

# Online Ad Procurement in Non-stationary Autobidding Worlds

Jason Cheuk Nam Liang

Operations Research Center, Massachusetts Institute of Technology, jcnliang@mit.edu

Haihao Lu

The University of Chicago, Booth School of Business, haihao.lu@chicagobooth.edu

Baoyu Zhou

The University of Chicago, Booth School of Business, baoyu.zhou@chicagobooth.edu

Today’s online advertisers procure digital ad impressions through interacting with autobidding platforms: advertisers convey high level procurement goals via setting levers such as budget, target return-on-investment, max cost per click, etc.. Then ads platforms subsequently procure impressions on advertisers’ behalf, and report final procurement conversions (e.g. click) to advertisers. In practice, advertisers may receive minimal information on platforms’ procurement details, and procurement outcomes are subject to non-stationary factors like seasonal patterns, occasional system corruptions, and market trends which make it difficult for advertisers to optimize lever decisions effectively. Motivated by this, we present an online learning framework that helps advertisers dynamically optimize ad platform lever decisions while subject to general long-term constraints in a realistic bandit feedback environment with non-stationary procurement outcomes. In particular, we introduce a primal-dual algorithm for online decision making with multi-dimension decision variables, bandit feedback and long-term uncertain constraints. We show that our algorithm achieves low regret in many worlds when procurement outcomes are generated through procedures that are stochastic, adversarial, adversarially corrupted, periodic, and ergodic, respectively, without having to know which procedure is the ground truth. Finally, we emphasize that our proposed algorithm and theoretical results extend beyond the applications of online advertising.

*Key words:* Online advertising, autobidding, return-on-investment, budget management, ad campaign management, bandit learning, online optimization

---

## 1. Introduction

Automated bidding (or autobidding for short) has become the dominant mode for advertisers to procure digital ad inventories and impressions, contributing to more than \$120 billion dollar ad spend in 2022 and more than 90% of total online ad transaction volumes [1, 2]. In an autobidding platform, advertisers only need to convey their high-level procurement goals for an ad campaign to the platform, which then takes charge of procuring ads on advertisers’ behalf. Such procurement goals are communicated to a platform through platform *levers*, which are adjustable parameters that advertisers can control to influence the bidding process and campaign performance. To exemplify, Figure 1 displays certain several levers presented on the Google Ads interface, where an advertiser can set per-campaign budgets, target cost-per-actions, campaign duration, campaign schedules, targeting, etc; similar examples are also shown in related literature [22].

The screenshot displays the Google Ads campaign setup interface with the following sections:

- Budget:** Set your average daily budget for this campaign. Currency: US Dollar (USD \$). Amount: \$100.00.
- Bidding:** Select your bid strategy: Target CPA. Pay for: Interactions. Target CPA: \$1.00.
- Start and end dates:** Start date: Jan 23, 2023. End date: Jan 25, 2023.
- Ad schedule:** Saturdays - Sundays, 12:00 to 19:00.
- Languages:** English, Chinese (simplified), French, German.
- Demographics:**

Gender	Age	Household income
<input checked="" type="checkbox"/> Female	<input checked="" type="checkbox"/> 18 - 24	<input checked="" type="checkbox"/> Top 10%
<input checked="" type="checkbox"/> Male	<input checked="" type="checkbox"/> 25 - 34	<input checked="" type="checkbox"/> 11 - 20%
<input type="checkbox"/> Unknown	<input checked="" type="checkbox"/> 35 - 44	<input checked="" type="checkbox"/> 21 - 30%
	<input checked="" type="checkbox"/> 45 - 54	<input checked="" type="checkbox"/> 31 - 40%
	<input checked="" type="checkbox"/> 55 - 64	<input checked="" type="checkbox"/> 41 - 50%
	<input checked="" type="checkbox"/> 65+	<input checked="" type="checkbox"/> Lower 50%

**Figure 1** Example levers on the Google Ads interface to create digital advertising campaigns.

As the primary avenue for advertisers to run ad campaigns on autobidding platforms and influence ad conversion outcomes (e.g. clicks), making efficient lever decisions is vital to advertisers to achieve their procurement objectives. However, advertisers face many challenges in practice when optimizing for lever decisions, namely high-dimensional decision making under long-term constraints, non-stationary autobidding environments, and limited procurement feedback.

**High-dimensional decision making and long-term constraint satisfaction.** Making multiple lever decisions involves evaluating numerous possible combinations of lever configurations, which is computationally intensive and time-consuming in real-time decision making setups. Also, advertisers need to understand the potential interactions and dependencies between levers, as adjusting one lever may have unintended consequences or interactions with other levers, making it challenging to predict the overall impact of adjustments accurately. Further, advertisers may need to satisfy certain long-term constraints over time, e.g. limiting total spend under a budget, or achieving certain return-on-investment targets. Hence in addition to analyzing the interactions between different levers, advertisers also need to concern long-term consequences of making certain lever decisions.

**Limited feedback on procurement outcome and constraints.** Despite the fact that autobidding via lever decisions greatly simplifies advertisers' ad procurement process as they no longer have to handle the intricacies of procuring individual ad impressions, the procurement procedure becomes a black box for advertisers, as advertisers only have limited visibility into the specific details of how the platform executes ad placement processes. This lack of transparency can make it challenging for advertisers to control the nuances of their campaign execution through lever decisions, and amplifies the complexity to conduct counterfactual analyses on outcomes for past lever decisions.

**Many non-stationary autobidding worlds.** Autobidding procurement environments are highly non-stationary, as a wide spectrum of latent factors in online ad marketplaces may greatly

vary procurement outcomes of the same lever decisions in different time periods. These latent factors include, but are not limited to changing user preferences, seasonality effects, shifts in market trends, occasional malfunctions in autobidding platforms, etc. These dynamics may influence how users interact with different types of ads, which necessitates continuous adjustment of lever decision strategies to adapt to current and future market conditions.

To address these challenges, in this work we answer the following questions: *How should an advertiser dynamically set multiple levers to optimize conversion outcomes subject to long-term constraints under limited information? And can she run a unified algorithm that can perform well while being agnostic to many types of non-stationary autobidding procurement environments?*

Motivated by these questions, in this work we study an advertiser’s online high-dimensional lever decision problem with long-term constraints under limited bandit feedback in many non-stationary worlds. We summarize our main contributions as followed:

**1. Modelling online lever decisions in many worlds using online constrained optimization with bandit function-valued feedback and uncertain constraints (Section 2).**

We cast the advertiser lever decision problem as an online constrained optimization with bandit function-valued feedback and uncertain constraints, where functions at which (lever) decisions are evaluated correspond to the conversion and constraint functions in our autobidding setup. Further, we model real-world non-stationarity in autobidding environments as possibly time-varying distributions from which conversion and constraint functions are sampled, and then further applied to lever decisions. Under this model, we discuss five different input procedures from which the sequence of functional distributions are sampled to model stochastic, adversarial, corrupted, periodic, and ergodic environments; see Section 2.2 for more details. To the best of our knowledge, from a modelling perspective this is the first work to model high-dimensional lever decision problem in practical non-stationary autobidding worlds.

**2. Proposing a constrained bandit optimization algorithm universally applicable to many worlds (Section 3).** We develop a unified bandit optimization algorithm with robust performance guarantees across different worlds. Our algorithm incorporates four key designs to handle bandit function-valued feedback and unknown non-stationary environments. 1. we utilize dual descent to update dual variables associated with long-term constraints and decouple decisions over time; 2. we employ a primal-ascent-based bandit online convex optimization (BOCO) approach to make (primal) lever decisions to cope with bandit function valued feedback; 3. we implement an exponential weights expert algorithm on top of primal-ascent BOCO, where each expert corresponds to a different primal ascent step size. This enables our algorithm to adapt to the optimal primal ascent step size for each world without prior knowledge on which world we are in; 4. our algorithm

dynamically checks for realized constraint violations and applies "safe lever decisions" to ensure long-term constraint satisfaction. For further details, refer to Algorithm 1 in Section 3.

**3. Analyzing the performance of proposed algorithms in many worlds (Section 4).** We present theoretical analysis (Theorem 4.2) on the regret bound of the proposed algorithm, and we show that our unified algorithm can achieve reasonable regret bounds with all five input procedures. These regret bounds are summarized in the following Table 1.

Stochastic	$\delta$ -corrupted	Adversarial	Periodic	Ergodic
$\mathcal{O}(T^{\frac{3}{4}})$	$\mathcal{O}(T^{\frac{3}{4}} + \delta)$	$(1 - \frac{1}{\xi})\text{OPT}(\mathcal{P}_{1:T}) + o(T)$	$\mathcal{O}(T^{\frac{3}{4}} + q\sqrt{T})$	$\mathcal{O}(T^{\frac{3}{4}} + \kappa\sqrt{T})$

**Table 1**  $T$ -period regret bounds for Algorithm 1 under different input models.

The parameters  $(\delta, \xi, \text{OPT}(\mathcal{P}_{1:T}), q, \kappa)$  are formally defined in Section 2.2 and Theorem 4.2. Finally, in Section 5 we remark all results are applicable to other problems.

### 1.1. Related works.

**Autobidding.** In an autobidding framework, there is a considerable body of works that study the price of anarchy, which aims to improve worst case individual or total advertiser welfare guarantees w.r.t. the optimal welfare via mechanism design frameworks; see e.g. [7, 21, 23, 24, 38]. We remark that this line of work does not concern developing online learning algorithms. On the other hand, numerous works have concentrated on developing online bidding algorithms for repeated ad auctions [9, 31, 32, 41], as well as designing repeated selling mechanisms for ad impression allocation [12, 28] and references therein. However, as discussed in the introduction, autobidding platforms conduct bidding on behalf of advertisers while keeping bidding and selling mechanism details undisclosed. In this study, we treat the bidding procedure and selling mechanism as a black box and directly model bidding and auction outcomes using conversion functions; see Section 2. To the best of our knowledge, the most pertinent work to this paper is [22], which explores a similar ad procurement problem by optimizing levers within a bandit environment. However, this work solely focuses on the stochastic world and optimizing a single lever (i.e., a 1-dimensional decision space). In contrast, our paper develops a unified algorithm capable of making high-dimensional lever decisions in many worlds.

**Online convex optimization.** In Section 2, we cast the advertiser's online learning problem of interest to a high-dimensional bandit online convex optimization problem with uncertain long-term constraints, and develop an algorithm that yields good performance guarantees under different procedures from which the objective and constraint functions are generated. There has been a rich line of works that study bandit convex optimization with no long-term constraints in stochastic and adversarial worlds [17, 26, 33, 42, 44], as well as works that study (full-information feedback) online convex optimization with long-term constraints [34, 36, 37, 43]. Further, works that study both bandit

feedback and long-term constraints either consider single-dimensional (such as multi-arm bandits with constraints) [4, 6, 25, 40], or consider static regret (i.e. benchmarking performance to that of a single optimal action) [13, 14, 16, 19]. This paper distinguishes itself from these two streams of works by considering high-dimensional decisions as well as dynamic regret, i.e. comparing to the best sequence of actions instead of a single action; see Eq.(1). Finally, all aforementioned works only study algorithms in stochastic or adversarial setups, whereas in this work we go beyond these two worlds and address more complex learning environment such as periodic, corrupted, ergodic, and finite switching. To the best of our knowledge, the only related work that develops a universal algorithm under “many world” setups is [9]. However, this work considers a full-information scenario where the online decisions in period can be made after observing realized objective and constraint functions during that period. In this work, we present an efficient algorithm to handle bandit feedback.

**Notation.** For any vector  $\mathbf{z} \in \mathbb{R}^d$ , let  $\|\mathbf{z}\|$  be its Euclidean norm. Denote  $\mathbb{B} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| \leq 1\}$  as the  $d$ -dimensional unit ball centered at  $\mathbf{0}$ , and let  $\mathbb{S} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| = 1\}$  be the unit sphere. For any set  $\mathcal{S}$ , let  $U(\mathcal{S})$  be the uniform distribution over  $\mathcal{S}$ . Let  $\mathbf{e}$  denote the vector whose components are all 1’s, and  $\mathbf{e}_j$  be the unit vector whose  $j$ ’th position is 1. We use  $\mathcal{O}$  notation to represent the asymptotic order of a term when the period  $T \rightarrow \infty$  and ignore the  $\log T$  terms.

## 2. Autobidding with bandit feedback in many worlds (input models)

### 2.1. Autobidding as a bandit online optimization with long-term uncertain constraints

Consider an advertiser repeatedly interacting with an ad platform over  $T \in \mathbb{N}$  periods, where each period can be interpreted as a single ad campaign that is run on the platform. During each period  $t \in [T]$ , the advertiser makes  $d \geq 1$  lever decisions denoted as  $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ , e.g. setting the per-campaign budget, campaign duration, per-campaign target return-on-investment, max spend per conversion, etc; see Figure 1 for example lever decisions in practice. Here,  $\mathcal{X}$  is some compact and convex decision set whose diameter is  $D = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$ . For simplicity, assume  $\mathbf{0} \in \mathcal{X}$  so  $\|\mathbf{x}\| \leq D$  for any  $\mathbf{x} \in \mathcal{X}$ . After the lever decisions  $\mathbf{x}_t$  are made the campaign is fully executed via autobidding, the advertiser observes bandit feedback for her campaign outcomes: she only observes her realized conversion  $f_t(\mathbf{x}_t) \in \mathbb{R}_{\geq 0}$  (e.g. number of clicks on her ads), as well as some constraint balance  $\mathbf{g}_t(\mathbf{x}_t) \in \mathbb{R}^K$ . The conversion and constraint functions  $(f_t, \mathbf{g}_t)$  are sampled from some (possibly infinite) support  $\mathcal{S}$  according to distribution  $\mathcal{P}_t \in \Delta(\mathcal{S})$  (we will discuss how  $\mathcal{P}_t$ ’s are generated by nature in Section 2.2). Using the notation  $\mathcal{P}_{1:T} = (\mathcal{P}_t)_{t \in [T]}$  the advertiser’s hindsight optimization problem is

$$\text{OPT}(\mathcal{P}_{1:T}) = \max_{\mathbf{x} \in \mathcal{X}^T} \sum_{t \in [T]} \mathbb{E}_{\mathcal{P}_t}[f_t(\mathbf{x}_t)] \quad \text{s.t.} \quad \sum_{t \in [T]} \mathbb{E}_{\mathcal{P}_t}[\mathbf{g}_{k,t}(\mathbf{x}_t)] \geq 0 \quad k = 1 \dots K. \quad (1)$$

Here, we use a constraint function  $\mathbf{g}_{k,t} : \mathcal{X} \rightarrow \mathbb{R}$  to characterize general performance metrics of the advertiser for her ad campaigns. In the following, we present several examples for constraint

functions that are widely used in practice or studied in literature, and for illustrative purposes assume  $x_{t,1} = \mathbf{e}_1^\top \mathbf{x}_t$  (i.e. the first lever) represents the per-campaign budget for the  $t$ 'th campaign.

