Online Learning under Budget and ROI Constraints via Weak Adaptivity

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

Abstract

We study online learning problems in which a decision maker has to make a sequence of costly decisions, with the goal of maximizing their expected reward while adhering to budget and return-on-investment (ROI) constraints. Existing primal-dual algorithms designed for constrained online learning problems under adversarial inputs rely on two fundamental assumptions. First, the decision maker must know beforehand the value of parameters related to the degree of strict feasibility of the problem (i.e. Slater parameters). Second, a strictly feasible solution to the offline optimization problem must exist at each round. Both requirements are unrealistic for practical applications such as bidding in online ad auctions. In this paper, we show how such assumptions can be circumvented by endowing standard primal-dual templates with *weakly adaptive* regret minimizers. This results in a "dual-balancing" framework which ensures that dual variables stay sufficiently small, even in the absence of knowledge about Slater's parameter. We prove the first *best-of-both-worlds* no-regret guarantees which hold in absence of the two aforementioned assumptions, under stochastic and adversarial inputs. Finally, we show how to instantiate the framework to optimally bid in various mechanisms of practical relevance, such as first-price auctions.

1. Introduction

A decision maker takes decisions over T rounds. At each round t, the decision $x_t \in \mathcal{X}$ is chosen before observing a reward function f_t together with a set of *time-varying* constraint functions. The decision maker is allowed to make decisions that are *not* feasible, provided that the overall sequence of decisions obeys the *long-term constraints* over the entire time horizon, up to a small cumulative violation across the T rounds. The goal of the decision maker is to maximize their cumulative reward, while satisfying the long-term constraints. This model was first proposed by Mannor et al. (2009) and later developed along various directions (Mahdavi et al., 2012; Jenatton et al., 2016; Liakopoulos et al., 2019; Yu et al., 2017; Castiglioni et al., 2022b).

Motivated by applications in online ad auctions, we consider the case in which the decision maker has a budget and a *return-on-investment* (ROI) constraint (Auerbach et al., 2008; Golrezaei et al., 2023; 2021). The decision maker is subject to bandit feedback: at each time t, the decision maker takes a decision x_t and then observes the realized reward $f_t(x_t)$ and a cost $c_t(x_t)$. Inputs (f_t, c_t) can either be generated i.i.d. or selected by an oblivious adversary.

A key challenge of our model is that ROI constraints are not *packing*, thereby preventing the direct application of known algorithms for *adversarial bandits with knapsacks* (ABwK) (Immorlica et al., 2022; Castiglioni et al., 2022a), or for online allocation problems with resource-consumption constraints (Balseiro et al., 2022). Moreover, previous work addressing the adversarial case with non-packing constraints makes the assumption that the "worst-case feasibility" with respect to all constraint functions observed up to T is strictly positive (Sun et al., 2017; Castiglioni et al., 2022b; Immorlica et al., 2022; Balseiro et al., 2022). In other words, there has to exist a "safe" policy guaranteeing that, at each t, the constraints can be satisfied by a margin at least $\alpha > 0$, which has to be known in advance by the learner. This can be problematic for at least two reasons: **i) α may not be known in advance to the decision maker, and ii) the safe policy may not exist in all rounds t.** For example, in the case of bidding in repeated ad auctions under budget and ROI constraints, such assumptions do not hold. In particular, the decision maker would be required to have an action yielding expected ROI strictly above their target for each round t. If we assume one ad placement is being allocated at each t then this assumption is equivalent to assuming that the bidders' value is always strictly higher than the highest competing bid, which clearly would not hold in practice. In this paper, we propose a general approach to enhance

[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 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

primal-dual frameworks based on the template by Immorlica et al. (2022), in order to obtain *best-of-both-worlds* no-regret guarantees while bypassing both assumptions.

1.1. Contributions

Previous primal-dual templates for adversarial inputs assume that the parameter $\alpha > 0$ is known (see, *e.g.,* (Immorlica et al., 2022; Balseiro et al., 2022)). This is crucial to ensure boundedness of Lagrange multipliers, which is typically obtained by requiring the ℓ_1-norm of the multipliers to be less than or equal to $1/\alpha$ (Castiglioni et al., 2022a; Nedić & Ozdaglar, 2009). The key contribution of the paper is showing that, in absence of any information on $\alpha > 0$, if we require the primal and dual regret minimizers to be *weakly adaptive* (*i.e.,* to guarantee sublinear regret on any interval $[t_1, t_2] \subseteq [T]$ (Hazan & Seshadhri, 2007)), then boundedness of multipliers automatically emerges as a byproduct of the interaction between the primal and dual algorithms (thereby **solving challenge (i)**). Moreover, our framework only requires the existence of a safe policy "frequently enough", and not at all time steps t (thereby **solving challenge (ii)**). This is the first time that the notion of adaptive regret minimization is used within primal-dual frameworks. Interestingly, this usage is significantly different from its original motivating applications (Hazan & Seshadhri, 2007; Adamskiy et al., 2016; Daniely et al., 2015).

We show that the resulting framework provides best-of-both-worlds no-regret guarantees while solving both limitations. We prove a tight $\tilde{O}(T^{1/2})$ regret upper bound in the stochastic setting, and an $\alpha/(\alpha+1)$ constant-factor competitive ratio in the adversarial setting, under the standard assumption that the budget is $\Omega(T)$ (Balseiro et al., 2022). In both settings, our framework guarantees vanishing cumulative ROI constraint violation, and cumulative expenditure less than or equal to the available budget. Best-of-both-worlds algorithms for problems with long-term constraints typically require different proof techniques for the two input models. We unify most of the analysis, with the only difference being in the characterization of a particular set of policies.

Finally, we show how our framework can be employed for bidding in any mechanism with finite types. In particular, we show that it can handle the case of repeated *non-truthful* auctions (*e.g.,* first-price auctions). Previous work could only handle budget- and ROI-constrained bidders in the simpler case of second-price auctions, in which truthfulness can be exploited (Feng et al., 2023; Golrezaei et al., 2023).

1.2. Related works

Standard primal-dual approaches for bandit problems with knapsack constraints cannot be applied in our setting, as they require as an input the Slater's parameter α (Balseiro & Gur, 2019; Immorlica et al., 2022; Balseiro et al., 2022; Castiglioni et al., 2022a;b). In these works, the knowledge of α is exploited to ensure that dual variables are "small" through an explicit projection step over a set which depends on α. This is not possible in our setting, due to the presence of non-packing ROI constraints. In our new analysis, we show how such frameworks can be suitably adapted to work in more complex scenarios than the standard one. The issue of not knowing α has been effectively addressed in stochastic settings (Yu et al., 2017; Wei et al., 2020; Castiglioni et al., 2022b; Lobos et al., 2021). Nonetheless, since our goal is to provide guarantees that hold also in the presence of adversarial inputs, such results do not extend to our setting.

Repeated auctions. The problem of online bidding under budget constraints has been studied in various settings (Balseiro & Gur, 2019; Ai et al., 2022). In the context of online allocation problems with an arbitrary number of constraints, Balseiro et al. (2020; 2022) propose a class of primal-dual algorithms attaining asymptotically optimal performance in the stochastic and adversarial case. In their setting, at each round, the input (f_t, c_t) is observed by the learner *before* they make a decision. This makes the problem substantially different from ours. In particular, their framework cannot handle *non-truthful* repeated auctions such as first-price auctions. Recent works have also examined settings similar to ours, involving bidders with constraints on their budget and ROI. The framework by Feng et al. (2023) can handle both ROI and budget constraints, but crucially relies on truthfulness of second-price auctions, and on the stochasticity of the environment. The recent work by Wang et al. (2023) considers the problem of bidding in repeated first-price auctions under only budget constraints and stationary competition. Their analysis cannot be extended to our setting for the same reasons mentioned at the beginning of the section.

Concurrent work. Slivkins et al. (2023) studies a stochastic setting with general long-term constraints similar to Castiglioni et al. (2022b). They provide a $\tilde{O}(T^{1/2})$ guarantee when α is known, and $\tilde{O}(T^{3/4})$ guarantees when α is not known. The latter result cannot be extended to the case of inputs generated by an adversary. The recent paper by Bernasconi et al. (2023) studies a different setting from ours, in which they only have budget constraints, and knowledge of α is hindered by the fact that resources can be replenished (*i.e.,* costs can be negative, as in Kumar & Kleinberg (2022)). They provide best-of-both-world guarantees by exploiting results presented in this paper. Further, related works are presented in Appendix A.

2. Preliminaries

At each round $t \in [T]$, the decision maker chooses an action $x_t \in \mathcal{X}$, where $\mathcal{X}$ is non-empty set of available actions, and subsequently observes reward $f_t(x_t)$ with $f_t : \mathcal{X} \to [0, 1]$,

and incurs a cost $c_t(x_t)$, with $c_t : \mathcal{X} \to [0, 1]$. We denote as $\mathcal{F}$, respectively $\mathcal{C}$, the set of all the possible functions f_t, respectively c_t (*e.g.,* $\mathcal{F}$ and $\mathcal{C}$ may contain all the Lipschitz-continuous functions defined over $\mathcal{X}$, or all the convex functions over $\mathcal{X}$). We assume that functions in $\mathcal{F}$ and $\mathcal{C}$ are measurable with respect to probability measures over $\mathcal{X}$. This ensures that expectations are well-defined, since the functions are assumed to be bounded above, and they are therefore integrable. Following previous work (Agarwal et al., 2014b; Badanidiyuru et al., 2018; Immorlica et al., 2022), we assume the existence of a *void action* $\varnothing$ such that, for any pair $(f, c) \in \mathcal{F} \times \mathcal{C}$, $f(\varnothing) = c(\varnothing) = 0$. The decision maker has an overall budget $B \in \mathbb{R}_+$, $B = \Omega(T)$, which limits the total expenditure throughout the T rounds. We denote by $\rho > 0$ the *per-iteration budget* defined as B/T. Moreover, the decision maker has a target *return-on-investments* (ROI) $\omega > 0$. In order to simplify the notation, throughout the paper we will assume $\omega := 1$. This comes without loss of generality: whenever $\omega > 1$ we can suitably scale down values of reward functions f_t. Then, the decision maker has the goal of maximizing their cumulative utility $\sum_{t=1}^T f_t(x_t)$, subject to the following constraints:

- **Budget constraints**: $\sum_{t=1}^T c_t(x_t) \leq \rho T$. Such constraints should be satisfied "no matter what," so we refer to them as *hard* constraints.
- **ROI constraints**: $\sum_{t=1}^T (c_t(x_t) - f_t(x_t)) \leq 0$. We say ROI constraints are *soft* meaning that we allow, in expectation, a small (vanishing in the limit) cumulative violation across the T rounds.

In the context of repeated ad auctions, as we will discuss in Section 8, this model can be easily instantiated to describe *any* mechanism with finite types beyond the well-studied case of second-price auctions.

Auxiliary LP. We endow $\mathcal{X}$ with the Lebesgue σ-algebra, and we denote by Π be the set of *randomized policies*, defined as the set of probability measures on the Borel sets of $\mathcal{X}$. At any $t \in [T]$ the decision maker will compute a policy $\boldsymbol{\pi}_t \in \Pi$ and play an action $x_t \sim \boldsymbol{\pi}_t$ accordingly.[1] Given a reward function f and a cost function c, let $g : \Pi \ni \boldsymbol{\pi} \mapsto \mathbb{E}_{x\sim\boldsymbol{\pi}}[c(x)] - \rho$ be the expected gap between the cost for policy $\boldsymbol{\pi}$ and the per-iteration budget ρ, and $h : \Pi \ni \boldsymbol{\pi} \mapsto \mathbb{E}_{x\sim\boldsymbol{\pi}}[c(x) - f(x)]$ be the expected ROI constraint violation for policy $\boldsymbol{\pi}$. We will denote by g_t, resp. h_t, the constraints defined for the pair (f_t, c_t) observed at round t. In order to simplify the notation, given $x \in \mathcal{X}$, the value of the reward function for the policy that deterministically plays action x (*i.e.,* the Dirac mass δ_x) will be denoted by $f_t(x)$ in place of $f_t(\delta_x)$. Analogously, we will write $c_t(x)$, $g_t(x)$, and $h_t(x)$ instead of using Dirac measures δ_x. Let $\mathcal{P}$ be an arbitrary probability measure over the space of possible inputs $\mathcal{F} \times \mathcal{C}$. Then, we define the linear program $\texttt{LP}_{\mathcal{P}}$ as follows:

$$\texttt{OPT}_{\mathcal{P}} := \begin{cases} \sup_{\boldsymbol{\pi}\in\Pi} & \mathbb{E}_{f\sim\mathcal{P}} f(\boldsymbol{\pi}) \\ \text{s.t.} & \mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}) \leq 0 \\ & \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}) \leq 0 \end{cases}. \tag{$\texttt{LP}_{\mathcal{P}}$}$$

$\texttt{LP}_{\mathcal{P}}$ selects the bidding policy $\boldsymbol{\pi}$ maximizing the expected reward according to $\mathcal{P}$, while ensuring that constraints g and h encoded by $\mathcal{P}$ are satisfied in expectation (both g and h are defined by $(f, c) \sim \mathcal{P}$). The *Lagrangian function* $\mathcal{L}_{\mathcal{P}} : \Pi \times \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ of the above LP is defined as

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\pi}, \lambda, \mu) := \mathbb{E}_{(f,c)\sim\mathcal{P}}[f(\boldsymbol{\pi}) - \lambda g(\boldsymbol{\pi}) - \mu h(\boldsymbol{\pi})].$$

3. Baselines

Our goal is to design online algorithms that output a sequence of policies $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T$ such that i) the *cumulative regret* with respect to the performance of the baseline grows sublinearly in T, ii) the budget constraint is (deterministically) satisfied, *i.e.,* $\sum_{t=1}^T c_t(x_t) \leq B$, and iii) the *cumulative ROI constraint violation* $\sum_{t=1}^T h_t(\boldsymbol{\pi}_t)$ grows sublinearly in the number of rounds T. The cumulative regret of the algorithm is defined as $R^T := T\,\texttt{OPT} - \sum_{t=1}^T f_t(x_t)$, where the baseline $\texttt{OPT}$ depends on how the input sequence $\gamma := (f_t, c_t)_{t=1}^T$ is generated. We consider the following two settings for which we define an appropriate value of the baseline, and a suitable problem-specific parameter $\alpha \in \mathbb{R}$ which is related to the feasibility of the offline problem.

