepesvári, C. Continuous time associative bandit problems. In *20th Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 830–835, 2007.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019*, pp. 202–219. IEEE Computer Society, 2019.

Kesselheim, T. and Singla, S. Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pp. 2286–2305. PMLR, 2020.

Luenberger, D. G. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Mahdavi, M., Jin, R., and Yang, T. Trading regret for efficiency: online convex optimization with long term constraints. *J. of Machine Learning Research (JMLR)*, 13 (Sep):2503–2528, 2012.

Mahdavi, M., Yang, T., and Jin, R. Stochastic convex optimization with multiple objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1115–1123, 2013.

Mehta, A. et al. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science*, 8 (4):265–368, 2013.

Neely, M. J. and Yu, H. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Rangi, A., Franceschetti, M., and Tran-Thanh, L. Unifying the stochastic and the adversarial bandits with knapsack. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3311–3317, 7 2019.

Sankararaman, K. A. and Slivkins, A. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1760–1770. PMLR, 2018.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pp. 1287–1295. PMLR, 2014.

Singla, A. and Krause, A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1167–1178, 2013.

Slivkins, A. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*, 2013.

Slivkins, A. Introduction to multi-armed bandits. *CoRR*, abs/1904.07272, 2019. URL http://arxiv.org/abs/1904.07272.

Tambe, M. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge university press, 2011.

Tran-Thanh, L., Chapman, A., De Cote, E. M., Rogers, A., and Jennings, N. R. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

von Stackelberg, H. *Marktform und Gleichgewicht*. Springer, Vienna, 1934.

Wang, Z., Deng, S., and Ye, Y. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

A. Omitted Proofs

A.1. Proofs for Section 4

Theorem 4.2. *Let $f : \mathcal{X} \to [0,1]$, $c : \mathcal{X} \to [0,1]^m$, and $(f,c) \in \mathcal{I}$. It holds:*

$$\sup_{\boldsymbol{\xi} \in \Xi} \inf_{\boldsymbol{\lambda} \geq 0} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \inf_{\boldsymbol{\lambda} \geq 0} \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \texttt{OPT}^{\texttt{LP}}_{f,c}.$$

Proof. **Auxiliary sets.** Let

$$\mathcal{V} \coloneqq \left\{(\boldsymbol{v}, t) \in \mathbb{R}^{m+1} : \exists \boldsymbol{\xi} \in \Xi \text{ s.t. } \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})[i]] - \rho \leq \boldsymbol{v}[i], \forall i \in [m],\, \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] \geq t\right\},$$

$$\mathcal{W} \coloneqq \left\{(\mathbf{0}, w) \in \mathbb{R}^{m+1} : w > \texttt{OPT}^{\texttt{LP}}_{f,c}\right\},$$

where each element of $\mathcal{W}$ is composed of an m-dimensional vector of zeros, and a scalar w. Notice that the dimension of the two sets does not depend on the dimensionality of Ξ. In particular, $\mathcal{V}, \mathcal{W}$ have finite dimension even when we have an infinite-dimensional space of strategy mixtures. We claim that $\mathcal{V}$ and $\mathcal{W}$ are convex, and $\mathcal{V} \cap \mathcal{W} = \varnothing$. Take any two points $(\boldsymbol{v}_1, t_1) \in \mathcal{V}$, $(\boldsymbol{v}_2, t_2) \in \mathcal{V}$, and $\alpha \in [0,1]$. Then, let $\boldsymbol{\xi}_\alpha = \alpha\, \boldsymbol{\xi}_1 + (1-\alpha)\, \boldsymbol{\xi}_2$, where $\boldsymbol{\xi}_1$ (resp., $\boldsymbol{\xi}_2$) is a point in Ξ for which the constraints of $\mathcal{V}$ for $(\boldsymbol{v}_1, t_1)$ (resp., $(\boldsymbol{v}_2, t_2)$) are satisfied. We have that $\boldsymbol{\xi}_\alpha \in \Xi$. Moreover, by linearity of expectation, we have, for each resource $i \in [m]$, $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}_\alpha}[c(\boldsymbol{x})[i]] - \rho \leq \alpha \boldsymbol{v}_1[i] + (1-\alpha)\boldsymbol{v}_2[i]$, and $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}_\alpha}[f(\boldsymbol{x})] \geq \alpha t_1 + (1-\alpha) t_2$. Then, $\alpha(\boldsymbol{v}_1, t_1) + (1-\alpha)\,(\boldsymbol{v}_2, t_2) \in \mathcal{V}$ (i.e., $\mathcal{V}$ is convex). It is immediate to check that $\mathcal{W}$ is convex. Finally, assume that there exists a point $(\boldsymbol{v}', t') \in \mathcal{V} \cap \mathcal{W}$. By definition of $\mathcal{V}$, there exists $\boldsymbol{\xi}' \in \Xi$ such that $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}'}[c(\boldsymbol{x})] - \boldsymbol{\rho} \leq \boldsymbol{v}'$, and $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}'}[f(\boldsymbol{x})] \geq t'$. Then, by definition of $\mathcal{W}$, $\boldsymbol{v}' = \mathbf{0}$, that is, $\boldsymbol{\xi}'$ is budget-feasible. However, the fact that $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}'}[f(\boldsymbol{x})] > \texttt{OPT}^{\texttt{LP}}_{f,c}$ is in contradiction with $\texttt{OPT}^{\texttt{LP}}_{f,c}$ being the optimal value of a feasible solution to LP (1).