- *Long-term budget constraint.* The advertiser has a total budget  $BT > 0$ . Then by letting constraint function  $g_{k,t}(\mathbf{x}) = B - x_{t,1}$ , we have  $\sum_{t \in [T]} g_{k,t}(\mathbf{x}_t) \geq 0 \implies \sum_{t \in [T]} x_{t,1} \leq BT$ , which means total campaign spend over  $T$  periods (assuming each campaign fully depletes the campaign budget) must be less than the advertiser's total budget  $B$ .
- *Long-term return-on-investment constraint.* The advertiser intends to safeguard a long-term return-on-investment  $\gamma > 0$ , i.e. she attains a long-term average of at least  $\gamma$  conversions per dollar spent. Then by considering  $g_{k,t}(\mathbf{x}) = f_t(\mathbf{x}) - \gamma x_{t,1}$ , we have  $\sum_{t \in [T]} g_{k,t}(\mathbf{x}_t) \geq 0 \implies \sum_{t \in [T]} f_t(\mathbf{x}_t) \geq \gamma \sum_{t \in [T]} x_{t,1}$ , which means total conversion over  $T$  periods is at least  $\gamma$  times total spend. Finally, we make the following mild assumptions on conversion and constraint functions.<sup>1</sup>

**Assumption 2.1 (Mild assumptions on conversion and constraint functions)** For any  $(f, \mathbf{g}) \in \mathcal{S}$ ,  $f, g_1 \dots g_K$  are all bounded concave functions, i.e. we assume  $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{g}(\mathbf{x})\|_\infty \leq \bar{G}$  and  $\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| \leq \bar{F}$  for some  $\bar{G}, \bar{F} < \infty$ . Further,  $f$  and  $\mathbf{g}$  are  $L$ -Lipschitz, i.e. for  $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$ , we have  $|f(\mathbf{x}) - f(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|$  and  $\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|$ . Further, there exists  $(f, \mathbf{g}) \in \mathcal{S}$  such that  $\min_{\mathbf{x} \in \mathbf{X}} \min_{k \in [K]} g_k(\mathbf{x}) < 0$ .

## 2.2. Five input models characterizing many autobidding worlds

In this subsection, we describe structural properties of the of input distribution sequence  $\mathcal{P}_{1:T}$ , and shed light on how we utilize various properties to model a wide spectrum of autobidding environments (called worlds) such as time-varying user preferences, seasonality, shifts in market trends, etc., that may potentially lead to different procurement outcomes for the same ad campaign lever decisions.

**Stochastic:** There exists some probability distribution  $\mathcal{P} \in \Delta(\mathcal{S})$  such that  $\mathcal{P}_1 = \dots = \mathcal{P}_T = \mathcal{P}$ . This stochastic world represents a stationary autobidding environment where the underlying latent factors influencing user behavior (and correspondingly conversion results) remain constant over time; see e.g. [8, 22, 32].

**$\delta$ -corrupted.** There exists  $\mathcal{P} \in \Delta(\mathcal{S})$  as well as  $\delta \in \mathbb{N}$  periods  $\mathcal{T} = \{\tau_1 \dots \tau_\delta\} \subset [T]$  such that  $\mathcal{P}_t = \mathcal{P}$  for all  $t \notin \mathcal{T}$ . This  $\delta$ -corrupted input sequence represents occasional anomalies in the autobidding environment that may be caused by systematic malfunctions in the autobidding platform, or deliberate attempts by malicious competitors to exploit the system for their own benefit; e.g. some competitors may engage in click fraud to inflate the number of clicks on our ads to exhaust our budget or generate false data to manipulate autobidding algorithms (see e.g. [30]).

<sup>1</sup> These assumptions are justified in many related works in autobidding; see e.g. [14, 22] and references therein. For example, [22] shows that the conversion functions is concave and piecewise-linear when autobidding platforms procure ad impressions on behalf of advertisers in standard second price or VCG auctions.

**Adversarial.**  $\mathcal{P}_{1:T}$  is adversarially chosen by nature before the process starts, and the distributions over time can possibly be non-identical and/or dependent. This adversarial world can be viewed a hypothetical extreme case for the  $\delta$ -corrupted world where each period the procurement outcomes can potentially be corrupted by an adversary. Adversarial input sequences have been widely studied in the literature to assess algorithmic performances in worst-case scenarios; see e.g. [3, 31]

**Periodic.** There exists period length  $q \in \mathbb{N}$  such that  $T = cq$  for some integer  $c \geq 2$  with  $\mathcal{P}_{1:T}$  satisfying  $\mathcal{P}_{1:q} = \mathcal{P}_{q+1:2q} = \dots = \mathcal{P}_{(c-1)q+1:T}$ . This periodic world captures regular cyclic patterns or fluctuations in user behavior over specific time intervals; e.g. seasonality, day-of-week patterns, time of day, etc.

**Ergodic.**  $\mathcal{P}_{1:T}$  is an ergodic process (e.g. an irreducible and aperiodic Markov chain or stationary autoregressive processes), where there exists some  $\kappa \in \mathbb{N}$  and a stationary distribution such that the distance between  $\kappa$ -step transition probabilities and this stationary distribution decreases exponentially fast in  $\kappa$ . An ergodic input sequence signifies that the procurement outcomes in close time proximity are correlated, which is commonly observed in real-world autobidding systems, as they often involve iterative processes that enable procurement algorithms (operated on behalf of advertisers) to converge to a stable state; see details in e.g. [8, 29].

### 2.3. Minimizing regret subject to long-term constraints

In this work, we take the perspective of an advertiser making repeated lever decisions as described in Section 2.1. We focus on designing an online algorithm that determines a lever decision  $\mathbf{x}_t \in [T]$  in each period  $t$  based solely on historical available information with the goal to minimize regret  $\mathcal{R}_T$  (defined as followed) under any input sequence  $\mathcal{P}_{1:T}$  while satisfying long-term constraints

$$\mathcal{R}_T = \text{OPT}(\mathcal{P}_{1:T}) - \sum_{t \in [T]} \mathbb{E}[f_t(\mathbf{x}_t)] \quad \text{and} \quad \sum_{t \in [T]} \mathbb{E}[\mathbf{g}_t(\mathbf{x}_t)] \geq \mathbf{0} \quad (2)$$

Here, the expectation is taken w.r.t. randomness from the input sequence as well as any randomness in our algorithm. We remark our desired policy should be agnostic to the input sequence  $\mathcal{P}_{1:T}$ , and can universally perform well in any autobidding world described in Section (2.2).

## 3. A universal constrained BOCO framework for many worlds

In this section, we focus on designing a algorithm that achieves low regret (per Eq. 2) while maintaining long-term constraint satisfaction in any world described Section 2.2, without having to know which world we are in. We first highlight three main challenges for our problem of interest.

**Dynamic benchmark lever sequence and unknown input sequence  $\mathcal{P}_{1:T}$ .** Recall the hindsight optimal problem in Eq.(1), with which we are comparing our algorithm's performance to the optimal sequence of lever decisions over time given any input sequence  $\mathcal{P}_{1:T}$ , instead of comparing to a single optimal lever decision as in many related works; see e.g. [15] and references therein. This

dynamic optimal sequence presents a very strong benchmark and makes learning very difficult as we (ideally) need to account for variations in the underlying non-stationary ground truth input sequence  $\mathcal{P}_{1:T}$ . Nevertheless, we do not know the input sequence  $\mathcal{P}_{1:T}$  in advance, nor do we know in which world the sequence belongs (Section 2.2).

**Bandit function-value (zero order) feedback.** Our setup concerns a bandit function-value feedback model, where the realized conversion and constraint functions in each period, namely  $f_t : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$  and  $\mathbf{g}_t : \mathcal{X} \rightarrow \mathbb{R}^K$  are never revealed to us, and we can only access their function values at our single lever decisions during the period. We also contrast our setup with a “two-point feedback” setup in online convex optimization literature, where in each period one can access function values two times (see e.g. [44] and reference therein). In the context of our setup, after nature samples  $f_t : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$  and  $\mathbf{g}_t : \mathcal{X} \rightarrow \mathbb{R}^K$  from  $\mathcal{P}_t$ , we can observe the values for two lever decisions  $\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)} \in \mathcal{X}$ , namely  $(f_t(\mathbf{x}_t^{(1)}), \mathbf{g}_t(\mathbf{x}_t^{(1)}))$  and  $(f_t(\mathbf{x}_t^{(2)}), \mathbf{g}_t(\mathbf{x}_t^{(2)}))$ .

**Universal algorithm for many worlds.** Instead of developing separate algorithms that cope with each individual world described in Section 2.1, we aim to develop a single universal algorithm that performs well in every world without exploiting any knowledge on the realized input sequence  $\mathcal{P}_{1:T}$  or its structural properties.

The key component of our algorithm to handle the aforementioned challenges is the following Lagrangian function: for any  $f_t : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$  and  $\mathbf{g}_t : \mathcal{X} \rightarrow \mathbb{R}^K$  define  $\mathcal{L}_t : \mathcal{X} \times \mathbb{R}_{\geq 0}^K \rightarrow \mathbb{R}$  such that

$$\mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda}) = f_t(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{g}_t(\mathbf{x}) . \quad (3)$$

Surrounding this Lagrangian function, our proposed algorithm is formally stated in Algorithm 1 with four key components described as follows.

**1. Dual descent to decouple decisions across time.** As our lever decisions over time should be somewhat intertwined with one another due to the presence of long term constraints, we decouple this cross-period dependency by lagrangifying the hindsight problem in Eq. (1) w.r.t. some benchmark dual variables. In each period  $t$ , we maintain an estimate of dual variables  $\boldsymbol{\lambda}_t \in \mathbb{R}_{\geq 0}^K$  corresponding to each of our  $K$  constraints, and adopt a standard dual descent approach to dynamically adjust these estimates; see Eq. (8). In particular, after observing  $(f_t(\mathbf{x}_t), \mathbf{g}_t(\mathbf{x}_t))$ , we update our dual variable by projecting each coordinate of  $\boldsymbol{\lambda}_t - \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t) = \boldsymbol{\lambda}_t - \eta \mathbf{g}_t(\mathbf{x}_t)$  onto some interval  $[0, C]$  where  $\eta > 0$  is the dual descent step size,  $\bar{F}$  is the maximum achievable conversion, and  $\beta > 0$  is some “safety buffer” to be defined later.

**2. Primal ascent to handle bandit function valued feedback in many worlds.** Given a dual variable  $\boldsymbol{\lambda}_t$  during period  $t$ , standard optimization frameworks suggest optimizing primal (i.e. lever) decisions by setting  $\mathbf{x}_t = \arg \max_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathcal{P}_t}[\mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda}_t)]$ . However, in our setting this is not possible due to the bandit function-valued feedback structure: we do not know  $f_t(\cdot)$ ,  $\mathbf{g}_t(\cdot)$  nor  $\mathcal{P}_t$ ,



and hence cannot optimize for  $\mathbb{E}[\mathcal{L}_t(\cdot, \boldsymbol{\lambda}_t)]$ . To handle this, we view the primal decision problem as a *bandit online convex optimization problem (BOCO)* where the adversarial objective functions are  $\mathcal{L}_t(\cdot, \boldsymbol{\lambda}_t)$ , and run a BOCO algorithm to make primal lever decisions in each period.

Mathematically, define the BOCO objective function  $h_t : \mathcal{X} \rightarrow \mathbb{R}$  as followed

$$h_t(\mathbf{x}) = \mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda}_t) = \mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda}) = f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}) \quad (4)$$

Then, the BOCO algorithm constructs an estimate  $\nabla_t$  of  $\nabla h_t(\mathbf{x})$  based on a random perturbation approach with perturbation parameter  $\rho > 0$  (see Eq.(6) and [26] for more details in random perturbations methods to estimate gradients). Next, BOCO algorithm makes primal a decision  $\mathbf{x}_t$  by an ascent type update, i.e. projecting  $\mathbf{x}_{t-1} + \gamma \nabla_t$  back onto  $\mathcal{X}$ , where  $\gamma > 0$  is the primal ascent step size; see Eq. (8). The loss we incur by running BOCO primal update instead of setting  $\arg \max_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathcal{P}_t}[h_t(\mathbf{x}_t)]$  in the full information setting depends on the specific type of world we are in, i.e. depends on the input sequence  $\mathcal{P}_{1:T}$ . We later show in Lemma 4.3 that input sequences in each world induces some comparator sequence  $\mathbf{y}_1 \dots \mathbf{y}_T \in \mathcal{X}$ , so the conversion loss due to BOCO primal ascent updates can be characterized as some BOCO “dynamic” regret  $\sum_{t \in [T]} h_t(\mathbf{y}_t) - h_t(\mathbf{x}_t^{BOCO})$ . Later in Lemma 4.6, we show that our BOCO algorithm achieves low dynamic regret against any comparator sequence  $\mathbf{y}_1 \dots \mathbf{y}_T \in \mathcal{X}$ .

