Online Resource Allocation With General Constraints

Eleonora Fidelia Chiefari
Politecnico di Milano
eleonorafidelia.chiefari@polimi.it

Francesco Emanuele Stradi
Politecnico di Milano
francescoemanuele.stradi@polimi.it

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

Alberto Marchesi
Politecnico di Milano
alberto.marchesi@polimi.it

Abstract

Online resource allocation (ORA) is a fundamental framework for sequential decision-making problems under *budget constraints*, with applications ranging from online advertising to revenue management. In this work, we study a broader setting that includes *both* budget constraints and *general constraints*, extending the classical budget-only model. This extension is essential for modeling critical economic requirements, such as Return-on-Investment (ROI) constraints. We develop an algorithm that achieves best-of-both-world guarantees within this generalized framework. In particular, against a dynamic benchmark, our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the *stochastic* regime and α-regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$ in the *adversarial* regime, where α depends on the feasibility margin of the corresponding offline problem. At the same time, our algorithm guarantees strict satisfaction of the budget constraints and $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative violation for the general ones. From a technical perspective, introducing general constraints alongside budgets precludes the use of standard budget-focus methods. While budget methods rely on a zero-consumption "safe" action to ensure feasibility, general constraints are much less "aligned" towards feasibility. We overcome these difficulties with a new analysis that exploits *weak adaptivity* to get boundedness of the Lagrangian multipliers and best-of-both-world guarantees.

1 Introduction

Online resource allocation (ORA) serves as a fundamental framework for sequential decision-making under *budget constraints*. In these settings, requests arrive sequentially, and, for each request, the decision maker must select an appropriate action. A defining characteristic of ORA, which distinguishes it from many online learning paradigms, is that the learner possesses exact knowledge of the current round *before* picking an action: upon a request's arrival, the resource consumption and reward associated with each available action are fully known. Hence, the challenge stems from the uncertainty about future requests, requiring a strategy that balances immediate rewards against the preservation of resources for future arrivals.

Such ORA problems arise in various real-world scenarios, such as, *e.g.*, online advertising [Mehta et al., 2007], revenue management [Talluri and van Ryzin, 2004, Ball and Queyranne, 2009], and online auctions [Zhou et al., 2008, Balseiro and Gur, 2019]. A wide range of works have addressed ORA problems in various stochastic settings [Devanur and Hayes, 2009, Feldman et al., 2010, Kesselheim et al., 2013, Agrawal et al., 2014, Devanur et al., 2019, Li et al., 2020, Sun et al., 2022]. Only a few works have gone beyond these settings: some consider non-stationary models [Ciocan and Farias, 2011, Esfandiari et al., 2017], while others address adversarial ones, albeit under specific structural assumptions [Mehta et al., 2007, Buchbinder et al., 2007, Balseiro and Gur,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

2019, Stradi et al., 2025]. Balseiro et al. [2023] are the first to consider general ORA problems in both stochastic and adversarial regimes by designing a *best-of-both-world* algorithm. Their algorithm employs a dual-based approach that simultaneously achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the stochastic regime and α-regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$ in the adversarial setting, where α is an instance-dependent constant parameter related to the per-round budget.

Despite these progresses, most existing results focus exclusively on budget constraints. A notable exception is the work of Feng et al. [2023], which, however, focuses only on Return-on-Investment (ROI) constraints within stochastic settings. Despite that, many real-world applications involve additional *long-term constraints*—such as fairness, risk, or performance requirements—that must be satisfied cumulatively over time. A fundamental difference between general and budget constraints is that while budget constraints come with an *a priori* known feasibility margin (the per-round budget), general constraints often have an *unknown* feasibility margin. Furthermore, such constraints are inherently less "aligned" with one another. For instance, an action that maximizes ROI may simultaneously violate fairness. This tension requires algorithms to allow violations in some rounds to be later compensated by strict satisfaction in others, thereby forcing learning algorithms to tolerate temporary violations. These dynamics introduce additional challenges that prevent results for budget-constrained settings from being directly applied to problems involving general constraints.

1.1 Original Contributions

In this work, we study generalized ORA problems that incorporate *both budget and general long-term constraints*. To the best of our knowledge, we propose the first algorithm that attains *best-of-both-world* guarantees in this setting. Notably, while the performance of our algorithm in adversarial environments depends on the Slater's parameter ρ—which encodes the feasibility margin of the problem—the algorithm operates without prior knowledge of this parameter. Our algorithm achieves the following performance guarantees:

- $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative expected regret in the *stochastic* setting, where requests are drawn i.i.d. across rounds from an unknown fixed probability distribution;
- $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative α-regret in the *adversarial* setting, where requests are chosen arbitrarily in advance by an adversary, with $\alpha := \rho/1+\rho$ an instance-dependent constant.

At the same time, regardless of the underlying environment, the algorithm:

- strictly satisfies all the budget constraints;
- guarantees $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative violation of the general constraints.

Our algorithm employs a *dual-based* approach, wherein Lagrangian multipliers are updated via an *online gradient descent* (OGD) scheme and primal actions are selected greedily in order to maximize the instantaneous Lagrangian. The central challenge in our analysis lies in bounding the magnitude of these Lagrangian multipliers throughout the primal-dual dynamics. Crucially, we do not impose an external upper bound on the dual variable, as such a bound depends on the unknown Slater's parameter ρ. Instead, we show that the multipliers are naturally bounded under the learning dynamics.

The crucial difference with respect to the classical budget-only setting [Balseiro et al., 2023] stems from the absence of a point-wise feasibility-maximizing action, a.k.a. the "void" action. In our more general setting, constraints allow for cross-constraint compensation, which can prevent the violations from vanishing and the multipliers from ceasing to grow. Indeed, since the constraints might not be aligned, it could be the case that multipliers are large, but the most profitable decision for the primal is to keep violating a constraint to satisfy another one by a huge margin. Despite that, exploiting the *weak adaptivity* of OGD, we are still able to prove that, when the multipliers exceed a certain threshold, the no-regret guarantees of the OGD dual update prevent the primal, which plays greedily, from indefinitely exploiting this cross-constraint compensation. This allows us to prove that the norm of the multipliers cannot exceed a threshold inversely proportional to the feasibility margin ρ with high probability, completely bypassing the need for a point-wise feasibility-maximizing action.

1.2 Related Works

A wide range of works on the ORA problem focuses on stochastic input models. Early contributions [Devanur and Hayes, 2009, Feldman et al., 2010] introduced algorithms that learn dual prices from an initial sample and exploit them. Although conceptually simple, these algorithms attained suboptimal guarantees. Subsequent works [Agrawal et al., 2009, Kesselheim et al., 2013, Devanur et al., 2019] improved performance by sequentially solving linear programming relaxations, achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. A growing line of research aims to design computationally efficient algorithms, avoiding heavy optimization steps. Recent works [Li et al., 2020, Sun et al., 2022] employ primal-dual or gradient-style updates, attaining strong theoretical guarantees although relying on specific problem structures or assumptions.

More recent works [Ciocan and Farias, 2011, Esfandiari et al., 2017] explore non-stationary models, filling the gap between stochastic and adversarial ones. For what concerns the adversarial setting, early works [Mehta et al., 2007, Buchbinder et al., 2007] establish optimal constant-factor guarantees for the specific cases of online matching and AdWords problem. These results assume linear relationship between reward and consumption and cannot be directly applied to more general allocation settings. The first work providing guarantees in both the stochastic and the adversarial setting is Balseiro et al. [2023]. The authors develop a dual-based algorithm that achieves simultaneously $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the stochastic setting, $\widetilde{\mathcal{O}}(\sqrt{T})$ α-regret in the adversarial setting and in the non-stationary setting.

In the specific domain of online bidding, Balseiro and Gur [2019] provide an adaptive pacing strategy achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound in the stochastic setting, under suitable assumptions. In the adversarial setting, it establish a fundamental limitation, showing that no algorithm can achieve a competitive ratio better than v/ρ, where v is the uniform upper bound on the advertiser's value. Similarly, Feng et al. [2023] consider an online bidding problem under a RoS constraint, where the goal is to maximize the total value obtained from ad placements while ensuring that the ratio between value and expenditure exceeds a given threshold. Their analysis relies on the structure of the RoS constraint to obtain optimal regret guarantees. Alternative approaches have also studied competitive guarantees under budget constraints. In particular, Zhou et al. [2008] design online algorithms whose performance depends on the structure of value-to-weight ratios of the incoming items.

A closely related line of research focuses on online decision-making subject to long-term constraints. Early works, such as [Mahdavi et al., 2012, Jenatton et al., 2015] consider fixed constraints, achieving sublinear regret while controlling cumulative violation. Subsequent works, such as [Mannor et al., 2009, Yu et al., 2017, Sun et al., 2017], extend these results to time-varying constraints. The *Bandits with Knapsacks* (BwK) framework,introduced by Badanidiyuru et al. [2013], focuses on budget constraints. Several works, such as [Agrawal and Devanur, 2019] have improved the regret guarantees in the stochastic setting. More recently, primal-dual algorithms for BwK have been extended to adversarial and non-stationary environments [Immorlica et al., 2023, Kesselheim and Singla, 2020, Castiglioni et al., 2022]. Most closely related to our research are Bernasconi et al. [2024], Castiglioni et al. [2024], which generalize the BwK framework to more general constraints. Also in their case, the analysis is built upon the weak adaptivity of the primal and dual regret minimizers.

2 Preliminaries

We study *online resource allocation* (ORA) problems in which a decision maker (hereafter, agent) selects actions over a finite horizon of T rounds. We denote the action space of the agent by $\mathcal{X} \subseteq \mathbb{R}^K$. At each round $t \in [T]$,[1] the agent observes an input tuple $\gamma_t = (f_t, \boldsymbol{g}_t, \boldsymbol{h}_t)$, where:

- $f_t : \mathcal{X} \to [0, 1]$ is the reward function;
- $\boldsymbol{g}_t : \mathcal{X} \to [-1, 1]^m$, whose i-th component $g_{t,i} : \mathcal{X} \to [-1, 1]$ encodes the cost function associated with constraint $i \in [m]$;
- $\boldsymbol{h}_t : \mathcal{X} \to [0, 1]^n$, whose j-th component $h_{t,j} : \mathcal{X} \to [0, 1]$ encodes the consumption function associated with resource $j \in [n]$.

[1] Throughout this paper, we denote by $[a] := \{1, \dots, a\}$ the set of the first $a \in \mathbb{N}$ natural numbers.

The input functions are fully revealed to the agent before they make any decision. After observing the input, the agent selects an action $x_t \in \mathcal{X}$, which consumes $h_{t,j}(x_t)$ of each resource $j \in [n]$, produces a reward $f_t(x_t)$ and incurs a constraint cost $g_{t,i}(x_t)$ for each $i \in [m]$, where strictly positive values of $g_{t,i}(x_t)$ indicate a violation of constraint i, while $g_{t,i}(x_t) \leq 0$ means that constraint i is satisfied. The input tuples may be either *stochastic*, *i.e.*, generated i.i.d. across rounds according to a fixed probability distribution, or *adversarial*, *i.e.*, chosen arbitrarily by an adversary. Throughout the paper, we denote by Γ the set of all the possible input tuples $\gamma = (f, \boldsymbol{g}, \boldsymbol{h})$.

The agent has an overall budget $B_j = \Omega(T)$ limiting the cumulative consumption of each resource $j \in [n]$ over the time horizon. We denote by $\boldsymbol{\beta} \in \mathbb{R}^n$ the vector representing the per-iteration budget for each resource, where $\beta_j > 0$ is defined as B_j/T. As it is common in the online resource allocation literature (see, *e.g.*, [Balseiro et al., 2023]), we assume that there exists a *void action* $\varnothing \in \mathcal{X}$ such that, for any input $(f, \boldsymbol{g}, \boldsymbol{h}) \in \Gamma$, it holds $f(\varnothing) = 0$, $g_i(\varnothing) = 0$ for $i \in [m]$ and $h_j(\varnothing) = 0$ for $j \in [n]$.