Separating $\mathcal{V}$ and $\mathcal{W}$. By assumption we have that the primal objective $\texttt{OPT}^{\texttt{LP}}_{f,c}$ is finite (otherwise, if $\texttt{OPT}^{\texttt{LP}}_{f,c} = +\infty$, we could immediately recover our result by weak duality). The sets $\mathcal{V}$ and $\mathcal{W}$ are convex and do not intersect. Therefore, by the separating hyperplane theorem, there exists a point $(\tilde{\boldsymbol{\lambda}}, \tilde{\mu}) \in \mathbb{R}^{m+1} \setminus \{\mathbf{0}\}$, and a scalar value $\beta \in \mathbb{R}$ such that

$$(\boldsymbol{v}, t) \in \mathcal{V} \implies \langle \tilde{\boldsymbol{\lambda}}, v \rangle + \tilde{\mu}\, t \leq \beta, \tag{10}$$

$$(\mathbf{0}, w) \in \mathcal{W} \implies \tilde{\mu}\, w \geq \beta. \tag{11}$$

First, it must be that $\tilde{\boldsymbol{\lambda}}[i] \leq 0$ for all $i \in [m]$, and $\tilde{\mu} \geq 0$, otherwise we would get unboundedness of the lefthand side of Equation (10). Moreover, since Equation (11) must hold for each $w > \texttt{OPT}^{\texttt{LP}}_{f,c}$, by continuity we have $\tilde{\mu}\, \texttt{OPT}^{\texttt{LP}}_{f,c} \geq \beta$. For each $\boldsymbol{\xi} \in \Xi$ there exists a pair $(\boldsymbol{v}_{\boldsymbol{\xi}}, t_{\boldsymbol{\xi}}) \in \mathcal{V}$ such that $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] - \boldsymbol{\rho} = \boldsymbol{v}_{\boldsymbol{\xi}}$, and $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] = t_{\boldsymbol{\xi}}$. Together with the fact that Equation (10) holds for each $(\boldsymbol{v}, t) \in \mathcal{V}$, this yields that for each $\boldsymbol{\xi} \in \Xi$,

$$\langle \tilde{\boldsymbol{\lambda}}, \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] - \boldsymbol{\rho} \rangle + \tilde{\mu}\, \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] \leq \beta \leq \tilde{\mu}\, \texttt{OPT}^{\texttt{LP}}_{f,c}. \tag{12}$$

If $\tilde{\mu} > 0$, then for each $\boldsymbol{\xi} \in \Xi$

$$\frac{1}{\tilde{\mu}} \langle \tilde{\boldsymbol{\lambda}}, \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] - \boldsymbol{\rho} \rangle + \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[f(\boldsymbol{x})] \leq \texttt{OPT}^{\texttt{LP}}_{f,c}.$$