Stochastic setting: at each $t \in [T]$, the pair (f_t, c_t) is independently drawn according to a fixed but unknown distribution $\mathcal{P}$ over $\mathcal{F} \times \mathcal{C}$. The baseline is $\texttt{OPT}_{\mathcal{P}}$, which is the standard baseline for stochastic BwK problems since its value is guaranteed to be closed to that of the best dynamic policy (Badanidiyuru et al., 2018, Lemma 3.1). In this setting, let

$$\alpha := -\inf_{\boldsymbol{\pi}\in\Pi} \max\{\mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}), \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi})\}.$$

Adversarial setting: the sequence of inputs γ is selected by an oblivious adversary. Given γ, we define the following distribution over inputs: for any pair $(f, c) \in \mathcal{F} \times \mathcal{C}$, $\bar{\gamma}[f, c] = \sum_{t=1}^T \mathbb{1}[f_t = f, c_t = c]/T$. Then, the baseline is the solution of $\texttt{LP}_{\bar{\gamma}}$ (*i.e.,* $\texttt{OPT}_{\bar{\gamma}}$), which is the standard baseline for the adversarial setting (see, *e.g.,* Balseiro et al. (2022); Immorlica et al. (2022)). Therefore, the baseline is obtained by solving the offline problem initialized with the average of the realizations observed over the T rounds. Moreover, our results will also hold with respect to the best

[1] The set $\{1, \ldots, n\}$, with $n \in \mathbb{N}$, is compactly denoted as $[n]$, and we let $[0]$ be equal to the empty set. Moreover, given a discrete set $\mathcal{X}$, we denote by $\Delta_{\mathcal{X}}$ the $|\mathcal{X}|$-simplex.

unconstrained policy. We define α as

$$\alpha := -\inf_{\boldsymbol{\pi}} \max_{t\in[T]} \max\{g_t(\boldsymbol{\pi}), h_t(\boldsymbol{\pi})\}.$$

In this setting, α represents the "worst-case feasiblity" with respect to functions observed up to T.

We remark that the parameter α measures the worst case feasibility of the problem by considering both budget and ROI constraints. In absence of the latter constraints, α would coincide with ρ. We start by developing our analysis under the following standard assumption.

Assumption 3.1. In the adversarial (resp., stochastic) setting, γ (resp., $\mathcal{P}$) is such that $\alpha > 0$.

This means that $\texttt{LP}_{\mathcal{P}}$ and $\texttt{LP}_{\bar{\gamma}}$ satisfy (stochastic) Slater's condition. In particular, in the adversarial setting we are requiring the existence of a randomized "safe" policy that, in expectation, strictly satisfies the constraints for each t. This is a frequent assumption in works focusing on settings similar to ours (see, *e.g.,* (Chen et al., 2017; Neely & Yu, 2017; Yi et al., 2020; Castiglioni et al., 2022b)). In Section 7 we show how this requirement can be relaxed.

When studying primal-dual algorithms, a key implication of Slater's condition is the existence and boundedness of Lagrange multipliers (see, *e.g.,* Nedić & Ozdaglar (2009)). Therefore, when $\alpha > 0$ is known, the set of dual multipliers can be safely bounded by requiring the ℓ_1-norm of the multiplier to be less than or equal to $1/\alpha$ (see, *e.g.,* Balseiro et al. (2022)). This is the case, for example, for problems with only budget constraints, in which $\alpha = \rho > 0$, which is achieved by bidding the void action $\varnothing$ at each round. However, ROI constraints complicate the problem as the decision maker does *not* know α beforehand.

4. Adaptivity in Primal-Dual Frameworks

In this section, we first provide a concise overview of a generic primal-dual template that adheres to the structure presented by Immorlica et al. (2022); Castiglioni et al. (2022a). Then, we provide a simple example demonstrating that a direct application of such framework would result in violations of the constraints which are linear in T. Finally, we describe the modifications needed to update the standard primal-dual template in order to achieve the desired behavior, and we show that online gradient descent already meets the new criteria for the dual regret minimizer.

4.1. A Standard Primal-Dual Template

Algorithm 1 summarizes the structure of a standard primal-dual framework. It assumes access to two regret minimizers with the following characteristics. The first one is the *primal regret minimizer* $\mathcal{R}^{\text{P}}$. It outputs policies in Π, and receives

Algorithm 1 Primal-dual framework.

Input: parameters B, T, δ; regret minimizers $\mathcal{R}^{\text{P}}$, $\mathcal{R}^{\text{D}}$
Initialization: $B_1 \leftarrow B$; initialize $\mathcal{R}^{\text{P}}, \mathcal{R}^{\text{D}}$
for $t = 1, 2, \dots, T$ **do**
  Dual decision: $(\lambda_t, \mu_t) \leftarrow$ output of $\mathcal{R}^{\text{D}}$
  Primal decision: $\Pi \ni \pi_t \leftarrow$ output of $\mathcal{R}^{\text{P}}$
  Select action as

$$x_t \leftarrow \begin{cases} x_t \sim \boldsymbol{\pi}_t & \text{if } B_t \geq 1 \\ \varnothing & \text{otherwise} \end{cases}.$$

  Observe: observe $f_t(x_t)$ and $c_t(x_t)$, and update available resources: $B_{t+1} \leftarrow B_t - c_t(x_t)$
  Primal update: update $\mathcal{R}^{\text{P}}$ using $\ell_t^{\text{P}}(x_t)$
  Dual update: update $\mathcal{R}^{\text{D}}$ using $u_t^{\text{D}}(\cdot)$
end for

as feedback the loss $\ell_t^{\text{P}} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ such that, for any x_t sampled according to policy $\boldsymbol{\pi}_t$,

$$\ell_t^{\text{P}}(x_t) = -f_t(x_t) + \lambda_t(g_t(x_t) + 1) + \mu_t(h_t(x_t) + 1) + 1.$$

The primal loss function is obtained from the Lagrangian relaxation of the problem at time t, plus the additive term $1 + \lambda_t + \mu_t$ to ensure $\ell_t^{\text{P}}(\cdot) \in \mathbb{R}_{\geq 0}$. For ease of presentation, whenever we write $\ell_t^{\text{P}}(\boldsymbol{\pi})$ we mean $\ell_t^{\text{P}}(\boldsymbol{\pi}) = \mathbb{E}_{x\sim\boldsymbol{\pi}}\ell_t^{\text{P}}(x)$. The second regret minimizer is the *dual regret minimizer* $\mathcal{R}^{\text{D}}$. It outputs points in the space of dual variables $\mathcal{D}_\alpha := \{\boldsymbol{y} \in \mathbb{R}^2_{\geq 0} : \|\boldsymbol{y}\|_1 \leq 1/\alpha\}$, and observes the linear utility

$$u_t^{\text{D}} : \mathcal{D}_\alpha \ni (\lambda, \mu) \mapsto \lambda g_t(x_t) + \mu h_t(x_t) \in \mathbb{R}.$$

The dual regret minimizer has full feedback by construction.

For each t, the algorithm first computes primal and dual actions. The action at time t is $x_t \sim \boldsymbol{\pi}_t$ unless the available budget B_t is less than 1, in which case we set x_t equal to the void action $\varnothing$. Then, $\ell_t^{\text{P}}(x_t)$ and u_t^{D} are observed, and the budget consumption is updated according to the realized cost c_t. Finally, the internal state of the two regret minimizers is updated on the basis of the feedback specified by $\ell_t^{\text{P}}, u_t^{\text{D}}$. We will denote by $\tau \in [T]$ the time in which the budget is fully depleted and the decision maker starts playing the void action $\varnothing$.

The traditional requirement on the primal and dual regret minimizers is that their cumulative regret should grow sublinearly in time (see, *e.g.,* Castiglioni et al. (2022a)).[2] Although sufficient for handling simpler problems, the following example shows that this requirement is not enough to provide guarantees in our setting.

[2] In general, the cumulative regret for losses ℓ_t is defined as $\inf_{x\in\mathcal{X}} \sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(x))$ (Cesa-Bianchi & Lugosi, 2006).

	Slot ①	Slot ②	Slot ③	b_t	μ_t
v	1	1/2	1/16		
$\mathcal{I}_1$	1/2	1/2	0	1/2	0
$\mathcal{I}_2$	1	1/2	0	0	0
$\mathcal{I}_3$	1	1/2	0	1	16

Table 1. Setup of the repeated GFP auctions.

4.2. When Standard Primal-Dual Algorithms Fail

We present a simple example in which the direct application of Algorithm 2 with the standard requirements presented in Section 4.1 does not yield the desired behavior, even if the learner knows α a priori. We observe that this rule out the direct application of known primal-templates for adversarial inputs such as those by Balseiro et al. (2022); Immorlica et al. (2022); Castiglioni et al. (2022a).

Suppose the learner is participating in a sequence of *generalized first price auctions* (GFP) under an ROI constraint. In this setting, there are multiple ad slots that have to be allocated at each t. The bidder with the i-th highest bid is allocated the i-th slot and, upon winning a slot, their payment is equal to their bid amount. Table 1 provides a summary of the instance being considered. There are three ad slots ①, ②, and ③ at each t. The valuation v_i for each slot i is fixed across the entire time horizon. Let $t_1, t_2 \in [T]$, $t_1 \leq t_2$. We denote by $\mathcal{I} \coloneqq [t_1, t_2]$ the set $\{t_1, t_1+1, \dots, t_2\}$, and we call $\mathcal{I}$ the *time interval* starting from round t_1 to round t_2. We consider three time intervals, denoted by $\mathcal{I}_1 = [t_1]$, $\mathcal{I}_2 = (t_1, t_2]$, $\mathcal{I}_3 = (t_2, T]$ for some t_1, t_2. The cells highlighted in gray provide the highest competing bid for the different slots, in the three different time intervals. As an illustration, within interval $\mathcal{I}_2$ the learner has to bid $1/2 \leq b_t < 1$ to win ②. Within $\mathcal{I}_1$, when the learner bids $b_t \geq 1/2$ they win ①. The learner has quasi-linear utilities: their utility for winning slot i at time t is $v_i - b_t$. The learner has a ROI constraint with target ROI 1.

The last two columns of Table 1 describe a possible sequence of primal actions b_t and dual actions μ_t, which are constant within each interval. We observe that the dual variable is always at most $1/\alpha = 16$. It is possible to show that, by setting the length of intervals so that $|\mathcal{I}_1| = 39|\mathcal{I}_3|$ and $|\mathcal{I}_2| \geq 12|\mathcal{I}_3|$, the primal and dual regret over T are ≤ 0, thereby matching the standard requirements on regret as per Section 4.1 (calculations are provided in Appendix B). However, the cumulative constraint violations is

$$\sum_{t=1}^{T}(c_t(b_t) - f_t(b_t)) = \sum_{t=1}^{T}(2b_t - v_t) = \Omega(T).$$

Therefore, the standard requirements for primal and dual regret minimizers are insufficient to ensure sublinear regret and constraint violations. The crucial problem here is that, since the primal player attains negative regret in $\mathcal{I}_1$ and $\mathcal{I}_2$, then it can afford to make decisions that significantly violate the ROI constraint in $\mathcal{I}_3$. We observe that frameworks employing a recovery phase, such as the one by Castiglioni et al. (2022b), are not viable for our stated goals, since they rely on knowledge of α to switch between phases.

4.3. New Requirements: Weak Adaptivity

Unlike previous work, we require $\mathcal{R}^{\mathrm{P}}$ and $\mathcal{R}^{\mathrm{D}}$ to be *weakly adaptive*, that is, they should guarantee sublinear *adaptive* (*a.k.a. interval*) regret (see, *e.g.,* (Hazan & Seshadhri, 2007; Luo et al., 2018)). This notion of regret is stronger than "standard" external regret, and it will be essential in our analysis. The primal regret minimizer must be such that, for $\delta \in (0, 1]$, with probability at least $1 - \delta$ it holds that, for any $\boldsymbol{\pi} \in \Pi$ and for any interval $\mathcal{I}$,

$$\sum_{t\in\mathcal{I}}(\ell_t^{\mathrm{P}}(x_t) - \ell_t^{\mathrm{P}}(\boldsymbol{\pi})) \leq M_{\mathcal{I}}^2 \mathcal{E}_{T,\delta}^{\mathrm{P}}, \tag{4.1}$$

where $M_{\mathcal{I}}$ is the maximum absolute value of the losses ℓ_t^{P} observed in interval $\mathcal{I}$, and $\mathcal{E}_{T,\delta}^{\mathrm{P}}$ is a term of order $\tilde{O}(\sqrt{T})$. We require a similar property for the dual regret minimizer. However, since the dual regret minimizer works under full-information feedback by construction, we can use a regret minimizer with deterministic regret guarantees. In particular, $\mathcal{R}^{\mathrm{D}}$ should guarantee that, for any time interval $\mathcal{I} = [t_1, t_2]$, and for any pair of dual variables $(\lambda, \mu) \in \mathbb{R}^2_{\geq 0}$ it holds

$$\sum_{t\in\mathcal{I}}(u_t^{\mathrm{D}}(\lambda, \mu) - u_t^{\mathrm{D}}(\lambda_t, \mu_t)) \leq \nu(T)(\mu - \mu_{t_1})^2 + \mathcal{E}_T^{\mathrm{D,B}} + \mathcal{E}_T^{\mathrm{D,R}},$$

where $\nu(T) \geq 0$ is such that $\nu(T) = o(T)$, and $\mathcal{E}_T^{\mathrm{D,B}}$ (resp., $\mathcal{E}_T^{\mathrm{D,R}}$) is a term sublinear in T related to the budget (resp., ROI) constraint.

The choice of the primal regret minimizer is primarily influenced by the specific problem being considered. On the other hand, the dual problem remains constant, and thus we proceed by presenting an appropriate dual regret minimizer. Section 8 will provide some examples of primal regret minimizers satisfying Equation (4.1) for relevant applications.

4.4. A Weakly Adaptive Dual Regret Minimizer

As a dual regret minimizer we employ the standard online gradient descent algorithm (OGD) (Zinkevich, 2003) on each of the two Lagrangian multipliers λ and μ. We initialize the algorithm by letting $\mu_1 = \lambda_1 = 0$. We employ two separate learning rates η_{B} and η_{R}, which will be specified in Lemma 4.1. At each round $t \in [T]$, the dual regret minimizer updates the Lagrangian multipliers as $\lambda_{t+1} \leftarrow P_{[0,1/\rho]}(\lambda_t + \eta_{\mathrm{B}} g_t(x_t))$, and $\mu_{t+1} \leftarrow P_{\mathbb{R}_+}[\mu_t + \eta_{\mathrm{R}} h_t(x_t)]$, where P denotes the projection operator. The former update performs a gradient step and then projects the result on the

interval $[0, 1/\rho]$. This is possible because we know that playing the void action $\varnothing$ would satisfy the budget constraints by a margin of at least ρ, and therefore we can safely consider as the set of λ the interval $[0, 1/\rho]$ (Castiglioni et al., 2022a). On the other hand, in the update of μ we perform a grandient step and then ensure that the value is in $\mathbb{R}_+$. Since the decision maker does not know the feasibility parameter of ROI constraints, bounding μ becomes more complex, and we show how to approach this problem in Section 5.

Given a time interval $\mathcal{I} = [t_1, t_2]$, and $\delta \in [0, 1]$, we let

$$\mathcal{E}_{T,\delta}^{\mathcal{I}} \coloneqq \begin{cases} 2\sqrt{(t_2 - t_1)\log(2T/\delta)} & \text{if } \delta \in (0, 1] \\ 0 & \text{if } \delta = 0 \end{cases},$$

and, when clear from context, we drop the dependency on $\mathcal{I}$ to denote $\mathcal{E}_{T,\delta}^{[T]}$. Let $\mathcal{E}_T^{\mathrm{D,B}}$ be a term of order $O(T^{1/2}/\rho)$, and $\mathcal{E}_T^{\mathrm{D,R}}$ be a term of order $O(T^{1/2})$. The regret guarantees of the dual regret minimizer follow from standard results on OGD (see Hazan et al. (2016, Chapter 10)).