Given a specific sequence of inputs $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_T)$, where $\gamma_t = (f_t, \boldsymbol{g}_t, \boldsymbol{h}_t)$ for all t, we define the *cumulative reward* of the agent over the time horizon T as:

$$\texttt{Rew}(T, \boldsymbol{\gamma}) := \sum_{t \in [T]} f_t(x_t).$$

Moreover, we define the *cumulative constraint violation* over the time horizon T for the general, i.e., non-budget, constraints as:

$$V_T := \max_{i \in [m]} \sum_{t \in [T]} g_{t,i}(x_t).$$

The goal of the agent is to maximize $\texttt{Rew}(T, \boldsymbol{\gamma})$, while ensuring:

- V_T grows sublinearly in T, namely, it is $o(T)$;
- $\sum_{t \in [T]} h_{t,j}(x_t) \leq \beta_j T$ for all $j \in [n]$, *i.e.*, the budget constraints are strictly enforced.

2.1 Stochastic Setting

In the stochastic setting, the inputs are drawn i.i.d. across rounds from a fixed but unknown probability distribution $\mathcal{P}$ supported on Γ. Formally, at each round $t \in [T]$, it holds $\gamma_t = (f_t, \boldsymbol{g}_t, \boldsymbol{h}_t) \sim \mathcal{P}$. We denote an input sequence by $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_T) \in \Gamma^T$ and compactly write $\boldsymbol{\gamma} \sim \mathcal{P}^T$ to indicate that each γ_t is drawn i.i.d. from $\mathcal{P}$. Then, we let $\text{OPT}_{\text{stoc}} := \mathbb{E}_{\boldsymbol{\gamma} \sim \mathcal{P}^T}[\text{OPT}(\boldsymbol{\gamma})]$ be the dynamic-optimum in the stochastic setting, where $\text{OPT}(\boldsymbol{\gamma})$ is defined as:

$$\text{OPT}(\boldsymbol{\gamma}) := \begin{cases} \max_{(x_1,...,x_T) \in \mathcal{X}^T} & \sum_{t \in [T]} f_t(x_t) \\ \text{s.t.} & \sum_{t \in [T]} g_{t,i}(x_t) \leq 0 \quad \forall i \in [m] \\ & \sum_{t \in [T]} h_{t,j}(x_t) \leq \beta_j T \quad \forall j \in [n]. \end{cases}$$

We define a key quantity, known as the *Slater's parameter*, related to the feasibility of the problem:

$$\rho_{\text{stoc}} = - \min_{\pi : \Gamma \to \mathcal{X}} \max \left\{ \max_{i \in [m]} \mathbb{E}_{\gamma \sim \mathcal{P}} \left[g_i(\pi(\gamma)) \right], \max_{j \in [n]} \mathbb{E}_{\gamma \sim \mathcal{P}} \left[h_j(\pi(\gamma)) - \beta_j \right] \right\},$$

where the minimum is taken over all the possible functions mapping inputs to actions. The Slater's parameter quantifies the strict feasibility margin of the problem. Intuitively, it represents how well constraints are satisfied by the "most feasible" solution.

We remark that the proposed definition of ρ_{stoc} naturally reduces to the standard strict feasibility margin when applied to classical budget-constrained problems (*e.g.*, Balseiro et al. [2023]). In such settings, evaluating our definition under the constant policy $\pi(\gamma) = \varnothing$ for all $\gamma \in \Gamma$, the expectation over the selected action trivially yields zero consumption. Consequently, for each resource $j \in [n]$, the expectation over the input distribution evaluates deterministically to $\mathbb{E}_{\gamma \sim \mathcal{P}}[0 - \beta_j] = -\beta_j$, which yields $\rho = \min_{j \in [n]} \beta_j$ Thus, the expectation over γ in our formulation provides the generalization required for settings without a deterministically safe action.

In the stochastic setting, we measure the performance of the agent over the time horizon T by means of the *expected cumulative regret* (or simply *regret* for short), which is defined as follows:

$$R_T := \text{OPT}_{\text{stoc}} - \mathbb{E}_{\boldsymbol{\gamma} \sim \mathcal{P}^T}[\texttt{Rew}(T, \boldsymbol{\gamma})].$$

[2] All the results presented in this work can be easily extended to baselines defined using $\sup_{(x_1,...,x_T) \in \mathcal{X}^T}$.

2.2 Adversarial Setting

In the adversarial setting, the dynamic-optimum is defined as $\text{OPT}_{\text{adv}} = \text{OPT}(\boldsymbol{\gamma})$ for a given fixed input sequence $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_T) \in \Gamma^T$. Moreover, in this setting, the *Slater's parameter* is:

$$\rho_{\text{adv}} = - \min_{(x_1,\dots,x_T)\in\mathcal{X}^T} \max_{t\in[T]} \max \left\{ \max_{i\in[m]} g_{t,i}(x_t), \max_{j\in[n]} h_{t,j}(x_t) - \beta_j \right\}.$$

As shown by Balseiro et al. [2023], it is impossible to obtain both sublinear regret and sublinear cumulative constraint violation when constraints are selected adversarially. Consequently, we cannot rely on the standard notion of regret for the adversarial setting. To overcome this limitation, we study the cumulative regret with respect to a fraction $\alpha \in (0, 1)$ of the optimum, called α-regret:

$$\alpha\text{-}R_T := \alpha \cdot \text{OPT}_{\text{adv}} - \texttt{Rew}(T),$$

where $\alpha := \rho_{\text{adv}}/1+\rho_{\text{adv}}$ and we let $\texttt{Rew}(T) := \texttt{Rew}(T, \boldsymbol{\gamma})$ since the dependence of (the deterministic) γ is clear from the context.

We remark that, both in the stochastic setting introduced in Section 2.1 and the adversarial one discussed in Section 2.2, the agent has no prior knowledge of the Slater's parameter.

2.3 Lagrangian Formulation

For the sake of exposition, it is convenient to group the general long-term constraints and the budget constraints into a single unified vector. Let $M := m+n$ be the total number of constraints. For each round $t \in [T]$ and action $x \in \mathcal{X}$, we define the unified constraint vector $\tilde{\boldsymbol{g}}_t(x) \in \mathbb{R}^M$ as follows:

$$\tilde{g}_{t,i}(x) = \begin{cases} g_{t,i}(x) & \text{if } i \in [m] \\ h_{t,i-m}(x) - \beta_{i-m} & \text{if } m < i \leq M. \end{cases}$$

A fundamental tool for addressing constrained optimization problems is the Lagrangian formulation. Given the instantaneous reward function f_t, and the unified constraint function $\tilde{\boldsymbol{g}}_t$, at round $t \in [T]$, we define the instantaneous Lagrangian function as:

$$\mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x, \boldsymbol{\lambda}) := f_t(x) - \sum_{i=1}^{M} \lambda_i \tilde{g}_{t,i}(x) = f_t(x) - \langle \boldsymbol{\lambda}, \tilde{\boldsymbol{g}}_t(x) \rangle,$$

where $\boldsymbol{\lambda} \in \mathbb{R}^M_{\geq 0}$ is the vector of dual multipliers.

3 Algorithm

In this section, we describe our algorithm, whose pseudo-code is provided in Algorithm 1.

Algorithm 1 DUAL GRADIENT DESCENT

Require: T, η, B_j for all $j \in [n]$.

1: $\boldsymbol{\lambda}_1 \leftarrow \mathbf{0}$
2: **for** $t = 1, 2, \dots, T$ **do**
3: Observe input tuple $\gamma_t = (f_t, \boldsymbol{g}_t, \boldsymbol{h}_t)$
4: $\widehat{x}_t \leftarrow \arg\max_{x\in\mathcal{X}} \mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x, \boldsymbol{\lambda}_t)$ ▷ Compute candidate action
5: **if** $\sum_{s=1}^{t-1} h_{s,j}(x_s) \leq \beta_j T - 1$ for every $j \in [n]$ **then**
6: $x_t \leftarrow \widehat{x}_t$
7: **else**
8: $x_t \leftarrow \varnothing$
9: **end if**
10: $\boldsymbol{\lambda}_{t+1} \leftarrow \Pi_{\mathbb{R}^M_{\geq 0}}(\boldsymbol{\lambda}_t + \eta \tilde{\boldsymbol{g}}_t(x_t))$ ▷ Dual update
11: **end for**

Algorithm 1 operates as follows. It receives as input the time horizon T, the learning rate for the dual update $\eta > 0$, and the budget B_j for each resource $j \in [n]$. At the beginning of each round $t \in [T]$, the input $\gamma_t = (f_t, \boldsymbol{g}_t, \boldsymbol{h}_t)$ is fully revealed (Line 3). Then, the algorithm computes the

candidate action $\widehat{x}_t$ by greedily maximizing the instantaneous Lagrangian function (Line 4). Since budget constraints must be strictly satisfied, the algorithm checks, for each resource $j \in [n]$, whether the cumulative resource consumption up to the previous round exceeds the budget (Line 5). If the budget constraints are satisfied for all resources, the algorithm plays the candidate action $\widehat{x}_t$ (Line 6). Once the budget is exceeded for at least one resource, the algorithm is forced to select the void action $\varnothing$ (Line 8).[3] Finally, after the primal action is played, the algorithm performs an *online gradient descent* (OGD) step to update and project the dual variable $\boldsymbol{\lambda}$ onto the non-negative orthant $\mathbb{R}^M_{\geq 0}$ for the next round (Line 10).

In the following, it will be useful to define the stopping time $\tau \leq T$, i.e., the last round in which $\sum_{t \in [\tau]} h_{t,j}(x_t) \geq \beta_j T - 1$ for each resource $j \in [n]$. For any $t > \tau$, the agent play the void action $x_t = \varnothing$.

4 Bounding the Lagrangian Multipliers

In this section, we present the primary technical contribution of our work. Specifically, we show that the Lagrange multipliers remain bounded during the learning dynamic. This analysis requires novel techniques with respect to the ones employed in budget-constrained problems, as general constraints are much more prone to conflict with one another, pushing the primal decision in opposing directions. For instance, while playing the void action is the trivial way to satisfy all budget constraints simultaneously, this property does not hold for general functions. We provide a detailed discussion of this distinction in Section 4.1 and establish the upper bounds for the multipliers in Section 4.2.

4.1 Difference Between Budget and General Constraints

Classical ORA problems, such as those studied in [Balseiro et al., 2023], focus exclusively on budget constraints. In such settings, bounding the dual variables follows directly from the definition of void action: playing $\varnothing$ yields zero resource consumption ($h_{t,j}(\varnothing) = 0$), guaranteeing a strictly positive and a priori known feasibility margin (the per-iteration budget β_j). In absence of other constraints, the overall feasibility margin of the problem simply coincides with the minimum budget parameter ($\rho = \min_j \beta_j > 0$). This margin allows previous works to analytically bound dual multipliers. Our setting, however, incorporates general constraints $\boldsymbol{g}_t \in [-1, 1]^m$ alongside budget constraints. For these general constraints, the void action yielding $g_{t,i}(\varnothing) = 0$ provides zero slack. Consequently, the overall Slater's parameter no longer coincides with the minimum β_j, rather, it becomes a problem-dependent unknown quantity. To highlight the fundamental technical difference between standard budget-constrained problems (such as [Balseiro et al., 2023]) and our setting with general constraints, we present a simplified single-round example with two constraints ($m = 2$).

