BANDITS WITH REPLENISHABLE KNAPSACKS: THE BEST OF BOTH WORLDS

Martino Bernasconi
Bocconi University
martino.bernasconi@unibocconi.it

Matteo Castiglioni
Politecnico di Milano
matteo.castiglioni@polimi.it

Andrea Celli
Bocconi University
andrea.celli2@unibocconi.it

Federico Fusco
Sapienza University
federico.fusco@uniroma1.it

ABSTRACT

The bandits with knapsacks (BwK) framework models online decision-making problems in which an agent makes a sequence of decisions subject to resource consumption constraints. The traditional model assumes that each action consumes a non-negative amount of resources and the process ends when the initial budgets are fully depleted. We study a natural generalization of the BwK framework which allows non-monotonic resource utilization, i.e., resources can be replenished by a positive amount. We propose a best-of-both-worlds primal-dual template that can handle any online learning problem with replenishment for which a suitable primal regret minimizer exists. In particular, we provide the first positive results for the case of adversarial inputs by showing that our framework guarantees a constant competitive ratio α when $B = \Omega(T)$ or when the possible per-round replenishment is a positive constant. Moreover, under a stochastic input model, our algorithm yields an instance-independent $\tilde{\mathcal{O}}(T^{1/2})$ regret bound which complements existing instance-dependent bounds for the same setting. Finally, we provide applications of our framework to some economic problems of practical relevance.

1 INTRODUCTION

We study online learning problems in which a decision maker tries to maximize their cumulative reward over a time horizon T, subject to a set of m resource-consumption constraints. At each t, the decision maker plays an action $x_t \in \mathcal{X}$, and subsequently observes a realized reward $f_t(x_t)$, with $f_t : \mathcal{X} \to [0, 1]$, and an m-dimensional vector of resource consumption $\boldsymbol{c}_t(x_t)$.

Our framework extends the well-known *Bandits with Knapsacks* (BwK) framework of Badanidiyuru et al. (2018). In the BwK model, the resource consumption is *monotonic* (*i.e.,* $\boldsymbol{c}_t(\cdot) \in [0, 1]^m$ for all $t \in [T]$). This framework has numerous motivating applications ranging from dynamic pricing to online ad allocation (see, *e.g.,* Besbes and Zeevi (2009); Babaioff et al. (2012); Wang et al. (2014); Badanidiyuru et al. (2012); Combes et al. (2015)), and it has been extended in numerous directions such as modeling adversarial inputs (Immorlica et al., 2022) and other non-stationary input models Celli et al. (2023); Fikioris and Tardos (2023), more general notions of resources and constraints Agrawal and Devanur (2019), contextual and combinatorial bandits Badanidiyuru et al. (2014); Agrawal et al. (2016); Sankararaman and Slivkins (2018).

Kumar and Kleinberg (2022) recently proposed a natural generalization of the BwK model in which resource consumption can be *non-monotonic*, *i.e.,* , resources can be replenished over time so costs are $\boldsymbol{c}_t(x_t) \in [-1, 1]^m$. We call such model *Bandits with Replenishable Knapsacks* (BwRK). Kumar and Kleinberg (2022) focus on the stochastic setting, while we are interested in providing *best-of-both-worlds* algorithms that provide guarantees under both stochastic and *adversarial* inputs.

Contributions. We propose a general primal-dual template that can handle online learning problems in which the decision maker has to guarantee some long-term resource-consumption constraints and resources can be renewed over time. We show that our framework provides *best-of-both-worlds* guarantees in the spirit of Balseiro et al. (2022): it guarantees a regret bound of $\tilde{O}(T^{1/2})$ in the case in which $(f_t, \boldsymbol{c}_t)$ are i.i.d. samples from a fixed but unknown distribution, and it guarantees a constant-factor competitive ratio in the case in which budgets grow at least linearly in T, or when the possible per-round replenishment is a positive constant. We remark that known best-of-both-worlds frameworks like the one by Balseiro et al. (2022) cannot be applied to this setting as they assume monotonic resource consumption. In that case, we show that our framework recovers the state-of-the-art rate of $1/\rho$ by Castiglioni et al. (2022a), where ρ is the per-iteration budget. Our primal-dual template is applicable to any online problem for which a suitable primal regret minimizer is available. Therefore, we first provide general guarantees for the framework without making any assumption on the primal and dual regret minimizers being employed (Section 4). Then, we show how such regret minimizers should be chosen depending on the information available to the learner about the intensity of the budget replenishment (Section 5). Moreover, we provide explicit bounds that only depend on the guarantees of the primal regret minimizer. In particular, we show how the primal and dual minimizers should be instantiated in the case in which the amount of resources that can be replenished at each time t is known, and in the more challenging case in which it is unknown a-priori to the decision maker. Finally, we demonstrate the flexibility of our framework by instantiating it in some relevant settings (Section 6). First, we instantiate the framework in the BwRK model by Kumar and Kleinberg (2022), thereby providing the first positive results for BwRK under adversarial inputs, and the first instance-independent regret bound for the stochastic setting. The latter complements the instance-dependent analysis by Kumar and Kleinberg (2022). Then, we apply the framework to a simple inventory management problem, and to revenue maximization in bilateral trade.

Related works. Primal-dual approaches for bandit problems with resource-consumption constraints popularized by the work of Immorlica et al. (2022) cannot be applied in our setting. Such primal-dual approaches (see also Castiglioni et al. (2022a); Balseiro et al. (2022); Balseiro and Gur (2019)) usually require as an input the Slater's parameter of the problem (*i.e.,* in their setting, the per-round budget). This is not the case in our setting since the amount by which each constraints is replenished is a priori unknown. This issue has been effectively addressed in stochastic settings where sublinear constraints violations are allowed, such as online optimization under *stochastic* inputs (see, *e.g.,* , Yu et al. (2017); Wei et al. (2020)), and *stochastic* online learning problems with long-term constraints Castiglioni et al. (2022b); Slivkins et al. (2023). However, in BwK problems constraints are required to be satisfied strictly at all rounds. Addressing this issue in the adversarial setting is considerably more challenging and this direction remains largely unexplored. A notable exception is the recent work by Castiglioni et al. (2023). Such works focuses on a setting with only *one* monotone budget constraint, for which Slater's parameter is known, and *one* "soft" return-on-investments (ROI) constraints (*i.e.,* ROI constraints can be violated up to a sublinear amount). Being allowed to violated the ROI constrain renders the problem significantly simpler than ours, where budget constraints must be strictly satisfied at each time t. Moreover, our framework can handle an arbitrary number of constraints, while Castiglioni et al. (2023) can only manage one unknown feasibilty parameter. Finally, we mention that in the case of stochastic inputs other approaches for budget-management problems are known, most notably *optimism-under-uncertainty* (Agrawal and Devanur, 2019; Badanidiyuru et al., 2018).

2 PRELIMINARIES

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

2.1 BASIC SET-UP

There are T rounds and m resources. The decision maker has an arbitrary non-empty set of available strategies $\mathcal{X}$. In each round $t \in [T]$, the decision maker chooses $x_t \in \mathcal{X}$, and subsequently observes a reward function $f_t : \mathcal{X} \to [0, 1]$, and a function $\boldsymbol{c}_t : \mathcal{X} \to [-1, 1]^m$ specifying the consumption or replenishment of each of the m resources. Each resource $i \in [m]$ is endowed with an initial budget of B to be spent over the T steps.[1] We denote by ρ the per-iteration budget, which is such that $B = T\rho$,

[1]For ease of notation we consider a uniform intial buget, but the case of different initial budgets easily follows from our results.

and we let $\boldsymbol{\rho} \coloneqq \rho \mathbf{1} \in \mathbb{R}^m_{>0}$, where $\mathbf{1} \in \mathbb{R}^m$ is the vector of all ones. For $i \in [m]$ and $x \in \mathcal{X}$, if $c_{t,i}(x) < 0$ we say that at time t action x *restores* a positive amount to the budget available for the i-th resource. If $c_{t,i}(x) > 0$, we say that action x at time t *depletes* some of the available budget.

Let $\gamma_t \coloneqq (f_t, c_t)$ be the input pair at time t, and $\boldsymbol{\gamma}_T \coloneqq (\gamma_1, \gamma_2, \ldots, \gamma_T)$ be the sequence of inputs up to time T. The repeated decision making process stops at the end of the time horizon T. The goal of the decision maker is to maximize the cumulative reward $\sum_{t=1}^T f_t(x_t)$ while satisfying the resource constraints $\sum_{t=1}^T c_{t,i}(x_t) \le T \cdot \rho$ for each $i \in [m]$. Given two functions $f : \mathcal{X} \to \mathbb{R}$ and $\boldsymbol{c} : \mathcal{X} \to \mathbb{R}^m$, we denote by $\mathcal{L}_{f,\boldsymbol{c}} : \mathcal{X} \times \mathbb{R}^m_{\ge 0} \to \mathbb{R}$ the Lagrangian function defined as

$$\mathcal{L}_{f,\boldsymbol{c}}(x, \boldsymbol{\lambda}) \coloneqq f(x) + \langle \boldsymbol{\lambda}, \boldsymbol{\rho} - \boldsymbol{c}(x) \rangle \quad \text{for all } x \in \mathcal{X}, \boldsymbol{\lambda} \in \mathbb{R}^m_{\ge 0}.$$

Given $\mathcal{X}$, the set of *strategy mixtures* Ξ is the set of probability measures on the Borel sets of $\mathcal{X}$.

2.2 Baseline adversarial setting

Given a sequence of inputs $\boldsymbol{\gamma}_T$ selected by an oblivious adversary, the baseline for the adversarial setting is $\mathtt{OPT}_{\boldsymbol{\gamma}} \coloneqq \sup_{x \in \mathcal{X}} \sum_{t=1}^T f_t(x)$, which is the total expected reward of the best *fixed unconstrained* strategy in hindsight belonging to $\mathcal{X}$. Moreover, for any sequence of inputs $\boldsymbol{\gamma}_T$, and $t \in [T]$, let $\tilde{f}_t : \mathcal{X} \to [0,1]$ and $\tilde{\boldsymbol{c}}_t : \mathcal{X} \to [-1,1]^m$ be such that:

$$\tilde{f}_t(x) \coloneqq \frac{1}{t} \sum_{s=1}^t f_s(x) \quad \text{and} \quad \tilde{\boldsymbol{c}}_t(x) \coloneqq \frac{1}{t} \sum_{s=1}^t \boldsymbol{c}_s(x), \qquad \forall x \in \mathcal{X}. \tag{2.1}$$

Then, for $t \in [T]$, we define $\mathtt{OPT}_{\tilde{f}_t} \coloneqq \sup_{x \in \mathcal{X}} \tilde{f}_t(x)$, and the baseline for the adversarial setting can be rewritten as $\mathtt{OPT}_{\boldsymbol{\gamma}} = T \cdot \mathtt{OPT}_{\tilde{f}_T}$. In the setting with monotonic resource utilization and adversarial inputs, previous work usually employs weaker baselines (see, *e.g.,* Immorlica et al. (2022); Castiglioni et al. (2022a;b)). For example, Castiglioni et al. (2022a) considers the reward attained by the best fixed strategy mixture until budget depletion, after which the void action is played. We show that, despite the stronger baseline, we match the state-of-the-art $1/\rho$ competitive-ratio by Castiglioni et al. (2022a) when constraints are monotonic. We will work under the following standard assumption.

