A Unifying Framework for Online Safe Optimization

Matteo Castiglioni[*] **Andrea Celli**[†] **Alberto Marchesi**[*] **Giulia Romano**[*]

Nicola Gatti[*]

Abstract

We study online learning problems in which a decision maker has to take a sequence of decisions subject to m *long-term constraints*. The goal of the decision maker is to maximize their total reward, while at the same time achieving small cumulative constraints violation across the T rounds. We present the first *best-of-both-world* type algorithm for this general class of problems, with no-regret guarantees both in the case in which rewards and constraints are selected according to an unknown stochastic model, and in the case in which they are selected at each round by an adversary. Our algorithm is the first to provide guarantees in the adversarial setting with respect to the optimal fixed strategy that satisfies the long-term constraints. In particular, it guarantees a $\rho/(1+\rho)$ fraction of the optimal reward and sublinear regret, where ρ is a feasibility parameter related to the existence of strictly feasible solutions. Our framework employs traditional regret minimizers as black-box components. Therefore, by instantiating it with an appropriate choice of regret minimizers it can handle the *full-feedback* as well as the *bandit-feedback* setting. Moreover, it allows the decision maker to seamlessly handle scenarios with non-convex rewards and constraints. We show how our framework can be applied in the context of budget-management mechanisms for repeated auctions in order to guarantee long-term constraints that are not *packing* (*e.g.*, ROI constraints).

1 Introduction

We study online learning problems where a decision maker takes decisions over T rounds. At each round t, the decision $\boldsymbol{x}_t \in \mathcal{X}$ is chosen before observing a reward function f_t together with a set of m *time-varying* constraint functions g_t. The decision maker is allowed to make decisions that are *not* feasible, provided that the overall sequence of decisions obeys the *long-term constraints* $\sum_{t=1}^{T} g_t(\boldsymbol{x}_t) \le \mathbf{0}$, up to a small cumulative violation across the T rounds. The problem becomes that of finding a sequence of decisions $\boldsymbol{x}_t$ which guarantees a reward close to that of the best fixed decision in hindsight while satisfying long-term constraints. This type of framework was first proposed by Mannor et al. [23], and it has numerous applications ranging from wireless communication [23] and multi-objective online classification [11], to *safe* online learning [2, 9, 10].

Mannor et al. [23] show that guaranteeing sublinear regret and sublinear cumulative constraints violation is impossible even when f_t and g_t are simple linear functions. Therefore, previous works either focus on the case in which constraints are generated i.i.d. according to some unknown stochastic model, without providing any guarantees for the adversarial case, or provide results for adversarially-generated constraints under some strong assumptions on the structure of the problem or using a weaker baseline. A few examples in the latter case are [25, 27, 15, 13]. In the former setting (*i.e.*, stochastic constraints), Wei et al. [26] consider a weaker baseline that is feasible for each constraint g_t, going against the basic idea of long-term constraints. A notable exception is the work by Yu

[*] DEIB, Politecnico di Milano, {matteo.castiglioni, alberto.marchesi, giulia.romano, nicola.gatti}@polimi.it.
[†] Computing Sciences Department, Bocconi University, andrea.celli2@unibocconi.it.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Algorithm	Constr.	Non-convex	Bound — constant ρ		Bound — arbitrary ρ	
		f_t and g_t	Reward	Violation	Reward	Violation
Yu et al. [29]	Stoc	✗	$\texttt{OPT} - \tilde{O}(T^{1/2})$	$\tilde{O}(T^{1/2})$	—	—
Ours	Stoc	✓	$\texttt{OPT} - \tilde{O}(T^{1/2})$	$\tilde{O}(T^{1/2})$	$\texttt{OPT} - \tilde{O}(T^{3/4})$	$\tilde{O}(T^{3/4})$
	Adv	✓	$\frac{\rho}{1+\rho}\texttt{OPT} - \tilde{O}\Big(T^{1/2}\Big)$	$\tilde{O}(T^{1/2})$	—	—

Table 1: Comparison between performances of our algorithm and previous work using the same baseline. Bounds for settings that were previously intractable are in gray. $\texttt{OPT}$ is the baseline reward.

et al. [29], who employ the same baseline as ours, and provide an upper bound of $\tilde{O}(T^{1/2})$ for both regret and constraints violation (see Table 1). We also mention that there are some works studying the problem in which constraints are *static* (see, *e.g.*, [20, 22, 28, 30]), or focus on specific types of constraints, such as *knapsack constraints* [5, 19]. Our framework differs from those works as we deal with *arbitrary* and *time-varying* constraints. Moreover, it also extends the *online convex optimization* framework introduced by Zinkevich [31] by allowing for general non-convex loss functions f_t, arbitrary feasibility sets $\mathcal{X}$, and arbitrary time-varying long-term constraints.