Example 4.1. Let us consider an ORA problem instance.[4] Suppose that the dual multipliers at the beginning of some round $t \in [T]$ are large and equal $\boldsymbol{\lambda}_t = [\lambda_{t,1}\ \lambda_{t,2}] \simeq [2/\rho\ \ 2/\rho]$. We analyze how a standard dual-based algorithm that maximizes the Lagrangian behaves in two different settings: the first with only budget constraints, and the second with only general constraints.

- In the classical setting with only budget constraints, the constraint function is $h_{t,i}(x) - \rho$, namely, it takes values in $[-\rho, 1-\rho]$. This means that the maximum extent to which an action can "satisfy" the budget constraint is bounded by $-\rho$. The agent has access to a deterministic void action $\varnothing$, which yields zero rewards and consumes zero budget. Consider any other action in $\mathcal{X}$ that incurs a violation greater than 0 in any of the constraints. For instance, action $x_A \in \mathcal{X}$ such that $f(x_A) = 1$ and $\boldsymbol{h}_t(x_A) = [\rho + \epsilon\ 0]^\top$. We compare the Lagrangian objective obtained by playing the two actions $\varnothing$ and x_A at round t:

$$\mathcal{L}_{f_t, \tilde{\boldsymbol{g}}_t}(\varnothing, \boldsymbol{\lambda}_t) = f_t(\varnothing) - \langle \boldsymbol{\lambda}_t, \boldsymbol{h}_t(\varnothing) - \rho \mathbf{1} \rangle = 0 + \langle \boldsymbol{\lambda}_t, \rho \mathbf{1} \rangle = \lambda_{t,1}\rho + \lambda_{t,2}\rho = 4 >$$

$$\mathcal{L}_{f_t, \tilde{\boldsymbol{g}}_t}(x_A, \boldsymbol{\lambda}_t) = f_t(x_A) - \langle \boldsymbol{\lambda}_t, \boldsymbol{h}_t(x_A) - \rho \mathbf{1} \rangle = 1 - \lambda_{t,1}(\rho + \epsilon - \rho) + \lambda_{t,2}\rho = 3 - \frac{2\epsilon}{\rho}.$$

  Although x_A yields the best per-round reward possible and satisfies one of the constraint as much as possible, the slight violation on the other constraint penalizes the Lagrangian objective with respect to satisfying both the constraints simultaneously. This holds for any action that violates

[3] Technically, we stop slightly before in order to guarantee that the budget constraints are exactly satisfied.

[4] We remark that while this reasoning holds for the adversarial setting, it can be trivially extended to the stochastic setting.

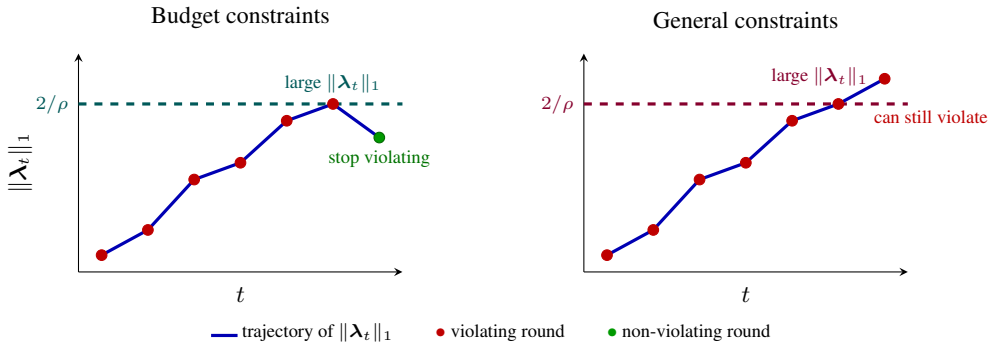


Figure 1: Illustration of the behavior of standard dual algorithms that *do not use a weakly adaptive regret minimizers*. Left: in the budget-constrained setting, once $\|\boldsymbol{\lambda}_t\|_1$ becomes sufficiently large, the algorithm stops violating, and the dual variable starts decreasing. Right: under general constraints, violations may persist even when $\|\boldsymbol{\lambda}_t\|_1$ is already large.

one of the constraints. Since no other action could yield a better Lagrangian value, the algorithm is encouraged to play $\varnothing$, and therefore it stops violating.

- In the setting with general constraints, constraints take values in $[-1, 1]$ and are not aligned. Indeed, extreme cross-constraint compensation in the Lagrangian may happen. The "safest" action in this setting is the strictly safe action leading to the computation of ρ, i.e. x° such that $\boldsymbol{g}_t(x^\circ) = [-\rho \quad -\rho]^\top$ by definition. Consider action $x_B \in \mathcal{X}$, yielding $f(x_B) = 1$ and $\boldsymbol{g}_t(x_B) = [\rho + \epsilon \quad -1]^\top$. Similarly to x_A, x_B provides the best reward possible and satisfies as much as possible one constraints while slightly violating the other.

  Comparing the Lagrangian value obtained by playing either of the two actions x° and x_B, we obtain the following inequality:

$$\mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x^\circ, \boldsymbol{\lambda}_t) = f_t(x^\circ) - \langle \boldsymbol{\lambda}_t, \boldsymbol{g}_t(x^\circ) \rangle \leq 1 + \lambda_{t,1}\rho + \lambda_{t,2}\rho = 5 <$$
$$\mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_B, \boldsymbol{\lambda}_t) = f_t(x_B) - \langle \boldsymbol{\lambda}_t, \boldsymbol{g}_t(x_B) \rangle = 1 - \lambda_{t,1} \cdot (\rho + \epsilon) + \lambda_{t,2} \cdot 1 \overset{(3\rho+\epsilon=\frac{1}{2})}{=} 5 + \frac{1}{\rho}.$$

  When $3\rho + \epsilon < 1$, since the satisfaction of the second constraint is much larger than ρ, its contribution to the Lagrangian function is large and strictly positive although the first constraint is violated. The strictly safe action, which only provides margin ρ, yields a worse contribution. Therefore, the algorithm will prefer to play x_B (*i.e.*, keep violating) with respect to the strictly safe action, even when x° leads to the maximum possible reward $f_t(x^\circ) = 1$.

In Figure 1, we provide a graphical representation of these aspects.

4.2 A Linear Upper Bound

In standard Lagrangian frameworks, the magnitude of dual variables is typically controlled through an explicit projection step onto a bounded domain. However, this operation requires the knowledge of the Slater's parameter ρ. To bypass this restrictive assumption, we leverage the property of *weak adaptivity* [Hazan, 2019], which we formally defined as follows.

Definition 4.1 (Weakly-adaptive regret minimizer)**.** An online learning algorithm is said to be a *weakly-adaptive regret minimizer* if, for any interval $I = [t_1, t_2] \subseteq [T]$, its regret R_I accumulated over the interval is sublinear in T. Formally:

$$\max_{I=[t_1,t_2]\subseteq[T]} R_I(\mathcal{X}) = \max_{I=[t_1,t_2]\subseteq[T]} \left\{ \max_{x\in\mathcal{X}} \sum_{t\in I} f_t(x) - \sum_{t\in I} f_t(x_t) \right\} \leq o(T).$$

Intuitively, a weakly-adaptive regret minimizer guarantees sublinear regret not just over the full time horizon T, but over any arbitrary sub-interval $[t_1, t_2] \subseteq [T]$. Online gradient descent, which we employ to update the Lagrangian multipliers, naturally satisfies this property for linear losses. In

this section, we use weak adaptivity to provide a theoretical bound on the ℓ_1-norm of the Lagrangian multipliers $\boldsymbol{\lambda}_t$, which allows us to subsequently bound the cumulative constraint violation.

Lemma 4.2. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, the ℓ_1-norm of the dual variables is bounded for all $t \in [T]$ as:*

$$\|\boldsymbol{\lambda}_t\|_1 \leq \frac{14M}{\rho}.$$

To prove Lemma 4.2, we proceed by contradiction. Supposing that the ℓ_1-norm of the dual variables $\|\boldsymbol{\lambda}_t\|_1$ exceeds a threshold of $\mathcal{O}(M/\rho)$ at some round t_2, we analyze the resulting primal-dual dynamics. Intuitively, the dual variables act as penalty weights for constraint violations. When these multipliers grow disproportionately large, the penalty term $\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x)\rangle$ in the instantaneous Lagrangian objective completely dominates the reward, which is bounded in $[0,1]$. Consequently, to maximize the Lagrangian, the greedy primal step is forced to select an action that yields a penalty at least as favorable as the strictly feasible baseline. Since the baseline guarantees a strictly negative value, the primal response systematically pulls the multipliers back down. We formalize this intuition by comparing the Lagrangian value of the selected action with that of the strictly feasible baseline, obtaining a lower bound on the cumulative Lagrangian $\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t)$. On the other hand, we derive an upper bound on the same quantity by exploiting the weak adaptivity of online gradient descent. Specifically, rather than evaluating the dual regret globally over the full horizon T, we evaluate it locally over the sub-interval $[t_1,t_2]$. This step highlights why our framework specifically requires OGD rather than a general online mirror descent (OMD) method. Algorithms guaranteeing only *global* no-regret over $[T]$ (such as standard OMD with entropy regularizer), can accumulate negative regret during early rounds and exploit it to "absorb" later large penalties. Our weak adaptivity property prevents the algorithm from relying on past negative regret. This represents a fundamental difference with respect to [Balseiro et al., 2023].

We show that for the dual variables to reach the assumed threshold, they would need to be updated at a rate strictly incompatible with our hypothesis. This means that the greedy primal is guaranteed to react and halt constraint violation long before the multiplier can ever reach that threshold. The extended proof can be found in Appendix A.2.

We note that for $t > \tau$, the algorithm can only play the void action $\varnothing$, for which it holds $\tilde{g}_{t,i}(\varnothing) \leq 0$ for all $i \in [M]$. This means that the dual update can only decrease the multipliers (or keep them unchanged) and thus the analysis can focus on the interval $[1,\tau]$.

5 Theoretical Results

In this section, we prove the theoretical guarantees attained by Algorithm 1. In particular, given the previous bound on the Lagrangian variables, we analyze the constraint violation, the cumulative regret in the stochastic setting, and the cumulative α-regret in the adversarial setting.

5.1 Constraint Violation

Using Lemma 4.2, we can now establish the worst-case bound on the constraint violation V_T.

Theorem 5.1. *Let $\delta \in (0,1)$. Both in the stochastic and the adversarial setting, Algorithm 1 with learning rate $\eta := 1/60M\sqrt{2T\ln(T^2/\delta)}$, attains with probability at least $1-\delta$:*

$$V_T \leq 840\frac{M^2}{\rho}\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},$$

where $\rho := \rho_{\text{stoc}}$ in the stochastic setting and $\rho := \rho_{\text{adv}}$ in the adversarial setting. Moreover, the budget constraints are deterministically satisfied, i.e.,

$$\sum_{t\in[T]} h_{t,j}(x_t) \leq \beta_j T \quad \forall j \in [n].$$

Theorem 5.1 states that our algorithm guarantees, in both the stochastic and the adversarial setting, a constraint violation of the order of $\widetilde{\mathcal{O}}(\sqrt{T})$ without knowing the Slater's parameter. Intuitively,

by construction of the algorithm, budget constraints are never violated, since once the cumulative consumption of any resource j reaches $\beta_j T - 1$, the algorithm starts selecting the void action $\varnothing$. Hence, it is sufficient to bound the violation of the general constraints $g_{t,i}$ for all $i \in [m]$. The bound is obtained by applying Lemma 4.2 and using the OGD update definition $\lambda_{t+1,i} = \max\{0, \lambda_{t,i} + \eta g_{t,i}(x_t)\}$. Telescoping the update rule over all rounds t, we obtain that the cumulative violation can be upper-bounded by a factor proportional to the final value of the dual multiplier. Intuitively, every violation increases the corresponding multiplier, and therefore, large cumulative violations imply large dual variables. Since the void action satisfies $g_{t,i}(\varnothing) = 0$ for all $t > \tau$, the bound is studied on the violation accumulated up to τ. The full proof can be found in Appendix A.3.