Assumption 2.1. *There exists a* void action $\varnothing \in \mathcal{X}$ *and a constant* $\beta \ge 0$ *such that* $c_{t,i}(\varnothing) \le -\beta$*, for all resources* $i \in [m]$ *and* $t \in [T]$.

Notice that when $\beta = 0$ we recover the standard assumption of BwK (see, *e.g.,* Badanidiyuru et al. (2018)). We will often parametrize regret bounds using $\nu \coloneqq \beta + \rho$. This parameter measures how much budget is available at each iteration, and how fast the available budget can be replenished.

2.3 Baseline stochastic setting

In the stochastic version of the problem, each input $\gamma_t = (f_t, \boldsymbol{c}_t)$ is drawn i.i.d. from some fixed but unknown distribution $\mathcal{P}$ over a set of possible input pairs. Let $\bar{f} : \mathcal{X} \to [0,1]$ be the expected reward function, and $\bar{\boldsymbol{c}} : \mathcal{X} \to [-1,1]^m$ be the expected resource-consumption function (where both expectations are taken with respect to $\mathcal{P}$).

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

$$\mathtt{OPT}^{\mathrm{LP}}_{f,c} \coloneqq \begin{cases} \sup\limits_{\xi \in \Xi} \mathbb{E}_{x \sim \xi}[f(x)] \\ \text{s.t. } \mathbb{E}_{x \sim \xi}[\boldsymbol{c}(x)] \preceq \boldsymbol{\rho} \end{cases}. \tag{2.2}$$

In the stochastic setting, our baseline is $\mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{\boldsymbol{c}}}$. It is well-known that $T \cdot \mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{\boldsymbol{c}}}$ is an upper bound on the expected reward of any algorithm (see, *e.g.,* , (Badanidiyuru et al., 2018, Lemma 3.1) and (Kumar and Kleinberg, 2022, Lemma 2.1)).

Lemma 2.2 (Lemma 2.1 of Kumar and Kleinberg (2022))**.** *In the stochastic setting, the total expected reward of any algorithm is at most* $T \cdot \mathtt{OPT}^{\mathrm{LP}}_{\bar{f},\bar{\boldsymbol{c}}}$.

In the stochastic setting we make the same "positive drift" assumption of Kumar and Kleinberg (2022), which is weaker than our Assumption 2.1 in the adversarial case.

Assumption 2.3. *There exists of a* void action $\varnothing \in \mathcal{X}$ *such that, for all resources* $i \in [m]$*, it holds that* $\mathbb{E}[c_i(\varnothing)] \le -\beta$ *and* $\mathbb{P}[c_i(\varnothing) \le 0] = 1$*, where* $\beta \ge 0$ *and the expectation is with respect to the draw of the* $\boldsymbol{c}$ *from* $\mathcal{P}$.

Algorithm 1 Primal-Dual template

Input: parameters B, T; regret minimizers $\mathcal{A}^{\text{P}}$ and $\mathcal{A}^{\text{D}}$
Initialization: $B_{1,i} \leftarrow B, \forall i \in [m]$; initialize $\mathcal{A}^{\text{P}}, \mathcal{A}^{\text{D}}$; $\mathcal{T}_{\text{G}} = \{\emptyset\}, \mathcal{T}_{\varnothing} = \{\emptyset\}$.
for $t = 1, 2, \ldots, T$ **do**
  if $\exists\, i \in [m] : B_{t,i} < 1$ **then**
    $\mathcal{T}_{\varnothing} \leftarrow \mathcal{T}_{\varnothing} \cup \{t\}$
    Primal action: $x_t \leftarrow \varnothing$
    Observe costs: Observe $\boldsymbol{c}_t(\varnothing)$ and update available resources: $\boldsymbol{B}_{t+1} \leftarrow \boldsymbol{B}_t - \boldsymbol{c}_t(\varnothing)$
  else
    $\mathcal{T}_{\text{G}} \leftarrow \mathcal{T}_{\text{G}} \cup \{t\}$
    Dual decision: $\boldsymbol{\lambda}_t \leftarrow \mathcal{A}^{\text{D}}$.NEXTELEMENT()
    Primal decision: $x_t \leftarrow \mathcal{A}^{\text{P}}$.NEXTELEMENT()
    Observe cost: Observe $\boldsymbol{c}_t(x_t)$ and update available resources: $\boldsymbol{B}_{t+1} \leftarrow \boldsymbol{B}_t - \boldsymbol{c}_t(x_t)$
    Primal update:
    - $u_t^{\text{P}}(x_t) \leftarrow f_t(x_t) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x_t) \rangle$
    - $\mathcal{A}^{\text{P}}$.OBSERVEUTILITY($u_t^{\text{P}}(x_t)$)

    Dual update:
    - $u_t^{\text{D}} : \mathbb{R}^d \ni \boldsymbol{\lambda} \mapsto \langle \boldsymbol{\lambda}, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho} \rangle$
    - $\mathcal{A}^{\text{D}}$.OBSERVEUTILITY(u_t^{D})

2.4 REGRET MINIMIZATION

We will consider regret minimizers for a set $\mathcal{W}$ as generic algorithms that implements two functions: (i) the function NEXTELEMENT() returns an element $w_t \in \mathcal{W}$, and (ii) the function OBSERVEUTILITY($\cdot$) which takes some feedback and updates the internal state of the regret minimizer. In the *full-feedback* model the regret minimizer observes as feedback a function $u_t : \mathcal{X} \to \mathbb{R}$, while in the *bandit-feedback* model it observes only the realized $u_t(w_t)$. The standard objective of a regret minimizer is to control the cumulative regret with respect to a set $\mathcal{Y} \subseteq \mathcal{W}$ defined as $\mathcal{R}_T(\mathcal{Y}) := \sup_{w \in \mathcal{Y}} \sum_{t=1}^{T} (u_t(w) - u_t(w_t))$. In the following, we will also exploit a more general notion of regret, in which the regret minimizer suffers regret only in specific rounds. In particular, given a subset of rounds $\mathcal{T} \subset [T]$ we define $\mathcal{R}_{\mathcal{T}}(\mathcal{Y}) := \sup_{w \in \mathcal{Y}} \sum_{t \in \mathcal{T}} (u_t(w) - u_t(w_t))$. Then, we can recover common notions of regret such as standard (external) regret for which $\mathcal{T} = [T]$, and *weakly-adaptive regret* for which $\mathcal{R}_T^{\text{I}}(\mathcal{Y}) := \sup_{\mathcal{I} = [t_1, t_2] \subseteq [T]} \mathcal{R}_{\mathcal{I}}(\mathcal{Y})$ Hazan and Seshadhri (2007). We remove the dependency from $\mathcal{Y}$ when $\mathcal{Y} = \mathcal{W}$ (*e.g.,* we write $\mathcal{R}_T$ in place of $\mathcal{R}_T(\mathcal{W})$).

3 PRIMAL-DUAL TEMPLATE

We assume to have access to two regret minimizers with the following characteristic. A bandit-feedback *primal* regret minimizer $\mathcal{A}^{\text{P}}$ which outputs a strategy $x_t \in \mathcal{X}$ at each t, and subsequently receives as feedback the realized utility function $u_t^{\text{P}}(x_t) = f_t(x_t) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x_t) \rangle$, and a full-feedback *dual* regret minimizer $\mathcal{A}^{\text{D}}$ that receives as input the utility function: $u_t^{\text{D}} : \boldsymbol{\lambda} \mapsto \langle \boldsymbol{\lambda}, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho} \rangle$. note that the dual regret minimizer always has full feedback by construction.[2]

Algorithm 1 summarizes the structure of our primal-dual template. For each t, if the available budget $B_{t,i}$ is less than 1 for some resource i, the algorithm plays the void action $\varnothing$ and updates the budget accordingly. This ensures that the budget will never fall below 0. Otherwise, the regret minimizer plays action x_t at time t, which is determined by invoking NEXTELEMENT(). Then, $u_t^{\text{P}}(x_t)$ and u_t^{D} are observed, and the budget consumption is updated according to the realized costs $\boldsymbol{c}_t$. If the budget was at least 1, the internal state of the two regret minimizers is updated via OBSERVEUTILITY($\cdot$), on the basis of the feedback specified by the primal loss $u_t^{\text{P}}(x_t)$, and the dual loss function u_t^{D}. The algorithm terminates when the time horizon T is reached.

We partition the set of rounds in two disjoint sets $\mathcal{T}_{\text{G}} \subseteq [T]$ and $\mathcal{T}_{\varnothing} \subseteq [T]$. The set $\mathcal{T}_{\text{G}} := \{t \in [T] : \forall i \in [m], B_{t,i} \geq 1\}$ includes all the rounds in which all the resources were at least 1, and hence

[2] We focus on the more challenging bandit-feedback setting. Our results easily extend to full-feedback.

the regret minimizers $\mathcal{A}^{\text{P}}$ and $\mathcal{A}^{\text{D}}$ were actually invoked. On the other hand, $\mathcal{T}_{\varnothing} := \{t \in [T] : \exists i \in [m], B_{t,i} < 1\}$ is the set of rounds in which at least one resource is smaller than 1. Clearly, we have $\mathcal{T}_{\text{G}} \cup \mathcal{T}_{\varnothing} = [T]$. Then, let $\tau \in \mathcal{T}_{\varnothing}$ be the last time in which the budget was strictly less then 1 for at least one resource, *i.e.,* $\tau = \max \mathcal{T}_{\varnothing}$. We partition $\mathcal{T}_{\text{G}}$ in two sets $\mathcal{T}_{\text{G},<\tau}$ and $\mathcal{T}_{\text{G},>\tau}$ which are the rounds in $\mathcal{T}_{\text{G}}$ before and after τ, respectively. Formally, $\mathcal{T}_{\text{G},<\tau} := [\tau] \setminus \mathcal{T}_{\varnothing}$, and $\mathcal{T}_{\text{G},>\tau} := [T] \setminus [\tau]$.

We denote by $\mathcal{R}^{\text{P}}_{T}$ (resp., $\mathcal{R}^{\text{D}}_{T}$) the cumulative regret incurred by $\mathcal{A}^{\text{P}}$ (resp., $\mathcal{A}^{\text{D}}$). Following the notation introduced in Section 2.4 we will write $\mathcal{R}^{\text{P}}_{\mathcal{T}_{\text{G}}}$ to denote the regret accumulated by the primal regret minizer over time steps in $\mathcal{T}_{\text{G}}$. Let $\mathcal{D} := \{\boldsymbol{\lambda} \in \mathbb{R}^d_+ : \|\boldsymbol{\lambda}\|_1 \leq 1/\nu\}$. We consider the regret of the dual in the set of rounds $\mathcal{T}_{\text{G},<\tau}$ and $\mathcal{T}_{\text{G},>\tau}$, with respect to the action set $\mathcal{D}$. Formally, $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D}) = \sup_{\boldsymbol{\lambda}\in\mathcal{D}} \sum_{t\in\mathcal{T}_{\text{G},<\tau}} u^{\text{D}}_t(\boldsymbol{\lambda}) - u^{\text{D}}_t(\boldsymbol{\lambda}_t)$. The term $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}$ is defined analogously. Finally, let $M := \max_t \|\boldsymbol{\lambda}_t\|_1$ be the largest value of the ℓ_1-norm of dual multipliers over the time horizon.

4 General guarantees of the primal-dual template

In this section, we provide no-regret guarantees of the general template described in Algorithm 1.