Given the negative result by Mannor et al. [23], a natural question is what kind of guarantees we can reach in the adversarial setting, when adopting the standard baseline of the best fixed decision in hindsight satisfying (in expectation) the long-term constraints. We provide the first positive result going in this direction, by designing a no-α-regret algorithm that guarantees a sublinear cumulative constraints violation. Moreover, we make a step forward in the line of work initiated by Bubeck and Slivkins [12], by showing that our algorithm is also the first *best-of-both-worlds* algorithm for problems with arbitrary long-term constraints. This allows our algorithm to guarantee good worst-case performance (adversarial case), while being able to exploit well-behaved problem instances (stochastic case). The only assumption which we require is the existence of a decision that is strictly feasible with respect to the sequence of constraints. We denote by ρ the “margin” by which this decision is strictly feasible. At the same time, we show that even without this assumption, we can recover sublinear regret and violation with stochastic constraints.

Previous work usually assumes that ρ is a given *constant*. In that case, our algorithm matches the guarantees by Yu et al. [29] when constraints are generated i.i.d. according to an unknown distribution, and has no-α-regret with $\alpha = \rho/(1+\rho)$ in the adversarial case (Table 1). Moreover, we argue that if ρ is allowed to depend on T and take arbitrarily small values, then there are certain values for which any regret bound depending on $1/\rho$ would be useless. This setting is usually overlooked by previous work, which assumes ρ to be a given constant. We show that, in the case of an arbitrary ρ, in the stochastic setting our algorithm guarantees $\tilde{O}(T^{3/4})$ regret and cumulative constraints violation.

Our framework employs traditional regret minimizers as black-box components. Therefore, by instantiating it with an appropriate choice of regret minimizers it can handle *full-feedback* as well as *bandit-feedback* settings. In the former case, after playing $\boldsymbol{x}_t$, the decision maker gets to observe f_t and g_t, while in the latter case only the realized values $f_t(\boldsymbol{x}_t)$ and $g_t(\boldsymbol{x}_t)$ are observed. Moreover, this allows the decision maker to seamlessly handle scenarios with non-convex reward and constraints, by employing a suitable regret minimizer for non-convex losses (see, *e.g.*, [24]). Our algorithm is based on a two-stage approach in which *primal* and *dual* players interact through *Lagrangian games*. In the first (*play*) phase, the primal player tries to balance out the maximization of their rewards with constraints violation. In the second (*recovery*) phase, the primal player only makes “safe decisions” to avoid violating constraints too much. In the case of stochastic rewards and constraints, the algorithm never enters phase two. This is particularly relevant for budget-pacing mechanisms in repeated auctions, being related to how budget is allocated. Our framework can also perform budget allocation subject to constraints *not* tractable by traditional mechanisms, such as ROI constraints [8, 16].[2]

2 Preliminaries

First, we introduce the set of probability measures on the Borel sets of $\mathcal{X}$. We refer to such a set as the set of *strategy mixtures*, denoted as Ξ. In the following, for the ease of presentation and with

[2] See [14] for an extended version of this work.

a slight abuse of notation, whenever we write a $\boldsymbol{\xi} \in \Xi$ in place of an $\boldsymbol{x} \in \mathcal{X}$, we mean that we are taking the expectation with respect to the probability measure $\boldsymbol{\xi}$. For instance, given $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we have that $f(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}} f(\boldsymbol{x})$ and $g(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\xi}} g(\boldsymbol{x})$.

Then, given two functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we define the following optimization problem, which chooses the strategy mixture $\boldsymbol{\xi} \in \Xi$ that maximizes the expected reward encoded by f, while guaranteeing that the constraints encoded by g are satisfied in expectation.

$$\texttt{OPT}_{f,g} \coloneqq \begin{cases} \sup\limits_{\boldsymbol{\xi} \in \Xi} \quad f(\boldsymbol{\xi}) \qquad \text{s.t.} \\ \qquad\quad g(\boldsymbol{\xi}) \leq 0. \end{cases} \tag{$\texttt{LP}_{f,g}$}$$

We denote by $d_g \in [-1, 1]$ the largest possible value for which there exists a strategy mixture $\boldsymbol{\xi} \in \Xi$ satisfying the constraints $g(\boldsymbol{\xi}) \leq 0$ by a margin of at least d_g. Formally, $d_g \coloneqq \sup_{\boldsymbol{\xi} \in \Xi} \min_{i \in [m]} -g_i(\boldsymbol{\xi})$. In order to ensure that $\texttt{OPT}_{f,g}$ is always well defined, we assume that it is always the case that $d_g \geq 0$. Notice that, if $d_g > 0$, then Problem $\texttt{LP}_{f,g}$ satisfies Slater's condition.