By letting $\hat{\boldsymbol{\lambda}} = -\tilde{\boldsymbol{\lambda}}/\tilde{\mu}$, we have $L(\boldsymbol{\xi}, \hat{\boldsymbol{\lambda}}, f, c) \leq \texttt{OPT}^{\texttt{LP}}_{f,c}$ for each $\boldsymbol{\xi}$. In particular, $\sup_{\boldsymbol{\xi}} L(\boldsymbol{\xi}, \hat{\boldsymbol{\lambda}}, f, c) \leq \texttt{OPT}^{\texttt{LP}}_{f,c}$. Then,

$$\inf_{\boldsymbol{\lambda} \geq 0} \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) \leq \texttt{OPT}^{\texttt{LP}}_{f,c}.$$

By weak duality we get $\inf_{\boldsymbol{\lambda} \geq 0} \sup_{\boldsymbol{\xi}} L(\boldsymbol{\xi}, \boldsymbol{\lambda}) = \texttt{OPT}^{\texttt{LP}}_{f,c}$, which proves our statement.

If $\tilde{\mu} = 0$, from Equation (12) we get that, for each $\boldsymbol{\xi} \in \Xi$, $\langle \tilde{\boldsymbol{\lambda}}, \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}}[c(\boldsymbol{x})] - \boldsymbol{\rho} \rangle \leq 0$. Let $\boldsymbol{\xi}_\varnothing \coloneqq \delta_\varnothing \in \Xi$ be the Dirac mass that plays the null action. We have that $\mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}_\varnothing}[c(\boldsymbol{x})] - \boldsymbol{\rho} < 0$. Then, since $\tilde{\boldsymbol{\lambda}} \leq 0$, it must be $\tilde{\boldsymbol{\lambda}} = 0$. That is in contradiction with $(\tilde{\boldsymbol{\lambda}}, \tilde{\mu}) \neq \mathbf{0}$. This concludes the proof. □

Lemma 4.3. *Let $\mathcal{D}$ be defined as in Equation (2). Given $f : \mathcal{X} \to [0,1]$, $c : \mathcal{X} \to [0,1]^m$, $(f,c) \in \mathcal{I}$, it holds*

$$\sup_{\boldsymbol{\xi} \in \Xi} \inf_{\boldsymbol{\lambda} \in \mathcal{D}} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \inf_{\boldsymbol{\lambda} \in \mathcal{D}} \sup_{\boldsymbol{\xi} \in \Xi} L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c) = \texttt{OPT}^{\texttt{LP}}_{f,c}.$$

Proof. As a first step, we show that

$$\inf_{\boldsymbol{\lambda}\in\mathcal{D}} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c) \leq \inf_{\boldsymbol{\lambda}\geq 0} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c).$$

To do so, notice that for any $\boldsymbol{\lambda}'$ with $||\boldsymbol{\lambda}'||_1 > 1/\rho$, we have

$$\sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda}',f,c) > 1 \geq \texttt{OPT}^{\text{LP}}_{f,c} = \inf_{\boldsymbol{\lambda}\geq 0} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c),$$

where the first inequality holds because the null action provides value at least $\langle\boldsymbol{\lambda}',\boldsymbol{\rho}\rangle > 1$. Then, we have:

$$\begin{aligned}
\sup_{\boldsymbol{\xi}\in\Xi} \inf_{\boldsymbol{\lambda}\in\mathcal{D}} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c) &\geq \sup_{\boldsymbol{\xi}\in\Xi} \inf_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c) \\
&= \texttt{OPT}^{\text{LP}}_{f,c} \\
&= \inf_{\boldsymbol{\lambda}\geq 0} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c) \\
&\geq \inf_{\boldsymbol{\lambda}\in\mathcal{D}} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c),
\end{aligned}$$

where the first inequality holds since on the lefthand side we have a more restrictive set of dual variables, and the second and the third inequalities hold by strong duality. Finally, by the max–min inequality

$$\sup_{\boldsymbol{\xi}\in\Xi} \inf_{\boldsymbol{\lambda}\in\mathcal{D}} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c) \leq \inf_{\boldsymbol{\lambda}\in\mathcal{D}} \sup_{\boldsymbol{\xi}\in\Xi} L(\boldsymbol{\xi},\boldsymbol{\lambda},f,c).$$

This proves our statement. □

Lemma 4.4. *Given a distribution over inputs $\mathcal{P}$, let $\bar{f}:\mathcal{X}\to[0,1]$ be the expected reward function, and $\bar{c}:\mathcal{X}\to[0,1]^m$ be the expected resource-consumption function. Then, $T\cdot \texttt{OPT}^{LP}_{\bar{f},\bar{c}} \geq \texttt{OPT}^{DP}$.*