**3. Choosing primal ascent step sizes from expert advice for different worlds.** Different worlds (and correspondingly different BOCO comparator sequences) may require different “optimal” BOCO primal ascent step sizes in order to achieve optimal regret guarantees. Since we are unaware of the world from which the input sequence  $\mathcal{P}_{1:T}$  are generated, on top of our primal ascent procedure we run a meta “expert” algorithm that allows us to adaptively approximate optimal step sizes in each world. In particular, we consider  $N$  experts, each corresponding to a primal ascent step size  $\gamma_i > 0$ , and produce independent lever decision based on primal ascent with her own step size; see Eq.(8). Then, we dynamically maintain exponential weights  $\mathbf{w}_t \in \Delta_N$  in the  $N$ -dimensional simplex over  $N$  experts, based on their past performance measured by some surrogate loss function defined in Eq. (7). Finally we set the weighted average primal decision over all experts as our ultimate primal decision; see Eq. (5). We later show in Lemma A.3 that such exponentially weighted average BOCO primal ascent decisions achieves low conversion loss compared to the “optimal” BOCO primal ascent expert (i.e. step size) in every world, without having to know which world we are in.

**4. Constraint violation check to ensure constraint satisfaction.** To ensure long-term constraint satisfaction, we maintain a constraint balance  $B_{k,t} = \sum_{\tau \in [t-1]} g_{k,t}(\mathbf{x}_\tau)$  for each constraint  $k \in [K]$  that keeps track of our deviation from satisfying the constraint. To make our problem

tractable, we make the assumption that there is always a “safe action” with which we are always guaranteed to attain some small positive constraint balance.<sup>2</sup>

**Assumption 3.1 (Safe action)** *Assume there exists some  $\bar{\beta} > 0$  and an action  $\tilde{\mathbf{x}}_{\beta} \in \mathcal{X}$  such that for any  $(f, \mathbf{g}) \in \mathcal{S}$ , we have  $g_k(\tilde{\mathbf{x}}_{\beta}) > \bar{\beta}$  for all  $k \in [K]$ .*

With the constraint balance, in each period before making a lever decision, we check the following: if by playing the safe action in all future rounds almost violates constraints violation (step 2 in Algorithm 1), then we hard stop our algorithm and play the safe action in all subsequent rounds. We remark that we do not necessarily know the constraint function value lower bound  $\bar{\beta} > 0$  in Assumption 3.1. However, we show later in Theorem 4.1 that by considering any safety buffer  $\beta < \bar{\beta}$  (e.g. taking  $\beta = 1/\log(T)$  for large enough  $T$ ), we are hard stopping more conservatively, and thus maintain constraint satisfaction. Note that our final regret scales with  $\frac{1}{\beta^2}$ ; see Lemma 4.6.

#### 4. Performance analysis of our constrained BOCO algorithm

In this section, we analyze the performance of Algorithm 1 under any input sequence  $\mathcal{P}_{1:T}$  which may be generated from any world described in Section 2.2. Our first result Lemma 4.1 shows that our algorithm maintains long-term constraint satisfaction almost surely for any input sequence.

**Lemma 4.1 (Strict constraint satisfaction)** *Suppose Assumption 3.1 hold, and  $T$  is large enough so that the safety buffer  $\beta = \frac{1}{\log(T)} < \bar{\beta}$ , where  $\bar{\beta}$  is defined in Assumption 3.1. Then, for any constraint  $k \in [K]$ , we have  $\sum_{t \in [T]} g_{k,t}(\mathbf{x}_t) > 0$ .*

The main result of this paper is the following Theorem 4.2, where we bound the regret of our proposed Algorithm 1 in all worlds specified in Section 2.2.

**Theorem 4.2 (Bounding regret in many worlds.)** *Let the safety buffer  $\beta = \frac{1}{\log(T)}$ , and the number of experts  $N = \left\lceil \log_2 \left( K^{-\frac{1}{6}} (1 + DT)^{\frac{1}{2}} T^{-\frac{3}{4}} \right) \right\rceil$ , where  $K$  is the total number of constraints and  $D$  is diameter of the decision set  $\mathcal{X}$ . Choose the corresponding primal ascent step sizes for the  $N$  experts as  $\{\gamma_1 \dots \gamma_N\} = \{2^{-i} K^{-\frac{1}{6}} (1 + DT)^{\frac{1}{2}} T^{-\frac{3}{4}} : i = 0 \dots N\}$ . Then by taking dual descent step size  $\eta = \frac{1}{\sqrt{KT}}$ , random perturbation parameter  $\rho = K^{\frac{1}{3}} T^{-\frac{1}{4}}$ , and exponential weighted expert learning rate  $\epsilon = T^{-\frac{1}{2}}$ , we obtain the following bounds on  $\mathcal{R}_T$  in each world:*

- **Stochastic:**  $\mathcal{R}_T \leq \mathcal{O}(T^{3/4})$
- **Adversarial:**  $\mathcal{R}_T \leq \left(1 - \frac{1}{\xi}\right) \text{OPT}(\mathcal{P}_{1:T}) + o(T)$ , where  $\xi = 1 - \frac{1}{\beta} \min_{(f, \mathbf{g}) \in \mathcal{S}} \min_{k \in [K], \mathbf{x} \in \mathcal{X}} g_k(\mathbf{x}) \in (0, 1)$  under Assumption 2.1, and  $\bar{\beta} > 0$  is defined in Assumption 3.1

<sup>2</sup> Existence of a safe action is not unnatural, and is studied in many literature; see [14, 22, 25]. For instance, recall our examples for constraint functions in Section 2.1 and assume we only have a long-term budget constraint. Then the corresponding safe action for the constraint function  $g_{k,t}(\mathbf{x}_t) = B - x_{t,1}$  would be any  $\mathbf{x}_t$  whose first entry  $x_{t,1} = 0$  so that we acquire positive constraint value  $B > 0$ .

---

**Algorithm 1**


---

**Input:** Initial dual variable  $\lambda_1 = \mathbf{0}$ , primal expert solutions  $\tilde{\mathbf{x}}_1^i = \mathbf{0}$  for any  $i \in [N]$ ; perturbation parameter  $\rho > 0$  and  $\alpha \in (0, 1)$ ; primal ascent step sizes for experts  $\{\gamma_1, \dots, \gamma_N\} > 0$ ; dual descent step size  $\eta$ ; learning rate of the meta-algorithm  $\epsilon$ ; Initial expert weights  $w_{i,1} = 1$  for all  $i \in [N]$ ; safety buffer  $\beta > 0$ .

- 1: Initialize constraint balance  $B_{k,1} = 0$  for all  $k \in [K]$
- 2: **while** {for all  $k \in [K]$ ,  $B_{k,t} - \bar{G} + \beta(T - t - 1) \geq 0$ } **do**
- 3:   **Compute exponentially weighted average forecaster:**

$$\tilde{\mathbf{x}}_t = \frac{1}{\sum_{i \in [N]} w_t^i} \sum_{i \in [N]} w_t^i \tilde{\mathbf{x}}_t^i \quad (5)$$

- 4:   Sample  $\mathbf{u}_t \sim U(\mathbb{S})$  uniformly at random from the unit sphere.
- 5:   Set  $\mathbf{x}_t = \tilde{\mathbf{x}}_t + \rho \mathbf{u}_t$  and observe  $f_t(\mathbf{x}_t)$  and  $\mathbf{g}_t(\mathbf{x}_t)$ . Update  $B_{k,t+1} = B_{k,t} + g_{k,t}(\mathbf{x}_t)$ .
- 6:   **Construct gradient estimate for  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \lambda_t)$**

$$\nabla_t = \frac{d}{\rho} (f_t(\mathbf{x}_t) + \lambda_t^\top \mathbf{g}_t(\mathbf{x}_t)) \cdot \mathbf{u}_t \quad (6)$$

- 7:   **Update exponential weights for experts:** Let  $\ell_t : \mathcal{X} \rightarrow \mathbb{R}$  be a surrogate loss function to measure the performance of each expert. Then we update expert weights by

$$w_{i,t+1} = w_{i,t} \exp(-\epsilon \ell_t(\tilde{\mathbf{x}}_t^i)) \quad \text{where} \quad \ell_t(\mathbf{z}) = \nabla_t^\top (\tilde{\mathbf{x}}_t - \mathbf{z}) \quad (7)$$

- 8:   **Primal ascent per expert and dual descent:**

$$\tilde{\mathbf{x}}_{t+1}^i = \Pi_{(1-\alpha)\mathcal{X}}(\tilde{\mathbf{x}}_t^i + \gamma_i \nabla_t) \quad \text{and} \quad \lambda_{t+1} = \Pi_{[0, C_e]}(\lambda_t - \eta \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t))_+ \quad (8)$$

- 9: **end while**

- 10: Set stopping time  $\tau_A = t$ . For  $t = \tau_A \dots T$  set safety option  $\mathbf{x}_t = \tilde{\mathbf{x}}_\beta$ .
- 

- **$\delta$ -corrupted:**  $\mathcal{R}_T \leq \mathcal{O}(T^{3/4} + \delta)$
- **Periodic:**  $\mathcal{R}_T \leq \mathcal{O}(T^{3/4} + q\sqrt{T})$
- **Ergodic:**  $\mathcal{R}_T \leq \mathcal{O}(T^{3/4} + \kappa\sqrt{T})$

We also summarize our regret bounds in Table 1. Although it is always challenging to identify which input procedure the data currently come from, Theorem 4.2 shows that the proposed Algorithm 1 achieves reasonable regret bound in five input settings without knowing which setting we are in. We remark that in the adversarial setting, our algorithm is  $(1 - \frac{1}{\xi})$ -competitive, namely, it can achieve at least a  $\frac{1}{\xi}$ -portion of the reward compared to OPT; later in Section 5 we comment that no sublinear regret is achievable. For the  $\delta$ -corrupted world, we corrupt  $\delta$  instances in the input sequence over time, so if  $\delta = 0$ , we recover the stochastic world and achieve diminishing  $\mathcal{O}(T^{3/4})$  regret; in the case we corrupt all samples, then we recover the adversarial case, and the regret becomes  $\mathcal{O}(T)$  as expected. In the periodic world, if period length  $q = 1$ , we recover the stochastic case.

In the rest of this section, we present a proof scratch of Theorem 4.2, and the formal proof of the results can be found in Appendix A. The high-level idea of the proof is that we decompose total

regret into three components as in Lemma 4.3 below, namely the conversion losses due to hard stopping, dual descent, and primal ascent, respectively. Then, we bound each component's regret.

**Lemma 4.3 (Regret decomposition)** *Let  $(\boldsymbol{\lambda}_t \in \mathbb{R}_{\geq 0}^K)_{t \in [T]}$  be the sequence of dual variables generated from Algorithm 1, and define  $h_t(\mathbf{x}_t) = \mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda}_t)$  where the Lagrangian function  $\mathcal{L}_t(\mathbf{x}, \boldsymbol{\lambda})$  is defined in Eq. (3). Then we have*

$$OPT(\mathcal{P}_{1:T}) - \sum_{t \in [T]} \mathbb{E}[f_t(\mathbf{x}_t)] \leq \bar{F}(T - \tau_A) + \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t) + \mathcal{R}_{BOCO}(\tau_A) \quad (9)$$

where  $\mathcal{R}_{BOCO}(\tau_A)$  admits the following bounds in each world:

1. **Stochastic:**  $\max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) - h_t(\mathbf{x}_t)$
2. **Adversarial:**  $\sum_{t \in [\tau_A]} h_t(\tilde{\mathbf{y}}_t) - h_t(\mathbf{x}_t) + \left(1 - \frac{1}{\xi}\right) OPT(\mathcal{P}_{1:T})$  where  $\tilde{\mathbf{y}}_t = \arg \max_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathcal{P}_t}[f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x})]$
3.  **$\delta$ -corrupted:**  $\max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) - h_t(\mathbf{x}_t) + \mathcal{O}(\delta)$
4. **Periodic:**  $\max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) - h_t(\mathbf{x}_t) + \mathcal{O}(\eta q T)$
5. **Ergodic:**  $\max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) - h_t(\mathbf{x}_t) + \mathcal{O}(\eta \kappa T)$ .

Here, we remark that the first term  $\bar{F}(T - \tau_A)$  is the loss due to hard stopping to maintain long-term constraint satisfaction; the second term  $\sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)$  is the loss due to dual descent; and the final term  $\mathcal{R}_{BOCO}(\tau_A)$  is the loss due to primal ascent. The first two components turn out to be identical in different settings, and the input model only affect the last term  $\mathcal{R}_{BOCO}(\tau_A)$ . We next present our key results with which we use to bound first two components (dual descent and hard stop loss), and the third component (primal ascent loss), respectively.

*Bounding dual descent and hard stop loss.* Our strategy to bound the loss due to dual descent and hard stopping relies on bounding some cumulative “complementary slackness” induced from dual descent in the following Lemma 4.4. We note that the bound in Lemma 4.4 holds for any  $\boldsymbol{\lambda} \in [\mathbf{0}, C\mathbf{e}]$ , and thus by choosing appropriate  $\boldsymbol{\lambda}$ , we can “internalize” the losses due to dual descent and hard stopping, as demonstrated in Lemma 4.5.