Lemma 4.1. *Let $\lambda_1 = \mu_1 = 0$. Then, OGD guarantees that, for any interval $\mathcal{I} = [t_1, t_2]$, it holds*

- $\sum_{t\in\mathcal{I}} \mu_t h_t(x_t) \leq (\mu - \mu_{t_1})^2/\eta_{\mathrm{R}} + \mathcal{E}_T^{D,R}$ *for* $\mu \in \mathbb{R}_+$,
- $\sum_{t\in\mathcal{I}} \lambda_t g_t(x_t) \leq \mathcal{E}_T^{D,B}$ *for all* $\lambda \in R_+$,

where learning rates are set as follows: $\eta_{\mathrm{B}} \coloneqq 1/\rho T^{1/2}$, *and* $\eta_{\mathrm{R}} \coloneqq 1/\left(6 + T^{1/2} + \mathcal{E}_T^{D,B} + 6\mathcal{E}_{T,\delta}^{\mathcal{I}} + 16\mathcal{E}_{T,\delta}^{P}\right)$.

The dependency on δ in the construction of η_{R} is resolved by Algorithm 1 taking δ as an input parameter, and the final guarantees of the framework are parametrized on δ. Next, we prove the following simple result characterizing the growth of the μ variables, which will be useful in the remainder of the paper (omitted proofs can be found in the appendix).

Lemma 4.2. *For all $t_1, t_2 \in [T]$, it holds*

$$\mu_{t_2} \geq \eta_{\mathrm{R}} \sum_{t' \in [t_1, t_2 - 1]} h_{t'}(x_{t'}) + \mu_{t_1}.$$

5. Bounding the Lagrange Multipliers

Previous work usually assumes knowledge, either exactly or via some upper bound, of the Slater's parameter α (Balseiro et al., 2020; Castiglioni et al., 2022a). This information is then used to bound the magnitude of dual multipliers, which is fundamental in order to obtain meaningful primal regret upper bounds since the magnitude of ℓ_t^{P} depends on dual multipliers. In our setting, the decision maker has no knowledge of the gap that the strictly feasible solution guarantees for the ROI constraint, which renders the traditional approach to bound μ_t not viable. We show that, even without a priori information on α, the primal-dual framework endowed with weakly adaptive regret minimizers guarantees that, with high probability, the Lagrange multiplier μ_t is bounded by $2/\alpha$ throughout the entire time horizon. We start by providing a general condition that we will prove to be satisfied both in the stochastic and adversarial setting.

Definition 5.1 (δ-safe policy)**.** Given $\delta \in (0, 1]$, a policy $\boldsymbol{\pi}^\circ$ is *δ-safe* if, for any interval $\mathcal{I} \coloneqq [t_1, t_2]$, with $t_1, t_2 \in [T]$, $t_1 < t_2$, it holds

$$\sum_{t\in\mathcal{I}} \lambda_t g_t(\boldsymbol{\pi}^\circ) + \mu_t h_t(\boldsymbol{\pi}^\circ) \leq (\mu_{\mathcal{I}} + 1/\alpha)\mathcal{E}_{T,\delta}^{\mathcal{I}} - \alpha \sum_{t\in\mathcal{I}} \mu_t,$$

where $\mu_{\mathcal{I}}$ is the largest multiplier in the interval $\mathcal{I}$.

A safe policy gives to the primal regret minimizer a way to limit the realized penalties imposed by the dual regret minimizer. In particular, we can show that if the dual regret minimizer increased the value of Lagrange multipliers μ_t too much, then the primal regret minimizer could "fight back" by playing the safe policy $\boldsymbol{\pi}^\circ$, thereby preventing the dual player from being no-regret. Indeed, the next result shows that whenever there exists a safe policy the Lagrange multipliers must be bounded.

Theorem 5.2. *If there exists a safe policy and the primal regret minimizer has regret at most $M_{\mathcal{I}}^2 \mathcal{E}_{T,\delta}^{P}$ for any time interval $\mathcal{I}$, then the Lagrange multipliers μ_t are such that $\mu_t \leq 2/\alpha$ for each $t \in [\tau]$.*

Then, we show that both in the stochastic and in the adversarial setting there exists a safe policy w.h.p., which implies that w.h.p. the Lagrange multipliers are bounded.

Lemma 5.3. *If inputs (f_t, c_t) are drawn i.i.d. from $\mathcal{P}$, and there exists a policy $\boldsymbol{\pi}$ such that $\mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}) \leq -\alpha$ and $\mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}) \leq -\alpha$, then there exists a δ-safe policy with probability at least $1 - \delta$.*

Lemma 5.4. *If inputs (f_t, g_t) are selected adversarially, and there exists a policy $\boldsymbol{\pi}$ such that $g_t(\boldsymbol{\pi}) \leq -\alpha$ and $h_t(\boldsymbol{\pi}) \leq -\alpha$ for each $t \in [T]$, then there exists a δ-safe policy for any $\delta \in [0, 1]$.*

6. Regret and Violations Guarantees

In this section, we describe guarantees for the stochastic and adversarial setting provided by Algorithm 1 equipped with a weakly adaptive primal and dual algorithm. Interestingly, we prove best-of-both-worlds guarantees through a unified argument which captures both cases. This is not the case in previous work, where the analysis of the stochastic case typically requires to study convergence to a Nash equilibrium of the *expected Lagrangian game* (Immorlica et al., 2022), which is not well defined in the adversarial setting.

We introduce the following event that holds w.h.p., most of our results will hold deterministically given this event.

Definition 6.1. We denote with $\mathbf{E}$ the event in which Algorithm 1 satisfies the following conditions: i) the primal

regret minimizer has regret upper bounded by $(3/\alpha+1)\mathcal{E}^{\mathrm{P}}_{T,\delta}$ for all time intervals $\mathcal{I}$, and ii) the dual multipliers for the ROI constraint are such that $\mu_t \leq 2/\alpha$ for each $t \in [T]$.

By applying a union bound to the events of Equation (4.1) and Lemma 5.3 or Lemma 5.4 (depending on the input model), we can use Theorem 5.2 to get the following result.

Lemma 6.2. *Event* $\mathbf{E}$ *holds with probability at least* $1-2\delta$.

We start by observing that the cumulative violation of ROI constraints must be sublinear in T with high probability, under both input models. This is a direct consequence of properties of the dual regret minimizer (see Section 4.4), and of the bound on dual multipliers implied by Lemma 6.2.

Lemma 6.3. *If* $\mathbf{E}$ *holds, then*

$$\sum_{t\in[\tau]} h_t(\boldsymbol{\pi}_t) \leq 1 + 2/(\eta_{\mathrm{R}}\alpha).$$

Then, we define the following class of policies.

Definition 6.4 ($(\delta, q, \texttt{OPT})$-optimal policy)**.** Given $\delta \in (0,1]$, $q \in (0,1]$, a sequence of T inputs $\{(f_t, c_t)\}_{t=1}^T$, and the value of a baseline $\texttt{OPT}$, we say that a policy $\boldsymbol{\pi}$ is $(\delta, q, \mathit{OPT})$-*optimal*, if

$$\text{i)} \sum_{t\in[T]} f_t(\boldsymbol{\pi}) \geq q \cdot T \cdot \texttt{OPT} - \mathcal{E}_{T,\delta}, \text{ and}$$

$$\text{ii)} \sum_{t\in[t']} \lambda_t g_t(\boldsymbol{\pi}) + \mu_t h_t(\boldsymbol{\pi}) \leq \left(\mu_{[t']} + \frac{1}{\alpha}\right)\mathcal{E}_{T,\delta}$$

for each $t' \in [T]$, where $\mu_{[t']}$ is the largest multiplier μ_t observed up to t'.

A $(\delta, q, \texttt{OPT})$-optimal policy guarantees a reward which is a fraction q of the reward of the baseline up to a sublinear term, and guarantees that the cumulative value of the penalty due to the Lagrangian relaxation is vanishing in time.

First, we need the following result that holds both in the stochastic and adversarial setting.

Lemma 6.5. *Algorithm 1 guarantees that*

$$\sum_{t\in[\tau]} \lambda_t g(x_t) \geq T - \tau - 1/\rho - \mathcal{E}^{\mathrm{D},B}_T.$$

Given primal (resp., dual) regret minimizer with guarantees $\mathcal{E}^{\mathrm{P}}_{T,\delta}$ (resp., $\mathcal{E}^{\mathrm{D},\mathrm{R}}_T$ and $\mathcal{E}^{\mathrm{D},\mathrm{B}}_T$), and $\delta \in (0,1]$, we let

$$\mathcal{C}(T,\alpha,\delta) \coloneqq {}^1\!/_\alpha + ({}^3\!/_\alpha + 1)\big(\mathcal{E}^{\mathrm{P}}_{T,\delta} + \mathcal{E}_{T,\delta}\big) + \mathcal{E}^{\mathrm{D},\mathrm{R}}_T + \mathcal{E}^{\mathrm{D},\mathrm{B}}_T.$$

The existence of a $(\delta, q, \texttt{OPT})$-optimal policy implies the following bound with respect to the generic baseline $\texttt{OPT}$.

Lemma 6.6. *Suppose event* $\mathbf{E}$ *holds and that there exists a* $(\delta, q, \mathit{OPT})$*-optimal policy. Then,*

$$\textstyle\sum_{t\in[\tau]} f_t(x_t) \geq q T \mathit{OPT} - \mathcal{C}(T,\alpha,\delta).$$

Next, we show that a suitable $(\delta, q, \texttt{OPT})$-optimal policy exists w.h.p. both in the stochastic and adversarial setting.

Lemma 6.7. *In the stochastic setting, with probability at least* $1-2\delta$ *there exists a* $(\delta, 1, \mathit{OPT}_{\mathcal{P}})$*-optimal policy (where* $\mathit{OPT}_{\mathcal{P}}$ *is the optimal value of* $\texttt{LP}_{\mathcal{P}}$*).*

This is saying that, given a distribution $\mathcal{P}$, there exists with high probability a policy satisfying Definition 6.4. Stochasticity of the environment is used to prove that the solution to $\texttt{LP}_{\mathcal{P}}$ satisfies the second condition of Definition 6.4 for *all* $t \in [T]$. If we tried a similar approach in the adversarial setting, the solution to $\texttt{LP}_{\bar{\gamma}}$ would guarantee that the second condition is satisfied over the whole time horizon, but *not* necessarily at earlier time steps $t < T$. Moreover, feasibility in expectation has no implications on feasibility of a policy under the adversarial sequence $(\lambda_t, f_t, \mu_t, c_t)$, in which dual variables are optimized to "punish violations". However, it is possible to show the existence of a policy satisfying Definition 6.4 even in the adversarial setting via a different approach. We build a suitable convex combination between a strictly feasible policy $\boldsymbol{\pi}^\circ$ guaranteeing that constraints are satisfied by at least $\alpha > 0$ for each $t \in [T]$, and the optimal unconstrained policy $\boldsymbol{\pi}^*$ maximizing $\sum_{t\in[T]} f_t(\boldsymbol{\pi})$. The following lemma employs a policy $\hat{\boldsymbol{\pi}}$ such that, for all $x \in \mathcal{X}$, $\hat{\boldsymbol{\pi}}_x \coloneqq \boldsymbol{\pi}^\circ_x/(1+\alpha) + \alpha\boldsymbol{\pi}^*_x/(1+\alpha)$.

Lemma 6.8. *In the adversarial setting, there always exists a* $(0, \alpha/(1+\alpha), \mathit{OPT}_{\bar{\gamma}})$*-optimal policy.*

Now, we provide the overall guarantees of the dual-balancing primal-dual algorithm (Algorithm 1).

Theorem 6.9 (Stochastic setting)**.** *In the stochastic setting, for* $\delta > 0$*, with probability at least* $1 - 4\delta$ *Algorithm 1 guarantees*

$$T\mathit{OPT}_{\mathcal{P}} - \textstyle\sum_{t\in[T]} f_t(x_t) \leq \mathcal{C}(T,\alpha,\delta).$$

Moreover, we have the following guarantees on constraint violations: $\sum_{t\in[T]} h_t(x_t) \leq 1 + 2/(\eta_{\mathrm{R}}\alpha)$ *and* $\sum_{t\in[T]} c_t(x_t) \leq B$.

Proof. The regret upper bound holds since event $\mathbf{E}$ holds with probability at least $1-2\delta$ (Lemma 6.2), and by combining Lemma 6.7 and Lemma 6.6. The ROI constraint is upper bounded by Lemma 6.3, and the budget constraint is strictly satisfied by construction of Algorithm 1. □

Since the primal regret minimizer guarantees a high-probability primal regret upper bound of order $T^{1/2}$ (see

Equation (4.1)), then the cumulative regret and the cumulative ROI constraint violation of Theorem 6.9 are of order $\tilde{O}(\sqrt{T})$, while the budget constraint is strictly satisfied.

Analogously, by exploiting Lemma 6.8, we have the following guarantees for the adversarial setting.

Theorem 6.10 (Adversarial setting)**.** *Suppose the sequence of inputs $\gamma = (f_t, c_t)_{t=1}^T$ is selected by an oblivious adversary. Then, for $\delta > 0$, with probability at least $1 - 2\delta$, Algorithm 1 guarantees*

$$\frac{\alpha}{1+\alpha} OPT_{\bar{\gamma}} - \sum_{t\in[T]} f_t(x_t) \leq \mathcal{C}(T, \alpha, \delta).$$ [3]

Moreover, it holds that $\sum_{t\in[T]} h_t(x_t) \leq 1 + 2/(\eta_{\mathbb{R}}\alpha)$ and $\sum_{t\in[T]} c_t(x_t) \leq B$.

Our competitive ratio matches that of Castiglioni et al. (2022b).

Remark 6.11. In our analysis, we provide a slightly suboptimal competitive ratio with respect to the budget constraint. Indeed, in our proofs we only use $g(x) \leq 1$ while it always holds $g(x) \leq 1 - \rho$. To optimize the competitive ratio, we can set

$$\alpha := -\inf_{\boldsymbol{\pi}} \max_{t\in[T]} \max\{g_t(\boldsymbol{\pi})/1-\rho, h_t(\boldsymbol{\pi})\}.$$

Our result continues to hold, and provides a better dependency on the budget constraint. In the case in which there are only budget constraints, it yields the state-of-the-art $1/\rho$ competitive ratio of Castiglioni et al. (2022a)

7. Relaxing the Safe-Policy Assumption

In the adversarial setting, the usual assumption for recovering Slater's condition is that there exists a policy guaranteeing that constraints are satisfied by at least $\alpha > 0$ for each t (Chen et al., 2017; Yi et al., 2020; Castiglioni et al., 2022b). Our analysis, up to this point, made the same assumption (Assumption 3.1), except that, unlike those past works, we do not need to know the value of α. Now, we show that our analysis carries over with the following looser requirement.