4.1 Adversarial setting

We start by describing the guarantees of Algorithm 1 in the adversarial setting. The idea is that, since τ corresponds to the last time in which at least one resource had $B_{t,i} < 1$, it must be the case that the primal "spent a lot" during the time intervals before τ in which it played (i.e., $\mathcal{T}_{\text{G},<\tau}$). Ideally, the dual regret minimizer should adapt to this behavior and play a large $\boldsymbol{\lambda}$ in $\mathcal{T}_{\text{G},<\tau}$, thereby attaining a large cumulative utility in $\mathcal{T}_{\text{G},<\tau}$. We show that the Lagrange multipliers in $\mathcal{D}$ are enough for this purpose. Then, as soon as we reach $t > \tau$, the dual should adapt and start to play a small $\boldsymbol{\lambda}$. Indeed, during these rounds the primal regret minimizer gains resources and therefore the optimal dual strategy would be setting $\boldsymbol{\lambda} = \mathbf{0}$. The effectiveness of the dual regret minimizer in understanding in which phase it is playing, and in adapting to it by setting high/small penalties, is measured by the size of the regret terms $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D})$ and $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}(\mathcal{D})$. In particular, $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D})$ (resp., $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}(\mathcal{D})$) is low if the dual regret minimizer behaves as expected before τ (resp., after τ). Intuitively, if $\mathcal{A}^{\text{D}}$ guarantees that those terms are small, then the dual regret minimizer is able to react quickly to the change in the behavior of the primal player before and after τ. As a byproduct of this, we show that the part of the primal's cumulative utility due to the Lagrangian penalties is sufficiently small. At the same time, we know that the primal regret minimizer has regret at most $\mathcal{R}^{\text{P}}_{\mathcal{T}_{\text{G}}}$ with respect to a strategy mixture that plays the optimal fixed unconstrained strategy with probability $\nu/(1+\beta)$, and the void action otherwise. This strategy mixture guarantees a $\nu/(1+\beta)$ fraction of the optimal utility without violating the constraints at any rounds. Formally, we can show the following.[3]

Theorem 4.1. *Let $\alpha := \nu/(1+\beta)$. In the adversarial setting, Algorithm 1 outputs a sequence of actions $(x_t)_{t=1}^T$ such that*

$$\sum_{t\in[T]} f_t(x_t) \geq \alpha \cdot \text{OPT}_{\boldsymbol{\gamma}} - \left(\frac{2}{\nu} + \mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D}) + \mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}(\mathcal{D}) + \mathcal{R}^{\text{P}}_{\mathcal{T}_{\text{G}}}\right).$$

Notice that, in the case of $\beta = 0$, we recover the standard guarantees of adversarial bandits with knapsacks for the case in which $B = \Omega(T)$ Castiglioni et al. (2022a), where the competitive ratio α is exactly ρ. The possibility of replenishing resources yields an improved competitive ratio.

Remark. In order for Theorem 4.1 to provide a meaningful bound, we need the three regret terms on the right-hand side to be suitably upperbounded by some term sublinear in T (see Section 5). Since the time steps in $\mathcal{T}_{\text{G}}$ are the only rounds in which $\mathcal{A}^{\text{P}}$ is invoked, any standard regret minimizer can be used to bound $\mathcal{R}^{\text{P}}_{\mathcal{T}_{\text{G}}}$. However, the same does not holds for $\mathcal{A}^{\text{D}}$. Indeed, we need a regret minimizer which can provide suitable regret upper bounds to $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D})$ and $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}(\mathcal{D})$, at the same time. One cannot simply bound $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},<\tau}}(\mathcal{D}) + \mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G},>\tau}}(\mathcal{D})$ by $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G}}}(\mathcal{D})$, since the best action in the sets $\mathcal{T}_{\text{G},<\tau}$ and $\mathcal{T}_{\text{G},>\tau}$ may differ. Standard regret minimizers usually do not provide this guarantee, so we need special care in choosing $\mathcal{A}^{\text{D}}$. In particular, we need a weakly adaptive dual regret minimizer.

[3] All omitted proofs can be found in the appendix.

4.2 STOCHASTIC SETTING

In this setting, we can exploit stochasticity of the environment to show that the expected utility of the primal under the sequence of realized inputs $\boldsymbol{\gamma} = (f_t, \boldsymbol{c}_t)_{t=1}^T$ is close to the primal expected utility at $(\bar{f}, \bar{\boldsymbol{c}})$, and to provide a suitable upperbound to the amount by which each resource is replenished during $\mathcal{T}_\varnothing$. Given $\delta \in (0, 1]$, let $\mathcal{E}_{T,\delta} := \sqrt{8T \log(4mT/\delta)}$.

Lemma 4.2. *For any $\xi \in \Xi$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that:*

$$\textstyle\sum_{t \in \mathcal{T}_\varnothing} c_{t,i}(\varnothing) \le -\beta |\mathcal{T}_\varnothing| + M\mathcal{E}_{T,\delta}, \ \forall i \in [m], \quad \text{and} \tag{4.1}$$

$$\mathbb{E}_{x\sim\xi} \left[\textstyle\sum_{t \in \mathcal{T}_G} f_t(x) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x)\rangle\right] \ge \mathop{\mathbb{E}}_{x\sim\xi} \left[\textstyle\sum_{t \in \mathcal{T}_G} \bar{f}(x) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \bar{\boldsymbol{c}}(x)\rangle\right] - M\mathcal{E}_{T,\delta}. \tag{4.2}$$

Then, we can prove the following regret bound.

Theorem 4.3. *Let the inputs (f_t, c_t) be i.i.d. samples from a fixed but unknown distribution $\mathcal{P}$. For $\delta \in (0, 1]$, we have that with probability at least $1 - \delta$, it holds*

$$\sum_{t=1}^{T} f_t(x_t) \ge T \cdot \texttt{OPT}^{\text{LP}}_{\bar{f},\bar{c}} - \left(\frac{2}{\nu} + \frac{1}{\nu}\mathcal{E}_{T,\delta} + \mathcal{R}^{\text{D}}_{\mathcal{T}_{G,<\tau}}(\mathcal{D}) + \mathcal{R}^{\text{D}}_{\mathcal{T}_{G,>\tau}}(\mathcal{D}) + \mathcal{R}^{\text{P}}_{\mathcal{T}_G}\right).$$

The proof follows a similar approach to the one of Theorem 4.1, with two main differences. First, we can now exploit standard concentration inequalities to relate realizations of random variables with their mean. Second, we exploit the fact that the primal has regret at most $\mathcal{R}^{\text{P}}_{\mathcal{T}_G}$ against an optimal solution to $\texttt{OPT}^{\text{LP}}_{\bar{f},\bar{\boldsymbol{c}}}$, which is feasible in expectation and has an expected utility that matches the value of our baseline. This allows us to obtain competitive-ratio equal to 1 in the stochastic setting.

5 CHOOSING APPROPRIATE REGRET MINIMIZERS

In order to have meaningful guarantees in both the adversarial and stochastic setting, we need to choose the regret minimizers $\mathcal{A}^{\text{P}}$ and $\mathcal{A}^{\text{D}}$ so that $\mathcal{R}^{\text{P}}_{\mathcal{T}_G}$, $\mathcal{R}^{\text{D}}_{\mathcal{T}_{G,<\tau}}(\mathcal{D})$ and $\mathcal{R}^{\text{D}}_{\mathcal{T}_{G,>\tau}}(\mathcal{D})$ all grow sublinearly in T. In the following section we will discuss two different scenarios, which differ in the amount of information which the decision maker is required to have. In Section 5.1 we are going to assume that the decision maker knows the per-round replenishment factor β. Then, in Section 5.2, we will show that this assumption can be removed by employing a primal regret minimizer $\mathcal{A}^{\text{P}}$ with slightly stronger regret guarantees. In both cases, the dual regret minimizer $\mathcal{A}^{\text{D}}$ has to be weakly adaptive, since both terms $\mathcal{R}^{\text{D}}_{\mathcal{T}_{G,<\tau}}(\mathcal{D})$ and $\mathcal{R}^{\text{D}}_{\mathcal{T}_{G,>\tau}}(\mathcal{D})$ need to be sublinear in T. On the other hand, we will make minimal assumptions on the primal regret minimizer $\mathcal{A}^{\text{P}}$. Its choice largely depends on the application considered, as we show in Section 6. In general, the primal regret minimizer must meet the minimal requirement of guaranteeing a sublinear regret upper bound $\mathcal{E}^{\text{P}}_{T,\delta}$ with probability at least $1 - \delta$, when the adversarial rewards are in $[0, 1]$.

5.1 IMPLEMENTING ALGORITHM 1 WITH KNOWN REPLENISHMENT FACTOR

We start by assuming that the decision maker knows β or, more generally, a lower bound $\tilde{\beta}$ on it. Therefore, the algorithm can compute ν, or its lower bound. When $\tilde{\beta}$ is known, we can instantiate the regret minimizer $\mathcal{A}^{\text{D}}$ to play on the set $\tilde{\mathcal{D}} := \{\boldsymbol{\lambda} \in \mathbb{R}^d_+ : \|\boldsymbol{\lambda}\|_1 \le 1/\tilde{\nu}\} \supseteq \mathcal{D}$, where $\tilde{\nu} := \tilde{\beta} + \rho \le \nu$.

As discussed above, the dual regret minimizer on the set of Lagrange multipliers $\mathcal{D}$ must be weakly adaptive. This can be achieved via variations of the *fixed share algorithm* proposed by Herbster and Warmuth (1998). We will employ the *generalized share algorithm* of Bousquet and Warmuth (2002) and the analysis of Cesa-Bianchi et al. (2012), as $\mathcal{A}^{\text{D}}$ has full feedback by construction.

The set $\tilde{\mathcal{D}}$ can be written as $\tilde{\mathcal{D}} = \text{co}\{\mathbf{0}, {}^1\!/\!_{\tilde{\nu}}\mathbf{1}_i \text{ with } i \in [m]\}$, where $\mathbf{1}_i$ is the i-th standard basis vector, and co denotes the convex hull. Since $\mathcal{D} \subseteq \tilde{\mathcal{D}}$, achieving no-weakly-adaptive regret with respect to $\tilde{\mathcal{D}}$ implies the same result for $\mathcal{D}$. Thus, instantiating the fixed-share algorithm on the $(m + 1)$-simplex, and since the losses of $\mathcal{A}^{\text{P}}$ are linear we can prove that:

Lemma 5.1. *(Cesa-Bianchi et al., 2012, Corollary 2) For any $0 < \tilde{\beta} \le \beta$, there exists an algorithm that guarantees* $\max\left(\mathcal{R}^{\text{D}}_{\mathcal{T}_{G,<\tau}}(\mathcal{D}), \mathcal{R}^{\text{D}}_{\mathcal{T}_{G,>\tau}}(\mathcal{D})\right) \le \frac{2}{\tilde{\nu}}\sqrt{T \log(2mT)}$.