**Lemma 4.4 (Bounding complementary slackness)** *For any  $\boldsymbol{\lambda} \in [\mathbf{0}, C\mathbf{e}]$  and  $t \in [T]$ , we have*  

$$\sum_{\tau \in [t]} (\boldsymbol{\lambda}_\tau - \boldsymbol{\lambda})^\top \mathbf{g}_\tau(\mathbf{x}_\tau) = \sum_{\tau \in [t]} (\boldsymbol{\lambda}_\tau - \boldsymbol{\lambda})^\top \nabla_{\boldsymbol{\lambda}} \mathcal{L}_\tau(\mathbf{x}_\tau, \boldsymbol{\lambda}_\tau) \leq \frac{\eta}{2} t K \bar{G}^2 + \frac{1}{2\eta} \|\boldsymbol{\lambda}\|_2^2$$

**Lemma 4.5 (Internalizing dual descent and hard stop losses.)** *Given the stopping time  $\tau_A \in [T]$ , w.p. 1 we have  $\bar{F}(T - \tau_A) + \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t) \leq \frac{\eta}{2} T K \bar{G}^2 + \frac{1}{2\eta} \left(C\right)^2 + \bar{F} + C\bar{G}$ .*

*Bounding primal ascent loss (dynamic BOCO regret).* We present our main result Lemma 4.6, which bounds the loss due to primal ascent for given any comparator sequence  $\mathbf{y}_1 \dots \mathbf{y}_T \in \mathcal{X}$ .

**Lemma 4.6 (Bounding primal ascent (dynamic BOCO regret))** *For any  $i \in [N]$ ,  $t \in [T]$  and any sequence  $\mathbf{y}_{1:t} \in \mathcal{X}^t$  we have  $\sum_{\tau \in [t]} h_\tau(\mathbf{y}_\tau) - h_\tau(\mathbf{x}_\tau) \leq \mathcal{O}\left(\frac{\rho T}{\beta} + \frac{1+P(\mathbf{y}_{1:T})}{\gamma_i} + \frac{\gamma_i K T}{\beta^2 \rho^2} + T\epsilon + \frac{1}{\epsilon}\right)$ , where  $P(\mathbf{y}_{1:T}) = \sum_{t \in [T-1]} \|\mathbf{y}_t - \mathbf{y}_{t+1}\|$ , and parameters  $(\alpha, \gamma_i, \epsilon, \rho)$  are specified in Algorithm 1.*

*Proof sketch.* We define the smoothed-versions of BOCO rewards  $h_t : \mathcal{X} \rightarrow \mathbb{R}$  as  $\hat{h}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})}[\mathcal{L}_t(\mathbf{x} + \rho \mathbf{v}, \boldsymbol{\lambda}_t)]$ . Then, we decompose the primal ascent loss (BOCO dynamic regret)  $\sum_{\tau \in [t]} h_\tau(\mathbf{y}_\tau) - h_\tau(\mathbf{x}_\tau)$  into 3 terms:

$$\sum_{\tau \in [t]} h_\tau(\mathbf{y}_\tau) - \hat{h}_\tau((1-\alpha)\mathbf{y}_\tau) + \sum_{\tau \in [t]} \hat{h}_\tau((1-\alpha)\mathbf{y}_\tau) - \hat{h}_\tau(\tilde{\mathbf{x}}_\tau) + \sum_{\tau \in [t]} \hat{h}_\tau(\tilde{\mathbf{x}}_\tau) - h_\tau(\mathbf{x}_\tau)$$

where we recall the  $\tilde{\mathbf{x}}_\tau$  is the un-perturbed version of our primal lever decision in Algorithm 1. The first and third summands can be directly be bounded via exploiting Lipschitz continuity properties for  $h_t$  (see Lemma A.1). The second summand can further be upper bounded (using Lemma A.2) by the surrogate loss difference between the primal decision weighted overall all expert decisions and the comparator decision, namely  $\ell_\tau(\tilde{\mathbf{x}}_\tau) - \ell_\tau((1-\alpha)\mathbf{y}_\tau)$ . Then, we choose any expert  $i \in [N]$  corresponding to primal ascent step size  $\gamma_i > 0$  as an “intermediary” and decompose cumulative surrogate loss as  $\sum_{t \in [T]} \ell_t(\tilde{\mathbf{x}}_t) - \ell_t((1-\alpha)\mathbf{y}_t) = \sum_{t \in [T]} \ell_t(\tilde{\mathbf{x}}_t) - \ell_t(\tilde{\mathbf{x}}_t^i) + \sum_{t \in [T]} \ell_t(\tilde{\mathbf{x}}_t^i) - \ell_t((1-\alpha)\mathbf{y}_t)$ . The first summand corresponds to the cumulative loss of our unperturbed primal decisions against any expert, and the second summand represents the cumulative loss of any expert w.r.t. the comparator sequence. Both summands are bounded in Lemma A.3.  $\square$

We note that choosing different expert as the intermediary may yield different bounds. But since Lemma A.3 holds for any expert  $i \in [N]$ , we can consider the expert with the “optimal” primal ascent step size in each world, respectively.

## 5. Additional discussions

**Lower bounds.** A natural question to ask is whether regret bounds in Theorem 4.2 are optimal. Here, we look at lower bounds in “relaxed” problem settings, e.g. bandit online convex optimization with no-constraints [11, 44], or online constrained optimization under full-information feedback [9, 10]:

Stochastic	$\delta$ -corrupted	Adversarial	Periodic	Ergodic
$\Omega(\sqrt{T})$ [20]	$\Omega(\sqrt{T} + \delta)$ [9]	$\Omega\left(\left(1 - \frac{1}{\xi}\right)T\right)$ [10]	$\Omega(\sqrt{qT})$ [9]	NA

These relaxed lower bounds are applicable to our setting (though they may not be the tightest bounds), and we can see that it may be possible to employ more complex machinery to attain better

regret performances for our bandit convex online optimization setup with uncertain constraints. Yet, to identify an algorithm that can achieve the optimal regret in many worlds is an extremely challenging open problem. For instance, even in the relaxed setting of bandit convex optimization with no-constraints, to the best of our knowledge there is no universal algorithm that achieves optimal regret in just the stochastic and adversarial worlds.

**Other applications.** Our proposed Algorithm 1 addresses a more general problem of online decision making with multi-dimensional decisions, bandit feedback and long-time uncertain constraints. That being said, our algorithm can also be applicable to other problem settings, for example:

*Personalized recommendation and assortment in online retailing.* Online retail platforms aim to optimize sales or revenue by recommending a limited assortment of products to customers. However, the likelihood of customer purchases given an assortment as well as their preferences are unknown, and different customer types may change over time due to evolving market trends. Platforms only observe bandit binary purchase decisions. Hence in our context, the decision set  $\mathcal{X}$  represents a probability simplex over  $d$  items. The online decisions  $\mathbf{x}_t \in \mathcal{X}$  correspond to recommendation probabilities for each item. The function  $f_t(\mathbf{x}_t)$  reflects the revenue generated from customer purchases, while  $\mathbf{g}_t(\mathbf{x}_t)$  captures assortment capacity constraints, product inventory limits, or fairness considerations to ensure equal visibility for each product [18].

*Real time posted pricing in E-commerce and cloud computing.* In online E-commerce platforms such as Amazon, eBay, etc., seller set prices to sell various items to sequentially arriving customers [5, 27, 35]; and in cloud computing, operators of cloud services such as Alibaba Cloud (Alicloud) or Amazon Web Services (AWS) set prices for renting out different computing capacities (virtual machines, VMs etc.) upon customer requests [39]. The arrival of customer types as well as their demand may differ significantly over time (think about demand for sanitary products during pandemics outbursts, or computing resource demands during periodic business hours). Further, decision makers only get to observe (bandit) demand at the realized pricing decisions. Thereby, we can view  $\mathbf{x}_t \in \mathcal{X}$  as the vector of prices for various sold products,  $f_t(\mathbf{x}_t)$  as the generated revenue over a given period,  $\mathbf{g}_t(\mathbf{x}_t)$  as real time product/resource capacity constraints, operating cost constraints, etc.

## References

- [1] Programmatic advertising trends, stats, & news. <https://roirevolution.com/blog/programmatic-advertising-trends-stats-news/>. Accessed: 2023-03-30.
- [2] Programmatic digital display advertising in 2022: Ad spend, formats, and forecast. <https://www.insiderintelligence.com/insights/programmatic-digital-display-ad-spending/>. Accessed: 2023-03-30.

- 
- [3] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.
  - [4] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.
  - [5] Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply, 2015.
  - [6] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.
  - [7] Santiago Balseiro, Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Robust auction design in the auto-bidding world. *Advances in Neural Information Processing Systems*, 34:17777–17788, 2021.
  - [8] Santiago Balseiro, Anthony Kim, Mohammad Mahdian, and Vahab Mirrokni. Budget management strategies in repeated auctions. *Operations Research*, 2021. forthcoming.
  - [9] 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.
  - [10] Santiago R Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.
  - [11] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.
  - [12] Mark Braverman, Jieming Mao, Jon Schneider, and Matt Weinberg. Selling to a no-regret buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 523–538, 2018.
  - [13] Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022.
  - [14] Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online bidding in repeated non-truthful auctions under budget and roi constraints. *arXiv preprint arXiv:2302.01203*, 2023.
  - [15] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *arXiv preprint arXiv:2209.07454*, 2022.
  - [16] Sapana Chaudhary and Dileep Kalathil. Safe online convex optimization with unknown linear safety constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6175–6182, 2022.
  - [17] Lin Chen, Mingrui Zhang, and Amin Karbasi. Projection-free bandit convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2047–2056. PMLR, 2019.
  - [18] Qinyi Chen, Negin Golrezaei, Fransisca Susan, and Edy Baskoro. Fair assortment planning. *arXiv preprint arXiv:2208.07341*, 2022.

- 
- [19] Qinyi Chen, Jason Cheuk Nam Liang, Negin Golrezaei, and Djallel Bouneffouf. Interpolating item and user fairness in recommendation systems. *arXiv preprint arXiv:2306.10050*, 2023.
  - [20] Varsha Dani, Sham M Kakade, and Thomas Hayes. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2007.
  - [21] Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. Fairness in the autobidding world with machine-learned advice. *arXiv preprint arXiv:2209.04748*, 2022.
  - [22] Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. Multi-channel autobidding with budget and roi constraints. *arXiv preprint arXiv:2302.01523*, 2023.
  - [23] Yuan Deng, Jieming Mao, Vahab Mirrokni, Hanrui Zhang, and Song Zuo. Efficiency of the first-price auction in the autobidding world. *arXiv preprint arXiv:2208.10650*, 2022.
  - [24] Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Towards efficient auctions in an auto-bidding world. In *Proceedings of the Web Conference 2021*, pages 3965–3973, 2021.
  - [25] Zhe Feng, Swati Padmanabhan, and Di Wang. Online bidding algorithms for return-on-spend constrained advertisers. *arXiv preprint arXiv:2208.13713*, 2022.
  - [26] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.
  - [27] Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. No-regret learning in price competitions under consumer reference effects. *Advances in Neural Information Processing Systems*, 33:21416–21427, 2020.
  - [28] Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9331–9361. PMLR, 2023.
  - [29] 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.
  - [30] Negin Golrezaei, Vahideh Manshadi, Jon Schneider, and Shreyas Sekar. Learning product rankings robust to fake users. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 560–561, 2021.
  - [31] Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*, 2020.
  - [32] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*, 2020.
  - [33] Elad Hazan and Yuanzhi Li. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.



- 
- [34] 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.
  - [35] Chia-Wei Kuo, Hyun-Soo Ahn, and Göker Aydın. Dynamic pricing of limited inventories when customers negotiate. *Operations research*, 59(4):882–897, 2011.
  - [36] Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In *International Conference on Machine Learning*, pages 3944–3952. PMLR, 2019.
  - [37] 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.
  - [38] Aranyak Mehta. Auction design in an auto-bidding setting: Randomization improves efficiency beyond vcg. In *Proceedings of the ACM Web Conference 2022*, pages 173–181, 2022.
  - [39] Huijie Peng, Yan Cheng, and Xingyuan Li. Real-time pricing method for spot cloud services with non-stationary excess capacity. *Sustainability*, 15(4):3363, 2023.
  - [40] 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.
  - [41] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.
  - [42] Scott Yang and Mehryar Mohri. Optimistic bandit convex optimization. *Advances in Neural Information Processing Systems*, 29, 2016.
  - [43] Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. *Advances in Neural Information Processing Systems*, 31, 2018.
  - [44] Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. *The Journal of Machine Learning Research*, 22(1):5562–5606, 2021.

# Appendices for Online Ad Procurement in Non-stationary Autobidding Worlds

## Appendix A: Proofs for Section 4

### A.1. Additional definitions for Section 4

**Definition A.1 (Total variation between probability distributions)** Consider two distributions  $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{S})$ . Then we define their total variation as  $\|\mathcal{P} - \mathcal{P}'\|_{TV} = \frac{1}{2} \int_{\mathcal{S}} |\mathcal{P}(s) - \mathcal{P}'(s)| ds$ .

We also define the smoothed version of  $h_t : \mathcal{X} \rightarrow \mathbb{R}$  (see Eq. (4)) for any  $t$  as followed:

$$\hat{h}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})} [\mathcal{L}_t(\mathbf{x} + \rho \mathbf{v}, \boldsymbol{\lambda}_t)] \quad (10)$$

where we recall the Lagrangian function  $\mathcal{L}_t$  is defined in Eq. (3).

### A.2. Additional lemmas for Section 4

**Lemma A.1 (Lipschitz continuity)** Let Assumption 2.1 hold, and recall the definitions  $h_t(\mathbf{x})$  and  $\hat{h}_t(\mathbf{x})$  from Eqs. (4) as well as (10), respectively, and recall  $\boldsymbol{\lambda}_1 \dots \boldsymbol{\lambda}_T$  are the dual variables generated from Algorithm 1. Then for any  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ , we have  $|h_t(\mathbf{x}) - h_t(\mathbf{x}')| \leq (1 + KC)L \cdot \|\mathbf{x} - \mathbf{x}'\|$  and  $|h_t(\mathbf{x}) - \hat{h}_t(\mathbf{x})| \leq (1 + KC)L\rho$ .