Proof. Let $\boldsymbol{\lambda}^* = (\lambda_1^*,\dots,\lambda_m^*) \in \mathcal{D}$ be an optimal dual vector for LP (1) with functions $\bar{f},\bar{c}$. By strong duality (Lemma 4.3), it holds

$$\bar{f}(\boldsymbol{x}) + \langle\boldsymbol{\lambda}^*, \boldsymbol{\rho} - \bar{c}(\boldsymbol{x})\rangle \leq \texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}},\ \forall \boldsymbol{x}\in\mathcal{X}. \tag{14}$$

Moreover, let $(\boldsymbol{\xi}_t)_{t=1}^T$ be the sequence of strategy mixtures specified by a given policy ψ, and denote by $\boldsymbol{x}_t\sim\boldsymbol{\xi}_t$ an action realization sampled according to the strategy mixture at t. Let

$$Z_t \coloneqq (T-t)\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} + \sum_{t'\in[t]} (f_{t'}(\boldsymbol{x}_{t'}) + \langle\boldsymbol{\lambda}, \boldsymbol{\rho} - c_{t'}(\boldsymbol{x}_{t'})\rangle).$$

Then, by Equation (14), and since

$$\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}\left[\bar{f}(\boldsymbol{x}) + \langle\boldsymbol{\lambda}, \boldsymbol{\rho} - \bar{c}(\boldsymbol{x})\rangle\right] = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t,\,(f_t,c_t)\sim\mathcal{P}}[f_t(\boldsymbol{x}) + \langle\boldsymbol{\lambda}, \boldsymbol{\rho} - c_t(\boldsymbol{x})\rangle],$$

the stochastic process $Z_0,\dots,Z_T$ is a supermartingale. Let $\tau\in[T]$ be the stopping time of the algorithm, that is, when the algorithm depletes the first resource. Then, the realized utility is such that $\sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t) \leq Z_\tau$. This holds since

$$\begin{aligned}
(T-\tau)\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} + \sum_{t=1}^{\tau}\langle\boldsymbol{\lambda}, \boldsymbol{\rho} - c_t(\boldsymbol{x}_t)\rangle &\geq (T-\tau)\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} + \sum_{i=1}^{m}\boldsymbol{\lambda}[i](\tau\rho - T\rho) \\
&\geq (T-\tau)\Big(\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} - \langle\boldsymbol{\rho},\boldsymbol{\lambda}\rangle\Big) \geq 0,
\end{aligned}$$

where the first inequality comes from the fact that $\sum_{t\in[\tau]} c_t(\boldsymbol{x}_t)[i] \leq B$ for each $i\in[m]$, and the last inequality comes from Equation (14) with $\boldsymbol{x}$ equal to the void action $\varnothing$. Then, taking the expectation on both sides, we get $\mathbb{E}[Z_\tau] \geq \mathbb{E}[\sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t)]$. Let v_ψ be the value obtained through policy ψ. By Doob's optional stopping theorem, $T\cdot\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} = Z_0 \geq \mathbb{E}[Z_\tau] \geq v_\psi$. Then, since this holds for every possible policy ψ, we have $T\cdot\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} \geq \sup_{\psi\in\Psi}\mathbb{E}[v_\psi] = \texttt{OPT}^{\text{DP}}$. This concludes the proof. □

A.2. Proofs for Section 6

In this section, we provide more details on why Remark 6.3 holds.

Consider a regret minimizer for bandit feedback guaranteeing with probability at least $1-\delta$ that

$$\hat{R}_T^{\mathrm{P}} \coloneqq \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{t=1}^{\tau} \left(\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}) - \hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}_t) \right) \leq \mathcal{E}_{T,\delta}^{\mathrm{P}}. \tag{15}$$

Then, let

$$w_t \coloneqq \hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}_t) - \ell_t^{\mathrm{P}}(\boldsymbol{\xi}_t) = \hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}_t) - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}\left[\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x})\right],$$

and observe that $|w_t| \leq 1 + 1/\rho$. By applying the Azuma-Hoeffding inequality we get that given a $\tau \in [T]$

$$\Pr\left(\sum_{t=1}^{\tau} w_t > \underbrace{(1+1/\rho)\sqrt{2T\ln(1/\delta)}}_{=:q(\delta)} \right) \leq \delta.$$