Since the dual regret minimizer $\mathcal{A}^{\mathrm{D}}$ can play any Lagrange multiplier $\boldsymbol{\lambda}_t \in \tilde{\mathcal{D}}$, we have that the rewards observed by the primal regret minimizer $\mathcal{A}^{\mathrm{P}}$ are in the range $[0, 1 + 2/\tilde{\nu}]$ because it holds

$$\sup_{x\in\mathcal{X}, t\in[T]} |u_t^{\mathrm{P}}(x)| \leq \sup_{x\in\mathcal{X}, t\in[T]} \left\{ |f_t(x)| + \|\boldsymbol{\lambda}_t\|_1 \cdot \|\boldsymbol{\rho} - \boldsymbol{c}_t(x)\|_\infty \right\} \leq 1 + \frac{2}{\tilde{\nu}} \leq \frac{4}{\tilde{\nu}}.$$

With probability at least δ, the primal regret minimizer $\mathcal{A}^{\mathrm{P}}$ guarantees a regret $\mathcal{E}^{\mathrm{P}}_{T,\delta}$ against rewards in $[0, 1]$. Then, by re-scaling the realized rewards before giving them in input to the regret minimizer, we get a regret bound of $\frac{4}{\tilde{\nu}}\mathcal{E}^{\mathrm{P}}_T$ against rewards $u_t^{\mathrm{P}}(\cdot)$ that are in $[0, 4/\tilde{\nu}]$. This simple construction is applicable because the range of the rewards is known. By combining these observations we can easily recover the following corollary of Theorem 4.1 and Theorem 4.3.

Corollary 5.2. *Assume that the dual regret minimizer is generalized fixed share on $\tilde{\mathcal{D}}$, and that the primal regret minimizer has regret at most $\mathcal{E}^{P}_{\mathcal{T}_G,\delta}$ against losses in $[0, 1]$ with probability at least $1 - \delta$, for $\delta \in (0, 1]$. In the adversarial setting, for any $\tilde{\beta} \leq \beta$, with probability at least $1 - \delta$ Algorithm 1 guarantees that*

$$\sum_{t\in[T]} f_t(x_t) \geq \alpha OPT_{\boldsymbol{\gamma}} - \left(\frac{2}{\nu} + \frac{1}{\tilde{\nu}}\sqrt{T \log(2mT)} + \frac{4}{\tilde{\nu}}\mathcal{E}^{P}_{\mathcal{T}_G,\delta} \right),$$

where $\alpha = \nu/1+\beta$. In the stochastic setting, with probability at least $1 - 2\delta$, Algorithm 1 guarantees

$$\sum_{t\in[T]} f_t(x_t) \geq T \cdot \mathrm{OPT}^{LP}_{\bar{f},\bar{c}} - \left(\frac{2}{\nu} + \frac{1}{\nu}\mathcal{E}_{T,\delta} + \frac{1}{\tilde{\nu}}\sqrt{T\log(2mT)} + \frac{4}{\tilde{\nu}}\mathcal{E}^{P}_{\mathcal{T}_G,\delta} \right).$$

5.2 Implementing Algorithm 1 with unknown replenishment factor

In this section we will show how to implement Algorithm 1 when no information about the per-round replenishment factor β is available. This impacts both the primal and the dual regret minimizer. Differently from the previous section, we cannot instantiate the regret minimizer $\mathcal{A}^{\mathrm{D}}$ directly on $\mathcal{D}$ (or on a larger set $\tilde{\mathcal{D}}$) since we do not know such set. Therefore, we need a dual regret minimizer that plays on $\mathbb{R}^m_+$, but has sublinear weakly-adaptive regret with respect to Lagrange multipliers in $\mathcal{D}$. To achieve this, we can use the Online Gradient Descent (OGD) algorithm, instantiated on $\mathbb{R}^m_+$ with starting point $\boldsymbol{\lambda}_0 = \mathbf{0}$. Indeed, it is well known that OGD guarantees that the regret on any interval of rounds $[t_1, t_2] \subseteq [T]$ is upper bounded by the ℓ_2 distance between the point played at t_1 and the comparator. In our setting this is equivalent to the following lemma (Hazan et al., 2016, Chapter 10).

Lemma 5.3. *For any $\mathcal{T}_G \subset [T]$ and any $t_1, t_2 \in \mathcal{T}_G$, if the dual regret minimizer is OGD with learning rate η, we have that the regret with respect to $\boldsymbol{\lambda}$ is upper bounded by*

$$\mathcal{R}^{D}_{\mathcal{T}_G\cap[t_1,\ldots,t_2]}(\{\boldsymbol{\lambda}\}) \leq \frac{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t_1}\|_2^2}{2\eta} + \frac{1}{2}\eta\, m\, T.$$

In this setting, we also need additional assumptions on the primal regret minimizer $\mathcal{A}^{\mathrm{P}}$. Formally we need an algorithm that satisfies the following condition.

Condition 5.4. *We assume that, for any $\mathcal{T} = [t_1, t_2] \subseteq [T]$, the weakly-adaptive regret of $\mathcal{A}^{P}$ facing adversarial losses with unknown range L is upper bounded by $L^2\mathcal{E}^{P}_{T,\delta}$ with probability at least $1 - \delta$, where $\mathcal{E}^{P}_{T,\delta}$ is independent from the range of payoffs.*

In order to provide the final regret bound for this setting, we need to show that the size of Lagrange multipliers remains bounded by a suitable term.

Lemma 5.5. *Assume that the dual regret minimizer is OGD on $\mathbb{R}^m_+$ with $\eta = (k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{P}_{T,\delta} + 2m\sqrt{T})^{-1}$, where k_1, k_2 are absolute constants, and the primal regret minimizer $\mathcal{A}^{P}$ satisfies Condition 5.4. Then, both in the adversarial and stochastic setting, the Lagrange multipliers $\boldsymbol{\lambda}_t$ played by the dual regret minimizer $\mathcal{A}^{D}$ are such that $M := \sup_{t\in[T]} \|\boldsymbol{\lambda}_t\|_1 \leq 8m/\nu$.*

This result extends the similar result of Castiglioni et al. (2023, Theorem 6.2) to the case of multiple constraints. Lemma 5.5 allows us bound the regret $\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_G}$ of the primal, and the regret terms $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_G<\tau}(\mathcal{D})$ and $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_G>\tau}(\mathcal{D})$ of the dual regret minimizer. In particular, for what concerns the dual regret minimizer

$\mathcal{A}^{\text{D}}$, we can bound the maximum distance between Lagrange multipliers which are played by $\mathcal{A}^{\text{D}}$. Then, we can bound the regret terms $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G}}<\tau}(\mathcal{D})$ and $\mathcal{R}^{\text{D}}_{\mathcal{T}_{\text{G}}>\tau}(\mathcal{D})$ via Lemma 5.3. Similarly, we can bound the regret of the primal through Condition 5.4. Using this observations, together with Theorem 4.1 and Theorem 4.3, we can extend Corollary 5.2 to the case of unknown β.

Corollary 5.6. *Assume that the dual regret minimizer is OGD on $\mathbb{R}^m_+$ with $\eta = (k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{\text{P}}_{T,\delta} + 2m\sqrt{T})^{-1}$, where k_1, k_2 are absolute constants, and the primal regret minimizer $\mathcal{A}^{\text{P}}$ satisfies Condition 5.4. Then, in the adversarial setting, Algorithm 1 guarantees with probability $1-\delta$ that*

$$\sum_{t\in[T]} f_t(x_t) \geq \alpha \cdot \texttt{OPT}_{\boldsymbol{\gamma}} - k_3\frac{m^4}{\nu^2}\left(\mathcal{E}^{\text{P}}_{T,\delta} + \mathcal{E}_{T,\delta}\right),$$

where k_3 is an absolute constant and $\alpha = \nu/1+\beta$. In the stochastic setting, there is an absolute constant k_4 such that Algorithm 1 guarantees that, with probability at least $1-2\delta$,

$$\sum_{t\in[T]} f_t(x_t) \geq T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{c}} - k_4\frac{m^4}{\nu^2}\left(\mathcal{E}^{\text{P}}_{T,\delta} + \mathcal{E}_{T,\delta}\right).$$

6 APPLICATIONS

This section demonstrates the flexibility of our framework by studying three well motivated models: BwRK, inventory management, and revenue maximization in bilateral trade.

6.1 BANDITS WITH REPLENISHABLE KNAPSACKS (BWRK)

Consider the standard BwK problem: at each time step t, the learner selects an action i_t out of K actions (thus $\mathcal{X} = [K]$), suffers a loss $\ell_t(i_t) \in [0,1]$ and incurs a cost vector $\boldsymbol{c}_t(i_t) \in [-1,1]^m$ that specifies the consumption of each one of its m resources. We focus on the most challenging scenario, in which the parameter β is not known. We instantiate the primal-dual framework (Algorithm 1) using EXP3-SIX Neu (2015) as the primal regret minimizer, while online gradient descent is employed as the dual regret minimizer. Castiglioni et al. (2023) show that EXP3-SIX achieve weakly-adaptive regret $L^2\tilde{\mathcal{O}}(\sqrt{KT})$, where L is the range of the observed losses.

First, let's consider the adversarial case, where losses and cost functions are generated by an oblivious adversary. Assumption 2.1 is verified when there exists a null action $\varnothing$ that always yields non-negative resource replenishment (*i.e.,* there exists $\beta \geq 0$ s.t. $c_{t,j}(\varnothing) \leq -\beta$ for each resource j and time t).

Theorem 6.1. *Consider the BwRK problem in the adversarial setting. There exists an algorithm satisfying the following bound on the regret:*

$$\alpha \cdot \texttt{OPT}_{\boldsymbol{\gamma}} - \sum_{t\in[T]} f_t(x_t) \leq \tilde{\mathcal{O}}\left(\frac{m^4}{\nu^2}\sqrt{KT}\log\left(\frac{1}{\delta}\right)\right),$$

with probability at least $1-\delta$, where $\alpha := \nu/(1+\beta)$.

This is the first positive result for the BwRK problem in the adversarial setting.

Second, we consider a stochastic version, in which losses and cost vectors are drawn i.i.d. from an unknown distribution. In this setting, the void action assumption (Assumption 2.3) requires the existence of a distribution $\varnothing \in \Delta_K$ over the actions such that, in expectation over the draws of the cost function, it verifies $\mathbb{E}[c_j(\varnothing)] \leq -\beta$ for all resources j[4]. The analysis of our primal-dual framework allows us to show that the same learning algorithm presented for the adversarial setting yields the following instance-independent results in the stochastic setting.

Theorem 6.2. *Consider the BwRK in the stochastic setting. There exists an algorithm satisfying, with probability at least $1-2\delta$, the following bound on the regret:*

$$T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{c}} - \sum_{t=1}^{T} f_t(x_t) \leq \tilde{\mathcal{O}}\left(\frac{m^4}{\nu^2}\sqrt{KT}\log\left(\frac{1}{\delta}\right)\right).$$

[4]Note, the role of β is played in Kumar and Kleinberg (2022) by the parameter δ_{slack}. Furthermore, there the authors make additional assumptions on the structure of the optimal LP and its solution.

Theorem 6.2 provides the first instance-independent regret bound under i.i.d. inputs, and it complements the instance-dependent analysis by Kumar and Kleinberg (2022). In particular, in the stochastic setting, they provide a logarithmic instance-dependent bound on the regret of order $O(Km^2/\Delta^2 \log T)$, where Δ is a notion of suboptimality gap (already present in Li et al. (2021)), which in principle may be arbitrarily small. Although our work does not offer instance-dependent logarithmic bounds on the regret, we advance the study of the BwRK problem along three main directions: i) we design an algorithm which handles at the same time both stochastic and adversarial inputs (while Kumar and Kleinberg (2022) only deal with the stochastic input model), ii) our primal-dual approach is arguably simpler, and iii) we provide the first worst-case dependence on the time horizon T.