**Lemma A.2 (Bounding BOCO dynamic regret with surrogate loss)** Recall the definition  $\hat{h}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})} [\mathcal{L}_t(\mathbf{x} + \rho \mathbf{v}, \boldsymbol{\lambda}_t)]$ . Then,  $\hat{h}_t(\mathbf{x})$  is concave. Further, For any  $\mathbf{y} \in (1 - \alpha)\mathcal{X}$ , we have  $\hat{h}_t(\mathbf{y}) - \hat{h}_t(\tilde{\mathbf{x}}_t) \leq \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} [\ell_t(\tilde{\mathbf{x}}_t) - \ell_t(\mathbf{y})]$ , where  $\tilde{\mathbf{x}}_t$  is defined in Eq. (5), and the surrogate loss function  $\ell_t : \mathcal{X} \rightarrow \mathbb{R}$  is defined in Eq. (7).

**Lemma A.3 (Bounding surrogate loss for each expert)** Recall the definition of individual forecasters  $\tilde{\mathbf{x}}_t^i$  defined in Eq. (8), and the surrogate loss function  $\ell_t : \mathcal{X} \rightarrow \mathbb{R}$  defined in Eq. (7). Then for any  $i \in [N]$  and any sequence  $\mathbf{y}_{1:T} \in \mathcal{X}^T$  we have (i)  $\sum_{t \in [T]} \ell_t(\tilde{\mathbf{x}}_t^i) - \ell_t((1 - \alpha)\mathbf{y}_t) \leq \mathcal{O}\left(\frac{1+P(\mathbf{y}_{1:T})}{\gamma_i} + \frac{\gamma_i}{\beta^2 \rho^2} T\right)$  and (ii)  $\sum_{t \in [T]} \ell_t(\tilde{\mathbf{x}}_t) - \ell_t(\tilde{\mathbf{x}}_t^i) \leq \mathcal{O}(T\epsilon + \frac{1}{\epsilon})$ . where the constant  $\beta$  is specified in Algorithm 1. Here, recall  $D$  is the diameter of the decision set  $\mathcal{X}$ .

The proofs of Lemmas A.1, A.2, A.3 are shown in Appendices A.9, A.10, and A.11, respectively.

### A.3. Proof for Lemma 4.1

For any  $k \in [K]$  we have

$$\begin{aligned} \sum_{t \in [T]} g_{k,t}(\mathbf{x}_t) &= \sum_{t \in [\tau_A - 1]} g_{k,t}(\mathbf{x}_t) + \sum_{t = \tau_A}^T g_{k,t}(\mathbf{x}_t) \stackrel{(a)}{\geq} \sum_{t \in [\tau_A - 1]} g_{k,t}(\mathbf{x}_t) + \bar{\beta}(T - \tau_A + 1) \\ &\geq \sum_{t \in [\tau_A - 1]} g_{k,t}(\mathbf{x}_t) + \beta(T - \tau_A) + \beta \stackrel{(b)}{\geq} \bar{G} + \beta > 0 \end{aligned} \quad (11)$$

where in (a) we set  $\mathbf{x}_t = \tilde{\mathbf{x}}_\beta$  for all  $t = \tau_A \dots T$  and  $g_{k,t}(\tilde{\mathbf{x}}_\beta) \geq \bar{\beta}$  for any  $k \in [K]$ ; (b) follows from the definition of the stopping time such that for any  $t' < \tau_A$  and  $k \in [K]$  we have  $\sum_{t \in [t']} g_{k,t}(\mathbf{x}_t) - \bar{G} + \beta(T - t' - 1) \geq 0$ .  $\square$

#### A.4. Proof for Lemma 4.4

It is easy to see  $\lambda_{t+1} = \Pi_{[0, Ce]}(\lambda_t - \eta \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t))_+ = \arg \min_{\lambda \in [0, Ce]} \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top \lambda + \frac{1}{2\eta} \|\lambda - \lambda_t\|^2$ . By the first-order stationary condition at  $\lambda_{t+1}$ , we have for any  $\lambda \in [0, Ce]$

$$\left( \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t) + \frac{1}{\eta} (\lambda_{t+1} - \lambda_t) \right)^\top (\lambda - \lambda_{t+1}) \geq 0$$

Then for all  $\lambda \in \mathbb{R}_{\geq 0}^K$ , it follows that

$$\begin{aligned} & \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top (\lambda_t - \lambda) \\ &= \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top (\lambda_t - \lambda_{t+1}) + \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top (\lambda_{t+1} - \lambda) \\ &\leq \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top (\lambda_t - \lambda_{t+1}) + \frac{1}{\eta} (\lambda_{t+1} - \lambda_t)^\top (\lambda - \lambda_{t+1}) \\ &\leq \nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)^\top (\lambda_t - \lambda_{t+1}) + \frac{1}{2\eta} \|\lambda - \lambda_t\|^2 - \frac{1}{2\eta} \|\lambda - \lambda_{t+1}\|^2 - \frac{1}{2\eta} \|\lambda_{t+1} - \lambda_t\|^2 \\ &\leq \frac{\eta}{2} \|\nabla_{\lambda} \mathcal{L}_t(\mathbf{x}_t, \lambda_t)\|^2 + \frac{1}{2\eta} \|\lambda - \lambda_t\|^2 - \frac{1}{2\eta} \|\lambda - \lambda_{t+1}\|^2 \end{aligned}$$

By a telescoping argument, we have

$$\begin{aligned} \sum_{\tau \in [t]} \nabla_{\lambda} \mathcal{L}_{\tau}(\mathbf{x}_{\tau}, \lambda_{\tau})^\top (\lambda_{\tau} - \lambda) &\leq \frac{\eta}{2} \sum_{\tau \in [t]} \|\nabla_{\lambda} \mathcal{L}_{\tau}(\mathbf{x}_{\tau}, \lambda_{\tau})\|^2 + \frac{1}{2\eta} \|\lambda - \lambda_1\|^2 \\ &= \frac{\eta}{2} \sum_{\tau \in [t]} \|\nabla_{\lambda} \mathcal{L}_{\tau}(\mathbf{x}_{\tau}, \lambda_{\tau})\|^2 + \frac{1}{2\eta} \|\lambda\|^2. \end{aligned} \tag{12}$$

where in the final equality we used  $\lambda_1 = \mathbf{0}$ . Also,

$$\|\nabla_{\lambda} \mathcal{L}_{\tau}(\mathbf{x}_{\tau}, \lambda_{\tau})\|^2 = \|\mathbf{g}_{\tau}(\mathbf{x}_{\tau})\|^2 \leq K \bar{G}^2 \tag{13}$$

Hence, combining Eqs. (12) and (13), we get the desired bound.  $\square$

#### A.5. Proof of Lemma 4.5

If  $\tau_A = T$ , taking  $\lambda = \mathbf{0}$  in Lemma 4.4 yields  $\sum_{t \in [T]} \lambda_t^\top \mathbf{g}_t(\mathbf{x}_t) \leq \frac{\eta}{2} TK \bar{G}^2$  and thus the desired inequality holds. If  $\tau_A < T$ , then there exists some  $k \in [K]$  such that  $\sum_{t \in [\tau_A]} g_{k,t}(\mathbf{x}_t) - \bar{G} + \beta(T - \tau_A - 1) < 0$ , so by taking  $\lambda = C \mathbf{e}_k$  ( $\mathbf{e}_k \in \mathbb{R}^K$  is the unit vector whose  $k$ th entry is 1) in Lemma 4.4 yields

$$\begin{aligned} & \sum_{t \in [\tau_A]} \lambda_t^\top \mathbf{g}_t(\mathbf{x}_t) \\ &\leq \sum_{t \in [\tau_A]} \lambda^\top \mathbf{g}_t(\mathbf{x}_t) + \frac{\eta}{2} TK \bar{G}^2 + \frac{1}{2\eta} \|\lambda\|^2 \\ &= C \sum_{t \in [\tau_A]} g_{k,t}(\mathbf{x}_t) + \frac{\eta}{2} TK \bar{G}^2 + \frac{1}{2\eta} (C)^2 \\ &\leq -C \cdot \beta(T - \tau_A - 1) + C \bar{G} + \frac{\eta}{2} TK \bar{G}^2 + \frac{1}{2\eta} (C)^2 \\ &= -\bar{F}(T - \tau_A) + \bar{F} + C \bar{G} + \frac{\eta}{2} TK \bar{G}^2 + \frac{1}{2\eta} (C)^2 \end{aligned}$$

Summing with  $\bar{F}(T - \tau_A)$  yields the desired result.  $\square$

### A.6. Proof of Lemma 4.6

Recall the definition of  $\hat{h}_t(\mathbf{x})$  in Eq. (4). Then, we have

$$\begin{aligned} & \sum_{\tau \in [t]} h_\tau(\mathbf{y}_\tau) - \sum_{\tau \in [t]} h_\tau(\mathbf{x}_\tau) \\ &= \sum_{\tau \in [t]} \left( \underbrace{h_\tau(\mathbf{y}_\tau) - \hat{h}_\tau((1-\alpha)\mathbf{y}_\tau)}_A + \underbrace{\hat{h}_\tau((1-\alpha)\mathbf{y}_\tau) - \hat{h}_\tau(\tilde{\mathbf{x}}_\tau)}_B + \underbrace{\hat{h}_\tau(\tilde{\mathbf{x}}_\tau) - h_\tau(\mathbf{x}_\tau)}_C \right) \end{aligned} \quad (14)$$

**Bounding A.**

$$\begin{aligned} h_\tau(\mathbf{y}_\tau) - \hat{h}_\tau((1-\alpha)\mathbf{y}_\tau) &= h_\tau(\mathbf{y}_\tau) - h_\tau((1-\alpha)\mathbf{y}_\tau) + h_\tau((1-\alpha)\mathbf{y}_\tau) - \hat{h}_\tau((1-\alpha)\mathbf{y}_\tau) \\ &\stackrel{(a)}{\leq} (1+KC)L\alpha\|\mathbf{y}_\tau\| + (1+KC)L\rho \\ &\stackrel{(b)}{\leq} (1+KC)L\alpha D + (1+KC)L\rho \end{aligned} \quad (15)$$

where (a) follows from Lemma A.1; (b) follows from  $\|\mathbf{y}_\tau\| = \|\mathbf{y}_\tau - \mathbf{0}\| \leq D$  since we assumed  $\mathbf{0} \in \mathcal{X}$ .

**Bounding B.**

$$\begin{aligned} & \sum_{t \in [T]} \hat{h}_t((1-\alpha)\mathbf{y}_t) - \hat{h}_t(\tilde{\mathbf{x}}_t) \\ &\stackrel{(a)}{\leq} \sum_{t \in [T]} \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} [\ell_t(\tilde{\mathbf{x}}_t) - \ell_t((1-\alpha)\mathbf{y}_t)] \\ &= \sum_{t \in [T]} \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} [\ell_t(\tilde{\mathbf{x}}_t) - \ell_t(\tilde{\mathbf{x}}_t^i) + \ell_t(\tilde{\mathbf{x}}_t^i) - \ell_t((1-\alpha)\mathbf{y}_t)] \\ &\stackrel{(b)}{\leq} \mathcal{O}\left(\frac{P(\mathbf{y}_{1:T})}{\gamma_i} + \frac{\gamma_i KCT}{\rho^2} + T\epsilon + \frac{1}{\epsilon}\right) \end{aligned} \quad (16)$$

where (a) follows from Lemma A.2 and (b) follows from Lemma A.3 (i) and (ii).

**Bounding C.**

$$\begin{aligned} \hat{h}_\tau(\tilde{\mathbf{x}}_\tau) - h_\tau(\mathbf{x}_\tau) &= \hat{h}_\tau(\tilde{\mathbf{x}}_\tau) - h_\tau(\tilde{\mathbf{x}}_\tau) + h_\tau(\tilde{\mathbf{x}}_\tau) - h_\tau(\mathbf{x}_\tau) \\ &\stackrel{(a)}{\leq} (1+KC)L\rho + (1+KC)L \cdot \|\tilde{\mathbf{x}}_\tau - \mathbf{x}_\tau\| \\ &\stackrel{(b)}{=} (1+KC)L\rho + (1+KC)L \cdot \|\rho\mathbf{u}_\tau\| \\ &\leq 2\rho(1+KC)L \end{aligned} \quad (17)$$

where (a) follows from Lemma A.1; (b) follows from the definition  $\mathbf{x}_\tau = \tilde{\mathbf{x}}_\tau + \rho\mathbf{u}_\tau$  in Algorithm 1.  $\square$

### A.7. Proof of Lemma 4.3

**Stochastic.**

In the stochastic regime, we have  $\mathcal{P} = \mathcal{P}_1 = \dots = \mathcal{P}_T$  for some  $\mathcal{P}$ , and therefore we can rewrite  $\text{OPT}(\mathcal{P}_{1:T})$  in Eq. (1) as followed

$$\text{OPT}(\mathcal{P}_{1:T}) = \max_{\mathbf{x}_{1:T} \in \mathcal{X}^T} \sum_{t \in [T]} F(\mathbf{x}_t) \quad \text{s.t.} \quad \sum_{t \in [T]} \mathbf{G}(\mathbf{x}_t) \geq \mathbf{0}.$$

where we defined  $F(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}}[f(\mathbf{x})]$ , and  $\mathbf{G}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}}[\mathbf{g}(\mathbf{x})]$  for any  $\mathbf{x} \in \mathcal{X}$ . Hence, for any  $\boldsymbol{\lambda} \geq \mathbf{0}$  we have

$$\begin{aligned}
\text{OPT}(\mathcal{P}_{1:T}) &= \frac{T - \tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) + \frac{\tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) \\
&\leq (T - \tau_A) \bar{F} + \frac{\tau_A}{T} \max_{\mathbf{x}_{1:T} \in \mathcal{X}^T} \sum_{t \in [T]} (F(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}(\mathbf{x}_t)) \\
&= (T - \tau_A) \bar{F} + \frac{\tau_A}{T} \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [T]} (F(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{G}(\mathbf{x})) \\
&= (T - \tau_A) \bar{F} + \tau_A \max_{\mathbf{x} \in \mathcal{X}} (F(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{G}(\mathbf{x}))
\end{aligned} \tag{18}$$

where in the inequality we applied Assumption 2.1 which states  $\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  for all  $(f, g) \in \mathcal{S}$ . Choosing  $\boldsymbol{\lambda} = \bar{\boldsymbol{\lambda}}_{\tau_A} := \frac{1}{\tau_A} \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t$  we have

$$\begin{aligned}
\text{OPT}(\mathcal{P}_{1:T}) &\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \tau_A \max_{\mathbf{x} \in \mathcal{X}} (F(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{G}(\mathbf{x})) \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (F(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{G}(\mathbf{x})) \right] \\
&\stackrel{(a)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \mathbb{E} \left[ f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}) \mid \sigma(\mathcal{H}_{t-1}) \right] \right] \\
&\stackrel{(b)}{=} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \mathbb{E} \left[ h_t(\mathbf{x}) \mid \sigma(\mathcal{H}_{t-1}) \right] \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) \right]
\end{aligned} \tag{19}$$

where in (a) we used the fact that  $\boldsymbol{\lambda}_t$  is  $\mathcal{H}_{t-1}$ -measurable; in (b) we used definitions  $h_t(\mathbf{x}) = \mathcal{L}_t(\mathbf{x}; \boldsymbol{\lambda}_t)$  and  $\mathcal{L}_t(\mathbf{x}; \boldsymbol{\lambda}) = f_t(\mathbf{x}) + \boldsymbol{\lambda}^\top \mathbf{g}_t(\mathbf{x})$  in Eqs. (3) and (4) respectively.