Then, by a standard application of the union bound,

$$\Pr\left(\forall \tau \in [T], \quad \sum_{t=1}^{\tau} w_t \leq q(\delta) \right) \leq 1 - T\cdot\delta.$$

Then, with probability at least $1-\delta$,

$$\sum_{t=1}^{\tau} \ell_t^{\mathrm{P}}(\boldsymbol{x}_t) \leq q(T/\delta) + \sum_{t=1}^{\tau} \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\xi}_t}\left[\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x})\right].$$

Then, by Remark 6.2 and by Equation (15), we obtain that, with probability at least $1-2\delta$, for each $\tau \in [T]$

$$\begin{aligned}
\sup_{\boldsymbol{\xi}\in\Xi} \sum_{t=1}^{\tau} \left(\ell_t^{\mathrm{P}}(\boldsymbol{\xi}) - \ell_t^{\mathrm{P}}(\boldsymbol{\xi}_t) \right) &= \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{t=1}^{\tau} \left(\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}) - \ell_t^{\mathrm{P}}(\boldsymbol{\xi}_t) \right) \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{t=1}^{\tau} \left(\hat{\ell}_t^{\mathrm{P}}(\boldsymbol{x}) - \ell_t^{\mathrm{P}}(\boldsymbol{x}_t) \right) + q(T/\delta) \\
&\leq \mathcal{E}_{T,\delta}^{\mathrm{P}} + q(T/\delta).
\end{aligned}$$

This show that Remark 6.3 is verified.

A.3. Proofs for Section 7

Theorem 7.1. *Consider Meta-Algorithm 1 equipped with two arbitrary regret minimizers $\mathcal{R}^P$ and $\mathcal{R}^D$ for the sets Ξ and $\mathcal{D}$, respectively. In particular, assume that they guarantee a cumulative regret up to time T which is upper bounded by $\mathcal{E}_T^P$ and $\mathcal{E}_T^D$, respectively. For each $t \in [T]$, let the inputs(f_t, c_t) be i.i.d. samples from a fixed but unknown distribution $\mathcal{P}$ over the set of possible requests $\mathcal{I}$. For $\delta > 0$, with probability at least $1-\delta$ we have*

$$\mathtt{OPT}^{DP} - \mathtt{REW}_{\boldsymbol{\gamma}} \leq O\left(\frac{1}{\rho}\sqrt{2T\log(mT/\delta)} \right) + \mathcal{E}_T^P + \mathcal{E}_T^D,$$

where $\mathtt{REW}_{\boldsymbol{\gamma}} \coloneqq \sum_t f_t(\boldsymbol{x}_t)$ is the reward of the algorithm for the sequence of inputs $\boldsymbol{\gamma}$.

Proof. Let $\tau \in [T]$ be the stopping time of Algorithm 1, and $\bar{\boldsymbol{\xi}} \in \Xi$ be such that, for all $\boldsymbol{x} \in \mathcal{X}$, $\bar{\boldsymbol{\xi}}(\boldsymbol{x}) \coloneqq \sum_{t=1}^{\tau} \boldsymbol{\xi}_t(\boldsymbol{x})/\tau$. The proof proceeds in two steps.

Step 1 Consider the first $\tau \in [T]$ rounds. By applying the Azuma-Hoeffding inequality we have that the average reward and cost for each resource i up to τ is *close*, with high probability, to $\mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}[\bar{f}(\boldsymbol{x})]$ and $\mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}[\bar{c}(\boldsymbol{x})[i]]$, respectively. Formally, given $\tau \in [T]$, by letting $\mathcal{E}^0_{\tau,\delta} \coloneqq O(\sqrt{\tau \log(m/\delta)})$ with probability at least $1-\delta$, we have

$$\frac{1}{\tau}\sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t) \geq \mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}\big[\bar{f}(\boldsymbol{x})\big] - \frac{1}{\tau}\mathcal{E}^0_{\tau,\delta} \tag{16}$$

$$\frac{1}{\tau}\sum_{t=1}^{\tau} c_t(\boldsymbol{x})[i] \leq \mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}[\bar{c}(\boldsymbol{x})[i]] + \frac{1}{\tau}\mathcal{E}^0_{\tau,\delta} \quad \forall i \in [m]. \tag{17}$$