Assumption 7.1. There exists a policy $\boldsymbol{\pi}^\circ \in \Pi$ such that, for each interval $\mathcal{I} = [t_1, t_2]$ with $t_2 - t_1 = k$, we have $\sum_{t\in\mathcal{I}} g_t(\boldsymbol{\pi}^\circ) \leq -\alpha k$ and $\sum_{t\in\mathcal{I}} h_t(\boldsymbol{\pi}^\circ) \leq -\alpha k$.

The traditional assumption of requiring a safe policy for *each* t would require the decision maker to have an action yielding expected ROI strictly above their target for each round t. This may not hold in practice. For example, in the case of repeated ad auctions, if we assume one ad placement is being allocated at each t, then the agent would be priced out by other bidders for at least some time steps. Next, we show that if the size of the intervals k is not too big (*i.e.,* if there exists a "safe" policy frequently enough), there exist the following policies.

[3] The same guarantees would hold with respect to the optimal unconstrained policy maximizing $\sum f_t(\boldsymbol{\pi})$.

Algorithm 2 Primal regret minimizer.

Input: parameters $\eta > 0, \xi > 0, \sigma > 0$
Initialization: $[0,1]^{n\times m} \ni \boldsymbol{w}_1 \leftarrow \mathbf{1}$
for $t = 1, 2, \ldots, T$ **do**
- **Observe** valuation $v_t \in \mathcal{V}$
- **Set** $\boldsymbol{\pi}(v_t)_x \leftarrow w_{t,v_t,x} / \sum_{x'\in\mathcal{X}} w_{t,v_t,x'}, \ \forall x \in \mathcal{X}$
- **Bid** $x_t \sim \boldsymbol{\pi}(v_t)$
- **Observe** loss $\ell_t^{\mathbb{P}}(x_t)$
- $\tilde{\ell}_t^{\mathbb{P}}(x) \leftarrow \ell_t^{\mathbb{P}}(x)\mathbb{1}[x = x_t]/(\boldsymbol{\pi}(v_t)_x + \xi) \ \forall x \in \mathcal{X}$
- For each $x \in \mathcal{X}$, set $w_{t+1,v_t,x} \leftarrow (1-\sigma) w_{t,v_t,x} \cdot e^{-\eta\tilde{\ell}_t^{\mathbb{P}}(x)} + \frac{\sigma}{m}\sum_{x'\in\mathcal{X}} w_{t,v_t,x'} \cdot e^{-\eta\tilde{\ell}_t^{\mathbb{P}}(x')}$

end for

Lemma 7.2. *Suppose Assumption 7.1 holds with $k < \mathcal{E}_{T,\delta}/(2T\eta_{\mathbb{B}})$. Then, for $\delta > 0$, there exists a δ-safe and a $(\delta, \alpha/(1+\alpha), OPT_{\bar{\gamma}})$-optimal policy.*

This allows us to balance the tightness of the required assumption with the final regret guarantees, by suitably choosing the learning rates rates $\eta_{\mathbb{B}}$ and $\eta_{\mathbb{R}}$. When Assumption 7.1 holds for $k = \log T$, we recover exactly the bounds of Theorem 6.10. As a further example, if $k = T^{1/4}$, then we can obtain regret guarantees of order $\tilde{O}(T^{3/4})$ by setting $\eta_{\mathbb{B}} = O(T^{-3/4})$ and by suitably updating the definition of $\mathcal{E}_{T,\delta}$. In the context of ad auctions, this allows us to make the milder assumption that the bidder sees an auction with ROI > 0 at least every k steps, instead of at every step.

8. Bidding in Repeated Non-Truthful Auctions

In automatic bidding systems advertisers usually have to specify some parameters like their overall budget and their targeting criteria. Then a *proxy bidder* operated by the platform places bids on their behalf. A popular autobidding strategy is *value maximization* subject to budget and ROI constrains (Auerbach et al., 2008; Golrezaei et al., 2021; Deng et al., 2023). Recently, many advertising platforms have been transitioning from the second-price auction format toward a first-price format (see, *e.g.,* (Bigler, 2019; Wong, 2021)), which is *not* truthful. In this context, existing results for truthful auctions cannot be applied (Balseiro & Gur, 2019; Feng et al., 2023).

We show that our framework can be used to manage bidding in repeated non-truthful auctions under budget and ROI constraints. We will focus on the case of repeated first-price auctions, and we will make the simplifying assumption of having a finite set of possible valuations and bids. In Appendix G, we provide further details on this application.

Set-up. At each round $t \in [T]$, the bidder observes their valuation v_t extracted from a finite set $\mathcal{V} \subset [0,1]$ of n possible valuations. The set $\mathcal{X} \subset [0,1]$ is a set of m possible bids. Let β_t be the highest-competing bid at time t. In the value-maximizing utility model for each t we have $f_t(x_t) \coloneqq v_t \mathbb{1}[x_t \geq \beta_t]$ (Babaioff et al., 2021; Balseiro et al., 2021), and the cost function is $c_t(x_t) \coloneqq x_t \mathbb{1}[x_t \geq \beta_t]$, where the indicator function specifies whether the bidder won the auction at time t. We extend the definition of policies from Section 3 to model *randomized bidding policies*. Each $\boldsymbol{\pi} \in \Pi$ is now a mapping $\boldsymbol{\pi} : \mathcal{V} \to \Delta_{\mathcal{X}}$. We denote by $\boldsymbol{\pi}(v)_x$ the probability of selecting x under valuation v.

Primal regret minimizer. Our primal regret minimizer is based on the EXP3-SIX algorithm by Neu (2015) and it is described in Algorithm 2. At each round t, the algorithm maintains a set of weights $\boldsymbol{w}_t \in [0,1]^{n \times m}$. The probability of playing x under valuation v_t is proportional to the weight $w_{t,v_t,x}$. After drawing x_t, the algorithm observes $\ell_t^{\mathrm{P}}(x_t)$ and builds the estimated loss $\tilde{\ell}_t^{\mathrm{P}}$, where $\xi > 0$ is the implicit exploration term. Then, the update of weights $\boldsymbol{w}$ is inspired by the Fixed Share algorithm by Herbster & Warmuth (1998). We start by showing Equation (4.1) holds in the single-valuation setting.

Theorem 8.1. *Let $n = 1$, $\eta \coloneqq 1/\sqrt{mT}$, $\xi \coloneqq 1/(2\sqrt{mT})$, $\sigma \coloneqq 1/T$. For any $\delta > 0$,* EXP3-SIX *guarantees that, w.p. at least $1 - \delta$, for any interval $\mathcal{I}$, and for any $x \in \mathcal{X}$,*

$$\textstyle\sum_{t \in \mathcal{I}} (\ell_t^{\mathrm{P}}(x_t) - \ell_t^{\mathrm{P}}(x)) \leq O\Big(\sqrt{mT} \log\big(\frac{mT}{\delta}\big)\Big).$$

Then, if we instantiate one independent instance of EXP3-SIX for each valuation in $\mathcal{V}$ with the choice of parameters of Theorem 8.1, we have that for any time interval $\mathcal{I}$ the regret accumulated by Algorithm 2 over $\mathcal{I}$ is upper bounded by $M_{\mathcal{I}}^2 \sqrt{n}\, \mathcal{E}_{T,\delta}^{\mathrm{P}}$ with probability at least $1 - n\delta$ (see Lattimore & Szepesvári (2020, Chapter 18.4)). It follows that Equation (4.1) is satisfied and the guarantees of Theorems 6.9 and 6.10 readily apply to the problem of bidding in repeated first-price auctions under budget and ROI constraints.

Acknowledgements

MC and AC 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). 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. CK is supported by the Office of Naval Research awards N00014-22-1-2530 and N00014-23-1-2374, and the National Science Foundation awards IIS-2147361 and IIS-2238960.

Impact Statement

This paper presents theoretical results and has the goal of advancing the field of Machine Learning. There are no potential societal consequences of our work which we feel must be highlighted here.

References

Adamskiy, D., Koolen, W. M., Chernov, A., and Vovk, V. A closer look at adaptive regret. *The Journal of Machine Learning Research*, 17(1):706–726, 2016.

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.

Ai, R., Wang, C., Li, C., Zhang, J., Huang, W., and Deng, X. No-regret learning in repeated first-price auctions with budget constraints. *arXiv preprint arXiv:2205.14572*, 2022.

Akbarpour, M. and Li, S. Credible mechanisms. In *EC*, pp. 371, 2018.

Auerbach, J., Galenson, J., and Sundararajan, M. An empirical analysis of return on investment maximization in sponsored search auctions. In *Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, pp. 1–9, 2008.

Babaioff, M., Cole, R., Hartline, J., Immorlica, N., and Lucier, B. Non-quasi-linear agents in quasi-linear mechanisms. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, 2021.

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., 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.

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

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

Balseiro, S. R., Deng, Y., Mao, J., Mirrokni, V. S., and Zuo, S. The landscape of auto-bidding auctions: Value versus utility maximization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 132–133, 2021.

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

Bernasconi, M., Castiglioni, M., Celli, A., and Fusco, F. Bandits with replenishable knapsacks: the best of both worlds. *arXiv preprint arXiv:2306.08470*, 2023.

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.

Bigler, J. Rolling out first price auctions to google ad manager partners. https://tinyurl.com/mvpfc97n, 2019. Accessed: 2024-01-24.

Borgs, C., Chayes, J., Immorlica, N., Jain, K., Etesami, O., and Mahdian, M. Dynamics of bid optimization in online advertisement auctions. In *Proceedings of the 16th international conference on World Wide Web*, pp. 531–540, 2007.

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.

Castiglioni, M., Celli, A., and Kroer, C. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning, ICML 2022*, pp. 2767–2783, 2022a.

Castiglioni, M., Celli, A., Marchesi, A., Romano, G., and Gatti, N. A unifying framework for online optimization with long-term constraints. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33589–33602, 2022b.

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

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 980–988, 2012.

Chen, T. and Giannakis, G. B. Bandit convex optimization for scalable and dynamic iot management. *IEEE Internet of Things Journal*, 6(1):1276–1286, 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.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 1405–1411. PMLR, 2015.

Deng, Y., Golrezaei, N., Jaillet, P., Liang, J. C. N., and Mirrokni, V. Multi-channel autobidding with budget and roi constraints. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Despotakis, S., Ravi, R., and Sayedi, A. First-price auctions in online display advertising. *Journal of Marketing Research*, 58(5):888–907, 2021.

Devanur, N. R., Jain, K., Sivan, B., and Wilkens, C. A. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 29–38. ACM, 2011.

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.

Feng, Z., Padmanabhan, S., and Wang, D. Online bidding algorithms for return-on-spend constrained advertisers. In *Proceedings of the ACM Web Conference 2023*, pp. 3550–3560, 2023.

Golrezaei, N., Lobel, I., and Paes Leme, R. Auction design for roi-constrained buyers. In *Proceedings of the Web Conference 2021*, pp. 3941–3952, 2021.

Golrezaei, N., Jaillet, P., Liang, J. C. N., and Mirrokni, V. Pricing against a budget and roi constrained buyer. In *International Conference on Artificial Intelligence and Statistics*, pp. 9282–9307. PMLR, 2023.

Han, Y., Zhou, Z., Flores, A., Ordentlich, E., and Weissman, T. Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*, 2020a.

Han, Y., Zhou, Z., and Weissman, T. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*, 2020b.

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

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

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

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022.

Jenatton, R., Huang, J., and Archambeau, C. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pp. 402–411. PMLR, 2016.

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

Kumar, R. and Kleinberg, R. Non-monotonic resource utilization in the bandits with knapsacks problem. *Advances in Neural Information Processing Systems*, 35: 19248–19259, 2022.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liakopoulos, N., Destounis, A., Paschos, G., Spyropoulos, T., and Mertikopoulos, P. Cautious regret minimization: Online optimization with long-term budget constraints. In *International Conference on Machine Learning*, pp. 3944–3952. PMLR, 2019.

Lobos, A., Grigas, P., and Wen, Z. Joint online learning and decision-making via dual mirror descent. In *International Conference on Machine Learning*, pp. 7080–7089. PMLR, 2021.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pp. 1739–1776. PMLR, 2018.

Mahdavi, M., Jin, R., and Yang, T. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1):2503–2528, 2012.

Mahdavi, M., Yang, T., and Jin, R. Stochastic convex optimization with multiple objectives. pp. 1115–1123, 2013.

Mannor, S., Tsitsiklis, J. N., and Yu, J. Y. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.

Nedelec, T., Calauzènes, C., El Karoui, N., Perchet, V., et al. Learning in repeated auctions. *Foundations and Trends® in Machine Learning*, 15(3):176–334, 2022.

Nedić, A. and Ozdaglar, A. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

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

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

Paes Leme, R., Sivan, B., and Teng, Y. Why do competitive markets converge to first-price auctions? In *Proceedings of The Web Conference 2020*, pp. 596–605, 2020.

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

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

Sun, W., Dey, D., and Kapoor, A. Safety-aware algorithms for adversarial contextual bandit. In *International Conference on Machine Learning*, pp. 3280–3288. PMLR, 2017.

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.

Wang, Q., Yang, Z., Deng, X., and Kong, Y. Learning to bid in repeated first-price auctions with budgets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 36494–36513, 2023.

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.

Weed, J., Perchet, V., and Rigollet, P. Online learning in repeated auctions. In *Conference on Learning Theory*, pp. 1562–1583. PMLR, 2016.

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

Wong, M. Moving adsense to a first-price auction. https://blog.google/products/adsense/our-move-to-a-first-price-auction/, 2021. Accessed: 2024-01-24.

Yi, X., Li, X., Xie, L., and Johansson, K. H. Distributed online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Signal Processing*, 68:731–746, 2020.

Yu, H., Neely, M., and Wei, X. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A. Further Related Works

We survey the most relevant works with respect to ours. For further background on online learning the reader can refer to the monograph by Cesa-Bianchi & Lugosi (2006).

1) Bandits with Knapsacks. The stochastic *Bandits with Knapsacks* (BwK) framework was introduced and first solved by Badanidiyuru et al. (2018). Other regret-optimal algorithms for stochastic BwK have been proposed by Agrawal & Devanur (2019), and by Immorlica et al. (2022). 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). Moreover, the BwK framework has been employed to model various applications with budget/supply constraints such as, for example, dynamic pricing (Besbes & Zeevi, 2009; 2012; Wang et al., 2014), dynamic procurement (Badanidiyuru et al., 2012), and dynamic ad allocation (Combes et al., 2015; Balseiro & Gur, 2019). The *Adversarial Bandits with Knapsacks* (ABwK) setting was first studied by Immorlica et al. (2022), who proved a $O(m \log T)$ competitive ratio for the case in which the sequence of rewards and costs is chosen by an oblivious adversary. Immorlica et al. (2022) also show that no algorithm can achieve a competitive ratio better than $O(\log T)$ on all problem instances, even in instances with only two arms and a single resource. Recently, Kesselheim & Singla (2020) refined the analysis for the general ABwK setting to obtain an $O(\log m \ \log T)$ competitive ratio. They also prove that such competitive ratio is optimal up to constant factors. Moreover, Castiglioni et al. (2022a) proved a constant-factor competitive ratio in the regime $B = \Omega(T)$. We mention that further results have been obtained in the simplified setting with one constrained resource (Rangi et al., 2019; Tran-Thanh et al., 2010; 2012). All the works mentioned in this paragraph can only handle packing constraints (*e.g.,* budget constraints). They cannot handle ROI constraints, and they need perfect knowledge of the feasibility parameter α.