5.2 Stochastic Setting

In this section, we evaluate the performance of Algorithm 1 in the stochastic regime, where inputs are drawn from a fixed but unknown distribution.

Theorem 5.2. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, in the stochastic setting it holds:*

$$R_T \leq \frac{1}{\beta_{min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}}.$$

Theorem 5.2 shows that Algorithm 1 achieves the desired $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound for the stochastic setting. The analysis proceeds in two main steps. First, we establish an upper bound on the baseline OPT_{stoc} using weak duality and the definition of expected Lagrangian $\mathcal{L}_{f,\tilde{\boldsymbol{g}}}(\boldsymbol{\lambda}|\mathcal{P}) := \mathbb{E}_{(f,\tilde{\boldsymbol{g}})\sim\mathcal{P}}[\mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x_t, \boldsymbol{\lambda})]$. Then, we analyze the algorithm's performance up to the random stopping time τ. We lower-bound the instantaneous algorithm reward $f_t(x_t)$ and consider its expected value conditional on the past $\Gamma_{t-1} = \{\gamma_1, ..., \gamma_{t-1}\}$. We then relate the cumulative reward to the instantaneous Lagrangian by applying a martingale argument on both the reward and the penalty term. Leveraging the regret guarantees of the OGD update, we bound the penalty term $-\sum_{t\in[\tau]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle$ by exploiting the fact the, when the budget of some resource k is exhausted, the cumulative consumption provides a lower bound on the dual penalty. The extended analysis can be found in Appendix A.4.

5.3 Adversarial Setting

Finally, in this section, we analyze the performance of the algorithm in the adversarial setting, when inputs may arbitrarily change at each round.

Theorem 5.3. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, in the adversarial setting it holds:*

$$\alpha\text{-}R_T \leq \frac{1}{\beta_{min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}}.$$

Theorem 5.3 proves that our algorithm attains a bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ for the α-regret in the adversarial setting. The result follows from the regret guarantees of the unconstrained OGD update. We first split the α-regret into the contribution up to the stopping time τ and the contribution after τ. After τ, the algorithm is forced to play $\varnothing$, and therefore the regret is trivially bounded by $T - \tau$. For the rounds up to τ, we exploit the fact that the primal decision maximizes the instantaneous Lagrangian. To define the α-regret, we evaluate the primal performance against a specifically tailored reference sequence $(\xi_t^\diamond)_{t=1}^T$. Specifically, for each round $t \in [T]$, we define the strategy $\xi_t^\diamond$ as the one which plays x_t^* with probability α and x_t° with probability $1-\alpha$, where $\alpha = \rho_{\text{adv}}/1+\rho_{\text{adv}}$, $(x_t^*)_{t=1}^T$ is the optimal sequence of actions and $(x_t^\circ)_{t=1}^T$ is the sequence of strictly feasible actions that leads to the computation of ρ_{adv}. By definition of α, the sequence $(\xi_t^\diamond)_{t=1}^T$ is feasible, in expectation. This allows us to reduce the analysis to the cumulative penalty term induced by the algorithm $-\sum_{t\in[\tau]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle$. As in Theorem 5.2, this dual penalty is bounded by exploiting the regret guarantees of the OGD update. The detailed proof can be found in Appendix A.5.

6 Discussion and Open Problems

In this work, we study a generalized online resource allocation setting that incorporates both budget constraints and general constraints. We develop an algorithm that achieves best-of-both-world guarantees in this broader framework. Specifically, against the dynamic benchmark, it attains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the stochastic regime and α-regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$ in the adversarial regime, while ensuring strict satisfaction of the budget constraints and sublinear cumulative violation for the general ones.

An interesting direction for future work is to understand whether the results provided by this work in the adversarial regime can be improved in non-stationary settings, where the environment may change over time but the amount of non-stationarity is assumed to be bounded.

Acknowledgments and Disclosure of Funding

This publication was funded with the contribution of Ministero dellUniversità e della ricerca pursuant to D.D. n. 18010 of 12 November 2025 BANDO FIS 3. Project FIS-2024-05736 (Starting Grant), title: Towards a trustworthy strategic use of data in machine learning pipelines (STRATDATA). CUP: D53C25002380001

References

Shipra Agrawal and Nikhil R. Devanur. Bandits with global convex constraints and objective. *Oper. Res.*, 67(5):14861502, September 2019. ISSN 0030-364X. doi: 10.1287/opre.2019.1840. URL https://doi.org/10.1287/opre.2019.1840.

Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *CoRR*, abs/0911.2974, 2009. URL http://arxiv.org/abs/0911.2974.

Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, page 207216. IEEE, October 2013. doi: 10.1109/focs.2013.30. URL http://dx.doi.org/10.1109/FOCS.2013.30.

Michael O. Ball and Maurice Queyranne. Toward robust revenue management: Competitive analysis of online booking. *Operations Research*, 57(4):950–963, 2009. doi: 10.1287/opre.1080.0654. URL https://doi.org/10.1287/opre.1080.0654.

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

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023.

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization, 2024. URL https://arxiv.org/abs/2405.06575.

Niv Buchbinder, Kamal Jain, and Joseph Seffi Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In *Proceedings of the 15th Annual European Conference on Algorithms*, ESA'07, page 253264, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540755195.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2767–2783. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/castiglioni22a.html.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning under budget and roi constraints via weak adaptivity. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 5792–5816. PMLR, 2024.

Dragos Florin Ciocan and Vivek F. Farias. Dynamic allocation problems with volatile demand. In *Semantic Scholar*, 2011. URL https://api.semanticscholar.org/CorpusID:17524307.

Nikhil Devanur and Thomas P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *ACM Conference on Electronic Commerce*. Association for Computing Machinery, Inc., January 2009.

Nikhil R. Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A. Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *CoRR*, abs/1903.03944, 2019. URL http://arxiv.org/abs/1903.03944.

Hossein Esfandiari, Nitish Korula, and Vahab S. Mirrokni. Online allocation with traffic spikes: Mixing adversarial and stochastic models. *CoRR*, abs/1711.05764, 2017. URL http://arxiv.org/abs/1711.05764.

Jon Feldman, Monika Henzinger, Nitish Korula, Vahab S. Mirrokni, and Clifford Stein. Online stochastic ad allocation: Efficiency and fairness. *CoRR*, abs/1001.5076, 2010. URL http://arxiv.org/abs/1001.5076.

Zhe Feng, Swati Padmanabhan, and Di Wang. Online bidding algorithms for return-on-spend constrained advertisers, 2023. URL https://arxiv.org/abs/2208.13713.

Elad Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019. URL http://arxiv.org/abs/1909.05207.

Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks, 2023. URL https://arxiv.org/abs/1811.11881.

Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints, 2015. URL https://arxiv.org/abs/1512.07422.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. *CoRR*, abs/2010.07346, 2020. URL https://arxiv.org/abs/2010.07346.

Thomas Kesselheim, Klaus Radke, Andreas Tönnis, and Berthold Vöcking. Primal beats dual on online packing lps in the random-order model. *CoRR*, abs/1311.2578, 2013. URL http://arxiv.org/abs/1311.2578.

Xiaocheng Li, Chunlin Sun, and Yinyu Ye. Simple and fast algorithm for binary integer and online linear programming. *CoRR*, abs/2003.02513, 2020. URL https://arxiv.org/abs/2003.02513.

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: Online convex optimization with long term constraints, 2012. URL https://arxiv.org/abs/1111.6082.

Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(20):569–590, 2009. URL http://jmlr.org/papers/v10/mannor09a.html.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5):22es, October 2007. ISSN 0004-5411. doi: 10.1145/1284320.1284321. URL https://doi.org/10.1145/1284320.1284321.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Christian Kroer. No-regret learning under adversarial resource constraints: A spending plan is all you need! In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Rui Sun, Xinshang Wang, and Zijie Zhou. Near-optimal primal-dual algorithms for quantity-based network revenue management, 2022. URL https://arxiv.org/abs/2011.06327.

Wen Sun, Debadeepta Dey, and Ashish Kapoor. Safety-aware algorithms for adversarial contextual bandit. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3280–3288. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sun17a.html.

Kalyan T. Talluri and Garrett J. van Ryzin. *The Theory and Practice of Revenue Management*. International Series in Operations Research & Management Science. Springer New York, NY, 2004. ISBN 978-1-4020-7701-2. doi: 10.1007/b139000.

Hao Yu, Michael J. Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints, 2017. URL https://arxiv.org/abs/1708.03741.

Yunhong Zhou, Deeparnab Chakrabarty, and Rajan Lukose. Budget constrained bidding in keyword auctions and online knapsack problems. In Christos Papadimitriou and Shuzhong Zhang, editors, *Internet and Network Economics*, pages 566–576, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-92185-1.

Contents

1 Introduction **1**

1.1 Original Contributions 2

1.2 Related Works 3

2 Preliminaries **3**

2.1 Stochastic Setting 4

2.2 Adversarial Setting 5

2.3 Lagrangian Formulation 5

3 Algorithm **5**

4 Bounding the Lagrangian Multipliers **6**

4.1 Difference Between Budget and General Constraints 6

4.2 A Linear Upper Bound 7

5 Theoretical Results **8**

5.1 Constraint Violation 8

5.2 Stochastic Setting 9

5.3 Adversarial Setting 9

6 Discussion and Open Problems **10**

A Omitted Proofs and Lemmas **13**

A.1 Preliminary Results 13

A.2 Omitted Proofs and Lemmas of Section 4.2 15

A.3 Omitted Proofs of Section 5.1 19

A.4 Omitted Proofs of Section 5.2 20

A.5 Omitted Proofs of Section 5.3 22

B Dual Dynamic Simulation **22**

B.1 Experimental setup 23

B.2 Stochastic reward and stochastic constraints 23

B.3 Adversarial reward and stochastic constraints 24

B.4 Adversarial reward and adversarial constraints 24

A Omitted Proofs and Lemmas

A.1 Preliminary Results

Lemma A.1. *Let the dual algorithm be online gradient descent on the positive orthant $\mathbb{R}_+^M$ with learning rate $\eta > 0$. Then for any comparator $\boldsymbol{\mu} \in \mathbb{R}_+^M$, for any interval $[t_1, t_2] \subseteq [T]$ it holds:*

$$\sum_{t \in [t_1, t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t \rangle \geq \sum_{t \in [t_1, t_2]} \langle \boldsymbol{\mu}, \tilde{\boldsymbol{g}}_t \rangle - \frac{1}{2\eta} \|\boldsymbol{\lambda}_{t_1} - \boldsymbol{\mu}\|_2^2 - \frac{\eta}{2} TM.$$

Proof. By the properties of Euclidean projection onto a convex set:

$$\begin{aligned}\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\mu}\|_2^2 &= \|\Pi_\Lambda(\boldsymbol{\lambda}_t + \eta\tilde{\boldsymbol{g}}_t) - \boldsymbol{\mu}\|_2^2 \\ &\leq \|\boldsymbol{\lambda}_t + \eta\tilde{\boldsymbol{g}}_t - \boldsymbol{\mu}\|_2^2 \\ &= \|\boldsymbol{\lambda}_t - \boldsymbol{\mu}\|_2^2 + \eta^2\|\tilde{\boldsymbol{g}}_t\|_2^2 + 2\eta\langle\boldsymbol{\lambda}_t - \boldsymbol{\mu}, \tilde{\boldsymbol{g}}_t\rangle.\end{aligned}$$