We consider several settings, differing in how functions f_t and g_t are selected, either *stochastically* or *adversarially*. We say that functions f_t (respectively g_t) are selected stochastically, when they are independently drawn according to a given probability measure $\mu_{\mathcal{F}}$ over $\mathcal{F}$ (respectively $\mu_{\mathcal{G}}$ over $\mathcal{G}$). Instead, we say that functions f_t (respectively g_t) are selected adversarially if each f_t (respectively g_t) is chosen by an adversary based on the sequence of prior decisions, namely $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$.

We compare the performance of the decision maker against the baseline $T\,\texttt{OPT}_{\bar{f},\bar{g}}$, where $\bar{f}$ and $\bar{g}$ are suitably-defined functions. When functions f_t, respectively g_t, are selected stochastically, then we define function $\bar{f}$, respectively $\bar{g}$, so that $\bar{f}(\boldsymbol{x}) \coloneqq \mathbb{E}_{f \sim \mu_{\mathcal{F}}}[f(\boldsymbol{x})]$, respectively $\bar{g}(\boldsymbol{x}) \coloneqq \mathbb{E}_{g \sim \mu_{\mathcal{G}}}[g(\boldsymbol{x})]$. When functions f_t, respectively g_t, are selected adversarially, then we define function $\bar{f}$, respectively $\bar{g}$, so that $\bar{f}(\boldsymbol{x}) \coloneqq \frac{1}{T}\sum_{t=1}^{T} f_t(\boldsymbol{x})$, respectively $\bar{g}(\boldsymbol{x}) \coloneqq \frac{1}{T}\sum_{t=1}^{T} g_t(\boldsymbol{x})$.

Our goal is to design online algorithms for the decision maker that output a sequence of decisions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ such that both the *cumulative regret* with respect to the performance of the baseline, defined as $R^T \coloneqq T\,\texttt{OPT}_{\bar{f},\bar{g}} - \sum_{t=1}^{T} f_t(\boldsymbol{x}_t)$, and the *cumulative constraints violation*, defined as $V^T \coloneqq \max_{i \in [m]} \sum_{t=1}^{T} g_{t,i}(\boldsymbol{x}_t)$, grow sublinearly in the number of rounds T.

In conclusion, we introduce a Problem-specific parameter that is strictly related to the feasibility of Problem $\texttt{LP}_{\bar{f},\bar{g}}$. We call it the *feasibility parameter* $\rho \in \mathbb{R}$, which is formally defined as follows. When functions g_t are selected stochastically, $\rho \coloneqq \sup_{\boldsymbol{\xi} \in \Xi} \min_{i \in [m]} -\bar{g}_i(\boldsymbol{\xi})$. When functions g_t are selected adversarially, $\rho \coloneqq \sup_{\boldsymbol{\xi} \in \Xi} \min_{t \in [T]} \min_{i \in [m]} -g_{t,i}(\boldsymbol{\xi})$.

3 A unifying meta-algorithm

In this section, we present our meta-algorithm. Its core idea is to instantiate suitable pairs of RMs, where one is working in the domain $\mathcal{X}$ of primal variables and the other in a suitable subset of the domain $\mathbb{R}_{+}^{m}$ of dual variables. At each round $t \in [T]$, the algorithm makes the RMs "play" against each other in a *Lagrangian game*, where the utility functions observed by them are related to the Lagrangian function $\mathcal{L}_{f_t,g_t}(\boldsymbol{x}, \boldsymbol{\lambda})$ of Problem $\texttt{LP}_{f_t,g_t}$.