2) Online packing problems. Various well-known online packing problems can be seen as special cases of ABwK, with a more permissive feedback model which allows the decision maker to observe the full feedback before choosing an action (see, e.g., Buchbinder & Naor (2009); Devanur et al. (2011)). In online packing settings, since the decision maker is endowed with more information at the time of taking decisions, it is possible to derive $O(\log T)$ competitive ratio guarantees against the optimal dynamic policy. In the context of online allocation problems with fixed per-iteration budget, Balseiro et al. (2020; 2022) propose a class of algorithms which attain asymptotically optimal performance in the stochastic case, and they attain an asymptotically optimal (parametric) constant-factor competitive ratio when the inputs are adversarial. In their setting, as we already mentioned, 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. Even in this case, these works cannot handle problems with ROI-constrained decision makers, since they can handle only packing constraints, and require knowledge of the feasibility parameter.

3) Online convex optimization with time-varying constraints. Another line of related work concerns online convex optimization with time-varying constraints (see, *e.g.*, (Mahdavi et al., 2012; 2013; Jenatton et al., 2016; Neely & Yu, 2017; Chen & Giannakis, 2018; Castiglioni et al., 2022b)), 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. In contrast, in our setting, budget constraints apply in all rounds. Moreover, guarantees are usually provided either for stochastic constraints (Yu et al., 2017; Wei et al., 2020), or for adversarial constraints (Mannor et al., 2009; Sun et al., 2017; Liakopoulos et al., 2019), typically by employing looser notions of regret. In contrast, our framework will provide best-of-both-worlds guarantees. Moreover, these frameworks typically require perfect knowledge of α ad Assumption 3.1 to hold, while our framework relaxes both assumptions.

4) Bidding in repeated auctions. The problem of online bidding in repeated auctions has been extensively studied using online learning approaches (see, *e.g.,* Borgs et al. (2007); Weed et al. (2016); Nedelec et al. (2022)). In particular, online bidding under budget constraints has been studied in various settings. Balseiro & Gur (2019) and Ai et al. (2022) focus on utility-maximizing agents with one resource-consumption constraint. In the context of online allocation problems with an arbitrary number of constraints, Balseiro et al. (2020; 2022) propose a class of primal-dual algorithms attaining asymptotically optimal performance in the stochastic and adversarial case. 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 substantially different from ours. In particular, their framework cannot handle non-truthful repeated auctions. Recent works have also examined settings similar

to ours, involving bidders with constraints on their budget and ROI. The framework by Feng et al. (2023) can handle ROI and "hard" budget constraints, but crucially relies on truthfulness of second-price auctions, and on the stochasticity of the environment. The framework by Castiglioni et al. (2022b) allows for general "soft" constraints under both stochastic and adversarial inputs. Their framework cannot be applied in our setting for three reasons: i) we have hard budget constraints, ii) we don't make the stringent assumption of knowing the parameter α beforehand, and iii) we relax the assumption of having one strictly feasible solution for each round in the adversarial setting. Finally, Golrezaei et al. (2023) studies the dynamic pricing problem faced by a seller who repeatedly sells items to a single budget and ROI constrained buyer.

B. Further Details on the Example of Section 4.2

We have to set the length of $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$ so that the primal and dual regret are ≤ 0, while the constraint violations are $\Omega(T)$. We observe that, by construction, the two candidate actions for being the best in hindsight in the primal problem are bidding either 0 or $1/2$. Therefore, we start by computing the primal regret with respect those two actions. For simplicity, we drop the rescaling factor from the definition of ℓ_t^{P} since that is only needed for technical reasons in the new construction, and write:

$$\ell_t^{\mathrm{P}}(b_t) = -f_t(b_t) + \mu_t h_t(b_t) = (c_t(b_t) - v_t) + \mu_t \left(2c_t(b_t) - v_t\right).$$

Then, the regret with respect to bid $1/2$ is

$$\mathcal{E}_T^{\mathrm{P}}(1/2) = \sum_{t=1}^{T} (\ell_t^{\mathrm{P}}(b_t) - \ell_t^{\mathrm{P}}(1/2)) = -\frac{|\mathcal{I}_1|}{2} - \frac{|\mathcal{I}_2|}{16} + 16|\mathcal{I}_3| + \frac{|\mathcal{I}_1|}{2} - \frac{|\mathcal{I}_2|}{4} - \frac{49|\mathcal{I}_3|}{4} = -\frac{5|\mathcal{I}_2|}{16} + \frac{15|\mathcal{I}_3|}{4},$$

and the regret with respect to bid 0 is

$$\mathcal{E}_T^{\mathrm{P}}(0) = \sum_{t=1}^{T} (\ell_t^{\mathrm{P}}(b_t) - \ell_t^{\mathrm{P}}(0)) = -\frac{|\mathcal{I}_1|}{2} - \frac{|\mathcal{I}_2|}{16} + 16|\mathcal{I}_3| + \frac{|\mathcal{I}_1|}{16} + \frac{|\mathcal{I}_2|}{16} + \frac{17|\mathcal{I}_3|}{16} = -\frac{7|\mathcal{I}_1|}{16} + \frac{273|\mathcal{I}_3|}{16}.$$

We observe that $\mathcal{E}_T^{\mathrm{P}}(0) \leq 0$ for $|\mathcal{I}_1|$ big enough. Interval $\mathcal{I}_1$ can arbitrarily long since during the first phase the ROI violation is 0, so it does not impact on the constraints. In particular, we can set $|\mathcal{I}_1| = 39|\mathcal{I}_3|$. Moreover, we observe that $\mathcal{E}_T^{\mathrm{P}}(1/2) \leq 0$ whenever $|\mathcal{I}_2| \geq 12|\mathcal{I}_3|$. By setting $|\mathcal{I}_2| = 12|\mathcal{I}_3|$, we obtain $|\mathcal{I}_3| = T/52$.

The sequence of dual multipliers μ_t depicted in Table 1 guarantees no-regret on the dual problem since the dual player is exactly best responding to the primal actions: $\mu_t = 0$ when violations are ≤ 0, and $\mu_t = 1/\alpha = 16$ when violations are strictly positive. We observe that in this example $\alpha = \frac{1}{16}$ since bidding 0 yields a strict feasibility gap of at least $1/16$. Therefore, as expected, the two best-response actions of the dual player are the two extreme points of the interval $[0, 1/\alpha]$ (see Section 4.1).

Now, we can write the cumulative violation as a function of T. We have

$$\sum_{t=1}^{T} h_t(b_t) = \sum_{t=1}^{T} (2c_t(b_t) - v_t) = -\frac{1}{16}|\mathcal{I}_2| + |\mathcal{I}_3| = \frac{1}{208}T = \Omega(T).$$

Therefore, ROI constraint violations grow linearly in T, thereby violating one of our desiderata.

C. Proofs for Section 4

Lemma 4.2. *For all $t_1, t_2 \in [T]$, it holds*

$$\mu_{t_2} \geq \eta_{\mathrm{R}} \sum_{t' \in [t_1, t_2 - 1]} h_{t'}(x_{t'}) + \mu_{t_1}.$$

Proof. We prove the result by induction. Fix a starting point $t_1 \in [T]$. First, it's easy to see that the result holds for $t_2 = t_1$.

Then, suppose that the statement holds for round $t_2 = t$. Then,

$$\begin{aligned}\mu_{t+1} &= [\mu_t + \eta_{\mathbb{R}}\, h_t(x_t)]^+ \\ &\geq \mu_t + \eta_{\mathbb{R}}\, h_t(x_t) \\ &\geq \eta_{\mathbb{R}} \sum_{t' \in [t_1, t-1]} h_{t'}(x_{t'}) + \mu_{t_1} + \eta_{\mathbb{R}}\, h_t(x_t) \\ &= \eta_{\mathbb{R}} \sum_{t' \in [t_1, t]} h_{t'}(x_{t'}) + \mu_{t_1},\end{aligned}$$

where $[x]^+ \coloneqq \max\{x, 0\}$. This implies that the statement holds for $t_2 = t + 1$. This concludes the proof. □

D. Proofs for Section 5

Theorem 5.2. *If there exists a safe policy and the primal regret minimizer has regret at most $M_{\mathcal{I}}^2 \mathcal{E}^{\mathrm{P}}_{T,\delta}$ for any time interval $\mathcal{I}$, then the Lagrange multipliers μ_t are such that $\mu_t \leq 2/\alpha$ for each $t \in [\tau]$.*

Proof. We consider two cases.

Case 1: $\alpha \leq 10/\sqrt{T}$. By construction of the dual regret minimizer, and by the choice of $\eta_{\mathbb{R}}$, the dual variable μ_t can reach at most value $\eta_{\mathbb{R}} T \leq \sqrt{T}/16$. Therefore, we have $\mu_t \leq \sqrt{T}/16 < 2/\alpha$.

Case 2: $\alpha > 10/\sqrt{T}$. Let $\mathcal{I} = [t_1, t_2]$, with $t_1, t_2 \in [T]$, $t_1 \leq t_2$. Moreover, assume that there exists a safe policy $\boldsymbol{\pi}^\circ$. We show that, if the Lagrangian multiplier μ_t is greater than $2/\alpha$, we reach a contradiction.

Suppose, by contradiction, that there exists a round t_2 such that $\mu_{t_2} \geq 2/\alpha$. Let t_1 be the the first round such that $\mu_t \geq 1/\alpha$ for any $t \in [t_1, t_2]$. Notice that the structure of the dual regret minimizer (see Section 4.4) implies that

$$\mu_{t_1} \leq 1/\alpha + \eta_{\mathbb{R}} \quad \text{and} \quad \mu_{t_2} \leq 2/\alpha + \eta_{\mathbb{R}}, \tag{D.1}$$

since the dual losses are in $[-1, 1]$. Therefore, we can upperbound the primal loss function as $M_{[T]} \leq (1 + 4/\alpha)$ (*i.e.,* we use $\mu_t \leq 3/\alpha$). This implies that, for $m \geq 2$,

$$\eta M_{[T]} = \frac{1}{\sqrt{mT}}\left(1 + \frac{4}{\alpha}\right) < \frac{1}{\sqrt{mT}}\left(1 + \frac{2\sqrt{T}}{5}\right) \leq 1.$$

Therefore, the primal regret minimizer satisfies the bound on the adaptive regret of Equation (4.1). Then, by the no-regret property of the primal we get:

$$\begin{aligned}&\sum_{t \in \mathcal{I}} (f_t(x_t) - \lambda_t g_t(x_t) - \mu_t h_t(x_t)) \\ &\qquad \geq \sum_{t \in \mathcal{I}} (f_t(\boldsymbol{\pi}^\circ) - \lambda_t g_t(\boldsymbol{\pi}^\circ) - \mu_t h_t(\boldsymbol{\pi}^\circ)) - M_{\mathcal{I}}^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} \\ &\qquad \geq \alpha \sum_{t \in \mathcal{I}} \mu_t - \left(\mu_{\mathcal{I}} + \frac{1}{\alpha}\right) \mathcal{E}^{\mathcal{I}}_{T,\delta} - M_{\mathcal{I}}^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} && \text{(by Definition 5.1)} \\ &\qquad \geq (t_2 - t_1) - \left(\mu_{[t_1, t_2 - 1]} + \frac{3}{\alpha} + \eta_{\mathbb{R}}\right) \mathcal{E}^{\mathcal{I}}_{T,\delta} - M_{\mathcal{I}}^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} && \text{(by Def. of } t_1 \text{ and Equation (D.1))} \\ &\qquad \geq (t_2 - t_1) - \left(\frac{5}{\alpha} + \eta_{\mathbb{R}}\right) \mathcal{E}^{\mathcal{I}}_{T,\delta} - \left(1 + \frac{2}{\alpha} + \frac{1}{\alpha}\right)^2 \mathcal{E}^{\mathrm{P}}_{T,\delta} && \text{(by Def. of } M_{\mathcal{I}} \text{ and } \ell_t^{\mathrm{P}}) \\ &\qquad \geq (t_2 - t_1) - \left(\frac{5}{\alpha} + \eta_{\mathbb{R}}\right) \mathcal{E}^{\mathcal{I}}_{T,\delta} - \frac{16}{\alpha^2} \mathcal{E}^{\mathrm{P}}_{T,\delta}. && \text{(D.2)}\end{aligned}$$

Since the Lagrangian multipliers μ_t is always at least $1/\alpha$ for $t \in [t_1, t_2]$, the dual regret minimizer never has to project over $\mathbb{R}_{\geq 0}$ during interval $[t_1, t_2]$. In particular, projecting the dual multiplier at t back onto $\mathbb{R}_{\geq 0}$ would yield $\mu_t = 0$. This cannot

happen for $t \in [t_1, t_2]$, since $\mu_t \geq 1/\alpha$. Then, since the dual regret minimizer does not perform any projection operation during $[t_1, t_2]$, we have that the statement of Lemma 4.2 holds with equality:

$$\mu_{t_2} = \eta_{\mathrm{R}} \sum_{t' \in [t_1, t_2 - 1]} h_{t'}(x_{t'}) + \mu_{t_1}.$$

Hence, by definition of t_2 and Equation (D.1),

$$\sum_{t' \in [t_1, t_2 - 1]} h_{t'}(x_{t'}) = \frac{\mu_{t_2} - \mu_{t_1}}{\eta_{\mathrm{R}}} \geq \frac{1}{\alpha \eta_{\mathrm{R}}} - 1.$$

Then, by the regret bound of the dual with respect to $\mu = \mu_t$ and $\lambda = 0$, we get

$$\begin{aligned}
\sum_{t \in [t_1, t_2 - 1]} (f_t(x_t) - \lambda_t g_t(x_t) - \mu_t h_t(x_t)) &\leq \sum_{t \in [t_1, t_2 - 1]} (f_t(x_t) - \mu_t h_t(x_t)) + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} \\
&\leq (t_2 - t_1) - \frac{1}{\alpha} \sum_{t \in [t_1, t_2 - 1]} h_t(x_t) + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} \\
&\leq (t_2 - t_1) - \frac{1}{\alpha^2 \eta} + \frac{1}{\alpha} + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}}. \qquad \text{(D.3)}
\end{aligned}$$

By putting Equation (D.2) and Equation (D.3) together we have that

$$(t_2 - t_1) - \frac{1}{\alpha^2 \eta_{\mathrm{R}}} + \frac{3}{\alpha} + 2 + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} \geq (t_2 - t_1) - \left(\frac{5}{\alpha} + \eta_{\mathrm{R}} \right) \mathcal{E}_{T,\delta}^{\mathcal{I}} - \frac{16}{\alpha^2} \mathcal{E}_{T,\delta}^{\mathrm{P}}.$$