On the other hand, we have

$$f_t(\mathbf{x}_t) = h_t(\mathbf{x}_t) - \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t), \tag{20}$$

so combining this with Eq. (19) we have

$$\text{OPT}(\mathcal{P}_{1:T}) - \sum_{t \in [T]} \mathbb{E}[f_t(\mathbf{x}_t)] \leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (h_t(\mathbf{x}) - h_t(\mathbf{x}_t)) + \sum_{t \in \tau_A} \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t) \right] \tag{21}$$

where we also used the fact that  $f_t(\mathbf{x}) \geq 0$  for all  $t = \tau_A + 1 \dots T$  and  $\mathbf{x} \in \mathcal{X}$ .  $\square$

### Adversarial.

Recall the definition of  $\xi$  is Theorem 4.2:

$$\xi = 1 - \frac{\min_{(f,g) \in \mathcal{S}} \min_{k \in [K], \mathbf{x} \in \mathcal{X}} g_k(\mathbf{x})}{\bar{\beta}} > 1 \tag{22}$$

For any  $t \in [T]$ , define  $\tilde{\mathbf{y}}_t = \arg \max_{\mathbf{x}} f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x})$ .

By comparing to the safety action  $\mathbf{x}_\beta \in \mathcal{X}$  which ensures  $g_k(\mathbf{x}_\beta) \geq \bar{\beta}$  for any  $k \in [K]$  and  $(f, g) \in \mathcal{S}$ , as well as the optimal hindsight action  $\mathbf{x}_t^* \in \mathcal{X}$  (i.e.  $\mathbf{x}_1^* \dots \mathbf{x}_T^*$  is the optimal decision sequence to  $\text{OPT}(\mathcal{P}_{1:T})$ ), we have

$$\begin{aligned}
f_t(\tilde{\mathbf{y}}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t) &\geq f_t(\mathbf{x}_\beta) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_\beta) \geq \bar{\beta} \boldsymbol{\lambda}_t^\top \mathbf{e} \\
f_t(\tilde{\mathbf{y}}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t) &\geq f_t(\mathbf{x}_t^*) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t^*).
\end{aligned} \tag{23}$$

We further have

$$\begin{aligned}
\xi f_t(\tilde{\mathbf{y}}_t) &= f_t(\tilde{\mathbf{y}}_t) + (\xi - 1)f_t(\tilde{\mathbf{y}}_t) \\
&\stackrel{(a)}{\geq} f_t(\mathbf{x}_t^*) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t^*) - \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t) + (\xi - 1)(-\boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t) + \bar{\beta} \boldsymbol{\lambda}_t^\top \mathbf{e}) \\
&= f_t(\mathbf{x}_t^*) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t^*) - \xi \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t) + (\xi - 1)\bar{\beta} \boldsymbol{\lambda}_t^\top \mathbf{e} \\
&\stackrel{(b)}{\geq} f_t(\mathbf{x}_t^*) - \xi \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t)
\end{aligned} \tag{24}$$

where (a) follows Eq.(23); in (b) we used the fact that  $g_{k,t}(\mathbf{x}_t^*) + (\xi - 1)\bar{\beta} \geq 0$  since we have  $\min_{(f,g) \in \mathcal{S}} \min_{k \in [K], \mathbf{x} \in \mathcal{X}} (g_{k,t}(\mathbf{x}) + (\xi - 1)\bar{\beta}) \geq 0$  (see Eq. (22)). Hence we have

$$\begin{aligned}
&\text{OPT}(\mathcal{P}_{1:T}) - \sum_{t \in [T]} \mathbb{E}[f_t(\mathbf{x}_t)] \\
&= \left(1 - \frac{1}{\xi}\right) \text{OPT}(\mathcal{P}_{1:T}) + \sum_{t \in [T]} \mathbb{E}\left[\frac{1}{\xi} f_t(\mathbf{x}_t^*) - f_t(\mathbf{x}_t)\right] \\
&\leq \left(1 - \frac{1}{\xi}\right) \text{OPT}(\mathcal{P}_{1:T}) + \sum_{t \in [T]} \mathbb{E}\left[f_t(\tilde{\mathbf{y}}_t) - f_t(\mathbf{x}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t)\right] \\
&\leq \left(1 - \frac{1}{\xi}\right) \text{OPT}(\mathcal{P}_{1:T}) + \mathbb{E}\left[(T - \tau_A) \bar{F} + \sum_{t \in \tau_A} \left(f_t(\tilde{\mathbf{y}}_t) - f_t(\mathbf{x}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\tilde{\mathbf{y}}_t)\right)\right]
\end{aligned} \tag{25}$$

□

### $\delta$ -corrupted.

Here, we will prove a more general  $\delta$ -corrupted model where the input distribution sequence  $\mathcal{P}_{1:T}$  satisfies the following:

$$\sum_{t \in [T]} \|\mathcal{P}_t - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} \leq \delta \tag{26}$$

where the total variation norm is defined in Definition A.1. In fact, the definition in Section 2.2 for the  $\delta$ -corrupted regime satisfies the above property: recall in the definition of Section 2.2, there exists  $\mathcal{P} \in \Delta(\mathcal{S})$  as well as  $\delta \in \mathbb{N}$  periods  $\mathcal{T} = \{\tau_1 \dots \tau_\delta\} \subset [T]$  such that  $\mathcal{P}_t = \mathcal{P}$  for all  $t \notin \mathcal{T}$ , hence for any  $t \notin \mathcal{T}$ , we have

$$\begin{aligned}
\|\mathcal{P}_t - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} &= \|\mathcal{P} - \frac{1}{T} \left(T\mathcal{P} + \sum_{s \in \mathcal{T}} (\mathcal{P} - \mathcal{P}_s)\right)\|_{TV} \\
&= \left\| \frac{1}{T} \sum_{s \in \mathcal{T}} (\mathcal{P} - \mathcal{P}_s) \right\|_{TV} \\
&\leq \frac{\delta}{2T}
\end{aligned} \tag{27}$$

On the other hand, we have for any  $\tau \in \mathcal{T}$ ,  $\|\mathcal{P}_\tau - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} \leq \frac{1}{2}$ . Hence, summing up we get

$$\begin{aligned}
\sum_{t \in [T]} \|\mathcal{P}_t - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} &= \sum_{t \in \mathcal{T}} \|\mathcal{P}_t - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} + \sum_{t \notin \mathcal{T}} \|\mathcal{P}_t - \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s\|_{TV} \\
&\leq \frac{\delta}{2} + (T - \delta) \frac{\delta}{2T} \leq \delta
\end{aligned}$$

which coincides with our general definition of  $\delta$ -corruption in Eq. (26).

We now prove the  $\delta$ -corruption regime under the general definition in Eq. (26). Define  $\tilde{\mathcal{P}} = \frac{1}{T} \sum_{s \in [T]} \mathcal{P}_s$ ,  $\tilde{F}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \tilde{\mathcal{P}}} [f(\mathbf{x})]$ ,  $\tilde{\mathbf{G}}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \tilde{\mathcal{P}}} [\mathbf{g}(\mathbf{x})]$ ,  $F_t(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_t} [f(\mathbf{x})]$  and  $\mathbf{G}_t(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_t} [\mathbf{g}(\mathbf{x})]$  for all  $t \in [T]$  and any  $\mathbf{x} \in \mathcal{X}$ . Then for any  $\boldsymbol{\lambda} \in [\mathbf{0}, C\mathbf{e}]$ , we have

$$\begin{aligned}
\text{OPT}(\mathcal{P}_{1:T}) &\leq \max_{\mathbf{x}_{1:T} \in \mathcal{X}^T} \sum_{t \in [T]} (F_t(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}_t(\mathbf{x}_t)) \\
&\leq \max_{\mathbf{x}_{1:T} \in \mathcal{X}} \sum_{t \in [T]} (\tilde{F}(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x}_t)) + (\bar{F} + \bar{G}KC)\delta \\
&= T \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + (\bar{F} + \bar{G}KC)\delta,
\end{aligned} \tag{28}$$

where the last inequality follows the definitions of  $(\tilde{F}, \tilde{\mathbf{G}})$ , Assumption 2.1, and the general definition of  $\delta$ -corruption in Eq. (26). After choosing  $\bar{\boldsymbol{\lambda}} = \frac{1}{\tau_A} \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t$ , similar to our proof in Eq. (19) for the stochastic case we have

$$\begin{aligned}
&\text{OPT}(\mathcal{P}_{1:T}) \\
&= \mathbb{E} \left[ \frac{T - \tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) + \frac{\tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) \right] \\
&\stackrel{(a)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \tau_A \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC)\delta \right] \\
&= \mathbb{E} \left[ (T - \tau_A) \bar{F} + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC)\delta + \max_{\mathbf{x} \in \mathcal{X}} \left( \sum_{t \in [\tau_A]} \tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) \right] \\
&\stackrel{(b)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \left( 1 + \frac{\tau_A}{T} \right) (\bar{F} + \bar{G}KC)\delta + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (F_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{G}_t(\mathbf{x})) \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \left( 1 + \frac{\tau_A}{T} \right) (\bar{F} + \bar{G}KC)\delta + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \mathbb{E} \left[ f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}) \mid \sigma(\mathcal{H}_{t-1}) \right] \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + 2(\bar{F} + \bar{G}KC)\delta + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) \right],
\end{aligned} \tag{29}$$

where (a) follows from Eq. (28); (b) follows from the definition of general  $\delta$ -corruption in Eq. (26).

Finally, we complete the proof by using the definition  $f_t(\mathbf{x}_t) = h_t(\mathbf{x}_t) - \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)$  and following the same argument as in Eq. (21) for the stochastic regime.

### Periodic.

Recall in Section 2.2 that in the periodic regime, there exists cycle length  $q \in \mathbb{N}$  such that  $T = cq$  for some integer  $c \geq 2$  with  $\mathcal{P}_{1:T}$  as  $\mathcal{P}_{1:q} = \mathcal{P}_{q+1:2q} = \dots = \mathcal{P}_{(c-1)q+1:T}$ . For any  $t \in [T]$ , define  $c_t \in [c]$  such that  $(c_t - 1)q + 1 \leq t \leq c_t q$ . After denoting  $\tilde{\mathcal{P}} = \frac{1}{q} \sum_{t \in [q]} \mathcal{P}_t$ , we define the mean deviation within a single cycle of length  $q$  as

$$MD(\mathcal{P}_{1:q}) = \sum_{1 \leq t \leq q} \|\mathcal{P}_t - \tilde{\mathcal{P}}\|_{TV} \quad \text{and} \quad \delta = c \cdot MD(\mathcal{P}_{1:q}). \tag{30}$$