6.2 Economic applications and discussion

In the following, we describe two additional economic applications where our framework can be applied in the stochastic setting.

Inventory management. As a simple application of our techniques, consider the following inventory management problem. Each day t, a shopkeeper is confronted with a decision: either open for business and attempt to sell the goods it has in the store, or travel to its supplier in order to restore its inventory. This simplified model captures many real world application that contemplates the trade-off between exploiting the available inventory, and "skipping a turn" to replenish it. For the sake of clarity we restrict ourselves to the case of a single resource and supplier, however this could be easily extended to more general instances. Formally, the goods in stock at day t are $B_t \geq 0$, and the shopkeeper has a set two actions $\mathcal{X} = \{o, s\}$. Action o corresponds to opening for business with reward $r_t(o) \in [0,1]$ and inventory consumption $c_t(o) \in [0,1]$, while action s corresponds to going to the supplier, with $r_t(s) \in [-1,0]$ and negative resource consumption $c_t(s) \in [-1,0]$ which both depend on the supplier's availability and current price. Clearly, it is possible to select action o only if $B_t \geq 1$, while action s plays the role of the void action $\varnothing$. We focus of the stochastic case in which $\mathbb{E}[c_t(s)] \leq -\beta$, where β is the expected amount of good available from the supplier. By employing EXP-SIX as in Section 6.1, this immediately yields a $\tilde{O}(\sqrt{T})$ regret bound via Corollary 5.6.

Bilateral trade. We consider the well known bilateral trade model (Vickrey, 1961; Myerson and Satterthwaite, 1983; Cesa-Bianchi et al., 2021): each day, the merchant posts two prices, a price $p_t \in \mathbb{R}_+$ and a price $q_t \in \mathbb{R}_+$ to a seller and a buyer of a good, respectively. The seller (resp., the buyer) has a private valuation of $s_t \in [0,1]$ (resp., $b_t \in [0,1]$). If the valuation s_t is smaller than p_t, then the merchant buys one unit of good from the seller, while if the valuation b_t is larger than the price q_t (and there is some inventory left) then the merchant sells a unit of the good to the buyer. The merchant can only sell a good if $B_t \geq 1$, *i.e.,* the inventory has at least one unit of it. Moreover, we assume that the merchant has no initial budget, *i.e.,* $B_1 = 0$. The merchant's revenue at time t is $\texttt{rev}_t(p_t, q_t) \coloneqq (q_t - p_t)\mathbb{1}[q_t \leq b_t]\mathbb{1}[s_t \leq p_t]$. Similarly, for any $t \in [T]$, the stock consumption is updated as follows: $B_{t+1} = B_t - \mathbb{1}[q_t \leq b_t] + \mathbb{1}[s_t \leq p_t]$. Therefore, the strategy of choosing $p_t = 1$ and $q_t > 1$ surely increases the budget by one unit, and it has revenue $\texttt{rev}_t(p_t, q_t) = -1$. This strategy plays the role of the void action $\varnothing$, with per-round replenishment $\beta = 1$. Similarly to the example above, in the stochastic setting it is enough to translate and re-scale the revenue to satisfy all the assumptions of our algorithm. This immediately yields the $\tilde{O}(\sqrt{T})$ regret guarantees. The adversarial case was addressed in a recent paper by Bernasconi et al. (2024) using techniques similar to the ones developed in this paper. When considering the same baseline of our paper, they achieve a constant competitive ratio of 2 through an ad hoc analysis.

Challenges of the adversarial setting. In the case in which rewards are allowed to be negative and inputs are adversarial, the simple trick of re-scaling and translating the utilities is not applicable. This is due to the fact that our algorithm guarantees a multiplicative approximation of the benchmark and this multiplicative factor is *not* invariant with respect to translations. We leave as an open problem the question of how to handle application scenarios in which rewards may be negative in the adversarial setting. We observe that the results for the adversarial setting would be preserved if we could provide a void action $\varnothing$ that has non-negative reward and negative cost, but this is not usually the case in practical applications. Finally, it would be interesting to extend the simplified inventory-management model which we discussed to more complex models which have been proposed in the operations research literature (see, *e.g.,* Chen et al. (2022); Chen and Simchi-Levi (2004)).

ACKNOWLEDGMENTS

MB, MC, AC, FF are partially supported by the FAIR (Future Artificial Intelligence Research) project PE0000013, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence). FF is also partially supported by ERC Advanced Grant 788893 AMDROMA "Algorithmic and Mechanism Design Research in Online Markets", and PNRR MUR project IR0000013-SoBigData.it. MC is also partially supported by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237). AC is partially supported by MUR - PRIN 2022 project 2022R45NBB funded by the NextGenerationEU program.

REFERENCES

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

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

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 74–91, 2012.

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

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), 2018.

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

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 2022.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. No-regret learning in bilateral trade via global budget balance. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024.

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

Olivier Bousquet and Manfred K Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3(Nov):363–396, 2002.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022a.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. In *Advances in Neural Information Processing Systems*, volume 35, pages 33589–33602, 2022b.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online bidding in repeated non-truthful auctions under budget and roi constraints. *arXiv preprint arXiv:2302.01203*, 2023.

Andrea Celli, Matteo Castiglioni, and Christian Kroer. Best of many worlds guarantees for online learning with knapsacks. *arXiv preprint arXiv:2202.13710*, 2023.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, page 980–988, 2012.

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 289–309, 2021.

Boxiao Chen, David Simchi-Levi, Yining Wang, and Yuan Zhou. Dynamic pricing and inventory control with fixed ordering cost and incomplete demand information. *Management Science*, 68(8): 5684–5703, 2022.

Xin Chen and David Simchi-Levi. Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. *Operations research*, 52(6):887–896, 2004.

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

Giannis Fikioris and Éva Tardos. Approximately stationary bandits with knapsacks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3758–3782. PMLR, 2023.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Elad Hazan et al. *Introduction to online convex optimization*, volume 2. Now Publishers, Inc., 2016.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), 2022. ISSN 0004-5411.

Raunak Kumar and Robert Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 6483–6492. PMLR, 2021.

Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pages 1760–1770. PMLR, 2018.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4633–4656. PMLR, 2023.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

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

Xiaohan Wei, Hao Yu, and Michael J Neely. Online primal-dual mirror descent under stochastic constraints. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2): 1–36, 2020.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

A PROOFS OMITTED FROM SECTION 4.1

Theorem 4.1. *Let $\alpha := \nu/(1+\beta)$. In the adversarial setting, Algorithm 1 outputs a sequence of actions $(x_t)_{t=1}^T$ such that*

$$\sum_{t\in[T]} f_t(x_t) \geq \alpha \cdot \texttt{OPT}_{\boldsymbol{\gamma}} - \left(\frac{2}{\nu} + \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}(\mathcal{D}) + \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},>\tau}}(\mathcal{D}) + \mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\right).$$

Proof. We divided the proof in two steps. First, we show that over the rounds $\mathcal{T}_{\mathrm{G}}$ in which the regret minimizers play the "lagrangified" cumulative costs is controlled. Then, in the second step, we show that over the same set of rounds $\mathcal{T}_{\mathrm{G}}$ the "lagrangified" utility is large. In order to simplify the notation, we will write $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}$ (resp., $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},>\tau}}$) in place of $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}(\mathcal{D})$ (resp., $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},>\tau}}(\mathcal{D})$).

Part 1. First, we show that in $\mathcal{T}_{\mathrm{G},<\tau}$ the budget spent by the decision maker is sufficiently high. Let i^* be the resource that had the lowest budget in round $\tau \in \mathcal{T}_\varnothing$.

Summing the budget equation $B_{t+1,i^*} = B_{t,i^*} - c_{t,i^*}(x_t)$ over $t = \{1, \ldots, \tau-1\}$, gives us

$$B_{\tau,i^*} - B_{i^*} = -\sum_{t\leq\tau-1} c_{t,i^*}(x_t).$$

Since by definition $B_{\tau,i^*} < 1$ and $B_{1,i^*} = B$ we get that:

$$\sum_{t\leq\tau-1} c_{t,i^*}(x_t) > B - 1. \tag{A.1}$$

Moreover, in rounds $t \in \mathcal{T}_\varnothing$, since the budget was strictly less we have by construction that $x_t = \varnothing$, and thus $c_{t,i^*}(x_t) \leq -\beta$. This readily implies that:

$$\sum_{t\in\mathcal{T}_\varnothing} c_{t,i^*}(x_t) \leq -\beta|\mathcal{T}_\varnothing|. \tag{A.2}$$

By combining Equation (A.1) and Equation (A.2) we obtain that:

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathrm{G},<\tau}} c_{t,i^*}(x_t) &= \sum_{t\in[\tau]} c_{t,i^*}(x_t) - \sum_{t\in\mathcal{T}_\varnothing} c_{t,i^*}(x_t)\\ &\geq B - 2 + \beta|\mathcal{T}_\varnothing|.\end{aligned}$$

Since the budget spent in rounds $\mathcal{T}_{\mathrm{G},<\tau}$ is large, it must be the case that the dual regret minimizer $\mathcal{A}^{\mathrm{D}}$ collects a large cumulative utility in those rounds. Formally, the regret $\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}$ of $\mathcal{A}^{\mathrm{D}}$ on rounds $\mathcal{T}_{\mathrm{G},<\tau}$ with respect to $\boldsymbol{\lambda} = \frac{1}{\nu}\mathbf{1}_{i^*}$ reads as follows

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathrm{G},<\tau}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho}\rangle &\geq \frac{1}{\nu}\sum_{t\in\mathcal{T}_{\mathrm{G},<\tau}} (c_{t,i^*}(x_t) - \boldsymbol{\rho}_{i^*}) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&\geq \frac{1}{\nu}(B - 2 + \beta|\mathcal{T}_\varnothing| - |\mathcal{T}_{\mathrm{G},<\tau}|\rho_{i^*}) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&= \frac{1}{\nu}(\rho_{i^*}T - 2 + \beta|\mathcal{T}_\varnothing| - |\mathcal{T}_{\mathrm{G},<\tau}|\rho_{i^*}) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&= \frac{1}{\nu}(\rho_{i^*}(|\mathcal{T}_{\mathrm{G},<\tau}| + |\mathcal{T}_\varnothing| + |\mathcal{T}_{\mathrm{G},>\tau}|) - 2 + \beta|\mathcal{T}_\varnothing| - |\mathcal{T}_{\mathrm{G},<\tau}|\rho_{i^*}) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&= \frac{1}{\nu}(\nu|\mathcal{T}_\varnothing| + \rho_{i^*}|\mathcal{T}_{\mathrm{G},>\tau}| - 2) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&\geq \frac{1}{\nu}(\nu|\mathcal{T}_\varnothing| - 2) - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}\\
&= |\mathcal{T}_\varnothing| - \frac{2}{\nu} - \mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}.\end{aligned} \tag{A.3}$$

Then, by considering $\boldsymbol{\lambda} = \mathbf{0}$ and the definition of the regret on $\mathcal{T}_{\mathrm{G},>\tau}$ we get that the utility of dual is bounded by:

$$\sum_{t\in\mathcal{T}_{\mathrm{G},>\tau}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho}\rangle \geq -\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},>\tau}}. \tag{A.4}$$

Thus, by combining the inequalities of Equation (A.3) and Equation (A.4) we can conclude that:

$$\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{c}_t(x_t)-\boldsymbol{\rho}\rangle\geq|\mathcal{T}_\varnothing|-\frac{2}{\nu}-\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},<\tau}}-\mathcal{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G},>\tau}}. \tag{A.5}$$