We observe that in Lemma 4.1 we set $\eta_{\mathrm{R}} \coloneqq \left(6 + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} + 6\mathcal{E}_{T,\delta}^{\mathcal{I}} + 16\mathcal{E}_T^{\mathrm{P}} \right)^{-1}$. Then, from the inequality above we have

$$\frac{1}{\alpha^2 \eta_{\mathrm{R}}} \leq \frac{3}{\alpha} + 2 + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} + \left(\frac{5}{\alpha} + 1 \right) \mathcal{E}_{T,\delta}^{\mathcal{I}} + \frac{16}{\alpha^2} \mathcal{E}_{T,\delta}^{\mathrm{P}}.$$

However, we reach a contradiction since

$$\begin{aligned}
\frac{1}{\alpha^2 \eta_{\mathrm{R}}} &\geq \frac{4}{\alpha} + 2 + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} + \left(\frac{5}{\alpha} + 1 \right) \mathcal{E}_{T,\delta}^{\mathcal{I}} + \frac{16}{\alpha^2} \mathcal{E}_T^{\mathrm{P}} \\
&> \frac{3}{\alpha} + 2 + \mathcal{E}_T^{\mathrm{D,R}} + \mathcal{E}_T^{\mathrm{D,B}} + \left(\frac{5}{\alpha} + 1 \right) \mathcal{E}_{T,\delta}^{\mathcal{I}} + \frac{16}{\alpha^2} \mathcal{E}_T^{\mathrm{P}},
\end{aligned}$$

where we used the fact that $\alpha \in (0, 1]$ by assumption ($\alpha > 0$), and by boundedness of g_t and h_t for all $t \in [T]$. This concludes the proof. □

Lemma 5.3. *If inputs (f_t, c_t) are drawn i.i.d. from $\mathcal{P}$, and there exists a policy $\boldsymbol{\pi}$ such that $\mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}) \leq -\alpha$ and $\mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}) \leq -\alpha$, then there exists a δ-safe policy with probability at least $1 - \delta$.*

Proof. By the definition of α, there exists a policy $\boldsymbol{\pi}$ such that $\mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}) \leq -\alpha$ and $\mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}) \leq -\alpha$. Then, given a time interval $\mathcal{I} = [t_1, t_2]$, $t_1, t_2 \in [T]$, by appling the Azuma–Hoeffding inequality to the martingale difference sequence $W_1, \ldots, W_T$ with

$$W_t \coloneqq \lambda_t g_t(\boldsymbol{\pi}) + \mu_t h_t(\boldsymbol{\pi}) - \lambda_t \mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}) - \mu_t \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}),$$

we obtain that

$$\left| \sum_{t \in \mathcal{I}} W_t \right| \leq \left(\mu_{\mathcal{I}} + \frac{1}{\alpha} \right) \sqrt{2(t_2 - t_1) \log\left(\frac{2}{\delta} \right)}$$

holds with probability at least $1-\delta$. By applying a union bound we get that the inequalities for each time interval $\mathcal{I}$ hold simultaneously with probability at least $1-T^2\delta$. Let $\mathcal{E}^{\mathcal{I}}_{T,\delta} \coloneqq 2\sqrt{(t_2-t_1)\log\left(\frac{2T}{\delta}\right)}$ as per Definition 5.1. Then, with probability at least $1-\delta$ it holds

$$\begin{aligned}\sum_{t\in\mathcal{I}}(\lambda_t g_t(\boldsymbol{\pi})+\mu_t h_t(\boldsymbol{\pi})) &\leq \left(\mu_{\mathcal{I}}+\frac{1}{\alpha}\right)\mathcal{E}^{\mathcal{I}}_{T,\delta}+\sum_{t\in\mathcal{I}}(\lambda_t \mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi})+\mu_t \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi})) \\ &\leq \left(\mu_{\mathcal{I}}+\frac{1}{\alpha}\right)\mathcal{E}^{\mathcal{I}}_{T,\delta}-\alpha\sum_{t\in\mathcal{I}}(\lambda_t+\mu_t) \\ &\leq \left(\mu_{\mathcal{I}}+\frac{1}{\alpha}\right)\mathcal{E}^{\mathcal{I}}_{T,\delta}-\alpha\sum_{t\in\mathcal{I}}\mu_t.\end{aligned}$$

This concludes the proof. □

Lemma 5.4. *If inputs (f_t, g_t) are selected adversarially, and there exists a policy $\boldsymbol{\pi}$ such that $g_t(\boldsymbol{\pi}) \leq -\alpha$ and $h_t(\boldsymbol{\pi}) \leq -\alpha$ for each $t \in [T]$, then there exists a δ-safe policy for any $\delta \in [0,1]$.*

Proof. By assumption there exists a policy $\boldsymbol{\pi}$ such that $g_t(\boldsymbol{\pi}) \leq -\alpha$ and $h_t(\boldsymbol{\pi}) \leq -\alpha$ for each $t \in [T]$. Then, for each $t_1, t_2 \in [T]$, with $t_1 < t_2$, it holds $\sum_{t\in[t_1,t_2]}(\lambda_t g_t(\boldsymbol{\pi})+\mu_t h_t(\boldsymbol{\pi})) \leq -\alpha\sum_{t\in[t_1,t_2]}(\lambda_t+\mu_t) \leq -\alpha\sum_{t\in[t_1,t_2]}\mu_t$, which implies that $\boldsymbol{\pi}$ is δ-safe for any $\delta \in [0,1]$. □

E. Proofs for Section 6

Lemma 6.2. *Event* $\mathbf{E}$ *holds with probability at least* $1-2\delta$.

Proof. By Lemmas 5.4 and 5.3, we have that in both settings there exists a safe policy with probability at least $1-\delta$. Moreover, by Equation (4.1) with probability at least $1-\delta$ the regret of the primal is upperbounded by $M_{\mathcal{I}}\mathcal{E}^{\mathtt{P}}_T$ for each interval $\mathcal{I}=[t_1,t_2]$, $t_1, t_2 \in [T]$. Applying a union bound sufficies to show that the two events hold simultaneously with probability at least $1-2\delta$. Then, the statement directly follows from Theorem 5.2. □

Lemma 6.3. *If* $\mathbf{E}$ *holds, then*

$$\sum_{t\in[\tau]} h_t(\boldsymbol{\pi}_t) \leq 1 + 2/(\eta_{\mathtt{R}}\alpha).$$

Proof. By the definition of event $\mathbf{E}$ we have that $\mu_\tau \leq 2/\alpha$. Moreover, by Lemma 4.2 it holds that $\mu_\tau \geq \eta_{\mathtt{R}}\sum_{t\in[\tau-1]} h_t(\boldsymbol{\pi}_t)$. Hence, $\sum_{t\in[\tau]} h_t(\boldsymbol{\pi}_t) \leq \mu_\tau/\eta_{\mathtt{R}} + 1 \leq 2/(\eta_{\mathtt{R}}\alpha) + 1$. □

Lemma 6.5. *Algorithm 1 guarantees that*

$$\sum_{t\in[\tau]} \lambda_t g(x_t) \geq T - \tau - 1/\rho - \mathcal{E}^{D,B}_T.$$

Proof. We consider two cases.

- If $\tau = T$, then

$$\sum_{t\in[\tau]} \lambda_t g_t(x_t) \geq -\mathcal{E}^{\mathtt{D},\mathtt{B}}_T \geq T - \tau - \frac{1}{\rho} - \mathcal{E}^{\mathtt{D},\mathtt{B}}_T.$$

- Otherwise, if $\tau < T$,

$$\begin{aligned}\sum_{t\in[\tau]} \lambda_t g_t(x_t) &\geq \frac{1}{\rho}\sum_{t\in[\tau]} g_t(x_t) - \mathcal{E}_\tau^{\mathrm{D,B}} \\ &= \frac{1}{\rho}\sum_{t\in[\tau]} (c_t(x_t) - \rho) - \mathcal{E}_\tau^{\mathrm{D,B}} \\ &= \frac{1}{\rho}(B - 1 - \tau\rho) - \mathcal{E}_\tau^{\mathrm{D,B}} \\ &= \left(T - \tau - \frac{1}{\rho}\right) - \mathcal{E}_\tau^{\mathrm{D,B}}.\end{aligned}$$

  where the first inequality follows by the no-regret guarantee of the dual regret minimizer with respect to the fixed choice of $\lambda = 1/\rho$, and then we use the definition of g_t and the fact that τ is the time at which the budget is depleted, that is the round in which the available budget becomes strictly smaller than 1 (see Algorithm 1).

This concludes the proof. □

Lemma 6.6. *Suppose event* $\mathbf{E}$ *holds and that there exists a* $(\delta, q, \mathtt{OPT})$*-optimal policy. Then,*

$$\textstyle\sum_{t\in[\tau]} f_t(x_t) \geq qT\mathtt{OPT} - \mathcal{C}(T, \alpha, \delta).$$

Proof. Let $\boldsymbol{\pi}^*$ be a $(\delta, q, \mathtt{OPT})$-optimal policy. Then, we have that

$$\begin{aligned}\sum_{t\in[\tau]} f_t(x_t) &\geq \sum_{t\in[\tau]} (f_t(\boldsymbol{\pi}^*) - \lambda_t g_t(\boldsymbol{\pi}^*) - \mu_t h_t(\boldsymbol{\pi}^*) + \lambda_t g_t(x_t) + \mu_t h_t(x_t)) - \left(\frac{3}{\alpha} + 1\right)\mathcal{E}_{\tau,\delta}^{\mathrm{P}} \\ &\geq \sum_{t\in[\tau]} (f_t(\boldsymbol{\pi}^*) + \lambda_t g_t(x_t) + \mu_t h_t(x_t)) - \frac{3}{\alpha}\mathcal{E}_{\tau,\delta} - \left(\frac{3}{\alpha} + 1\right)\mathcal{E}_{\tau,\delta}^{\mathrm{P}} \\ &\geq \sum_{t\in[\tau]} f_t(\boldsymbol{\pi}^*) + \sum_{t\in[\tau]} \lambda_t g_t(x_t) - \frac{3}{\alpha}\mathcal{E}_{\tau,\delta} - \left(\frac{3}{\alpha} + 1\right)\mathcal{E}_{\tau,\delta}^{\mathrm{P}} - \mathcal{E}_\tau^{\mathrm{D,R}} \\ &\geq \sum_{t\in[\tau]} f_t(\boldsymbol{\pi}^*) + T - \tau - \frac{1}{\rho} - \frac{3}{\alpha}\mathcal{E}_{\tau,\delta} - \left(\frac{3}{\alpha} + 1\right)\mathcal{E}_{\tau,\delta}^{\mathrm{P}} - \mathcal{E}_\tau^{\mathrm{D,R}} - \mathcal{E}_T^{\mathrm{D,B}} \\ &\geq \sum_{t\in[T]} f_t(\boldsymbol{\pi}^*) - \frac{1}{\rho} - \frac{3}{\alpha}\mathcal{E}_{\tau,\delta} - \left(\frac{3}{\alpha} + 1\right)\mathcal{E}_{\tau,\delta}^{\mathrm{P}} - \mathcal{E}_\tau^{\mathrm{D,R}} - \mathcal{E}_T^{\mathrm{D,B}} \\ &\geq qT\mathtt{OPT} - \frac{1}{\rho} - \left(\frac{3}{\alpha} + 1\right)(\mathcal{E}_{T,\delta}^{\mathrm{P}} + \mathcal{E}_{T,\delta}) - \mathcal{E}_T^{\mathrm{D,R}} - \mathcal{E}_T^{\mathrm{D,B}},\end{aligned}$$

where the first inequality comes from the regret bound of the primal regret minimizer, the second follows by the definition of $(\delta, q, \mathtt{OPT})$-optimal policy, the third follows by the no-regret guarantee of the dual regret minimizer with respect to action $\mu = 0$, the fourth one follows from Lemma 6.5. Finally, the fifth inequality follows from the fact that $f_t(\cdot) \in [0, 1]$, and the last one is by definition of $(\delta, q, \mathtt{OPT})$-optimal policy. This proves our statement. □

Lemma 6.7. *In the stochastic setting, with probability at least* $1 - 2\delta$ *there exists a* $(\delta, 1, \mathtt{OPT}_{\mathcal{P}})$*-optimal policy (where* $\mathtt{OPT}_{\mathcal{P}}$ *is the optimal value of* $\mathtt{LP}_{\mathcal{P}}$*).*

Proof. Let $\boldsymbol{\pi}^*$ be an optimal solution to $\mathtt{LP}_\gamma$. We show that, with probability at least $1 - \delta$, the policy $\boldsymbol{\pi}^*$ is $(\delta, 1, \mathtt{OPT}_\gamma)$-optimal, proving the statement.

First, by Azuma–Hoeffding inequality we have that, for $t' \in [T]$, with probability at least $1 - \delta$

$$\sum_{t\in[t']} (\lambda_t g_t(\boldsymbol{\pi}^*) + \mu_t h_t(\boldsymbol{\pi}^*) - \lambda_t \mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}^*) - \mu_t \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}^*)) \leq \left(\mu_{[t']} + \frac{1}{\alpha}\right)\sqrt{2T\log\left(\frac{1}{\delta}\right)},$$

where $\mu_{[t']}$ is the largest dual multiplier μ_t observed up to t'. Notice that we cannot upper bound it righ away as $2/\alpha$ because here we are not requiring event $\mathbf{E}$ (Definition 6.1) to hold. Then, assuming $T > 2$, by taking a union bound over all possible rounds t', we get that the following inequality holds with probability at least $1 - \delta$ simultaneously for all $t' \in [T]$,

$$\sum_{t\in[t']} (\lambda_t g_t(\boldsymbol{\pi}^*) + \mu_t h_t(\boldsymbol{\pi}^*) - \lambda_t \mathbb{E}_{\mathcal{P}} g(\boldsymbol{\pi}^*) - \mu_t \mathbb{E}_{\mathcal{P}} h(\boldsymbol{\pi}^*)) \;\leq\; \left(\mu_{[t']} + \frac{1}{\alpha}\right)\sqrt{2T\log\left(\frac{T}{\delta}\right)} \;\leq\; \left(\mu_{[t']} + \frac{1}{\alpha}\right)\mathcal{E}_{T,\delta}.$$

Similarly, we can prove that

$$\left|\sum_{t\in[T]} (f_t(\boldsymbol{\pi}^*) - \mathbb{E}_{\mathcal{P}} f(\boldsymbol{\pi}^*))\right| \leq 2\sqrt{T\log\left(\frac{2T}{\delta}\right)} = \mathcal{E}_{T,\delta}$$

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

$$\sum_{t\in[T]} f_t(\boldsymbol{\pi}^*) \geq \sum_{t\in[T]} \mathbb{E}_{\mathcal{P}} f(\boldsymbol{\pi}^*) - \mathcal{E}^{\mathcal{I}}_{T,\delta} = \texttt{OPT}_\gamma - \mathcal{E}_{T,\delta}.$$

Assuming $T > 2$ and applying an union bound, the statement follows. □

Lemma 6.8. *In the adversarial setting, there always exists a $(0, \alpha/(1+\alpha), \texttt{OPT}_{\bar{\gamma}})$-optimal policy.*

Proof. Let $\boldsymbol{\pi}^\circ$ be a strictly feasible policy such that $\alpha = -\max_{t\in[T]} \max\{g_t(\boldsymbol{\pi}^\circ), h_t(\boldsymbol{\pi}^\circ)\}$, with $\alpha > 0$, and let $\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}\in\Pi} \sum_{t\in[T]} f_t(\boldsymbol{\pi})$ be an optimal unconstrained policy. It holds $\sum_{t\in[T]} f_t(\boldsymbol{\pi}^*) \geq T\texttt{OPT}_{\bar{\gamma}}$ since the optimal uncostrained policy is better than the optimal constrained policy, which is a solution to $\texttt{LP}_{\bar{\gamma}}$.