Step 2 Consider a sequence of repeated two-player, zero-sum games up to a given time $\tau \in [T]$, in which Player 1 (i.e., the *primal player*) chooses as their action $\boldsymbol{\xi} \in \Xi$, and Player 2 (i.e., the *dual player*) chooses $\boldsymbol{\lambda} \in \mathcal{D}$. For each $t \in [\tau]$, for a pair of actions $(\boldsymbol{\xi}, \boldsymbol{\lambda})$, Player 1 (resp., Player 2) observes utility function $L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f_t, c_t)$ (resp., $-L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f_t, c_t)$). When (f_t, c_t) are drawn i.i.d. from some fixed distribution $\mathcal{P}$, we can define $\bar{L}(\boldsymbol{\xi}, \boldsymbol{\lambda}) \coloneqq \mathbb{E}_{(f,c)\sim\mathcal{P}}[L(\boldsymbol{\xi}, \boldsymbol{\lambda}, f, c)]$. We say that $(\Xi, \mathcal{D}, \bar{L})$ is the *expected Lagrangian game*. Then, the following result holds.

Lemma A.1. *Given $\tau \in [T]$, for $\delta \in (0,1)$, with probability at least $1-2\delta$, the average strategy mixture $\bar{\boldsymbol{\xi}} \in \Xi$ up to τ is such that, for any $\boldsymbol{\lambda} \in \mathcal{D}$, $\bar{L}(\bar{\boldsymbol{\xi}}, \boldsymbol{\lambda}) \geq \mathtt{OPT}^{LP}_{\bar{f},\bar{c}} - \frac{1}{\tau}\Big(\mathcal{E}^P_T + \mathcal{E}^D_T + 4(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big)$.*

Let us condition on the fact that Equation (16), Equation (17), and Lemma A.1 hold for each $\tau \in [T]$. This event holds with probability $1-3\delta T$ (by a standard application of the union bound). Then, let $\tau \in [T]$ be the stopping time of the algorithm. By definition of stopping time, there exists a resource $i^{\mathrm{s}} \in [m]$ such that $\sum_{t=1}^{\tau} c_t(\boldsymbol{x}_t)[i^{\mathrm{s}}] > B - 1$ (see the primal decision in Algorithm 1). By taking $\hat{\boldsymbol{\lambda}}$ such that $\hat{\boldsymbol{\lambda}}[i^{\mathrm{s}}] = 1/\rho$, and $\hat{\boldsymbol{\lambda}}[i] = 0$ for $i \neq i^{\mathrm{s}}$, and by Equations (16) and (17), we have

$$\begin{aligned}
\bar{L}\Big(\bar{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}}\Big) &= \mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}\big[\bar{f}(\boldsymbol{x})\big] + \langle \hat{\boldsymbol{\lambda}}, \boldsymbol{\rho} - \mathbb{E}_{\boldsymbol{x}\sim\bar{\boldsymbol{\xi}}}[\bar{c}(\boldsymbol{x})]\rangle \\
&\leq \frac{1}{\tau}\left(\sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t) + \tau - \frac{1}{\rho}\sum_{t=1}^{\tau} c_t(\boldsymbol{x}_t)[i^{\mathrm{s}}] + 2\mathcal{E}^0_{\tau,\delta}\right) \\
&\leq \frac{1}{\tau}\left(\sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t) + \tau - T + \frac{1}{\rho} + 2\mathcal{E}^0_{\tau,\delta}\right).
\end{aligned}$$

Then, plugging the above expression in Lemma A.1 yields the following

$$\mathtt{REW}_{\boldsymbol{\gamma}} = \sum_{t=1}^{\tau} f_t(\boldsymbol{x}_t) \geq \tau\,\mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{c}} + T - \tau - \frac{1}{\rho} - \Big(2\mathcal{E}^0_{\tau,\delta} + \mathcal{E}^{\mathrm{P}}_T + \mathcal{E}^{\mathrm{D}}_T + 4(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big).$$

Then, by Lemma 4.4, and since $\mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{c}} \leq 1$,

$$\begin{aligned}
\mathtt{OPT}^{\mathrm{DP}} - \mathtt{REW}_{\boldsymbol{\gamma}} &\leq T\,\mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{c}} - \tau\,\mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{c}} - T + \tau + \frac{1}{\rho} + 2\mathcal{E}^0_{\tau,\delta} + \mathcal{E}^{\mathrm{P}}_T + \mathcal{E}^{\mathrm{D}}_T + 4(1+1/\rho)\sqrt{2T\log(T/\delta)} \\
&\leq \mathcal{E}^{\mathrm{P}}_T + \mathcal{E}^{\mathrm{D}}_T + O\Big(1/\rho\sqrt{2T\log(mT/\delta)}\Big).
\end{aligned}$$