Part 2. Now, let x^* be the best unconstrained strategy, *i.e.*, $x^*\in\arg\max_{x\in\mathcal{X}}\sum_{t\in[T]}f_t(x)$ and thus $\sum_{t\leq T}f_t(x^*):=\texttt{OPT}_{\boldsymbol{\gamma}}$.[5] Consider the mixed strategy $\xi^*\in\Xi$ that randomizes between x^* and $\varnothing$, with probability $\frac{\rho+\beta}{1+\beta}$ and $\frac{1-\rho}{1+\beta}$, respetively. Then $\mathbb{E}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x)\rangle\right]\geq 0$. This can be easily proved via the following chain of inequalities:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x)\rangle\right]&=\frac{\rho+\beta}{1+\beta}\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x^*)\rangle+\frac{1-\rho}{1+\beta}\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(\varnothing)\rangle\\
&=\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}\rangle-\frac{\rho+\beta}{1+\beta}\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{c}_t(x^*)\rangle-\frac{1-\rho}{1+\beta}\sum_{t\in\mathcal{T}_{\mathrm{G}}}\langle\boldsymbol{\lambda}_t,\boldsymbol{c}_t(\varnothing)\rangle\\
&\geq\rho\sum_{t\in\mathcal{T}_{\mathrm{G}}}\|\boldsymbol{\lambda}_t\|_1-\frac{\rho+\beta}{1+\beta}\sum_{t\in\mathcal{T}_{\mathrm{G}}}\|\boldsymbol{\lambda}_t\|_1+\frac{1-\rho}{1+\beta}\beta\sum_{t\in\mathcal{T}_{\mathrm{G}}}\|\boldsymbol{\lambda}_t\|_1\\
&=\left(\rho-\frac{\rho+\beta}{1+\beta}+\beta\frac{1-\rho}{1+\beta}\right)\sum_{t\in\mathcal{T}_{\mathrm{G}}}\|\boldsymbol{\lambda}_t\|_1\\
&=0.\end{aligned}\tag{A.6}$$

Then, we can use the defintion of the regret of the primal regret minimizer to find that:

$$\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}}f_t(x)+\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x)\rangle\right]-\sum_{t\in\mathcal{T}_{\mathrm{G}}}\left(f_t(x_t)+\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x_t)\rangle\right)\leq\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}},$$

and by rearranging and using the inequality of Equation (A.6) we have that:

$$\sum_{t\in\mathcal{T}_{\mathrm{G}}}f_t(x_t)+\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x_t)\rangle\geq\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}}f_t(x)\right]-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}.$$

By building upon this inequality we obtain the following:

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathrm{G}}}f_t(x_t)+\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x_t)\rangle&\geq\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}}f_t(x)\right]-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\\
&=\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\leq T}f_t(x)-\sum_{t\in\mathcal{T}_\varnothing}f_t(x)\right]-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\\
&\geq\mathop{\mathbb{E}}_{x\sim\xi^*}\left[\sum_{t\leq T}f_t(x)\right]-|\mathcal{T}_\varnothing|-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\\
&=\frac{\rho+\beta}{1+\beta}\sum_{t\leq T}f_t(x^*)+\frac{1-\rho}{1+\beta}\sum_{t\leq T}f_t(\varnothing)-|\mathcal{T}_\varnothing|-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\\
&\geq\frac{\rho+\beta}{1+\beta}\sum_{t\leq T}f_t(x^*)-|\mathcal{T}_\varnothing|-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}\\
&=\frac{\rho+\beta}{1+\beta}\texttt{OPT}_{\boldsymbol{\gamma}}-|\mathcal{T}_\varnothing|-\mathcal{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}}.\end{aligned}\tag{A.7}$$

[5]For simplicity we replaced the supremum with the maximum. Results would continue to hold with sup with minor modifications.

Concluding. Using the inequality of Equation (A.4) and the inequality of Equation (A.7) we have

$$\begin{aligned}
\sum_{t\le T} f_t(x_t) &\ge \sum_{t\in\mathcal{T}_{\mathrm{G}}} f_t(x_t) \\
&\ge \sum_{t\in\mathcal{T}_{\mathrm{G}}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(x_t)-\boldsymbol{\rho}\rangle + \frac{\rho+\beta}{1+\beta}\mathtt{OPT}_{\boldsymbol{\gamma}} - |\mathcal{T}_{\varnothing}| - \mathfrak{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}} \\
&\ge |\mathcal{T}_{\varnothing}| - \frac{2}{\nu} - \mathfrak{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G}},<\tau} - \mathfrak{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G}},>\tau} + \frac{\rho+\beta}{1+\beta}\mathtt{OPT}_{\boldsymbol{\gamma}} - |\mathcal{T}_{\varnothing}| - \mathfrak{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}} \\
&= \frac{\rho+\beta}{1+\beta}\mathtt{OPT}_{\boldsymbol{\gamma}} - \frac{2}{\nu} - \mathfrak{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G}},<\tau} - \mathfrak{R}^{\mathrm{D}}_{\mathcal{T}_{\mathrm{G}},>\tau} - \mathfrak{R}^{\mathrm{P}}_{\mathcal{T}_{\mathrm{G}}},
\end{aligned}$$

which concludes the proof. □

B PROOFS OMITTED FROM SECTION 4.2

Lemma 4.2. *For any $\xi \in \Xi$ and $\delta \in (0,1]$, with probability at least $1-\delta$, it holds that:*

$$\textstyle\sum_{t\in\mathcal{T}_{\varnothing}} c_{t,i}(\varnothing) \le -\beta|\mathcal{T}_{\varnothing}| + M\mathcal{E}_{T,\delta},\ \forall i\in[m], \quad \text{and} \tag{4.1}$$

$$\mathbb{E}_{x\sim\xi}\left[\textstyle\sum_{t\in\mathcal{T}_{G}} f_t(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\boldsymbol{c}_t(x)\rangle\right] \ge \mathop{\mathbb{E}}_{x\sim\xi}\left[\textstyle\sum_{t\in\mathcal{T}_{G}} \bar{f}(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\bar{\boldsymbol{c}}(x)\rangle\right] - M\mathcal{E}_{T,\delta}. \tag{4.2}$$

Proof. We start by proving that the inequality of Equation (4.1) holds with probability $1-\delta/2$. Let $K = |\mathcal{T}_{\varnothing}|$. Then, we can easily see that, for $i\in[m]$, with probability $1 - {}^{\delta}/_{(2mT)}$

$$\sum_{k\in[K]} (c_{k,i}(\varnothing) - \bar{c}_i(\varnothing)) \le \sqrt{2K\log\left(\frac{4mT}{\delta}\right)} \le \sqrt{2T\log\left(\frac{4mT}{\delta}\right)},$$

where the first inequality holds by Hoeffding's bound. By taking a union bound over all possible lengths of $\mathcal{T}_{\varnothing}$ (which are T) and $i\in[m]$, we obtain that for all possible sets $\mathcal{T}_{\varnothing}$ it holds:

$$\sum_{t\in\mathcal{T}_{\varnothing}} (c_{t,i}(\varnothing) - \bar{c}_i(\varnothing)) \le \sqrt{2T\log\left(\frac{4mT}{\delta}\right)}$$

with probability at least $1-\delta/2$. Then the proof of the inequality of Equation (4.1) is concluded by observing that:

$$\sum_{t\in\mathcal{T}_{\varnothing}} \bar{c}_i(\varnothing) \le -\beta|\mathcal{T}_{\varnothing}|,$$

and that $\sqrt{2T\log({}^{4mT}/_{\delta})} \le M\mathcal{E}_{T,\delta}$. Equation (4.2) can be proved in a similar way. Indeed, for any fixed $\mathcal{T}_{\mathrm{G}}$ of size $K = |\mathcal{T}_{\mathrm{G}}|$, and for any strategy mixture ξ, by Hoeffding we have that with probability at least $1 - {}^{\delta}/_{2T}$ the following holds

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\xi}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}} f_t(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\boldsymbol{c}_t(x)\rangle\right] &\ge \mathop{\mathbb{E}}_{x\sim\xi}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}} \bar{f}(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\bar{\boldsymbol{c}}(x)\rangle\right] - (1+2M)\sqrt{\frac{K}{2}\log\left(\frac{4T}{\delta}\right)} \\
&\ge \mathop{\mathbb{E}}_{x\sim\xi}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}} \bar{f}(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\bar{\boldsymbol{c}}(x)\rangle\right] - (1+2M)\sqrt{\frac{T}{2}\log\left(\frac{4T}{\delta}\right)} \\
&\ge \mathop{\mathbb{E}}_{x\sim\xi}\left[\sum_{t\in\mathcal{T}_{\mathrm{G}}} \bar{f}(x) + \langle\boldsymbol{\lambda}_t, \boldsymbol{\rho}-\bar{\boldsymbol{c}}(x)\rangle\right] - M\mathcal{E}_{T,\delta}.
\end{aligned}$$

By taking a union bound over all possible T lengths of $\mathcal{T}_{\mathrm{G}}$, we obtain that for all possible sets $\mathcal{T}_{\mathrm{G}}$, the equation above holds with probability $1-\delta/2$.

The Lemma follows by a union bound on the two equations above, which hold separately with probability $1-\delta/2$. □

Theorem 4.3. *Let the inputs (f_t, c_t) be i.i.d. samples from a fixed but unknown distribution $\mathcal{P}$. For $\delta \in (0, 1]$, we have that with probability at least $1-\delta$, it holds*

$$\sum_{t=1}^{T} f_t(x_t) \geq T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{c}} - \left(\frac{2}{\nu} + \frac{1}{\nu}\mathcal{E}_{T,\delta} + \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau}(\mathcal{D}) + \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},>\tau}(\mathcal{D}) + \mathcal{R}^{\texttt{P}}_{\mathcal{T}_{\texttt{G}}} \right).$$

Proof. The proof follows a similar approach to Theorem 4.1 with extra details regarding concentration inequalities to exploit stochasticity of the environment. We sketch here the proof for the sake of clarity. In order to simplify the notation of the proof, we will write $\mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau}$ (resp., $\mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},>\tau}$) in place of $\mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau}(\mathcal{D})$ (resp., $\mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},>\tau}(\mathcal{D})$).