Algorithm description. The algorithm works in two phases. In the first one, called *play phase*, the algorithm builds a primal RM, called $\mathcal{R}_{\mathrm{I}}^{\mathrm{P}}$, working in the primal domain $\mathcal{X}$ and a dual RM, called $\mathcal{R}_{\mathrm{I}}^{\mathrm{D}}$, operating on the subset $\mathcal{D}_{\tilde{\rho}}$ of the dual domain $\mathbb{R}_{+}^{m}$, where $\tilde{\rho}$ is set in Line 1. The algorithm makes the two RMs playing against each other (see the call $\textsc{LagrangianGame}(\mathcal{R}_{\mathrm{I}}^{\mathrm{P}}, \mathcal{R}_{\mathrm{I}}^{\mathrm{D}}, 1)$) until either the cumulative violation V^t incurred by the algorithm exceeds a given threshold or round T is reached. Then, in the second phase, called *recovery phase*, the algorithm constructs a new pair of primal, dual RMs, with the latter working on the $(m-1)$-dimensional simplex Δ_m. The recovery phase uses the remaining rounds to make these new RMs play against each other, with the primal RM observing modified utility functions that do *not* account for functions f_t (see the call $\textsc{LagrangianGame}(\mathcal{R}_{\mathrm{II}}^{\mathrm{P}}, \mathcal{R}_{\mathrm{II}}^{\mathrm{D}}, 0)$). The pseudo-code describing one "play" between two RMs, called $\mathcal{R}^{\mathrm{P}}$ and $\mathcal{R}^{\mathrm{D}}$, is defined by the sub-procedure $\textsc{LagrangianGame}(\mathcal{R}^{\mathrm{P}}, \mathcal{R}^{\mathrm{D}}, v)$ in Algorithm 2. The additional parameter $v \in \{0, 1\}$ is used to control the feedback fed into the primal RM $\mathcal{R}^{\mathrm{P}}$; specifically, if $v = 1$, then $\mathcal{R}^{\mathrm{P}}$ observes a utility function that also accounts for f_t (play phase), otherwise, if $v = 0$, the observed utility function only accounts for the term depending on g_t.

Algorithm 1 META-ALGORITHM$(T, \delta, \hat{\rho})$

```
$\tilde{\rho} \leftarrow \max\left\{\hat{\rho}/2, T^{-1/4}\right\}, \eta \leftarrow \delta/3, t \leftarrow 1$
▷ Phase I: Play
$\mathcal{R}^{\mathtt{P}}_{\mathrm{I}} \leftarrow \text{INIT}^{\mathtt{P}}\left(\mathcal{X}, \left[-1/\tilde{\rho}, 1 + 1/\tilde{\rho}\right], \eta\right)$
$\mathcal{R}^{\mathtt{D}}_{\mathrm{I}} \leftarrow \text{INIT}^{\mathtt{D}}\left(\mathcal{D}_{\tilde{\rho}}, \left[-1/\tilde{\rho}, 1/\tilde{\rho}\right], 0\right)$
while $V^t \leq (T - t)\tilde{\rho} + M_{\tilde{\rho}} - 1 \wedge t \leq T$ do
    $\boldsymbol{x}_t \leftarrow \text{LAGRANGIANGAME}(\mathcal{R}^{\mathtt{P}}_{\mathrm{I}}, \mathcal{R}^{\mathtt{D}}_{\mathrm{I}}, 1)$
$T_1 \leftarrow t - 1$
▷ Phase II: Recovery
$\mathcal{R}^{\mathtt{P}}_{\mathrm{II}} \leftarrow \text{INIT}^{\mathtt{P}}\left(\mathcal{X}, [-1, 1], \eta\right)$
$\mathcal{R}^{\mathtt{D}}_{\mathrm{II}} \leftarrow \text{INIT}^{\mathtt{D}}\left(\Delta_m, [-1, 1], 0\right)$
while $t \leq T$ do
    $\boldsymbol{x}_t \leftarrow \text{LAGRANGIANGAME}(\mathcal{R}^{\mathtt{P}}_{\mathrm{II}}, \mathcal{R}^{\mathtt{D}}_{\mathrm{II}}, 0)$
```