This concludes the proof. □

Lemma A.1. *Given $\tau \in [T]$, for $\delta \in (0,1)$, with probability at least $1-2\delta$, the average strategy mixture $\bar{\boldsymbol{\xi}} \in \Xi$ up to τ is such that, for any $\boldsymbol{\lambda} \in \mathcal{D}$, $\bar{L}(\bar{\boldsymbol{\xi}}, \boldsymbol{\lambda}) \geq \mathtt{OPT}^{LP}_{\bar{f},\bar{c}} - \frac{1}{\tau}\Big(\mathcal{E}^P_T + \mathcal{E}^D_T + 4(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big)$.*

Proof. We proceed in two steps:

(1) Let $\tau \in [T]$, and $\boldsymbol{\xi}^*$ be the optimal strategy up to τ for the primal player in hindsight (*i.e.*, with knowledge of the sequence of $\boldsymbol{\lambda}_t$ up to τ, while taking expectations over f_t and c_t at each t). Moreover, let v^* be the minimax value for the expected Lagrangian game $(\Xi, \mathcal{D}, \bar{L})$. Then,

$$\begin{aligned}
\frac{1}{\tau}\sum_{t=1}^{\tau} L(\boldsymbol{\xi}_t, \boldsymbol{\lambda}_t, f_t, c_t) &\geq \frac{1}{\tau}\sum_{t=1}^{\tau} L(\boldsymbol{\xi}^*, \boldsymbol{\lambda}_t, f_t, c_t) - \frac{1}{\tau}\mathcal{E}_T^{\mathtt{P}} \\
&\geq_{\text{WHP}} \frac{1}{\tau}\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{\xi}^*, \boldsymbol{\lambda}_t) - \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{P}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big) \\
&= \sup_{\boldsymbol{\xi}} \bar{L}(\boldsymbol{\xi}, \bar{\boldsymbol{\lambda}}) - \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{P}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big) \\
&\geq \inf_{\boldsymbol{\lambda}}\sup_{\boldsymbol{\xi}} \bar{L}(\boldsymbol{\xi}, \boldsymbol{\lambda}) - \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{P}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big) \\
&= v^* - \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{P}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big),
\end{aligned}$$

where $\geq_{\text{WHP}}$ denotes statements that hold with probability at least $1-\delta$.

(2) Fix $\boldsymbol{\lambda} \in \mathcal{D}$. We have,

$$\begin{aligned}
\frac{1}{\tau}\sum_{t=1}^{\tau} L(\boldsymbol{\xi}, \boldsymbol{\lambda}_t, f_t, c_t) &\leq \frac{1}{\tau}\sum_{t=1}^{\tau} L(\boldsymbol{\xi}_t, \boldsymbol{\lambda}, f_t, c_t) + \frac{1}{\tau}\mathcal{E}_T^{\mathtt{D}} \\
&\leq_{\text{WHP}} \frac{1}{\tau}\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{\xi}_t, \boldsymbol{\lambda}) + \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{D}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big) \\
&= \bar{L}(\bar{\boldsymbol{\xi}}, \boldsymbol{\lambda}) + \frac{1}{\tau}\Big(\mathcal{E}_T^{\mathtt{D}} + 2(1+1/\rho)\sqrt{2T\log(T/\delta)}\Big).
\end{aligned}$$

By Lemma 4.3, we have $v^* = \mathtt{OPT}^{\mathtt{LP}}_{\bar{f},\bar{c}}$. Then, by combining the inequalities from Step (1) and (2), and by taking a union bound we get the result. □