Then, consider the policy $\hat{\boldsymbol{\pi}}$ such that, for each $v \in \mathcal{V}, b \in \mathcal{B}$,

$$\hat{\boldsymbol{\pi}}(v)_b = \frac{1}{1+\alpha}\boldsymbol{\pi}^\circ(v)_b + \frac{\alpha}{1+\alpha}\boldsymbol{\pi}^*(v)_b,$$

where, given a policy $\boldsymbol{\pi}$, we denote by $\boldsymbol{\pi}(v)_b$ the probability of bidding b under valuation v.

At each iteration we have that both the budget and the ROI constraints are satisfied by the policy $\hat{\boldsymbol{\pi}}$ (in expectation with respect to $\hat{\boldsymbol{\pi}}$). Indeed, for each $t \in [T]$, we have thath $g_t(\hat{\boldsymbol{\pi}}) = \frac{1}{1+\alpha} g_t(\boldsymbol{\pi}^\circ) + \frac{\alpha}{1+\alpha} g_t(\boldsymbol{\pi}^*) \leq \frac{-\alpha}{1+\alpha} + \frac{\alpha}{1+\alpha} \leq 0$. Similarly, we can prove that for each $t \in [T]$ it holds $h_t(\hat{\boldsymbol{\pi}}) \leq 0$. Then, the policy $\hat{\boldsymbol{\pi}}$ satisfies the condition $\sum_{t\in[t']}(\lambda_t g_t(\hat{\boldsymbol{\pi}}) + \mu_t h_t(\hat{\boldsymbol{\pi}})) \leq 0$ for each $t' \in [T]$. Moreover,

$$\begin{aligned}
\sum_{t\in[T]} f_t(\hat{\boldsymbol{\pi}}) &= \sum_{t\in[T]} \left(\frac{1}{1+\alpha} f_t(\boldsymbol{\pi}^\circ) + \frac{\alpha}{1+\alpha} f_t(\boldsymbol{\pi}^*)\right) \\
&\geq \sum_{t\in[T]} \frac{\alpha}{1+\alpha} f_t(\boldsymbol{\pi}^*) \\
&\geq \frac{\alpha}{1+\alpha}\texttt{OPT}_{\bar{\gamma}},
\end{aligned}$$

which satisfies the first condition of Definition 6.4. This concludes the proof. □

F. Proofs for Section 7

Lemma 7.2. *Suppose Assumption 7.1 holds with $k < \mathcal{E}_{T,\delta}/(2T\eta_{\mathcal{B}})$. Then, for $\delta > 0$, there exists a δ-safe and a $(\delta, \alpha/(1+\alpha), \texttt{OPT}_{\bar{\gamma}})$-optimal policy.*

Proof. First, we need to show that there exists a δ-safe policy. In particular, we show that there exists a policy $\boldsymbol{\pi}^\circ$ such that, for any time interval $\mathcal{I} = [t_1, t_2]$, it holds

$$\sum_{t\in\mathcal{I}} (\lambda_t g_t(\boldsymbol{\pi}^\circ) + \mu_t h_t(\boldsymbol{\pi}^\circ)) \leq \mathcal{E}_{T,\delta} - \alpha \sum_{t\in\mathcal{I}} (\mu_t + \lambda_t). \tag{F.1}$$

To do that, we show that the interval $\mathcal{I}$ can be split in smaller intervals of length k, and for each of such smaller intervals $\mathcal{I}'$, it holds

$$\sum_{t\in\mathcal{I}'}(\lambda_t g_t(\boldsymbol{\pi}^\circ)+\mu_t h_t(\boldsymbol{\pi}^\circ)) \le 2k^2\eta_B - \alpha\sum_{t\in\mathcal{I}'}(\mu_t+\lambda_t).$$

We show that this holds for any $\mathcal{I}'$ of length k in Lemma F.1. Then, the cumulative sum on the original interval $\mathcal{I}$ is at most

$$\begin{aligned}\sum_{t\in\mathcal{I}}(\lambda_t g_t(\boldsymbol{\pi}^\circ)+\mu_t h_t(\boldsymbol{\pi}^\circ)) &\le \left\lceil\frac{|\mathcal{I}|}{k}\right\rceil\left(2k^2\eta_B - \alpha\sum_{t\in\hat{\mathcal{I}}}(\mu_t+\lambda_t)\right)\\ &\le 2Tk\eta_{\text{B}} - \alpha\sum_{t\in\mathcal{I}}(\mu_t+\lambda_t) \le \mathcal{E}_{T,\delta} - \alpha\sum_{t\in\mathcal{I}}(\mu_t+\lambda_t),\end{aligned}$$

where we set $\hat{\mathcal{I}}\in\arg\max_{\mathcal{I}':|\mathcal{I}'|=k}\sum_{t\in\mathcal{I}'}(\mu_t+\lambda_t)$. This shows that Equation (F.1) holds for any interval $\mathcal{I}$.

Then, we can show that a $(\delta,\alpha/(1+\alpha),\texttt{OPT}_{\bar{\gamma}})$-optimal policy exists. In particular, by defining a policy $\hat{\boldsymbol{\pi}}$ as in the proof of Lemma 6.8, we have

$$\begin{aligned}\sum_{t\in\mathcal{I}}(\lambda_t g_t(\hat{\boldsymbol{\pi}})+\mu_t h_t(\hat{\boldsymbol{\pi}})) &= \frac{1}{1+\alpha}\left(\sum_{t\in\mathcal{I}}(\lambda_t g_t(\boldsymbol{\pi}^\circ)+\mu_t h_t(\boldsymbol{\pi}^\circ))\right) + \frac{\alpha}{1+\alpha}\left(\sum_{t\in[t']}(\lambda_t g_t(\boldsymbol{\pi}^*)+\mu_t h_t(\boldsymbol{\pi}^*))\right)\\ &\le \mathcal{E}_{T,\delta} - \frac{\alpha}{1+\alpha}\sum_{t\in\mathcal{I}}(\mu_t+\lambda_t) + \frac{\alpha}{1+\alpha}\sum_{t\in\mathcal{I}}(\lambda_t+\mu_t)\\ &\le \mathcal{E}_{T,\delta} \le \frac{3}{\alpha}\mathcal{E}_{T,\delta},\end{aligned}$$

where the first inequality is by Equation (F.1). The first condition of Definition 6.4 can be shown to hold with the same steps of Lemma 6.8. This concludes the proof. □

Lemma F.1. *For any time interval $\mathcal{I}$ of length k, there exist a policy $\boldsymbol{\pi}^\circ\in\Pi$ for which it holds*

$$\sum_{t\in\mathcal{I}}(\lambda_t g_t(\boldsymbol{\pi}^\circ)+\mu_t h_t(\boldsymbol{\pi}^\circ)) \le 2k^2\eta_B - \alpha\sum_{t\in\mathcal{I}}(\mu_t+\lambda_t).$$

Proof. Given an interval $\mathcal{I}$ of lenght k, let $(\bar{\lambda},\bar{\mu})$ be the largest Lagrangian multipliers in the interval $\mathcal{I}$, and let $(\underline{\lambda},\underline{\mu})$ be the smallest Lagrangian multipliers in such interval. Let $G \coloneqq \max\{\bar{\lambda}-\underline{\lambda},\bar{\mu}-\underline{\mu}\}$. Then, we have $G\le k\eta_{\text{B}}$ since η_{B} is more aggressive than $\eta_{\mathbb{R}}$ (see Section 4.4), and there are at most k gradient updates in the interval. Then,

$$\begin{aligned}\sum_{t\in\mathcal{I}}(\lambda_t g_t(\boldsymbol{\pi}^\circ)+\mu_t h_t(\boldsymbol{\pi}^\circ)) \le{}& \sum_{\substack{t\in\mathcal{I}:\\ g_t(\boldsymbol{\pi}^\circ)>0}}\bar{\lambda}g_t(\boldsymbol{\pi}^\circ) + \sum_{\substack{t\in\mathcal{I}:\\ g_t(\boldsymbol{\pi}^\circ)\le 0}}\underline{\lambda}g_t(\boldsymbol{\pi}^\circ) + \sum_{\substack{t\in\mathcal{I}\\ :h_t(\boldsymbol{\pi}^\circ)>0}}\bar{\mu}h_t(\boldsymbol{\pi}^\circ) + \sum_{\substack{t\in\mathcal{I}:\\ h_t(\boldsymbol{\pi}^\circ)\le 0}}\underline{\mu}g_t(\boldsymbol{\pi}^\circ)\\ \le{}& -\bar{\lambda}\left(\sum_{\substack{t\in\mathcal{I}:\\ g_t(\boldsymbol{\pi}^\circ)\le 0}} g_t(\boldsymbol{\pi}^\circ)+\alpha k\right) + \sum_{\substack{t\in\mathcal{I}:\\ g_t(\boldsymbol{\pi}^\circ)\le 0}}\underline{\lambda}g_t(\boldsymbol{\pi}^\circ)\\ &- \bar{\mu}\left(\sum_{\substack{t\in\mathcal{I}:\\ h_t(\boldsymbol{\pi}^\circ)\le 0}} h_t(\boldsymbol{\pi}^\circ)+\alpha k\right) + \sum_{\substack{t\in\mathcal{I}:\\ h_t(\boldsymbol{\pi}^\circ)\le 0}}\underline{\mu}g_t(\boldsymbol{\pi}^\circ)\\ \le{}& kG - \alpha k\bar{\lambda} + kG - \alpha k\bar{\mu}\\ \le{}& 2kG - \alpha\sum_{t\in\mathcal{I}}(\lambda_t+\mu_t)\\ \le{}& 2k^2\eta_{\text{B}} - \sum_{t\in\mathcal{I}}(\lambda_t+\mu_t),\end{aligned}$$

where the second inequality comes from $\sum_{t\in\mathcal{I}} g_t(\boldsymbol{\pi}^\circ)\le -\alpha k$ and $\sum_{t\in\mathcal{I}} h_t(\boldsymbol{\pi}^\circ)\le -\alpha k$. This concludes the proof. □

G. Application to Non-truthful Auctions

Recently, many advertising platforms have been transitioning from the second-price auction format toward a first-price format (Akbarpour & Li, 2018; Despotakis et al., 2021; Paes Leme et al., 2020).This is the case, for example, for Google's Ad Manager and AdSense platforms (Bigler, 2019; Wong, 2021). It is not clear what is an appropriate online bidding strategy for a budget- and ROI-constrained bidder participating in a series of non-truthful auctions. While second-price auctions are a truthful mechanism, meaning that bidders can bid their true value and maximize their utility, this is not the case for first-price auctions. This makes existing results for the second-price setting inapplicable to the non-truthful setting (Balseiro & Gur, 2019; Feng et al., 2023).

We show that our framework can be used to manage bidding in repeated non-truthful auctions under budget and ROI constraints. We will focus on the case of repeated first-price auctions, and we will make the simplifying assumption of having a finite set of possible valuations and bids. Extending our results to the continuous-bid setting is an interesting open problem, and it would amount to designing a suitable primal regret minimizer to plug into our framework. One option to accomplish this would be to adapt techniques designed for the unconstrained setting by Han et al. (2020a;b).

At each round $t \in [T]$, the bidder observes their valuation v_t extracted from a finite set $\mathcal{V} \subset [0,1]$ of n possible valuations. The set $\mathcal{X} \subset [0,1]$ is interpreted as the finite set of m possible bids. Let β_t be the highest-competing bid at time t. In the value-maximizing utility model for each t we have $f_t(x_t) \coloneqq v_t \mathbb{1}[x_t \geq \beta_t]$ (Babaioff et al., 2021; Balseiro et al., 2021), and the cost function is $c_t(x_t) \coloneqq x_t \mathbb{1}[x_t \geq \beta_t]$, where the indicator function $\mathbb{1}[x_t \geq \beta_t]$ specifies whether the bidder won the auction at time t. In general, we can handle any reward of the form $f_t(x_t) \coloneqq (v_t - \omega x_t)\mathbb{1}[x_t \geq \beta_t]$, with $\omega \in [0,1]$. We extend the definition of policies from Section 3 to model *randomized bidding policies*. Each policy $\boldsymbol{\pi} \in \Pi$ is now a mapping $\boldsymbol{\pi} : \mathcal{V} \to \Delta_{\mathcal{X}}$. We denote by $\boldsymbol{\pi}(v)_x$ the probability of selecting x under valuation v.

In order to apply Algorithm 1 and obtain the guarantees of Theorem 6.9 and Theorem 6.10 we have to design a suitable primal regret minimizer satisfying Equation (4.1). The following result is a rewriting of Theorem 8.1 providing an explicit regret bound.

Theorem G.1. *Let $n = 1$, $\eta \coloneqq 1/\sqrt{mT}$, $\xi \coloneqq 1/(2\sqrt{mT})$, $\sigma \coloneqq 1/T$, and assume that $\eta \leq 1/M_{[T]}$. For any $\delta > 0$,* EXP3-SIX *guarantees that, w.p. at least $1-\delta$, for any interval $\mathcal{I}$, and for any $x \in \mathcal{X}$,*

$$\textstyle\sum_{t\in\mathcal{I}}(\ell_t^{\mathrm{P}}(x_t) - \ell_t^{\mathrm{P}}(x)) \leq M_{\mathcal{I}}^2 \mathcal{E}_{T,\delta}^{\mathrm{P}}, \text{ where}$$

$$\mathcal{E}_{T,\delta}^{\mathrm{P}} \coloneqq \left(\frac{3}{2} + 4\log\left(\frac{mT}{\delta}\right) M_{\mathcal{I}}^{-1} + (\log(T)+1) M_{\mathcal{I}}^{-2}\right)\sqrt{mT}.$$

G.1. Proof of Theorem 8.1

In order to proceed with the analysis of Algorithm 2, let $\boldsymbol{p}_{t+1} \in [0,1]^m$ be the vector of *pre-weights* for time $t+1$, which is defined as

$$p_{t+1,x} \coloneqq \frac{\boldsymbol{\pi}_t(v_t)_x \exp\{-\eta \tilde{\ell}_t^{\mathrm{P}}(x)\}}{\sum_{x'\in\mathcal{X}} \boldsymbol{\pi}_t(v_t)_{x'} \exp\{-\eta \tilde{\ell}_t^{\mathrm{P}}(x')\}} \qquad \text{for all } x \in \mathcal{X}.$$

Then, we have the following intermediate result.