Algorithm 2 LAGRANGIANGAME$(\mathcal{R}^{\mathtt{P}}, \mathcal{R}^{\mathtt{D}}, v)$

```
$\boldsymbol{x}_t \leftarrow \mathcal{R}^{\mathtt{P}}.\text{NEXTELEMENT}()$
$\boldsymbol{\lambda}_t \leftarrow \mathcal{R}^{\mathtt{D}}.\text{NEXTELEMENT}()$
Play $\boldsymbol{x}_t$ and get $f_t$ and $g_t$                                      ▷ Full f.
Play $\boldsymbol{x}_t$ and get $f_t(\boldsymbol{x}_t)$ and $g_t(\boldsymbol{x}_t)$                  ▷ Bandit f.
▷ Primal RM update
Let $u^{\mathtt{P}}_t : \boldsymbol{x} \mapsto v f_t(\boldsymbol{x}) - \langle \boldsymbol{\lambda}_t, g_t(\boldsymbol{x}) \rangle$                  ▷ Full f.
$u^{\mathtt{P}}_t(\boldsymbol{x}_t) \leftarrow v f_t(\boldsymbol{x}_t) - \langle \boldsymbol{\lambda}_t, g_t(\boldsymbol{x}_t) \rangle$               ▷ Bandit f.
$\mathcal{R}^{\mathtt{P}}.\text{OBSERVEUTILITY}(u^{\mathtt{P}}_t)$                                 ▷ Full f.
$\mathcal{R}^{\mathtt{P}}.\text{OBSERVEUTILITY}(u^{\mathtt{P}}_t(\boldsymbol{x}_t))$                        ▷ Bandit f.
▷ Dual RM update
Let $u^{\mathtt{D}}_t : \boldsymbol{\lambda} \mapsto -\langle \boldsymbol{\lambda}, g_t(\boldsymbol{x}) \rangle$
$\mathcal{R}^{\mathtt{D}}.\text{OBSERVEUTILITY}(u^{\mathtt{D}}_t)$
```

4 Applications to repeated auctions settings

Internet advertising platforms usually operationalize large auction markets by using *proxy bidders* that place bids in repeated auctions on the advertisers' behalf. A proxy-bidder selects bids according to a *budget-pacing mechanism*, which manages the usage of the advertisers' budget over time [1, 16, 7]. In this section, we discuss the application of our framework to budget-management in auctions, arguing that it can deal with more general constraints on ad slots allocation with respect to what is currently achievable with multiplicative pacing algorithms, which manage only *knapsack constraints*.

We consider the problem faced by a bidder who takes part in a sequence of repeated auctions. We focus on the case of *second-price* and *first-price* auctions, since they are the *de facto* standard in large Internet advertising platforms. At each round $t \in [T]$, the bidder observes their valuation v_t from a finite set of $n_{\mathtt{V}}$ possible valuations $\mathcal{V} \subset [0, 1]$. Such valuation models targeting preferences of the advertiser. Then, the bidder chooses a bid $b_t \in \mathcal{B}$, where $\mathcal{B} \subset [0, 1]$ is a finite set of $n_{\mathtt{B}}$ possible bids such that $0 \in \mathcal{B}$ (*i.e.*, the bidder is allowed to skip items without incurring in any cost). The utility of the bidder depends on the largest among competing bids, denoted by β_t. In particular, the utility is computed as $f_t(b_t) = (v_t - c_t(b_t))\mathbb{1}\{b_t \geq \beta_t\}$, where the cost c_t is such that $c_t(b_t) = \beta \mathbb{1}\{b_t \geq \beta_t\}$ in second-price auctions, and $c_t(b_t) = b_t \mathbb{1}\{b_t \geq \beta_t\}$ for first-price ones. Finally, the bidder has a target *per-round* budget of $\rho > 0$, which yields an overall budget $B \coloneqq \rho T$ that limits the total spending over the T rounds. In the case of budget-constrained bidding, a strictly feasible solution can be easily achieved by always bidding 0. Using the target per-round budget $\rho = B/T$ we can write the budget constraint as $\sum_{t \in [T]} g_t(b_t) \leq 0$, with $g_t(b) = c_t(b) - \rho$ for any $b \in \mathcal{B}$. As a benchmark to evaluate the algorithm, we consider the best feasible static policy $\pi : \mathcal{V} \to \mathcal{B}$. The set of static policies can be represented by $\mathcal{X} \coloneqq \mathcal{B}^{n_{\mathtt{V}}}$, where a vector $\boldsymbol{b} \in \mathcal{B}^{n_{\mathtt{V}}}$ encodes the policy's bids for each possible valuation. To apply our framework to this problem, it is sufficient to design a primal regret minimizer constructor (recall that, in order to design dual RMs, we can employ OMD). This can be implemented by instantiating a regret minimizer EXP3.P [3] for each possible valuation in $\mathcal{V}$. Given a failure probability $\nu \in (0, 1)$, each RM guarantees a regret bound of $\sqrt{T n_{\mathtt{B}} log(n_{\mathtt{B}}/\nu)}$ with probability at least $1 - \nu$. Thus, given a desired failure probability $\eta \in (0, 1)$, by setting $\nu = \eta/n_{\mathtt{V}}$ we get that, with probability at least $1 - \eta$, the bounds of all the RMs hold. Hence, by a union bound, we get that the regret of a primal RM is $\mathcal{E}^{\mathtt{P}}_{T,\eta} = O(n_{\mathtt{V}} \sqrt{T n_{\mathtt{B}} log(n_{\mathtt{B}} n_{\mathtt{V}}/\eta)})$.