Part 1. In the same fashion as the proof of Theorem 4.1, we can define i^* to be the index of the resource that had budget less then 1 at round τ, and show that

$$\sum_{t\leq\tau-1} c_{t,i^*}(x_t) > B - 1, \tag{B.1}$$

which is exactly Equation (A.1) from Theorem 4.1. Moreover, in rounds $t \in \mathcal{T}_{\varnothing} \subset [\tau]$ we have that $x_t = \varnothing$, and by Equation (4.1) we obtain

$$\sum_{t\in\mathcal{T}_{\varnothing}} c_{t,i^*}(\varnothing) \leq -\beta|\mathcal{T}_{\varnothing}| + \mathcal{E}_{T,\delta}, \tag{B.2}$$

which follows Equation (A.1) from Theorem 4.1, with the addition of the $\mathcal{E}_{T,\delta}$ term (see definition in Section 4.2), due to the concentration inequality. Then, since $[\tau] = \mathcal{T}_{\varnothing} \cup \mathcal{T}_{\texttt{G},<\tau}$, by combining Equation (B.1) and Equation (B.2) we can conclude that:

$$\sum_{t\in\mathcal{T}_{\texttt{G},<\tau}} c_{t,i^*}(x_t) = \sum_{t\leq\tau} c_{t,i^*}(x_t) - \sum_{t\in\mathcal{T}_{\varnothing}} c_{t,i^*}(x_t) \geq B - 2 + \beta|\mathcal{T}_{\varnothing}| - \mathcal{E}_{T,\delta}.$$

Then, through a chain of inequalities similar to the one of Equation (A.3) we can conclude that

$$\sum_{t\in\mathcal{T}_{\texttt{G},<\tau}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho} \rangle \geq |\mathcal{T}_{\varnothing}| - \frac{2}{\nu} - \frac{1}{\nu}\mathcal{E}_{T,\delta} - \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau}.$$

The same arguments used for Equation A.4 readily imply that

$$\sum_{t\in\mathcal{T}_{\texttt{G}}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(x_t) - \boldsymbol{\rho} \rangle \geq |\mathcal{T}_{\varnothing}| - \frac{2}{\nu} - \frac{1}{\nu}\mathcal{E}_{T,\delta} - \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau} - \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},>\tau}. \tag{B.3}$$

Part 2. Define ξ^* to be the optimal stochastic policy that achieves $\texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{\boldsymbol{c}}} = \sup_{\xi\in\Xi} \mathbb{E}_{x\sim\xi}[\bar{f}(x)]$ and $\mathbb{E}_{x\sim\xi}[\boldsymbol{c}(x)] \preceq \boldsymbol{\rho}$. By the definition of regret of $\mathcal{A}^{\texttt{P}}$ on the rounds $\mathcal{T}_{\texttt{G}}$ with respect to ξ^* the following holds

$$\begin{aligned}
\sum_{t\in\mathcal{T}_{\texttt{G}}} f_t(x_t) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x_t) \rangle &\geq \mathbb{E}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\texttt{G}}} f_t(x) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x) \rangle\right] - \mathcal{R}^{\texttt{P}}_{\mathcal{T}_{\texttt{G}}} \\
&\geq \mathbb{E}_{x\sim\xi^*}\left[\sum_{t\in\mathcal{T}_{\texttt{G}}} \bar{f}(x) + \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \bar{\boldsymbol{c}}(x) \rangle\right] - \mathcal{R}^{\texttt{P}}_{\mathcal{T}_{\texttt{G}}} - \mathcal{E}_{T,\delta} \\
&\geq T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{\boldsymbol{c}}} - |\mathcal{T}_{\varnothing}| - \mathcal{R}^{\texttt{P}}_{\mathcal{T}_{\texttt{G}}} - \mathcal{E}_{T,\delta},
\end{aligned} \tag{B.4}$$

where we used Equation (4.1) in the first inequality, and $[T] = \mathcal{T}_{\texttt{G}} \cup \mathcal{T}_{\varnothing}$ in the third one.

Concluding. By combining Equation (B.3) and Equation (B.4) we can conclude that:

$$\sum_{t\leq T} f_t(x_t) \geq T \cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{\boldsymbol{c}}} - \mathcal{R}^{\texttt{P}}_{\mathcal{T}_{\texttt{G}}} - \frac{2}{\nu} - \frac{1}{\nu}\mathcal{E}_{T,\delta} - \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},<\tau} - \mathcal{R}^{\texttt{D}}_{\mathcal{T}_{\texttt{G}},>\tau}.$$

which concludes the proof. □

C PROOFS OMITTED FROM SECTION 5

Lemma 5.5. *Assume that the dual regret minimizer is OGD on $\mathbb{R}^m_+$ with $\eta = (k_1 \mathcal{E}_{T,\delta} + k_2 m \mathcal{E}^{\mathrm{P}}_{T,\delta} + 2m\sqrt{T})^{-1}$, where k_1, k_2 are absolute constants, and the primal regret minimizer $\mathcal{A}^{\mathrm{P}}$ satisfies Condition 5.4. Then, both in the adversarial and stochastic setting, the Lagrange multipliers $\boldsymbol{\lambda}_t$ played by the dual regret minimizer $\mathcal{A}^{\mathrm{D}}$ are such that $M := \sup_{t\in[T]} \|\boldsymbol{\lambda}_t\|_1 \le 8m/\nu$.*

Proof. We will address the stochastic and adversarial case separately. First, we focus on the stochastic case. The adversarial case will follow via minor modifications.

Stochastic setting. We prove the theorem by contradiction. Suppose that there exists a round t_2 such that $\|\boldsymbol{\lambda}_{t_2}\|_1 \ge 8m/\nu$, and let $t_1 \in [t_2]$ be the the first round such that $\|\boldsymbol{\lambda}_{t_1}\|_1 \ge 1/\nu$. Notice that the dual regret minimizer $\mathcal{A}^{\mathrm{D}}$ (*i.e.,* OGD) guarantees that:

$$\|\boldsymbol{\lambda}_{t_1}\|_1 \le \frac{1}{\nu} + m\eta \le \frac{2}{\nu} \quad \text{and} \quad \|\boldsymbol{\lambda}_{t_2}\|_1 \le \frac{8m}{\nu} + m\eta \le \frac{9m}{\nu},$$

since the dual losses are in $[-1,1]^m$, and by assumption $\|\boldsymbol{\lambda}_{t_1-1}\|_1 \le 1/\nu$ and $\|\boldsymbol{\lambda}_{t_2-1}\|_1 \le 8m/\nu$. Hence, the range of payoffs of the primal regret minimizer $|u^{\mathrm{P}}_t|$ in the interval $\mathcal{T} := \{t_1, \ldots, t_2\}$ can be bounded as follows

$$\sup_{x\in\mathcal{X}, t\in\mathcal{T}} |u^{\mathrm{P}}_t(x)| \le \sup_{x\in\mathcal{X}, t\in\mathcal{T}} \left\{ |f_t(x)| + \|\boldsymbol{\lambda}_t\|_1 \cdot \|\boldsymbol{\rho} - \boldsymbol{c}_t(x)\|_\infty \right\} \le 1 + 2\frac{9m}{\nu} \le \frac{19m}{\nu}.$$

Therefore, by assumption, the regret of the primal regret minimizer is at most:

$$\mathcal{R}^{\mathrm{P}}_{\mathcal{T}} \le \left(\frac{19m}{\nu}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta}.$$

Similarly to the proof of Lemma 4.2, by applying a Hoeffding's bound to all the intervals and a union bound, we get that, with probability at least $1-\delta$, it holds

$$\begin{aligned}\sum_{t\in\mathcal{T}} \langle \boldsymbol{\lambda}_t, \boldsymbol{c}_t(\varnothing)\rangle &\le \sum_{t\in\mathcal{T}} \langle \boldsymbol{\lambda}_t, \bar{\boldsymbol{c}}(\varnothing)\rangle + \frac{9m}{\nu}\mathcal{E}_{T,\delta} \\ &\le -\beta \sum_{t\in\mathcal{T}} \|\boldsymbol{\lambda}_t\|_1 + \frac{9m}{\nu}\mathcal{E}_{T,\delta}. \qquad \text{(C.1)}\end{aligned}$$

Then, by the no-regret property of the primal regret minimizer we have

$$\begin{aligned}\sum_{t\in\mathcal{T}} (f_t(x_t) - \langle \boldsymbol{\lambda}_t, c_t(x_t) - \boldsymbol{\rho}\rangle) &\ge \sum_{t\in\mathcal{T}} (f_t(\varnothing) - \langle \boldsymbol{\lambda}_t, c_t(\varnothing) - \boldsymbol{\rho}\rangle) - \left(\frac{19m}{\nu}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} \\ &\ge \sum_{t\in\mathcal{T}} f_t(\varnothing) + \beta \sum_{t\in\mathcal{T}} \|\boldsymbol{\lambda}_t\|_1 + \sum_{t\in\mathcal{T}} \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho}\rangle - \left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta} - \left(\frac{19m}{\nu}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} \\ &\ge \nu \sum_{t\in\mathcal{T}} \|\boldsymbol{\lambda}_t\|_1 - \left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta} - \left(\frac{19m}{\nu}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} \\ &\ge (t_2 - t_1) - \left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta} - \left(\frac{19m}{\nu}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta}, \qquad \text{(C.2)}\end{aligned}$$

where in the second inequality we use Equation (C.1) and in the last one we use that $\|\boldsymbol{\lambda}_t\|_1 \ge 1/\nu$ for $t \in \mathcal{T}$.

For each resource $i \in [m]$, we consider two cases: i) the dual regret minimizer never has to perform a projection operation during $\mathcal{T}$, and ii) $\tilde{t}_i \in \mathcal{T}$ is the last time in which $\lambda_{t_i,i} = 0$. In both cases, we show that

$$\sum_{t\in\mathcal{T}} \langle \boldsymbol{\lambda}_t, \boldsymbol{\rho} - \boldsymbol{c}_t(x_t)\rangle \le \sum_{i\in[m]} \left[\frac{\lambda_{t_1,i} - \lambda_{t_2,i}}{\eta\nu}\right]^- + \frac{5m}{\nu^2\eta}. \qquad \text{(C.3)}$$

In the first case, since we are using OGD as the dual regret minimizer and, by assumption, it never has to perform projections during $\mathcal{T}$, it holds that for all resources $i \in [m]$:

$$\lambda_{t_2,i} = \eta \sum_{t\in\mathcal{T}} (c_{t,i}(x_t) - \rho) + \lambda_{t_1,i}.$$

Now consider the Lagrange multiplier $\boldsymbol{\lambda}^*$ such that, for each $i \in [m]$,

$$\lambda_i^* := \begin{cases} 1/\nu & \text{if} \quad \sum_{t\in\mathcal{T}} (c_{t,i}(x_t) - \rho) \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

By exploiting Lemma 5.3 for a single component $i \in [m]$, we have that:

$$\sum_{t\in\mathcal{T}} (\lambda_i^* - \lambda_{t,i})(c_{t,i}(x_t) - \rho) \leq \frac{(\lambda_i^* - \lambda_{t_1,i})^2}{2\eta} + \frac{1}{2}\eta T, \tag{C.4}$$

which yields the following

$$\begin{aligned} \sum_{t\in\mathcal{T}} \lambda_{t,i} \cdot (\rho - c_{t,i}(x_t)) &\leq \sum_{t\in\mathcal{T}} \lambda_i^* \cdot (\rho - c_{t,i}(x_t)) + \frac{(\lambda_i^* - \lambda_{t_1,i})^2}{2\eta} + \frac{1}{2}\eta T \\ &\leq \frac{1}{\nu}\left[\frac{\lambda_{t_1,i} - \lambda_{t_2,i}}{\eta}\right]^- + \frac{(\lambda_i^* - \lambda_{t_1,i})^2}{2\eta} + \frac{1}{2}\eta T. \end{aligned}$$

Then, since $\|\boldsymbol{\lambda}_{t_1-1}\|_1 \leq 1/\nu$ by construction, it holds $\lambda_{t_1-1,i} \leq 1/\nu$. Hence, since the dual regret minimizer is OGD and its utilities are in $[-1,1]^m$, it holds

$$\lambda_{t_1,i} \leq \frac{1}{\nu} + \eta. \tag{C.5}$$

Then, we have

$$\frac{(\lambda_i^* - \lambda_{t_1,i})^2}{2\eta} \leq \frac{(\max\{\lambda_i^*, \lambda_{t_1,i}\})^2}{2\eta} \leq \frac{1}{2\eta}\left(\frac{1}{\nu} + \eta\right)^2.$$