Lemma G.2. *Let $\eta > 0$ be such that $\eta\, \mathbb{E}_{\boldsymbol{\pi}} \tilde{\ell}_t^{\mathrm{P}}(x) < 1$ for all $t \in [T]$ and $\pi \in \Pi$. Then, for any $t \in [T]$, and $x' \in \mathcal{X}$, it holds*

$$\mathbb{E}_{\boldsymbol{\pi}_t(v_t)}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] - \tilde{\ell}_t^{\mathrm{P}}(x') \leq \frac{1}{\eta}\log\left(\frac{p_{t+1,x'}}{\boldsymbol{\pi}_t(v_t)_{x'}}\right) + \frac{\eta}{2}\mathbb{E}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right].$$

Proof. Let $\boldsymbol{y} \coloneqq \boldsymbol{\pi}_t(v_t) \in \Delta_m$. By the fact that, for any $n \geq 0$, $e^{-n} \leq 1 - n + n^2/2$, we have that

$$-\log \mathbb{E}_{\boldsymbol{y}}\left[e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)}\right] \geq -\log\left(1 - \eta\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] + \frac{\eta^2}{2}\mathbb{E}_{\boldsymbol{y}}\left[\ell_t^{\mathrm{P}}(x)^2\right]\right)$$

$$\log \mathbb{E}_{\boldsymbol{y}}\left[e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)}\right] \leq -\eta\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] + \frac{\eta^2}{2}\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right],$$

where the second inequality holds since, by assumption, $\eta\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] < 1$, which implies that the argument of the logarithm is strictly greater than 0. Then, by definition of the preweights $\boldsymbol{p}_{t+1}$, we have that, for any $x' \in \mathcal{X}$,

$$\begin{aligned}\eta\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] &\leq -\log \mathbb{E}_{\boldsymbol{y}}\left[e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)}\right] + \frac{\eta^2}{2}\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right] \\ &= -\log\left(\frac{\pi_{t,x'}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x')}}{p_{t+1,x'}}\right) + \frac{\eta^2}{2}\mathbb{E}_{\boldsymbol{y}}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right].\end{aligned}$$

This yields

$$\mathbb{E}_{\boldsymbol{\pi}_t(v_t)}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] - \tilde{\ell}_t^{\mathrm{P}}(x') \leq \frac{1}{\eta}\log\left(\frac{p_{t+1,x'}}{\pi_{t,x'}(v_t)}\right) + \frac{\eta}{2}\mathbb{E}_{\boldsymbol{\pi}_t(v_t)}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right],$$

for any possible alternative bid $x' \in \mathcal{X}$. □

Theorem 8.1. *Let $n = 1$, $\eta \coloneqq 1/\sqrt{mT}$, $\xi \coloneqq 1/(2\sqrt{mT})$, $\sigma \coloneqq 1/T$. For any $\delta > 0$,* EXP3-SIX *guarantees that, w.p. at least $1-\delta$, for any interval $\mathcal{I}$, and for any $x \in \mathcal{X}$,*

$$\textstyle\sum_{t\in\mathcal{I}}(\ell_t^{\mathrm{P}}(x_t) - \ell_t^{\mathrm{P}}(x)) \leq O\left(\sqrt{mT}\log\left(\frac{mT}{\delta}\right)\right).$$

Proof. In order to increase the readability, we will write $\boldsymbol{\pi}_t$ in place of $\boldsymbol{\pi}_t(v)$ since $v \in \mathcal{V}$ is constant throughout the proof (*i.e.,* $n = 1$).

By definition of $\tilde{\ell}_t^{\mathrm{P}}$, we have that for any $x \in \mathcal{X}$ and $\boldsymbol{\pi} \in \Delta_{\mathcal{X}}$, $\mathbb{E}\tilde{\ell}_t^{\mathrm{P}}(x) \leq \mathbb{E}[\mathbb{1}[x = x_t]\ell_t^{\mathrm{P}}(x)/\boldsymbol{\pi}_x] = \ell_t^{\mathrm{P}}(x)$. Therefore, since by assumption we have $\eta < 1/M_{[T]}$, where $M_{[T]}$ is the maximum range of the loss functions ℓ_t^{P} over the time horizon, the assumption of Lemma G.2 holds. Then, for any interval $[t_1, t_2]$, with $t_1, t_2 \in [T]$, $t_1 < t_2$, by Lemma G.2 we have that for any $x' \in \mathcal{X}$,

$$\sum_{t\in[t_1,t_2]}\left(\mathbb{E}_{\boldsymbol{\pi}_t}\left[\tilde{\ell}_t^{\mathrm{P}}(x)\right] - \tilde{\ell}_t^{\mathrm{P}}(x')\right) \leq \sum_{t\in[t_1,t_2]}\left(\frac{1}{\eta}\log\left(\frac{p_{t+1,x'}}{\pi_{t,x'}}\right) + \frac{\eta}{2}\mathbb{E}_{\boldsymbol{\pi}_t}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right]\right).$$

Moreover we have that

$$\begin{aligned}\sum_{t\in[t_1,t_2]}\log\left(\frac{p_{t+1,x'}}{\boldsymbol{\pi}_{t,x'}}\right) &= \log\left(\frac{1}{\boldsymbol{\pi}_{t_1,x'}}\right) + \sum_{t\in[t_1+1,t_2]}\log\left(\frac{p_{t,x'}}{\boldsymbol{\pi}_{t,x'}}\right) + \log(p_{t_2+1,x'}) \\ &\leq \log\left(\frac{m}{\sigma}\right) + \sum_{t\in[t_1+1,t_2]}\log\left(\frac{1}{1-\sigma}\right).\end{aligned}$$

The last inequality holds since, for any $t \in [T]$ and $x \in \mathcal{X}$,

$$\begin{aligned}\boldsymbol{\pi}_{t+1,x} &= \frac{(1-\sigma)w_{t,x}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)} + \frac{\sigma}{m}\sum_{i\in\mathcal{X}} w_{t,i}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)}}{\sum_{i\in\mathcal{X}}\left((1-\sigma)w_{t,i}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)} + \frac{\sigma}{m}\sum_{j\in\mathcal{X}} w_{t,j}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(j)}\right)} \\ &= \frac{(1-\sigma)w_{t,x}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)} + \frac{\sigma}{m}\sum_{i\in\mathcal{X}} w_{t,i}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)}}{\sum_{i\in\mathcal{X}} w_{t,i}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)}} \\ &\geq (1-\sigma)\frac{w_{t,x}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)}}{\sum_{i\in\mathcal{X}} w_{t,i}}\frac{\sum_{i\in\mathcal{X}} w_{t,i}}{\sum_{i\in\mathcal{X}} w_{t,i}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)}} \\ &= (1-\sigma)\frac{\boldsymbol{\pi}_{t,x}e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(x)}}{\mathbb{E}_{\boldsymbol{\pi}_t}\left[e^{-\eta\tilde{\ell}_t^{\mathrm{P}}(i)}\right]} \\ &= (1-\sigma)p_{t+1,x},\end{aligned}$$

where we used the definition of $\boldsymbol{\pi}_t$ and $\boldsymbol{p}_t$ as per Algorithm 2.

Then,

$$\sum_{t\in[t_1,t_2]}\left(\mathbb{E}_{\boldsymbol{\pi}_t}\tilde{\ell}_t^{\mathrm{P}}(x)-\tilde{\ell}_t^{\mathrm{P}}(x')\right)\leq\frac{1}{\eta}\left(\log\left(\frac{m}{\sigma}\right)+(t_2-t_1-1)\log\left(\frac{1}{1-\sigma}\right)\right)+\frac{\eta}{2}\sum_{t\in[t_1,t_2]}\mathbb{E}_{\boldsymbol{\pi}_t}\left[\tilde{\ell}_t^{\mathrm{P}}(x)^2\right]. \tag{G.1}$$

Neu (2015, Lemma 1) states that given a fixed non-increasing sequence (ξ_t) with $\xi_t\geq 0$, and by letting $\beta_{t,i}$ be a nonnegative random variable such that $\beta_{t,i}\leq 2\xi_t$ for all t and $i\in\mathcal{X}$, then with probability at least $1-\delta$,

$$\sum_{t\in[T]}\sum_{i\in\mathcal{X}}\beta_{t,i}\left(\tilde{\ell}_t^{\mathrm{P}}(i)-\ell_t^{\mathrm{P}}(i)\right)\leq\log(1/\delta).$$

Then, for any bid $i\in\mathcal{X}$, by setting

$$\beta_{t,j}=\begin{cases}2\xi_t\mathbb{1}[i=j] & \text{if } t\in\mathcal{I}\\ 0 & \text{otherwise}\end{cases},$$

and by applying a union bound we obtain that, with probability at least $1-\delta$,

$$\sum_{t\in\mathcal{I}}\left(\tilde{\ell}_t^{\mathrm{P}}(i)-\ell_t^{\mathrm{P}}(i)\right)\leq\frac{M_{\mathcal{I}}\log(m/\delta)}{2\xi}, \tag{G.2}$$

where $M_{\mathcal{I}}$ is the maximum range of the loss functions ℓ_t^{P} over time interval $\mathcal{I}$.

Moreover, from the definition of $\tilde{\ell}_t^{\mathrm{P}}$ (see Algorithm 2), we have that

$$\sum_{t\in\mathcal{I}}\mathbb{E}_{x\sim\boldsymbol{\pi}_t}\tilde{\ell}_t^{\mathrm{P}}(x)=\sum_{t\in\mathcal{I}}\left(\ell_t^{\mathrm{P}}(x_t)-\sum_{x\in\mathcal{X}}\xi\tilde{\ell}_t^{\mathrm{P}}(x)\right). \tag{G.3}$$

Finally, given $t\in\mathcal{I}$, we observe that

$$\mathbb{E}_{x\sim\boldsymbol{\pi}_t}\tilde{\ell}_t^{\mathrm{P}}(x)^2=\sum_{x\in\mathcal{X}}(\boldsymbol{\pi}_{t,x}\,\tilde{\ell}_t^{\mathrm{P}}(x))\,\tilde{\ell}_t^{\mathrm{P}}(x)\leq M_{\mathcal{I}}\sum_{x\in\mathcal{X}}\tilde{\ell}_t^{\mathrm{P}}(x). \tag{G.4}$$

Finally, we conclude by showing that, for any $x\in\mathcal{X}$, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t\in[t_1,t_2]}(\ell_t^{\mathrm{P}}(x_t)-\ell_t^{\mathrm{P}}(x))&\leq\sum_{t\in[t_1,t_2]}\ell_t^{\mathrm{P}}(x_t)+\frac{M_{\mathcal{I}}\log(m/\delta)}{2\xi}-\sum_{t\in[t_1,t_2]}\tilde{\ell}_t^{\mathrm{P}}(x)\qquad\text{(by Equation (G.2))}\\
&=\frac{M_{\mathcal{I}}\log(m/\delta)}{2\xi}+\sum_{t\in[t_1,t_2]}\sum_{i\in\mathcal{X}}\xi\tilde{\ell}_t^{\mathrm{P}}(i)+\sum_{t\in[t_1,t_2]}\left(\mathbb{E}_{\boldsymbol{\pi}_t}\tilde{\ell}_t^{\mathrm{P}}(j)-\tilde{\ell}_t^{\mathrm{P}}(x)\right)\\
&\qquad\qquad\text{(by Equation (G.3))}\\
&\leq\frac{M_{\mathcal{I}}\log(m/\delta)}{2\xi}+\sum_{t\in[t_1,t_2]}\sum_{i\in\mathcal{X}}\xi\tilde{\ell}_t^{\mathrm{P}}(i)+\frac{\eta}{2}M_{\mathcal{I}}\sum_{t\in[t_1,t_2]}\sum_{i\in\mathcal{X}}\tilde{\ell}_t^{\mathrm{P}}(i)\\
&\qquad+\frac{1}{\eta}\left(\log\left(\frac{m}{\sigma}\right)+(t_2-t_1-1)\log\left(\frac{1}{1-\sigma}\right)\right)\\
&\qquad\qquad\text{(by Equation (G.1) and Equation (G.4))}\\
&\leq\frac{M_{\mathcal{I}}\log(m/\delta)}{2\xi}+(t_2-t_1)m\xi M_{\mathcal{I}}+(t_2-t_1)\frac{\eta M_{\mathcal{I}}^2 m}{2}\\
&\qquad+\frac{1}{\eta}\left(\log\left(\frac{m}{\sigma}\right)+(t_2-t_1-1)\log\left(\frac{1}{1-\sigma}\right)\right).
\end{aligned}$$

By setting $\eta = \frac{1}{\sqrt{mT}}$, $\xi = \frac{1}{2\sqrt{mT}}$, with probability at least $1-\delta$,

$$\sum_{t\in[t_1,t_2]} (\ell_t^{\mathbb{P}}(x_t) - \ell_t^{\mathbb{P}}(x)) \leq \frac{3}{2} M_{\mathcal{I}}^2 (t_2 - t_1)\sqrt{\frac{m}{T}} + \sqrt{mT}\left(M_{\mathcal{I}} \log\left(\frac{m}{\delta}\right) + \log\left(\frac{m}{\sigma}\right) + (T-1)\log\left(\frac{1}{1-\sigma}\right)\right)$$
$$\leq \frac{3}{2} M_{\mathcal{I}}^2 (t_2 - t_1)\sqrt{\frac{m}{T}} + \sqrt{mT}\left(2M_{\mathcal{I}} \log\left(\frac{m}{\delta}\right) - \log\left(\sigma(1-\sigma)^{T-1}\right)\right).$$

By letting $h(z) \coloneqq -z\log z - (1-z)\log(1-z)$ be the binary entropy function for $z \in [0,1]$, we have that, for $z \in [0,1]$, $h(z) \leq z\log(e/z)$ (see, *e.g.*, Cesa-Bianchi et al. (2012, Corollary 1)). Then, for $\sigma = 1/T$, we have that $-\log \sigma(1-\sigma)^{T-1} \leq \log(eT)$. This yields

$$\sum_{t\in[t_1,t_2]} (\ell_t^{\mathbb{P}}(x_t) - \ell_t^{\mathbb{P}}(x)) \leq \frac{3}{2} M_{\mathcal{I}}^2 (t_2 - t_1)\sqrt{\frac{m}{T}} + \sqrt{mT}\left(2M_{\mathcal{I}} \log\left(\frac{m}{\delta}\right) + \log(eT)\right).$$

By taking a union bound over all possible intervals $[t_1, t_2]$ we obtain that, with probability at least $1-\delta$,

$$\sum_{t\in[t_1,t_2]} (\ell_t^{\mathbb{P}}(x_t) - \ell_t^{\mathbb{P}}(x)) \leq \frac{3}{2} M_{\mathcal{I}}^2 (t_2 - t_1)\sqrt{\frac{m}{T}} + \sqrt{mT}\left(4M_{\mathcal{I}} \log\left(\frac{mT}{\delta}\right) + \log(T) + 1\right)$$
$$\leq M_{\mathcal{I}}^2 \left(\frac{3}{2} + \frac{4}{M_{\mathcal{I}}} \log\left(\frac{mT}{\delta}\right) + \frac{\log(T)+1}{M_{\mathcal{I}}^2}\right)\sqrt{mT},$$

which proves our statement. □