Rearranging the terms:

$$\begin{aligned}2\eta\langle\boldsymbol{\mu} - \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t\rangle &\leq \|\boldsymbol{\lambda}_t - \boldsymbol{\mu}\|_2^2 - \|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\mu}\|_2^2 + \eta^2\|\tilde{\boldsymbol{g}}_t\|_2^2 \\ \langle\boldsymbol{\mu} - \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t\rangle &\leq \frac{1}{2\eta}\|\boldsymbol{\lambda}_t - \boldsymbol{\mu}\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}\|\tilde{\boldsymbol{g}}_t\|_2^2.\end{aligned}$$

Summing over t, we obtain:

$$\begin{aligned}\sum_{t\in[t_1,t_2]}\langle\boldsymbol{\mu} - \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t\rangle &\leq \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t_1} - \boldsymbol{\mu}\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t_2+1} - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}\sum_{t\in[t_1,t_2]}\|\tilde{\boldsymbol{g}}_t\|_2^2 \\ &\leq \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t_1} - \boldsymbol{\mu}\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t_2+1} - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}TM \\ &\leq \frac{1}{2\eta}\|\boldsymbol{\lambda}_{t_1} - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}TM.\end{aligned}$$

Rearranging the terms gives the stated bound. □

Lemma A.2. *Let $\{\boldsymbol{\lambda}_t\}_{t=1}^\tau$ be the sequence of dual multipliers generated by the OGD update rule (Lemma A.1) with learning rate $\eta > 0$ and $\boldsymbol{\lambda}_1 = \mathbf{0}$. Let $\tau \in [T]$ be the first round such that there exists a resource triggering the stopping condition, or $\tau = T$ if this condition is never met. Then, the dual penalty is bounded as:*

$$-\sum_{t\in[\tau]}\langle\boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \leq -(T-\tau) + \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM.$$

Proof. We analyze two cases by invoking Lemma A.1 on the interval $[\tau]$ with different comparators $\mu \in \mathbb{R}_+^M$:

- ($\tau = T$) We chose the comparator $\boldsymbol{\mu} = \mathbf{0}$. Since $\boldsymbol{\lambda}_1 = \mathbf{0}$, Lemma A.1 directly yields:

$$\begin{aligned}-\sum_{t\in[\tau]}\langle\boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle &\leq -\sum_{t\in[\tau]}\langle\boldsymbol{\mu}, \tilde{\boldsymbol{g}}_t(x_t)\rangle + \frac{1}{2\eta}\|\boldsymbol{\lambda}_1 - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}TM \\ &\leq \frac{\eta}{2}TM\end{aligned}$$

- ($\tau < T$) In this case, there exist a resource $k \in [n]$ that is exhausted at round τ. By definition of the unified constraint vector, we have:

$$\begin{aligned}\sum_{t\in[\tau]}\tilde{g}_{t,m+k}(x_t) &= \sum_{t\in[\tau]}(h_{t,k}(x_t) - \beta_k) \\ &= \sum_{t\in[\tau]}h_{t,k}(x_t) - \tau\beta_k \\ &\geq (\beta_k T - 1) - \tau\beta_k \\ &= \beta_k(T-\tau) - 1. \qquad (1)\end{aligned}$$

  We invoke Lemma A.1 choosing the comparator $\boldsymbol{\mu}$ such that $\mu_{m+k} = \frac{1}{\beta_k}$ and $\mu_j = 0$ for each $j \neq m+k$. Thus, we have:

$$-\sum_{t\in[\tau]}\langle\boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \leq -\sum_{t\in[\tau]}\langle\boldsymbol{\mu}, \tilde{\boldsymbol{g}}_t(x_t)\rangle + \frac{1}{2\eta}\|\boldsymbol{\lambda}_1 - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}TM$$

$$= -\frac{1}{\beta_k}\sum_{t\in[\tau]}\tilde{g}_{t,m+k} + \frac{1}{2\eta\beta_k^2} + \frac{\eta}{2}TM$$
$$\leq -(T-\tau) + \frac{1}{\beta_k} + \frac{1}{2\eta\beta_k^2} + \frac{\eta}{2}TM,$$

where the last step follows from Inequality (1). Therefore, in all cases it holds:

$$-\sum_{t\in[\tau]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \leq -(T-\tau) + \frac{1}{\beta_k} + \frac{1}{2\eta\beta_k^2} + \frac{\eta}{2}TM.$$

Defining $\beta_{\min} := \min_{j\in[n]}\beta_j$, it holds $\beta_k \geq \beta_{\min}$. Thus, we can write:

$$-\sum_{t\in[\tau]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \leq -(T-\tau) + \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM.$$

This concludes the proof. □

A.2 Omitted Proofs and Lemmas of Section 4.2

Lemma A.3. *Let $\boldsymbol{\lambda}_t \in \mathbb{R}^m_{\geq 0}$ be generated by OGD with learning rate η and utilities $\boldsymbol{\lambda} \to \langle \boldsymbol{\lambda}, \tilde{\boldsymbol{g}}_t\rangle$ where $\|\tilde{\boldsymbol{g}}_t\|_\infty \leq 1$ for all $t \in [T]$. Then:*

$$\left|\|\boldsymbol{\lambda}_{t+1}\|_1 - \|\boldsymbol{\lambda}_t\|_1\right| \leq \eta \cdot M.$$

Proof. By definition, the OGD update can be written as:

$$\lambda_{t+1,i} := \max\{0, \lambda_{t,i} + \eta\tilde{g}_{t,i}\} \quad \forall i \in [M].$$

For all $i \in [M]$, $\lambda_{t+1,i} = \lambda_{t,i} + \eta\tilde{g}_{t,i} \leq \lambda_{t,i} + \eta$ when $\tilde{g}_{t,i} \geq 0$, while $\lambda_{t+1,i} \geq \lambda_{t,i} + \eta\tilde{g}_{t,i} \geq \lambda_{t,i} - \eta$ when $\tilde{g}_{t,i} < 0$. Therefore, in all cases it holds:

$$|\lambda_{t+1,i} - \lambda_{t,i}| \leq \eta \quad \forall i \in [M].$$

Taking the sum over all components we obtain:

$$\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\|_1 \leq \eta \cdot M.$$

By the triangular inequality, we get:

$$\left|\|\boldsymbol{\lambda}_{t+1}\|_1 - \|\boldsymbol{\lambda}_t\|_1\right| \leq \eta \cdot M.$$

This concludes the proof. □

Lemma A.4. *In the stochastic setting, for any $\pi(\cdot) : \Gamma \to \mathcal{X}$ and $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\sum_{t\in[t_1,t_2]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle \leq \sum_{t\in[t_1,t_2]}\mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle\right] + 2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},$$

where $L = \max_{t\in[T]}\|\boldsymbol{\lambda}_t\|_1$ and $t_1, t_2 \in [T] : t_1 < t_2$.

Proof. Let $\mathcal{F}_{t-1}$ be the filtration representing the history up to the end of round $t-1$. Since:

$$\mathbb{E}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle\right],$$

it follows that the quantity $\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle - \mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle\right]$ is a martingale difference. By Azuma-Hoeffding inequality, given an interval $[t_1, t_2] \subseteq [T]$, with probability at least $1-\delta'$ it holds:

$$\sum_{t\in[t_1,t_2]}\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle - \sum_{t\in[t_1,t_2]}\mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle\right]$$
$$\leq 2L\sqrt{2(t_2-t_1)\ln\left(\frac{1}{\delta'}\right)} \leq 2L\sqrt{2T\ln\left(\frac{1}{\delta'}\right)}.$$

Taking $\delta = T^2\delta'$, we obtain that with probability at least $1-\delta$, by union bound over all possible intervals it holds:

$$\sum_{t\in[t_1,t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle - \sum_{t\in[t_1,t_2]} \mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi(\gamma_t))\rangle\right] \leq 2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}.$$

This concludes the proof. □

Lemma A.5. *Given an interval $[t_1,t_2]\subseteq[T]$ such that $\|\boldsymbol{\lambda}_t\|_1 \in [{}^{c_1}/_{\rho}, {}^{c_2}/_{\rho}]$ on $[t_1,t_2]$, $\|\boldsymbol{\lambda}_{t_2}\|_1 \geq {}^{c_2}/_{\rho}$ and $\|\boldsymbol{\lambda}_{t_1}\|_1 \leq {}^{c_1}/_{\rho}$. Given a learning rate $\eta>0$, it holds:*

$$\sum_{t\in[t_1,t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \geq \frac{M}{2\rho^2\eta}.$$

Proof. We analyze the cumulative penalty term during the interval $[t_1,t_2]$. For each constraint index $i\in[M]$, let $\bar{t}_i \in [t_1,t_2]$ be the last round such that the dual multiplier is zero. If $\lambda_{t,i}>0$ for the entire interval, we simply set $\bar{t}_i = t_1$. Formally:

$$\bar{t}_i = \max\left\{t_1, \max_{\tau\in[t_2]:\lambda_{\tau,i}=0} \tau\right\}.$$

We can now split the analysis into two separate sub-intervals: $[t_1,\bar{t}_i]$ and $[\bar{t}_i,t_2]$.

- $(t\in[t_1,\bar{t}_i])$: By definition, it can be either that $\lambda_{\bar{t}_i}=0$ or $\bar{t}_i = t_1$. In the latter case, $[t_1,\bar{t}_i]=\varnothing$ and the dual regret is trivially zero. In the former case, by Lemma A.1 with comparator $\boldsymbol{\mu}=\mathbf{0}$ we obtain:

$$0 \leq \sum_{t\in[t_1,\bar{t}_i]} \lambda_{t,i}\tilde{g}_{t,i}(x_t) + \frac{\lambda_{t_1}^2}{2\eta} + \frac{1}{2}\eta T. \tag{2}$$

- $(t\in[\bar{t}_i,t_2])$: Due to the definition of $\bar{t}_i$, gradient descent never projects the multiplier relative to constraint i, and we can write that

$$\sum_{t\in[\bar{t}_i,t_2]} \tilde{g}_{t,i}(x_t) = \frac{\lambda_{t_2,i}-\lambda_{\bar{t}_i,i}}{\eta}. \tag{3}$$

  By Lemma A.1, choosing the comparator $\boldsymbol{\mu}$ such that $\mu_i = {}^1/_{\rho}$ for all $i\in[M]$ we obtain:

$$\begin{aligned}\sum_{t\in[\bar{t}_i,t_2]} \lambda_{t,i}\tilde{g}_{t,i}(x_t) &\geq \sum_{t\in[\bar{t}_i,t_2]} \mu_i\tilde{g}_{t,i}(x_t) - \frac{(\mu_i-\lambda_{\bar{t}_i,i})^2}{2\eta} - \frac{1}{2}\eta T \\ &= \frac{\lambda_{t_2,i}-\lambda_{\bar{t}_i,i}}{\rho\eta} - \frac{(\mu_i-\lambda_{\bar{t}_i,i})^2}{2\eta} - \frac{1}{2}\eta T, \end{aligned} \tag{4}$$

  where we used Equation (3).