It is easy to verify that for all ν, T, and $\eta \leq \frac{1}{2\sqrt{T}}$:[6]

$$\frac{1}{2\eta}\left(\frac{1}{\nu} + \eta\right)^2 + \frac{1}{2}\eta T \leq \frac{5}{\nu^2 \eta},$$

which implies that

$$\sum_{t\in\mathcal{T}} \lambda_{t,i}, \rho - c_{t,i}(x_t) \leq \frac{1}{\nu}\left[\frac{\lambda_{t_1,i} - \lambda_{t_2,i}}{\eta}\right]^- + \frac{5}{\nu^2 \eta}.$$

In the second case we define $\tilde{t}_i$ as the last time step $t \in [t_1, t_2]$ in which $\lambda_t = 0$. Thus, the following holds:

$$\lambda_{t_2,i} = \eta \sum_{t\in[\tilde{t}_i, t_2]} (c_{t,i}(x_t) - \rho). \tag{C.6}$$

Now, consider the Lagrange multipliers $\lambda_{i,1} = 0$ and $\lambda_{i,2} = \frac{1}{\nu}$. By Lemma 5.3, we have that the regret of the dual regret minimizer with respect to 0 over $[t_1, \tilde{t}_i]$ is bounded by:

$$\sum_{t\in[t_1,\tilde{t}_i-1]} \lambda_{t,i}(\rho - c_{t,i}(x_t)) \leq \frac{\left(\lambda_{1,i} - \lambda_{i,1}^*\right)^2}{2\eta} + \frac{1}{2}\eta T \leq \frac{\left(\frac{1}{\nu} + \eta\right)^2}{2\eta} + \frac{1}{2}\eta T, \tag{C.7}$$

where in the last inequality we use Equation (C.5).

Similarly, on the interval $[\tilde{t}_i + 1, t_2]$ we have that the regret of the dual regret minimizer with respect to $1/\nu$ is bounded by

$$\sum_{t\in[\tilde{t}_i,t_2]} (\lambda_{i,2} - \lambda_{t,i}) \cdot (c_{t,i}(x_t) - \rho) \leq \frac{\left(\lambda_{i,2}^* - \lambda_{\tilde{t}_i,i}\right)^2}{2\eta} + \frac{1}{2}\eta T \leq \frac{1}{2\eta}\frac{1}{\nu^2} + \frac{1}{2}\eta T,$$

[6] Notice that the definition of η satisfies $\eta \leq \frac{1}{2\sqrt{T}}$.

which can be rearranged into

$$\sum_{t\in[\tilde{t}_i,t_2]} \lambda_{t,i}\cdot(\rho-c_{t,i}(x_t)) \le \frac{1}{\nu}\sum_{t\in[\tilde{t}_i,t_2]}(\rho-c_{t,i}(x_t))+\frac{1}{2\eta}\frac{1}{\nu^2}+\frac{1}{2}\eta T$$
$$= -\frac{\lambda_{t_2,i}}{\eta\nu}+\frac{1}{2\eta}\frac{1}{\nu^2}+\frac{1}{2}\eta T$$
$$\le \left[\frac{\lambda_{t_1,i}-\lambda_{t_2,i}}{\eta\nu}\right]^{-}+\frac{1}{2\eta}\frac{1}{\nu^2}+\frac{1}{2}\eta T \tag{C.8}$$

where the equality follows from Equation (C.6) and the last inequality from $-x \le \min(y-x,0)$ for $x, y \ge 0$.

Then by summing Equation (C.7) and Equation (C.8) we obtain

$$\sum_{t\in[t_1,t_2]} \lambda_{t,i}\cdot(\rho-c_{t,i}(x_t)) \le \left[\frac{\lambda_{t_1,i}-\lambda_{t_2,i}}{\eta\nu}\right]^{-}+\frac{\left(\frac{1}{\nu}+\eta\right)^2}{2\eta}+\frac{1}{2\eta\nu^2}+\eta T$$

If we take $\eta \le \frac{1}{2\sqrt{T}}$, then the following inequality holds for all ν and T:[7]

$$\frac{\left(\frac{1}{\nu}+\eta\right)^2}{2\eta}+\frac{1}{2\eta\nu^2}+\eta T \le \frac{5}{\nu^2\eta}.$$

This concludes the second case and concludes the proof of Equation (C.3).

Now we can sum over all resources $i \in [m]$ and conclude that:

$$\sum_{t\in\mathcal{T}}\langle\boldsymbol{\lambda}_t,\boldsymbol{\rho}-\boldsymbol{c}_t(x_t)\rangle \le \sum_{i\in[m]}\left[\frac{\lambda_{t_1,i}-\lambda_{t_2,i}}{\eta\nu}\right]^{-}+\frac{5m}{\eta\nu^2}$$
$$\le \frac{\|\boldsymbol{\lambda}_{t_1}\|_1-\|\boldsymbol{\lambda}_{t_2}\|_1}{\eta\nu}+\frac{5m}{\eta\nu^2}$$
$$\le -\frac{6m}{\nu^2\eta}+\frac{5m}{\eta\nu^2}$$
$$= -\frac{m}{\nu^2\eta}.$$

Finally, the cumulative utility of the primal regret minimizer over $\mathcal{T}$ is bounded by

$$\sum_{t\in[t_1,t_2]}(f_t(x_t)+\langle\lambda_t,\boldsymbol{\rho}-c_t(x_t)\rangle) \le (t_2-t_1)-\frac{m}{\nu^2\eta}. \tag{C.9}$$

By putting Equation (C.2) and Equation (C.9) together we have that

$$(t_2-t_1)-\frac{m}{\nu^2\eta} \ge (t_2-t_1)-\left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta}-\left(\frac{19m}{\nu}\right)^2\mathcal{E}^{\mathbb{P}}_{T,\delta}$$

Hence,

$$\frac{m}{\nu^2\eta} \le \left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta}+\left(\frac{19m}{\nu}\right)^2\mathcal{E}^{\mathbb{P}}_{T,\delta}. \tag{C.10}$$

which holds by the assumption that there exists a time t_2 such that $\|\boldsymbol{\lambda}_{t_2}\|_1 \ge \frac{8m}{\nu}$.

Thus, if we set

$$\eta := \left(18\mathcal{E}_{T,\delta}+361m\mathcal{E}^{\mathbb{P}}_{T,\delta}+2m\sqrt{T}\right)^{-1},$$

[7] Notice that the definition of η satisfies $\eta \le \frac{1}{2\sqrt{T}}$.

we reach a contradiction with Equation (C.10) by observing that

$$\frac{m}{\nu^2\eta} = \frac{m}{\nu^2}\left(18\mathcal{E}_{T,\delta} + 361m\mathcal{E}^{\mathbb{P}}_{T,\delta} + m\nu\sqrt{T}\right) > \left(\frac{9m}{\nu}\right)\mathcal{E}_{T,\delta} + \left(\frac{19m}{\nu}\right)^2\mathcal{E}^{\mathbb{P}}_{T,\delta}.$$

This concludes the proof for the stochastic setting.

Adversarial setting. In the adversarial setting all the passages above still apply by setting $\mathcal{E}_{T,\delta} = 0$. This is because, using Assumption 2.1, we can refrain from using the concentration inequality originating the term $\mathcal{E}_{T,\delta}$.

This concludes the proof. □

Corollary 5.6. *Assume that the dual regret minimizer is OGD on $\mathbb{R}^m_+$ with $\eta = (k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{\mathbb{P}}_{T,\delta} + 2m\sqrt{T})^{-1}$, where k_1, k_2 are absolute constants, and the primal regret minimizer $\mathcal{A}^{\mathbb{P}}$ satisfies Condition 5.4. Then, in the adversarial setting, Algorithm 1 guarantees with probability $1-\delta$ that*

$$\sum_{t\in[T]} f_t(x_t) \geq \alpha \cdot \texttt{OPT}_{\boldsymbol{\gamma}} - k_3\frac{m^4}{\nu^2}\left(\mathcal{E}^{\mathbb{P}}_{T,\delta} + \mathcal{E}_{T,\delta}\right),$$

where k_3 is an absolute constant and $\alpha = \nu/1+\beta$. In the stochastic setting, there is an absolute constant k_4 such that Algorithm 1 guarantees that, with probability at least $1-2\delta$,

$$\sum_{t\in[T]} f_t(x_t) \geq T\cdot \texttt{OPT}^{\texttt{LP}}_{\bar{f},\bar{c}} - k_4\frac{m^4}{\nu^2}\left(\mathcal{E}^{\mathbb{P}}_{T,\delta} + \mathcal{E}_{T,\delta}\right).$$

Proof. Lemma 5.5 allows us to bound the ℓ_1-norm of the Lagrange multipliers which are played by $\|\boldsymbol{\lambda}_t\|_1 \leq 8m/\nu$. This fact, by using Lemma 5.3, readily gives a bound on the terms $\mathcal{R}^{\mathbb{D}}_{\mathcal{T}_{\mathbb{G}}<\tau}(\mathcal{D})$ and $\mathcal{R}^{\mathbb{D}}_{\mathcal{T}_{\mathbb{G}}>\tau}(\mathcal{D})$ by observing that

$$\sup_{\substack{\boldsymbol{\lambda},\boldsymbol{\lambda}',\\ \|\boldsymbol{\lambda}\|_1,\|\boldsymbol{\lambda}'\|_1\leq\frac{8m}{\nu},}} \|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\|_2^2 \leq \frac{64m^3}{\nu^2}.$$

This, together with the definition of $\eta = (k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{\mathbb{P}}_{T,\delta} + 2m\sqrt{T})^{-1}$ yields the bound

$$\begin{aligned}
\max(\mathcal{R}^{\mathbb{D}}_{\mathcal{T}_{\mathbb{G}}<\tau}, \mathcal{R}^{\mathbb{D}}_{\mathcal{T}_{\mathbb{G}}>\tau}) &\leq \frac{32m^3}{\nu^2}(k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{\mathbb{P}}_{T,\delta} + 2m\sqrt{T}) + \frac{m}{2}\frac{T}{k_1\mathcal{E}_{T,\delta} + k_2 m\mathcal{E}^{\mathbb{P}}_{T,\delta} + 2m\sqrt{T}} \\
&= k_1\frac{32m^3}{\nu^2}\mathcal{E}_{T,\delta} + k_2\frac{32m^4}{\nu^2}\mathcal{E}^{\mathbb{P}}_{T,\delta} + \frac{32m^4}{\nu}\sqrt{T} + \frac{1}{4}\sqrt{T} \\
&\leq \frac{k_3}{2}\frac{m^4}{\nu^2}\left(\mathcal{E}_{T,\delta} + \mathcal{E}^{\mathbb{P}}_{T,\delta} + \sqrt{T}\right).
\end{aligned}$$

Then, by Condition 5.4 and Lemma 5.5, the regret of the primal regret minimizer $\mathcal{A}^{\mathbb{P}}$ is bounded by:

$$\begin{aligned}
\mathcal{R}^{\mathbb{P}}_{\mathcal{T}_{\mathbb{G}}} &\leq \left(1 + \frac{16m}{\nu}\right)^2\mathcal{E}^{\mathbb{P}}_T \\
&= k_4\frac{m^2}{\nu^2}\mathcal{E}^{\mathbb{P}}_T.
\end{aligned}$$

This concludes the proof by leveraging Theorem 4.1 and Theorem 4.3. □