We define  $\tilde{F}(\mathbf{x}) = \mathbb{E}_{(f, \mathbf{g}) \sim \tilde{\mathcal{P}}} [f(\mathbf{x})]$ ,  $\tilde{\mathbf{G}}(\mathbf{x}) = \mathbb{E}_{(f, \mathbf{g}) \sim \tilde{\mathcal{P}}} [\mathbf{g}(\mathbf{x})]$ ,  $F_t(\mathbf{x}) = \mathbb{E}_{(f, \mathbf{g}) \sim \mathcal{P}_t} [f(\mathbf{x})]$  and  $\mathbf{G}_t(\mathbf{x}) = \mathbb{E}_{(f, \mathbf{g}) \sim \mathcal{P}_t} [\mathbf{g}(\mathbf{x})]$  for all  $t \in [T]$  and any  $\mathbf{x} \in \mathcal{X}$ . Then for any  $\boldsymbol{\lambda} \in [0, Ce]$ , we have

$$\begin{aligned}
\text{OPT}(\mathcal{P}_{1:T}) &\leq \max_{\mathbf{x}_{1:T} \in \mathcal{X}^T} \sum_{t \in [T]} (F_t(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}_t(\mathbf{x}_t)) \\
&= c \cdot \max_{\mathbf{x}_{1:q} \in \mathcal{X}^q} \sum_{t \in [q]} (F_t(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}_t(\mathbf{x}_t)) \\
&\leq cq \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + (\bar{F} + \bar{G}KC)c \cdot MD(\mathcal{P}_{1:q}) \\
&\leq cq \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + (\bar{F} + \bar{G}KC)\delta,
\end{aligned}$$

where the equality follows the nature of periodic setting and the last inequality follows the definitions of  $(\tilde{F}, \tilde{G})$ , Assumption 2.1, and (30). After choosing  $\boldsymbol{\lambda} = \sum_{\hat{c} \in [c_{\tau_A}-1]} \frac{q}{\tau_A} \boldsymbol{\lambda}_{(\hat{c}-1)q+1} + \frac{\tau_A - (c_{\tau_A}-1)q}{\tau_A} \boldsymbol{\lambda}_{(c_{\tau_A}-1)q+1}$ , we further have that

$$\begin{aligned}
& \text{OPT}(\mathcal{P}_{1:T}) \\
&= \frac{T - \tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) + \frac{\tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) \\
&\leq (T - \tau_A) \bar{F} + \tau_A \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC) \delta \\
&= (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \left( \tau_A \tilde{F}(\mathbf{x}) + \left( \sum_{\hat{c} \in [c_{\tau_A}-1]} q \boldsymbol{\lambda}_{(\hat{c}-1)q+1} + (\tau_A - (c_{\tau_A}-1)q) \boldsymbol{\lambda}_{(c_{\tau_A}-1)q+1} \right)^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) \\
&\quad + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC) \delta \\
&= (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \left( q \cdot \sum_{\hat{c} \in [c_{\tau_A}-1]} \left( \tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_{(\hat{c}-1)q+1}^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) \right. \\
&\quad \left. + (\tau_A - (c_{\tau_A}-1)q) \cdot \left( \tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_{(c_{\tau_A}-1)q+1}^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) \right) + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC) \delta \\
&\leq (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \left( \tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) + \bar{G} \cdot \sum_{t \in [\tau_A]} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{(c_t-1)q+1}\|_1 + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC) \delta.
\end{aligned}$$

From (8) in Algorithm 1, we know that  $\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\|_1 \leq \eta \bar{G}K$ , which further implies  $\|\boldsymbol{\lambda}_{t+i} - \boldsymbol{\lambda}_t\|_1 \leq \eta \bar{G}Ki$  for any  $i \in [q-1]$  and thus

$$\sum_{t \in [\tau_A]} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{(c_t-1)q+1}\|_1 \leq c_{\tau_A} \eta \bar{G}K \sum_{i \in [q-1]} i \leq \frac{1}{2} \bar{G}K \eta c_{\tau_A} q^2. \quad (31)$$

After combining the two equations above, it follows that

$$\begin{aligned}
& \text{OPT}(\mathcal{P}_{1:T}) \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \left( \tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \tilde{\mathbf{G}}(\mathbf{x}) \right) + \frac{1}{2} \bar{G}^2 K \eta c_{\tau_A} q^2 + \frac{\tau_A}{T} (\bar{F} + \bar{G}KC) \delta \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \left( F_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{G}_t(\mathbf{x}) \right) + \frac{1}{2} \bar{G}^2 K \eta c_{\tau_A} q^2 + 2(\bar{F} + \bar{G}KC) \delta \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + 2(\bar{F} + \bar{G}KC) \delta + \frac{1}{2} \bar{G}^2 K \eta q T + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} \mathbb{E} \left[ h_t(\mathbf{x}) \mid \sigma(\mathcal{H}_{t-1}) \right] \right] \\
&\leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + 2(\bar{F} + \bar{G}KC) \delta + \frac{1}{2} \bar{G}^2 K \eta q T + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) \right]
\end{aligned}$$

where the second last inequality follows from  $c_{\tau_A} q \leq cq = T$ .

Finally, we complete the proof by using the definition  $f_t(\mathbf{x}_t) = h_t(\mathbf{x}_t) - \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)$  and following the same argument as in Eq. (21) for the stochastic regime.

### Ergodic.

Consider some  $\kappa \geq \log(T)$ . Given the input distribution sequence  $\mathcal{P}_{1:T}$ , denote  $\mathcal{P}_{(t+\kappa)[t-1]}$  as the conditional distribution of  $(f_{t+\kappa}, \mathbf{g}_{t+\kappa})$  conditioned on the  $\{(f_\tau, \mathbf{g}_\tau)\}_{\tau \in [t]}$ . Then, in the ergodic regime, there exists a stationary distribution  $\tilde{\mathcal{P}} \in \Delta(\mathcal{S})$  and absolute constant  $R > 0$  such that

$$\sup_{\{(f_t, \mathbf{g}_t)\}_{t \in [T]} \in \mathcal{S}^T} \sup_{t \in [T-\kappa]} \|\mathcal{P}_{(t+\kappa)[t-1]} - \tilde{\mathcal{P}}\|_{TV} \leq \delta := R \exp(-\kappa) \quad (32)$$



By defining  $\tilde{F}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \tilde{\mathcal{P}}}[f(\mathbf{x})]$ ,  $\tilde{\mathbf{G}}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \tilde{\mathcal{P}}}[\mathbf{g}(\mathbf{x})]$ ,  $\hat{F}_{t+\kappa}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_{(t+\kappa)|[t-1]}}[f(\mathbf{x})]$ ,  $\hat{\mathbf{G}}_{t+\kappa}(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_{(t+\kappa)|[t-1]}}[\mathbf{g}(\mathbf{x})]$ ,  $F_t(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_t}[f(\mathbf{x})]$  and  $\mathbf{G}_t(\mathbf{x}) = \mathbb{E}_{(f,g) \sim \mathcal{P}_t}[\mathbf{g}(\mathbf{x})]$  for all  $t \in [T]$  and any  $\mathbf{x} \in \mathcal{X}$ , we know that for any  $\boldsymbol{\lambda} \in [\mathbf{0}, C\mathbf{e}]$ , it follows that

$$\begin{aligned}
& \text{OPT}(\mathcal{P}_{1:T}) \\
& \leq \max_{\mathbf{x}_{1:T} \in \mathcal{X}^T} \mathbb{E} \left[ \sum_{t \in [T]} (F_t(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}_t(\mathbf{x}_t)) \right] \\
& = \max_{\mathbf{x}_{1:\kappa} \in \mathcal{X}^\kappa} \mathbb{E} \left[ \sum_{t \in [\kappa]} (F_t(\mathbf{x}_t) + \boldsymbol{\lambda}^\top \mathbf{G}_t(\mathbf{x}_t)) \right] + \max_{\mathbf{x}_{\kappa+1:T} \in \mathcal{X}^{T-\kappa}} \mathbb{E} \left[ \sum_{t=1}^{T-\kappa} (\hat{F}_{t+\kappa}(\mathbf{x}_{t+\kappa}) + \boldsymbol{\lambda}^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x}_{t+\kappa})) \right] \quad (33) \\
& \leq (\bar{F} + \bar{G}KC)\kappa + \max_{\mathbf{x}_{\kappa+1:T} \in \mathcal{X}^{T-\kappa}} \sum_{t=1}^{T-\kappa} (\tilde{F}(\mathbf{x}_{t+\kappa}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x}_{t+\kappa})) + (\bar{F} + \bar{G}KC) \cdot (T - \kappa)\delta \\
& \leq T \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) + (\bar{F} + \bar{G}KC)\kappa + (\bar{F} + \bar{G}KC) \cdot T\delta
\end{aligned}$$

By choosing  $\boldsymbol{\lambda} = \frac{1}{\tau_A} \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t$ , we further have

$$\begin{aligned}
& \text{OPT}(\mathcal{P}_{1:T}) \\
& = \mathbb{E} \left[ \frac{T - \tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) + \frac{\tau_A}{T} \text{OPT}(\mathcal{P}_{1:T}) \right] \\
& \leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \tau_A \cdot \max_{\mathbf{x} \in \mathcal{X}} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}^\top \tilde{\mathbf{G}}(\mathbf{x})) \right] + (\bar{F} + \bar{G}KC)\kappa + (\bar{F} + \bar{G}KC) \cdot T\delta \\
& = \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (\tilde{F}(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \tilde{\mathbf{G}}(\mathbf{x})) \right] + (\bar{F} + \bar{G}KC)\kappa + (\bar{F} + \bar{G}KC) \cdot T\delta \\
& \leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (\hat{F}_{t+\kappa}(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x})) \right] + (\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& = \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (\hat{F}_{t+\kappa}(\mathbf{x}) + \boldsymbol{\lambda}_{t+\kappa}^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x}) + (\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t+\kappa})^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x})) \right] \\
& \quad + (\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& \stackrel{(a)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} (\hat{F}_{t+\kappa}(\mathbf{x}) + \boldsymbol{\lambda}_{t+\kappa}^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x})) \right] + \kappa\eta TK\bar{G}^2 \quad (34) \\
& \quad + (\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& \leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A - \kappa]} (\hat{F}_{t+\kappa}(\mathbf{x}) + \boldsymbol{\lambda}_{t+\kappa}^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x})) \right] + \kappa\eta TK\bar{G}^2 \\
& \quad + 2(\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& \stackrel{(b)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t=\kappa+1}^{\tau_A} h_t(\mathbf{x}) \right] + \kappa\eta TK\bar{G}^2 + 2(\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& \leq \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) \right] + \kappa\eta TK\bar{G}^2 + 2(\bar{F} + \bar{G}KC)\kappa + 2(\bar{F} + \bar{G}KC) \cdot T\delta \\
& \stackrel{(c)}{\leq} \mathbb{E} \left[ (T - \tau_A) \bar{F} + \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) \right] + \kappa\eta TK\bar{G}^2 + 2(\bar{F} + \bar{G}KC)\kappa + 2R(\bar{F} + \bar{G}KC),
\end{aligned}$$

where in (a), from (8) in Algorithm 1, we know that  $\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\|_1 \leq \eta\bar{G}K$ , which further implies  $\|\boldsymbol{\lambda}_{t+\kappa} - \boldsymbol{\lambda}_t\|_1 \leq \kappa\eta\bar{G}K$  and thus

$$(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t+\kappa})^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x}_t) \leq \kappa\eta K\bar{G}^2 \quad (35)$$

In (b), we used the fact that for any  $t \geq \kappa + 1$ , we have

$$\begin{aligned}
& \mathbb{E} \left[ \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A - \kappa]} (\hat{F}_{t+\kappa}(\mathbf{x}) + \boldsymbol{\lambda}_{t+\kappa}^\top \hat{\mathbf{G}}_{t+\kappa}(\mathbf{x}_t)) \right] \\
&= \mathbb{E} \left[ \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A - \kappa]} \mathbb{E} [h_{t+\kappa}(\mathbf{x}) \mid (f_\tau, \mathbf{g}_\tau)_{\tau \in [t-1]}] \right] \\
&\leq \mathbb{E} \left[ \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A - \kappa]} h_{t+\kappa}(\mathbf{x}) \right]
\end{aligned} \tag{36}$$

In (c) we used the fact that  $\kappa \geq \log(T)$ , so  $\delta = R \exp(-\kappa) \geq R$ .

Finally, we complete the proof by using the definition  $f_t(\mathbf{x}_t) = h_t(\mathbf{x}_t) - \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)$  and following the same argument as in Eq. (21) for the stochastic regime.

### A.8. Proof of Theorem 4.2

We bound the regret in every world as followed

$$\begin{aligned}
\mathcal{R}_T &= \text{OPT}(\mathcal{P}_{1:T}) - \sum_{t \in [T]} \mathbb{E} [f_t(\mathbf{x}_t)] \\
&\stackrel{(a)}{\leq} \mathbb{E} \left[ \bar{F}(T - \tau_A) + \sum_{t \in [\tau_A]} \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t) + \mathcal{R}_{\text{BOCO}}(\tau_A) \right] \\
&\stackrel{(b)}{\leq} \mathbb{E} [\mathcal{R}_{\text{BOCO}}(\tau_A)]
\end{aligned}$$

where (a) follows from Lemma 4.3, and (b) follows from Lemma 4.5. Recall  $\mathcal{R}_{\text{BOCO}}(\tau_A)$  is specified in Lemma 4.3 for each world.

In the following we bound  $\mathcal{R}_{\text{BOCO}}(\tau_A)$  for each world.

#### Stochastic.

$$\mathbb{E} [\mathcal{R}_{\text{BOCO}}(\tau_A)] = \mathbb{E} \left[ \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x}) - h_t(\mathbf{x}_t) \right] \stackrel{(a)}{\leq} \mathcal{O} \left( \frac{\rho T}{\beta} + \frac{1}{\gamma_i} + \frac{\gamma_i K T}{\beta^2 \rho^2} + T\epsilon + \frac{1}{\epsilon} \right) \stackrel{(b)}{=} \mathcal{O} \left( T^{\frac{3}{4}} \right) \tag{37}$$

where (a) follows from Lemma 4.6 by taking the comparator sequence  $\mathbf{y}_t = \arg \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x})$  for all  $t \in [\tau_A]$  such that  $P(\mathbf{y}_{1:T}) = 1$ , as well as any primal ascent expert  $i \in [N]$ ; (b) follows from taking  $\eta = \frac{1}{\sqrt{KT}}$ ,  $\rho = K^{\frac{1}{3}} T^{-\frac{1}{4}}$ ,  $\epsilon = T^{-\frac{1}{2}}$ ,  $\beta = \frac{1}{\log(T)}$ , and finally choosing  $\gamma_i = K^{-\frac{1}{6}} (1 + DT)^{\frac{1}{2}} T^{-\frac{3}{4}}$ . Recall all primal ascent expert stepsizes are  $\{\gamma_1 \dots \gamma_N\} = \{2^{-i} K^{-\frac{1}{6}} (1 + DT)^{\frac{1}{2}} T^{-\frac{3}{4}} : i = 0 \dots N\}$ .