A.4. Proofs for Section 8

Lemma 8.2. *Let $\ell_{\mathtt{L},t}(\boldsymbol{x}, \boldsymbol{\lambda}) \coloneqq f_t(\boldsymbol{x}) - \langle \boldsymbol{\lambda}, \boldsymbol{x}^\top C_t\rangle$ for all pairs $(\boldsymbol{x}, \boldsymbol{\lambda})$. Then, for each $\tau \in [T]$, each sequence of receiver's types $(k_t)_{t=1}^{\tau}$, and each sequence $(\boldsymbol{\lambda}_t)_{t=1}^{\tau}$, it holds:*

$$\max_{\boldsymbol{x}^* \in \mathcal{X}^*} \sum_{t=1}^{\tau} \ell_{\mathtt{L},t}(\boldsymbol{x}^*, \boldsymbol{\lambda}_t) = \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{\tau} \ell_{\mathtt{L},t}(\boldsymbol{x}, \boldsymbol{\lambda}_t).$$

Proof. We show that given an optimal strategy $\boldsymbol{x} \in \mathcal{X}$, we can build a strategy $\boldsymbol{x}^* \in \mathcal{X}^*$ with the same utility. Let $\boldsymbol{a} \in \mathcal{A}_{\mathtt{L}}^{|\mathcal{K}|}$ be the tuple specifying one action per type such that $\boldsymbol{e}_{\boldsymbol{a}[k]} = \boldsymbol{y}_{\boldsymbol{x}}^k$ for each $k \in \mathcal{K}$, *i.e.*, each follower's type plays the best response for $\boldsymbol{x}$. Once we fix the best response for all the types, the objective is a linear function. Hence, it is linear on $\mathcal{X}^{\boldsymbol{a}}$. Then, there exists a vertex of $\mathcal{X}^{\boldsymbol{a}} \subseteq \mathcal{X}^*$ in which the objective is maximized. Notice that in the vertex the follower could play a different best response. However, by the optimistic tie breaking assumption, with the best response the leader's utility increases, while the costs do not change. This concludes the proof. □

Lemma 8.3. *It holds $|\mathcal{X}^*| \leq (|\mathcal{K}| n_{\mathtt{F}}^2)^{n_{\mathtt{L}}-1}$*

Proof. Each polytope $\mathcal{X}^{\boldsymbol{a}}$, $\boldsymbol{a} \in \mathcal{A}_{\mathtt{L}}^{|\mathcal{K}|}$ is defined by the following inequalities over $\boldsymbol{x}$ that imply the optimality of the tuple of best responses $\boldsymbol{a}$:

$$\boldsymbol{x}^\top U_k \boldsymbol{e}_{\boldsymbol{a}[k]} \geq \boldsymbol{x}^\top U_k \boldsymbol{e}_a \;\; \forall k \in \mathcal{K}, a \in \mathcal{A}_{\mathtt{F}}.$$

Thus, each vertex $V(\mathcal{X}^{\boldsymbol{a}})$, $\boldsymbol{a} \in \mathcal{A}_{\text{L}}^{|\mathcal{K}|}$, is the intersection of $n_{\text{L}} - 1$ equalities belonging to the following set:

$$\boldsymbol{x}^\top U_k \boldsymbol{e}_a = \boldsymbol{x}^\top U_k \boldsymbol{e}_{a'} \;\; \forall k \in \mathcal{K},\, a, a' \in \mathcal{A}_{\text{F}},$$

and the simplex constraint. Hence, there are at most $(|\mathcal{K}| n_{\text{F}}^2)^{n_{\text{L}}-1}$ vertices. □

B. On Strong Duality in Semi-Infinite LPs

We provide a simple example in which a semi-infinite linear optimization problem does not admit strong duality.

Let $\mathcal{X} = [0, 1]$. Define $f : \mathcal{X} \to \mathbb{R}$ by

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \in (0, 1] \end{cases}.$$

Then, let $\Xi = \Delta^{\mathcal{X}}$ and consider the linear program

$$\begin{cases} \inf\limits_{\xi \in \Xi} \mathbb{E}_{x \sim \xi}[f(x)] \\ \text{s.t. } \mathbb{E}_{x \sim \xi}[x] \leq 0 \end{cases}. \tag{18}$$

Since $x \in [0, 1]$, the only way in which the constraint can be satisfied is always selecting $x = 0$. Then, the primal optimal value is $p^* = 1$. Now, the Lagrangian dual of the problem is

$$g(\lambda) = \inf_{\xi \in \Xi} \{\mathbb{E}[f(x)] + \lambda \mathbb{E}[x]\}.$$

We have $d^* = \sup_{\lambda \geq 0} g(\lambda) = 0$. Therefore, we have a duality gap of 1.