Handling ROI constraints. Traditional budget-pacing mechanisms (see, *e.g.*, [8, 6]) are based on primal-dual algorithms that are near optimal in settings with knapsack constraints only, and they cannot be generalized to deal with other types of long-term constraints. However, there are many real-world situations in which guaranteeing other types of constraints is crucial for practical applications (see, *e.g.*, [18, 17]). One example is the case of *return on investment* (ROI) constraints [4, 18, 21].The recent work by Golrezaei et al. [17] presents a threshold-based algorithm for repeated second-price auctions under budget and ROI constraints. Our framework allows advertisers to reach a target ROI while keeping expenses under control also in the setting of repeated first-price auctions. In particular, given a target ROI ω, we define the ROI constraints as $g_t(b_t) = (\omega - v_t/b_t)\mathbb{1}\{b_t \geq \beta_t\} \leq 0$. Then, it is enough to instantiate our framework as described before to immediately get that the cumulative violation of the budget and ROI constraints are upper bounded by $\tilde{O}(T^{1/2})$. This holds both in the fully stochastic and in the fully adversarial setting.

References

[1] Deepak Agarwal, Souvik Ghosh, Kai Wei, and Siyu You. Budget pacing for targeted online advertisements at linkedin. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1613–1619, 2014.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[4] Jason Auerbach, Joel Galenson, and Mukund Sundararajan. An empirical analysis of return on investment maximization in sponsored search auctions. In *Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 1–9, 2008.

[5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[6] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *arXiv preprint arXiv:2011.10124*, 2020.

[7] Santiago Balseiro, Anthony Kim, Mohammad Mahdian, and Vahab Mirrokni. Budget-management strategies in repeated auctions. *Operations Research*, 2021.

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

[9] Martino Bernasconi, Federico Cacciamani, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Safe learning in tree-form sequential decision making: Handling hard and soft constraints. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1854–1873. PMLR, 2022.

[10] Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Sequential information design: Learning to persuade in the dark. *Advances in Neural Information Processing Systems*, 35, 2022.

[11] Andrey Bernstein, Shie Mannor, and Nahum Shimkin. Online classification with specificity constraints. *Advances in Neural Information Processing Systems*, 23, 2010.

[12] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

[13] Xuanyu Cao and KJ Ray Liu. Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on automatic control*, 64(7):2665–2680, 2018.

[14] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35, 2022.

[15] Tianyi Chen, Qing Ling, and Georgios B Giannakis. An online convex optimization approach to proactive network resource allocation. *IEEE Transactions on Signal Processing*, 65(24): 6350–6364, 2017.

[16] Vincent Conitzer, Christian Kroer, Eric Sodomka, and Nicolas E Stier-Moses. Multiplicative pacing equilibria in auction markets. *Operations Research*, 2021.

[17] Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. Bidding and pricing in budget and roi constrained markets. *arXiv preprint arXiv:2107.07725*, 2021.

[18] Negin Golrezaei, Ilan Lobel, and Renato Paes Leme. Auction design for roi-constrained buyers. In *Proceedings of the Web Conference 2021*, pages 3941–3952, 2021.

[19] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019*, pages 202–219. IEEE Computer Society, 2019.

[20] Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pages 402–411. PMLR, 2016.

[21] Bin Li, Xiao Yang, Daren Sun, Zhi Ji, Zhen Jiang, Cong Han, and Dong Hao. Incentive mechanism design for roi-constrained auto-bidding. *arXiv preprint arXiv:2012.02652*, 2020.

[22] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1): 2503–2528, 2012.

[23] Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.

[24] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 845–861. PMLR, 08 Feb–11 Feb 2020.

[25] Wen Sun, Debadeepta Dey, and Ashish Kapoor. Safety-aware algorithms for adversarial contextual bandit. In *International Conference on Machine Learning*, pages 3280–3288. PMLR, 2017.

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

[27] Xinlei Yi, Xiuxian Li, Lihua Xie, and Karl H Johansson. Distributed online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Signal Processing*, 68: 731–746, 2020.

[28] Hao Yu and Michael J. Neely. A low complexity algorithm with o($\sqrt{T}$) regret and o(1) constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*, 21(1):1–24, 2020.

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

[30] Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

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