We can now combine the results from both sub-intervals. Summing Inequalities (2)-(4) we obtain:

$$\begin{aligned}\sum_{t\in[t_1,t_2]} \lambda_{t,i}\tilde{g}_{t,i}(x_t) &\geq \frac{\lambda_{t_2,i}-\lambda_{\bar{t}_i,i}}{\rho\eta} - \frac{(\mu_i-\lambda_{\bar{t}_i,i})^2}{2\eta} - \frac{\lambda_{t_1}^2}{2\eta} - \eta T \\ &\geq \frac{\lambda_{t_2,i}-\lambda_{\bar{t}_i,i}}{\rho\eta} - \frac{(\mu_i)^2+\lambda_{\bar{t}_i,i}^2}{2\eta} - \frac{\lambda_{t_1}^2}{2\eta} - \eta T.\end{aligned}$$

Finally, by summing over all $i\in[M]$, and defining $\boldsymbol{\lambda}_{\bar{t}}$ as the vector that has $\lambda_{\bar{t}_i}$ as its i-th component, we get:

$$\sum_{t\in[t_1,t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle \geq \frac{\|\boldsymbol{\lambda}_{t_2}\|_1 - \|\boldsymbol{\lambda}_{\bar{t}}\|_1}{\rho\eta} - \frac{1}{2\eta}\left(\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\lambda}_{\bar{t}}\|_2^2 + \|\boldsymbol{\lambda}_{t_1}\|_2^2\right) - \frac{M}{\eta} \tag{5}$$

$$
\begin{aligned}
&\geq \frac{c_2}{\rho^2\eta} - \frac{1}{\rho\eta}\|\boldsymbol{\lambda}_{t_1}\|_1 - \frac{1}{2\eta}\left(\|\boldsymbol{\mu}\|_2^2 + 2\|\boldsymbol{\lambda}_{t_1}\|_2^2\right) - \frac{M}{\eta} && (6)\\
&\geq \frac{c_2}{\rho^2\eta} - \frac{1}{\rho\eta}\left(\frac{c_1}{\rho} + M\eta\right) - \frac{1}{2\eta}\left(\frac{M}{\rho^2} + 2\left(\frac{c_1}{\rho} + M\eta\right)^2\right) - \frac{M}{\eta}\\
&\geq \frac{c_2}{\rho^2\eta} - \frac{c_1+1}{\rho^2\eta} - \frac{M}{2\rho^2\eta} - \frac{2(c_1+1)^2}{2\rho^2\eta} - \frac{M}{\eta} && (7)\\
&\geq \frac{2c_2 - 6 - M - 18 - 2M}{2\rho^2\eta}\\
&\geq \frac{M}{2\rho^2\eta},
\end{aligned}
$$

where Inequality (5) holds by definition of $\eta \leq 1/\sqrt{T}$, Inequality (6) is due to the fact that $\|\boldsymbol{\lambda}\|_1 \geq c_2/\rho$ and $\|\boldsymbol{\lambda}_{\bar{t}}\|_1 \leq \|\boldsymbol{\lambda}_{t_1}\|_1$ and Inequality (7) follows from the condition that $\eta \leq 1/\rho M$. The last two inequalities hold due to the choice of parameters, that is, $c_1 = 2$ and $c_2 = 14M$. This concludes the proof. □

Lemma 4.2. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, the ℓ_1-norm of the dual variables is bounded for all $t \in [T]$ as:*

$$\|\boldsymbol{\lambda}_t\|_1 \leq \frac{14M}{\rho}.$$

Proof. Let $c_1 := 2$, $c_2 := 14M$ and $\eta > 0$ such that $\eta \leq \eta_{\text{OGD}}$. Moreover, let $\rho = \rho_{\text{stoc}}$ and $\rho = \rho_{\text{adv}}$, respectively in the stochastic and in the adversarial setting. By contradiction, let us suppose that at some point $\|\boldsymbol{\lambda}_t\|_1 \geq \frac{c_2}{\rho}$. Let $t_2 \in [T]$ be the first round in which this condition happens. Moreover, let us define t_1 as the last round such that for all $t \in [t_1, t_2]$, it holds $\|\boldsymbol{\lambda}_t\|_1 \in \left[\frac{c_1}{\rho}, \frac{c_2}{\rho}\right]$.

From Lemma A.3, we have:

$$\left|\|\boldsymbol{\lambda}_{t+1}\|_1 - \|\boldsymbol{\lambda}_t\|_1\right| \leq M\eta.$$

Telescoping the sum for all $t \in [t_1, t_2]$ we obtain:

$$\|\boldsymbol{\lambda}_{t_2}\|_1 - \|\boldsymbol{\lambda}_{t_1}\|_1 \leq M\eta(t_2 - t_1).$$

Moreover, by definition of t_1 and t_2 we can write the following inequalities:

$$
\begin{aligned}
\frac{c_1}{\rho} \leq \|\boldsymbol{\lambda}_{t_1}\|_1 \leq \|\boldsymbol{\lambda}_{t_1-1}\|_1 + M\eta \leq \frac{c_1}{\rho} + M\eta,\\
\frac{c_2}{\rho} \leq \|\boldsymbol{\lambda}_{t_2}\|_1 \leq \|\boldsymbol{\lambda}_{t_2-1}\|_1 + M\eta \leq \frac{c_2}{\rho} + M\eta.
\end{aligned}
$$

Putting everything together, we obtain:

$$
\begin{aligned}
\frac{c_2}{\rho} - \frac{c_1}{\rho} - M\eta \leq \|\boldsymbol{\lambda}_{t_2}\|_1 - \|\boldsymbol{\lambda}_{t_1}\|_1 \leq M\eta(t_2 - t_1)\\
\frac{c_2 - c_1}{\rho} - M\eta \leq M\eta(t_2 - t_1)\\
t_2 - t_1 \geq \frac{c_2 - c_1}{\rho M\eta} - 1 \geq \frac{c_2 - c_1}{2\rho M\eta}. \qquad (8)
\end{aligned}
$$

We now provide a lower bound for the quantity $\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t, \boldsymbol{\lambda}_t)$, holding for both the stochastic and the adversarial setting.

Since the primal decision x_t is such that $x_t = \arg\max_{x\in\mathcal{X}} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x, \boldsymbol{\lambda}_t)$, it point-wise dominates any other action. Let $\pi^\circ : \Gamma \to \mathcal{X}$ be the function leading to the computation of ρ_{stoc}, i.e. $\sum_{t\in[t_1,t_2]} \mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\tilde{\boldsymbol{g}}_t(\pi^\circ(\gamma_t))\right] \leq -\rho_{\text{stoc}}$ for all $i \in [M]$ Thus, in the stochastic setting, we can write:

$$\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t, \boldsymbol{\lambda}_t) \qquad (9)$$

$$
\begin{aligned}
&\geq \sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(\pi^\circ(\gamma_t),\boldsymbol{\lambda}_t) \\
&= \sum_{t\in[t_1,t_2]} f_t(\pi^\circ(\gamma_t)) - \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi^\circ(\gamma_t))\rangle \\
&\geq - \sum_{t\in[t_1,t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi^\circ(\gamma_t))\rangle \\
&\geq - \sum_{t\in[t_1,t_2]} \mathbb{E}_{\gamma_t=(f_t,\tilde{\boldsymbol{g}}_t)\sim\mathcal{P}}\left[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(\pi^\circ(\gamma_t))\rangle\right] - 2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \qquad (10)\\
&\geq \rho_{\text{stoc}} \sum_{t\in[t_1,t_2]} \|\boldsymbol{\lambda}_t\|_1 - 2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \qquad (11)\\
&\geq c_1(t_2-t_1) - 2\left(\frac{c_2}{\rho_{\text{stoc}}} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},
\end{aligned}
$$

where Inequality (10) holds by Lemma A.4, and Inequality (11) holds by definition of π°.

In the adversarial setting, let $(x_t^\circ)_{t=1}^T$ be the sequence of strictly feasible actions that leads to the definition of ρ_{adv}, meaning $\tilde{g}_{t,i}(x_t^\circ) \leq -\rho_{\text{adv}}$ for all $i \in [M]$. It holds:

$$
\begin{aligned}
\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t) &\geq \sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t^\circ,\boldsymbol{\lambda}_t) \qquad (12)\\
&= \sum_{t\in[t_1,t_2]} f_t(x_t^\circ) - \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t^\circ)\rangle \\
&\geq - \sum_{t\in[t_1,t_2]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t^\circ)\rangle \\
&\geq \rho_{\text{adv}} \sum_{t\in[t_1,t_2]} \|\boldsymbol{\lambda}_t\|_1 \qquad (13)\\
&\geq c_1(t_2-t_1),
\end{aligned}
$$

where Inequality (13) holds by definition of $(x_t^\circ)_{t=1}^T$.

Thus, for both settings it holds:

$$
\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t) \geq c_1(t_2-t_1) - 2\left(\frac{c_2}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}. \qquad (14)
$$

We employ Lemma A.5 to upper bound the following term $\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t)$:

$$
\begin{aligned}
\sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t) &= \sum_{t\in[t_1,t_2]} \left(f_t(x_t) - \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right) \\
&\leq \sum_{t\in[t_1,t_2]} f_t(x_t) - \frac{M}{2\rho^2\eta} \\
&\leq (t_2-t_1) - \frac{M}{2\rho^2\eta}. \qquad (15)
\end{aligned}
$$

Putting together Inequalities (14)-(15) we obtain:

$$
c_1(t_2-t_1) - 2\left(\frac{c_2}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \leq \sum_{t\in[t_1,t_2]} \mathcal{L}_{f_t,\tilde{\boldsymbol{g}}_t}(x_t,\boldsymbol{\lambda}_t) \leq (t_2-t_1) - \frac{M}{2\rho^2\eta}
$$

$$
c_1(t_2-t_1) - 2\left(\frac{c_2}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \leq (t_2-t_1) - \frac{M}{2\rho^2\eta}
$$

$$t_2 - t_1 \leq \frac{1}{c_1 - 1}\left[2\left(\frac{c_2}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} - \frac{M}{2\rho^2\eta}\right]$$
$$\frac{c_2 - c_1}{2\rho M\eta} \leq \frac{1}{c_1 - 1}\left[2\left(\frac{c_2}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} - \frac{M}{2\rho^2\eta}\right], \tag{16}$$

where the last inequality holds by Inequality (8). By substituting the constants $c_1 = 2$ and $c_2 = 14M$, we get:

$$\frac{14M - 2}{2\rho M\eta} \leq 2\left(\frac{14M}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} - \frac{M}{2\rho^2\eta}$$
$$\frac{\rho(14M - 2) + M^2}{2\rho^2 M\eta} \leq 2\left(\frac{14M}{\rho} + M\eta\right)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}$$
$$\frac{M^2 + 14\rho M - 2\rho}{2\rho^2 M\eta} \leq \frac{2(14M + 1)}{\rho}\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},$$

where we used the standard assumption that $\eta \leq \frac{1}{\rho M}$.

Rearranging, we obtain:

$$\eta \geq \frac{M^2 + 14\rho M - 2\rho}{4\rho M(14M + 1)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}},$$

which yields a contradiction since, by hypothesis, it holds:

$$\eta \leq \eta_{\text{OGD}} := \frac{1}{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}} < \frac{M^2 + 14\rho M - 2\rho}{4\rho M(14M + 1)\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}$$

Thus, we conclude that the assumption must be false, implying $\|\boldsymbol{\lambda}_t\|_1 \leq \frac{c_2}{\rho} = \frac{14M}{\rho}$ for each $t \in [T]$. This concludes the proof. □

A.3 Omitted Proofs of Section 5.1

Theorem 5.1. *Let $\delta \in (0, 1)$. Both in the stochastic and the adversarial setting, Algorithm 1 with learning rate $\eta := 1/60M\sqrt{2T\ln(T^2/\delta)}$, attains with probability at least $1 - \delta$:*

$$V_T \leq 840\frac{M^2}{\rho}\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},$$

where $\rho := \rho_{\text{stoc}}$ in the stochastic setting and $\rho := \rho_{\text{adv}}$ in the adversarial setting. Moreover, the budget constraints are deterministically satisfied, i.e.,

$$\sum_{t\in[T]} h_{t,j}(x_t) \leq \beta_j T \quad \forall j \in [n].$$

Proof. Let $\tau \leq T$ be the stopping time at which the budget is exhausted. By design of the algorithm, if playing the candidate action exceeds the budget $\beta_j T$ for any resource $j \in [n]$, the algorithm is forced to play the void action $x_t = \varnothing$, which yields zero consumption ($h_{t,j}(\varnothing) = 0$ for all $j \in [n]$). Thus, the budget constraints are never violated, i.e., it holds $\sum_{t\in[T]} h_{t,j}(x_t) = \sum_{t\in[\tau]} h_{t,j}(x_t) \leq \beta_j T$ for all $j \in [n]$. For the general constraints, by definition of the OGD update, it holds:

$$\lambda_{t+1,i} \geq \lambda_{t,i} + \eta g_{t,i} \quad \forall t \in [\tau], \forall i \in [m].$$

Telescoping the sum for all $t \in [\tau]$ we obtain:

$$\sum_{t\in[\tau]} \lambda_{t+1,i} \geq \sum_{t\in[\tau]} \lambda_{t,i} + \eta \sum_{t\in[\tau]} g_{t,i}$$

$$\lambda_{\tau+1,i} \geq \lambda_{1,i} + \eta \sum_{t\in[\tau]} g_{t,i}.$$

Therefore, we can write:

$$\begin{aligned}
\sum_{t\in[\tau]} g_{t,i} &\leq \frac{1}{\eta}\lambda_{\tau+1,i} \\
&\leq \frac{1}{\eta}\|\boldsymbol{\lambda}_{\tau+1}\|_1 \\
&\leq \frac{1}{\eta}\left(\frac{14M}{\rho}\right) \qquad (17) \\
&= 60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}\cdot\left(\frac{14M}{\rho}\right) \\
&= 840\frac{M^2}{\rho}\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)},
\end{aligned}$$

where Inequality (17) holds by Lemma 4.2 with probability at least $1-\delta$. Since from round $\tau+1$ the algorithm only plays $\varnothing$, we have:

$$V_T := \max_{i\in[m]} \sum_{t\in[T]} g_{t,i}(x_t) = \max_{i\in[m]} \sum_{t\in[\tau]} g_{t,i}(x_t).$$

Thus, with probability at least $1-\delta$ it holds:

$$V_T \leq 840\frac{M^2}{\rho}\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}.$$

This concludes the proof. □

A.4 Omitted Proofs of Section 5.2

Theorem 5.2. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, in the stochastic setting it holds:*

$$R_T \leq \frac{1}{\beta_{min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}}.$$

Proof. First, we provide an upper bound on the hindsight stochastic optimum. We define the expected Lagrangian evaluated with respect to the input distribution as:

$$\bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\boldsymbol{\lambda}|\mathcal{P}) := \mathbb{E}_{(f,\tilde{\boldsymbol{g}})\sim\mathcal{P}}\left[\max_{x\in\mathcal{X}} \mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x,\boldsymbol{\lambda})\right].$$

By weak duality, $\mathrm{OPT}(\boldsymbol{\gamma}) \leq \max_{x\in\mathcal{X}_t} \sum_{t\in T} \mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x,\boldsymbol{\lambda})$ for any dual variable $\boldsymbol{\lambda} \in \mathbb{R}_+^M$. Let $\bar{\boldsymbol{\lambda}}_\tau := \frac{1}{\tau}\sum_{t\in[\tau]} \boldsymbol{\lambda}_t$. We can bound the stochastic optimum as:

$$\begin{aligned}
\mathrm{OPT}_{stoc} = \mathbb{E}_{\boldsymbol{\gamma}}[\mathrm{OPT}(\boldsymbol{\gamma})] &= \frac{\tau}{T}\mathbb{E}_{\boldsymbol{\gamma}}[\mathrm{OPT}(\boldsymbol{\gamma})] + \frac{T-\tau}{T}\mathbb{E}_{\boldsymbol{\gamma}}[\mathrm{OPT}(\boldsymbol{\gamma})] \\
&\leq \tau\bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\bar{\boldsymbol{\lambda}}_\tau|\mathcal{P}) + (T-\tau), \qquad (18)
\end{aligned}$$

where we used that rewards are in $[0,1]$.

Now, we analyze the performance of the algorithm up to the stopping time τ. At each step $t \in [\tau]$, Algorithm 1 maximizes the instantaneous Lagrangian choosing $f_t(x_t) = \max_{x\in\mathcal{X}} \mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x,\boldsymbol{\lambda}_t) + \langle\boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x)\rangle$. Therefore, we can write the instantaneous reward as:

$$f_t(x_t) = \mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x_t,\boldsymbol{\lambda}_t) + \langle\boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle.$$

Let $\Gamma_t = (\gamma_1, ..., \gamma_t)$ denote the history of observed inputs up to time t for each $t \in [\tau]$. Since $\boldsymbol{\lambda}_t$ is Γ_{t-1}-measurable, taking the conditional expectation we obtain:

$$\mathbb{E}[f_t(x_t)|\Gamma_{t-1}] = \mathbb{E}[\mathcal{L}_{f,\tilde{\boldsymbol{g}}}(x_t, \boldsymbol{\lambda}_t)|\Gamma_{t-1}] + \mathbb{E}[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle|\Gamma_{t-1}]$$
$$\mathbb{E}[f_t(x_t)|\Gamma_{t-1}] = \bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\boldsymbol{\lambda}_t|\mathcal{P}) + \mathbb{E}[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle|\Gamma_{t-1}]. \tag{19}$$

To analyze the performance up to the stopping time τ, we consider the martingales $Z_t = \sum_{s\in[t]} (\langle \boldsymbol{\lambda}_s, \tilde{\boldsymbol{g}}_s(x_s)\rangle - \mathbb{E}[\langle \boldsymbol{\lambda}_s, \tilde{\boldsymbol{g}}_s(x_s)\rangle|\Gamma_{s-1}])$ and $Q_t = \sum_{s\in[t]} (f_s(x_s) - \mathbb{E}[f_s(x_s)|\Gamma_{s-1}])$. Since τ is a bounded stopping time with respect to $\{\Gamma_t\}_{t=1}^T$, from the Optimal Stopping Theorem we have:

$$\mathbb{E}\left[\sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right] = \mathbb{E}\left[\sum_{t\in[\tau]} \mathbb{E}[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle|\Gamma_{t-1}]\right] \tag{20}$$

and

$$\mathbb{E}\left[\sum_{t\in[\tau]} f_t(x_t)\right] = \mathbb{E}\left[\sum_{t\in[\tau]} \mathbb{E}[f_t(x_t)|\Gamma_{t-1}]\right]. \tag{21}$$

Thus, summing Equation (19) over t and taking the expectation on both sides, we have:

$$\mathbb{E}\left[\sum_{t\in[\tau]} \mathbb{E}[f_t(x_t)|\Gamma_{t-1}]\right] = \mathbb{E}\left[\sum_{t\in[\tau]} \bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\boldsymbol{\lambda}_t|\mathcal{P})\right] + \mathbb{E}\left[\sum_{t\in[\tau]} \mathbb{E}[\langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle|\Gamma_{t-1}]\right].$$

From Inequalities (20)-(21) we can write:

$$\mathbb{E}\left[\sum_{t\in[\tau]} f_t(x_t)\right] = \mathbb{E}\left[\sum_{t\in[\tau]} \bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\boldsymbol{\lambda}_t|\mathcal{P})\right] + \mathbb{E}\left[\sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right]$$
$$\geq \mathbb{E}\left[\tau \bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\bar{\boldsymbol{\lambda}}_\tau|\mathcal{P})\right] + \mathbb{E}\left[\sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right],$$

where the inequality holds by convexity of the expected Lagrangian. Combining this result with Inequality (18), we can bound the regret as:

$$R_T = \text{OPT}_{stoc} - \mathbb{E}_{\boldsymbol{\gamma}}\left[\sum_{t\in[\tau]} f_t(x_t)\right]$$
$$\leq \text{OPT}_{stoc} - \mathbb{E}\left[\tau \bar{\mathcal{L}}_{f,\tilde{\boldsymbol{g}}}(\bar{\boldsymbol{\lambda}}_\tau|\mathcal{P})\right] - \mathbb{E}\left[\sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right]$$
$$\leq \mathbb{E}\left[(T-\tau)\right] - \mathbb{E}\left[\sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right]$$
$$\leq \mathbb{E}\left[(T-\tau)\right] + \mathbb{E}\left[-\sum_{t\in[\tau]} \langle \boldsymbol{\mu}, \tilde{\boldsymbol{g}}_t(x_t)\rangle + \frac{1}{2\eta}\|\boldsymbol{\lambda}_1 - \boldsymbol{\mu}\|_2^2 + \frac{\eta}{2}TM\right] \tag{22}$$
$$\leq (T-\tau) - (T-\tau) + \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM \tag{23}$$
$$\leq \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM$$
$$\leq \frac{1}{\beta_{\min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{\min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}}.$$

where Inequality (23) holds by Lemma A.2. This concludes the proof. □

A.5 Omitted Proofs of Section 5.3

Theorem 5.3. *Let $\delta \in (0,1)$. Consider Algorithm 1 with learning rate $\eta = 1/60M\sqrt{2T\ln(T^2/\delta)}$. With probability at least $1-\delta$, in the adversarial setting it holds:*

$$\alpha\text{-}R_T \leq \frac{1}{\beta_{min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}}.$$

Proof. Let $(x_t^*)_{t=1}^T$ be the optimal sequence of actions and let $(x_t^\circ)_{t=1}^T$ be the sequence of strictly feasible actions that leads to the definition of ρ_{adv}, meaning $\tilde{g}_{t,i}(x_t^\circ) \leq -\rho_{\text{adv}}$ for all $i \in [M]$. Let $\alpha = \frac{\rho_{\text{adv}}}{1+\rho_{\text{adv}}}$. For each round $t \in [T]$, we define $\xi_t^\diamond$ as the strategy that plays x_t^* with probability α and x_t° with probability $1-\alpha$. By construction, the sequence $(\xi_t^\diamond)_{t=1}^T$ is always feasible in expectation. Indeed, for each $i \in [M]$, it holds:

$$\mathbb{E}_{x\sim\xi_t^\diamond}[\tilde{g}_{t,i}(x)] = \alpha \cdot \tilde{g}_{t,i}(x_t^*) + (1-\alpha)\tilde{g}_{t,i}(x_t^\circ) \leq \alpha \cdot 1 + (1-\alpha)(-\rho_{\text{adv}}) = 0. \tag{24}$$

Moreover, since rewards are non-negative, it holds:

$$\mathbb{E}_{x\sim\xi_t^\diamond}[f_t(x)] = \alpha \cdot f_t(x_t^*) + (1-\alpha)f_t(x_t^\circ) \geq \alpha \cdot f_t(x_t^*). \tag{25}$$

Defining $\tau \in [T]$ as the first time that a resource exceeds the budget, we can decompose the α-regret:

$$\begin{aligned}
\alpha\text{-}R_T &= \alpha \sum_{t\in[T]} f_t(x_t^*) - \sum_{t\in[\tau]} f_t(x_t) \\
&\leq \sum_{t\in[T]} \mathbb{E}_{x\sim\xi_t^\diamond}[f_t(x)] - \sum_{t\in[\tau]} f_t(x_t) && (26) \\
&= \sum_{t=\tau+1}^{T} \mathbb{E}_{x\sim\xi_t^\diamond}[f_t(x)] + \sum_{t\in[\tau]} \mathbb{E}_{x\sim\xi_t^\diamond}[f_t(x)] - \sum_{t\in[\tau]} f_t(x_t) \\
&\leq (T-\tau) + \sum_{t\in[\tau]} \left(\mathbb{E}_{x\sim\xi_t^\diamond}[f_t(x)] - f_t(x_t)\right) \\
&\leq (T-\tau) + \sum_{t\in[\tau]} \left(\langle \boldsymbol{\lambda}_t, \mathbb{E}_{x\sim\xi_t^\diamond}[\tilde{\boldsymbol{g}}_t(x)]\rangle - \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle\right) && (27) \\
&\leq (T-\tau) - \sum_{t\in[\tau]} \langle \boldsymbol{\lambda}_t, \tilde{\boldsymbol{g}}_t(x_t)\rangle, && (28)
\end{aligned}$$

where Inequality (26) follows from Inequality (25), Inequality (27) holds since x_t deterministically maximizes the instantaneous Lagrangian over $\mathcal{X}$ and Inequality (28) follows from Inequality (24). From Lemma A.2, similarly to the stochastic case, we have:

$$\begin{aligned}
\alpha\text{-}R_T &\leq (T-\tau) - (T-\tau) + \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM \\
&\leq \frac{1}{\beta_{\min}} + \frac{1}{2\eta\beta_{\min}^2} + \frac{\eta}{2}TM \\
&\leq \frac{1}{\beta_{\min}} + \frac{60M\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)}}{2\beta_{\min}^2} + \frac{\sqrt{T}}{120\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}},
\end{aligned}$$

which concludes the proof. □

B Dual Dynamic Simulation

In this section, we provide a series of synthetic simulations designed to study the evolution of the dual variables during the learning dynamics. Our goal is to illustrate how the algorithm effectively bounds the dual multipliers without requiring prior knowledge of the Slater's parameter ρ. The complete implementation is available at https://github.com/efc05/ORA.

B.1 Experimental setup

Throughout all experiments, the confidence parameter δ is fixed to 0.01. In the stochastic setting, the values of rewarda and constraints are sampled i.i.d. from a Bernoulli distribution (rescaled to $[-1, 1]$ in the case of general constraints). In the adversarial setting, rewards and constraints are generated by an adversary running an Online Gradient Descent (OGD) algorithm [Orabona, 2019] which uses the learner's previously selected action, with a uniform policy as initial update.

To ensure the statistical robustness of our empirical findings, each experiment is repeated over multiple independent runs ($n \simeq 10$). In the subsequent plots, we report the average ℓ_1-norm of the dual variables, denoted as $\|\boldsymbol{\lambda}\|$ and the maximum cumulative general-constraint violation V_T together with their 95% confidence intervals.

We structure our experimental evaluation across three different settings: (i) stochastic reward and stochastic constraints, (ii) adversarial reward and stochastic constraints, (iii) adversarial reward and adversarial constraints.

B.2 Stochastic reward and stochastic constraints

In this first scenario, both rewards and general/budget constraints are generated stochastically and independently across rounds from fixed distributions. This setting provides a stationary baseline for observing the natural dual response to sampled constraint costs.

Figure 2 reports the individual dual components together with their ℓ_1-norm $\|\boldsymbol{\lambda}\|$ and the cumulative general-constraint violation V_T over time. As expected from the dual update rule, the dual variables rise proportionally while positive cumulative costs exert pressure. Once the multipliers grow sufficiently large, the penalty term in the instantaneous Lagrangian objective completely dominates the bounded rewards. Consequently, the greedy primal step is forced to select actions that better satisfy the constraints, causing the dual variables to stabilize or decrease. Furthermore, the violation curve levels off after the void action $\varnothing$ is selected as the null action contributes zero general cost.

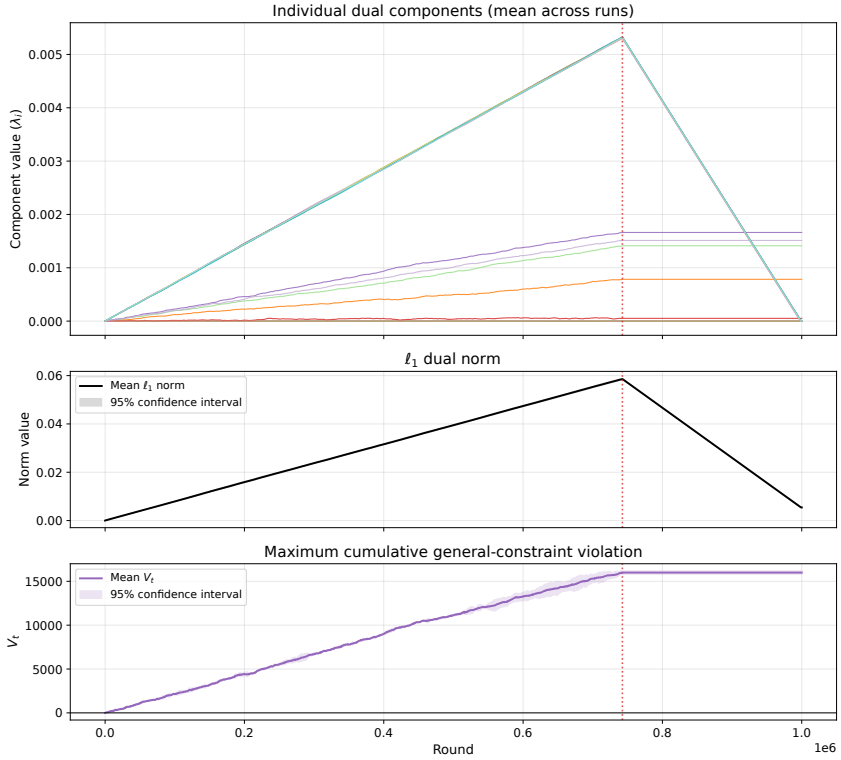


Figure 2: Stochastic rewards and stochastic constraints: individual dual components, overall ℓ_1-norm and cumulative constraint violation, averaged over runs.

B.3 Adversarial reward and stochastic constraints

In this intermediate setting, the constraint vectors remain stochastic, while the rewards are generated adversarially via OGD. This configuration isolates the effect of an adaptive reward sequence while retaining the stability of stationary general and budget constraints. As illustrated in Figure 3, the dual and violation trajectories demonstrate the learner's capability to adjusts its constraint prices dynamically. The algorithm effectively balances the observed constraint costs against the shifting values of the adversarial reward.

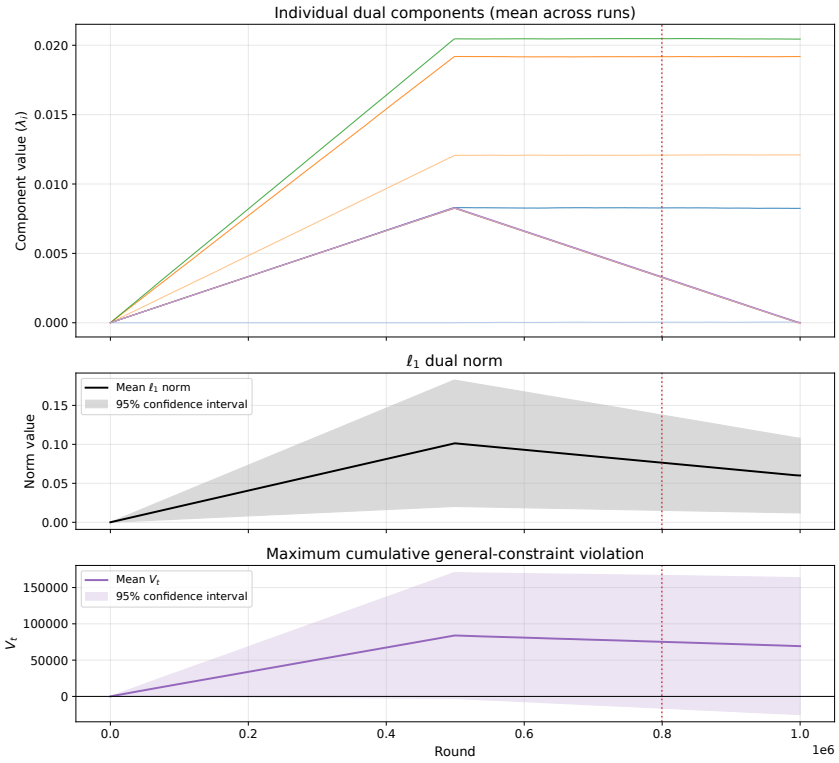


Figure 3: Stochastic constraints and adversarial rewards: individual dual components, overall ℓ_1-norm and cumulative constraint violation, averaged over runs.

B.4 Adversarial reward and adversarial constraints

Finally, we evaluate the algorithm in a fully adversarial regime. Here, both rewards and constraints are generated adversarially by OGD, with updates depending on the learner's previously selected action. This represents the most challenging scenario, testing the coupled evolution of the reward and constraint sequences under adaptive feedback. Figure 4 shows the robust response of the dual variables and the cumulative violations. Notably, despite the adversary actively pushing the constraints to induce violations, the weak adaptivity of the OGD dual updates successfully prevents the dual variables from diverging indefinitely. As theoretically proven, $\|\boldsymbol{\lambda}\|_1$ peak at a bounded threshold and subsequently decrease. Correspondingly, V_T flattens out over time, guaranteeing sublinear growth even in this worst-case scenario.

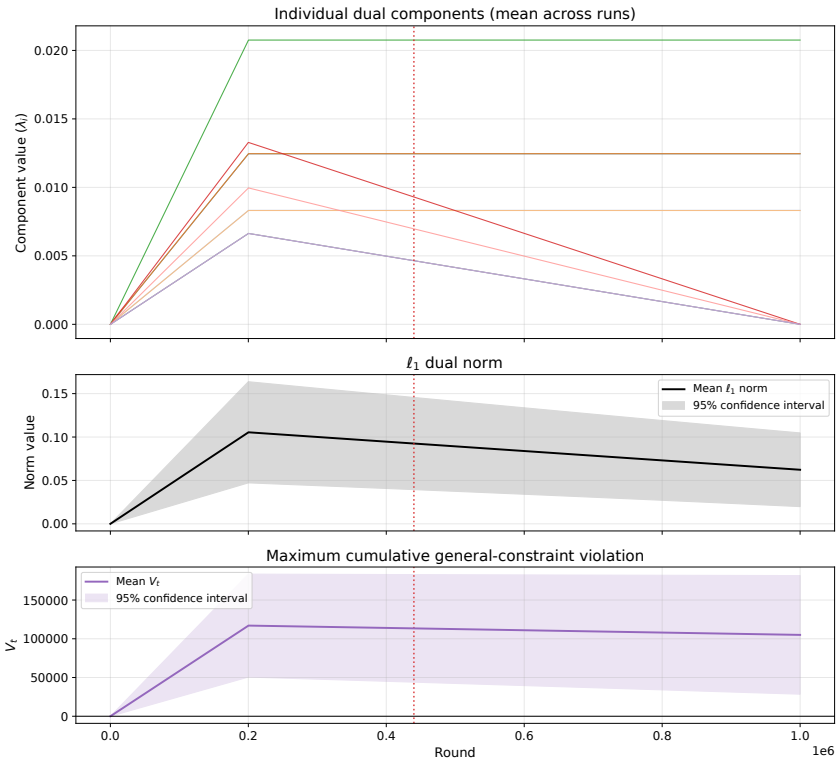


Figure 4: Adversarial rewards and adversarial constraints: individual dual components, overall ℓ_1-norm and cumulative constraint violation, averaged over runs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state all the main contributions made by the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All the assumptions are clearly stated in Section 2.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the theoretical results clearly state their assumptions, while all their proofs are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed, since the work is mainly theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.