**$\delta$ -corrupted, Periodic, and Ergodic.** The proof is nearly identical with that of the stochastic world in Eq. (37) given that we still consider the comparator sequence  $\mathbf{y}_t = \arg \max_{\mathbf{x} \in \mathcal{X}} \sum_{t \in [\tau_A]} h_t(\mathbf{x})$  for all  $t \in [\tau_A]$  such that  $P(\mathbf{y}_{1:T}) = 1$ . Hence we will omit the proof.

**Adversarial.** Recall the definition  $\tilde{\mathbf{y}}_t = \arg \max_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)$ . Then we have

$$\begin{aligned}
\mathbb{E} [\mathcal{R}_{\text{BOCO}}(\tau_A)] &= \left( 1 - \frac{1}{\xi} \right) \text{OPT}(\mathcal{P}_{1:T}) + \sum_{t \in [\tau_A]} \mathbb{E} [h_t(\tilde{\mathbf{y}}_t) - h_t(\mathbf{x}_t)] \\
&\leq \mathcal{O} \left( \frac{\rho T}{\beta} + \frac{1 + P(\tilde{\mathbf{y}}_{1:T})}{\gamma_i} + \frac{\gamma_i K T}{\beta^2 \rho^2} + T\epsilon + \frac{1}{\epsilon} \right) = \left( 1 - \frac{1}{\xi} \right) \text{OPT}(\mathcal{P}_{1:T}) + o(T)
\end{aligned} \tag{38}$$

where we chose the primal ascent stepsize  $\gamma_i$  s.t.

$$\frac{1}{2} K^{-\frac{1}{6}} (1 + P(\tilde{\mathbf{y}}_{1:T}))^{\frac{1}{2}} T^{-\frac{3}{4}} \leq \gamma_i \leq K^{-\frac{1}{6}} (1 + P(\tilde{\mathbf{y}}_{1:T}))^{\frac{1}{2}} T^{-\frac{3}{4}} \tag{39}$$

We note that such a  $\gamma_i$  must exist because  $P(\tilde{\mathbf{y}}_{1:T}) \leq DT$  given all  $\tilde{\mathbf{y}}_t \in \mathcal{X}$ , so that the largest element in the primal ascent stepsize set, namely  $K^{-\frac{1}{6}}(1+DT)^{\frac{1}{2}}T^{-\frac{3}{4}}$  is larger than the upper bound above, namely  $K^{-\frac{1}{6}}(1+P(\tilde{\mathbf{y}}_{1:T}))^{\frac{1}{2}}T^{-\frac{3}{4}}$ .

□

### A.9. Proof for Lemma A.1

Recall the definition  $h_t(\mathbf{x}) = f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x})$  in Eq. (4). Then we have

$$\begin{aligned} \left| h_t(\mathbf{x}) - h_t(\mathbf{x}') \right| &\leq \left| f_t(\mathbf{x}) - f_t(\mathbf{x}') \right| + \|\boldsymbol{\lambda}_t\| \cdot \|\mathbf{g}_t(\mathbf{x}) - \mathbf{g}_t(\mathbf{x}')\| \\ &\stackrel{(a)}{\leq} L\|\mathbf{x} - \mathbf{x}'\| + KCL\|\mathbf{x} - \mathbf{x}'\| = (1+KC)L \cdot \|\mathbf{x} - \mathbf{x}'\| \end{aligned} \quad (40)$$

where (a) follows from the fact that any  $(f, \mathbf{g}) \in \mathcal{S}$  are  $L$ -lipschitz under Assumption 2.1.

On the other hand, recall the definition  $\hat{h}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})}[\mathcal{L}_t(\mathbf{x} + \rho\mathbf{v}, \boldsymbol{\lambda}_t)]$  in Eq. (10). Then we have

$$\left| h_t(\mathbf{x}) - \hat{h}_t(\mathbf{x}) \right| = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})} \left[ h_t(\mathbf{x}) - h_t(\mathbf{x} + \rho\mathbf{v}) \right] \leq (1+KC)L\rho \cdot \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})}[\|\mathbf{v}\|] = (1+KC)L\rho \quad (41)$$

where the inequality follows from the first part of this lemma. □

### A.10. Proof of Lemma A.2

Recall the definitions  $h_t(\mathbf{x}) = f_t(\mathbf{x}) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x})$  in Eq. (4), and  $\hat{h}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})}[\mathcal{L}_t(\mathbf{x} + \rho\mathbf{v}, \boldsymbol{\lambda}_t)]$  in Eq. (10). Then, we have

$$\begin{aligned} \hat{h}_t(\mathbf{y}) - \hat{h}_t(\tilde{\mathbf{x}}_t) &\stackrel{(a)}{\leq} \langle \nabla \hat{h}_t(\tilde{\mathbf{x}}_t), \mathbf{y} - \tilde{\mathbf{x}}_t \rangle \\ &\stackrel{(b)}{=} \left\langle \frac{d}{\rho} \cdot \mathbb{E}_{\mathbf{u} \sim U(\mathbb{S})} [h_t(\tilde{\mathbf{x}}_t + \rho\mathbf{u}) \cdot \mathbf{u}] , \mathbf{y} - \tilde{\mathbf{x}}_t \right\rangle \\ &= \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} \left[ \left\langle \frac{d}{\rho} \cdot h_t(\tilde{\mathbf{x}}_t + \rho\mathbf{u}_t) \cdot \mathbf{u}_t , \mathbf{y} - \tilde{\mathbf{x}}_t \right\rangle \right] \\ &\stackrel{(c)}{=} \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} [\langle \nabla_t, \mathbf{y} - \tilde{\mathbf{x}}_t \rangle] \\ &\stackrel{(d)}{=} \mathbb{E}_{\mathbf{u}_t \sim U(\mathbb{S})} [\ell_t(\tilde{\mathbf{x}}_t) - \ell_t(\mathbf{y})] \end{aligned} \quad (42)$$

where (a) follows from concavity of  $\hat{h}_t(\cdot)$ ; (b) follows from Lemma B.2 by taking  $h = -h_t$ , so that in the lemma  $-\nabla_{\mathbf{x}} \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})}[h(\mathbf{x} + \rho\mathbf{v})] = \nabla \hat{h}_t(\mathbf{x})$  and  $-\mathbb{E}_{\mathbf{u} \sim U(\mathbb{S})}[h(\mathbf{x} + \rho\mathbf{u}) \cdot \mathbf{u}] = \mathbb{E}_{\mathbf{u} \sim U(\mathbb{S})}[h_t(\mathbf{x} + \rho\mathbf{u}) \cdot \mathbf{u}]$ ; (c) follows from the gradient estimate in Eq. (6) where

$$\nabla_t = \frac{d}{\rho} (f_t(\mathbf{x}_t) + \boldsymbol{\lambda}_t^\top \mathbf{g}_t(\mathbf{x}_t)) \cdot \mathbf{u}_t = \frac{d}{\rho} \cdot h_t(\mathbf{x}_t) \cdot \mathbf{u}_t = \frac{d}{\rho} \cdot h_t(\tilde{\mathbf{x}}_t + \rho\mathbf{u}_t) \cdot \mathbf{u}_t$$

Finally, (d) follows from the definition of surrogate loss functions in Eq. (7). □

### A.11. Proof of Lemma A.3

**Proving (i):**

Since  $\tilde{\mathbf{x}}_{t+1}^i = \Pi_{(1-\alpha)\mathcal{X}}(\tilde{\mathbf{x}}_t^i + \gamma_i \nabla_t)$  we have  $\|\mathbf{y} - \tilde{\mathbf{x}}_{t+1}^i\| \leq \|\mathbf{y} - (\tilde{\mathbf{x}}_t^i + \gamma_i \nabla_t)\|$  for any  $\mathbf{y} \in (1-\alpha)\mathcal{X}$ . Then

$$\begin{aligned} \|\mathbf{y} - \tilde{\mathbf{x}}_{t+1}^i\|^2 &\leq \|\mathbf{y} - \tilde{\mathbf{x}}_t^i\|^2 - 2\gamma_i \nabla_t^\top (\mathbf{y} - \tilde{\mathbf{x}}_t^i) + \gamma_i^2 \nabla_t^2 \\ \implies \|\tilde{\mathbf{x}}_{t+1}^i\|^2 &\leq \|\tilde{\mathbf{x}}_t^i\|^2 + 2\mathbf{y}^\top (\tilde{\mathbf{x}}_{t+1}^i - \tilde{\mathbf{x}}_t^i) - 2\gamma_i \nabla_t^\top (\mathbf{y} - \tilde{\mathbf{x}}_t^i) + \gamma_i^2 \nabla_t^2 \end{aligned}$$

Hence by taking  $\mathbf{y} = (1 - \alpha)\mathbf{y}_t \in (1 - \alpha)\mathcal{X}$  and rearranging we get

$$\begin{aligned}
& 2\gamma_i (\ell_t(\tilde{\mathbf{x}}_t^i) - \ell_t((1 - \alpha)\mathbf{y}_t)) \\
&= 2\gamma_i \nabla_t^\top ((1 - \alpha)\mathbf{y}_t - \tilde{\mathbf{x}}_t^i) \\
&\leq \|\tilde{\mathbf{x}}_t^i\|^2 - \|\tilde{\mathbf{x}}_{t+1}^i\|^2 + 2(1 - \alpha)\mathbf{y}_t^\top (\tilde{\mathbf{x}}_{t+1}^i - \tilde{\mathbf{x}}_t^i) + \gamma_i^2 \nabla_t^2
\end{aligned} \tag{43}$$

Telescoping with  $\tau = 1 \dots t$  we get

$$\begin{aligned}
& \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i) - \sum_{\tau \in [t]} \ell_\tau((1 - \alpha)\mathbf{y}_\tau) \\
&= \frac{1}{2\gamma_i} \|\tilde{\mathbf{x}}_1^i\|^2 + \frac{1 - \alpha}{\gamma_i} \sum_{t \in [T]} \mathbf{y}_\tau^\top (\tilde{\mathbf{x}}_{\tau+1}^i - \tilde{\mathbf{x}}_\tau^i) + \frac{\gamma_i}{2} \sum_{\tau \in [t]} \nabla_\tau^2 \\
&= \frac{1}{2\gamma_i} \|\tilde{\mathbf{x}}_1^i\|^2 + \frac{1 - \alpha}{\gamma_i} \left( \sum_{\tau \in [t-1]} (\mathbf{y}_\tau - \mathbf{y}_{\tau+1})^\top \tilde{\mathbf{x}}_{\tau+1}^i + \mathbf{y}_\tau^\top \tilde{\mathbf{x}}_{\tau+1}^i \right) + \frac{\gamma_i}{2} \sum_{\tau \in [t]} \nabla_\tau^2 \\
&\leq \frac{1}{2\gamma_i} \|\tilde{\mathbf{x}}_1^i\|^2 + \frac{1 - \alpha}{\gamma_i} \sum_{\tau \in [t-1]} (\|\mathbf{y}_\tau - \mathbf{y}_{\tau+1}\| \cdot \|\tilde{\mathbf{x}}_{\tau+1}^i\| + \|\mathbf{y}_\tau\| \cdot \|\tilde{\mathbf{x}}_{\tau+1}^i\|) + \frac{\gamma_i}{2} \sum_{\tau \in [t]} \nabla_\tau^2 \\
&\leq \frac{(1 - \alpha)^2 D^2}{2\gamma_i} + \frac{(1 - \alpha)^2 D}{\gamma_i} (P(\mathbf{y}_{1:T}) + D) + \frac{\gamma_i d^2}{2\rho^2} (\bar{F} + KCG)^2 t
\end{aligned} \tag{44}$$

□

**Proving (ii):**

First, we have for any  $t \in [T]$ ,  $i \in [N]$

$$|\ell_t(\tilde{\mathbf{x}}_t^i)| = |\nabla_t^\top (\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_t^i)| \leq \|\nabla_t\| \cdot \|\tilde{\mathbf{x}}_t^i - \tilde{\mathbf{x}}_t\| \leq \frac{d}{\rho} (\bar{F} + KCG) \cdot (1 - \alpha)D \tag{45}$$

where we recall  $D = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$  is the diameter of  $\mathcal{X}$ , and both  $\tilde{\mathbf{x}}_t^i, \tilde{\mathbf{x}}_t \in (1 - \alpha)\mathcal{X}$ .

Define  $W_t = \sum_{i \in [N]} w_{i,t}$  for all  $t \in [T]$ , then

$$\begin{aligned}
\log \left( \frac{W_{t+1}}{W_t} \right) &= \log \left( \sum_{i \in [N]} \frac{w_{i,t} \exp(-\epsilon \ell_t(\tilde{\mathbf{x}}_t^i))}{W_t} \right) \\
&= \log \left( \mathbb{E}_{I_t \sim \mathbf{w}_t / W_t} [\exp(-\epsilon \ell_t(\tilde{\mathbf{x}}_t^{I_t}))] \right) \\
&\stackrel{(a)}{\leq} -\epsilon \mathbb{E}_{I_t \sim \mathbf{w}_t / W_t} [\ell_t(\tilde{\mathbf{x}}_t^{I_t})] + \frac{\epsilon^2}{8} \\
&\stackrel{(b)}{=} -\epsilon \ell_t \left( \mathbb{E}_{I_t \sim \mathbf{w}_t / W_t} [\tilde{\mathbf{x}}_t^{I_t}] \right) + \frac{\epsilon^2}{8} \\
&\stackrel{(c)}{=} -\epsilon \ell_t(\tilde{\mathbf{x}}_t) + \frac{\epsilon^2}{8}
\end{aligned} \tag{46}$$

Here (a) follows from Hoeffding's Lemma as described in Lemma B.1 where we take  $X = \ell_t(\tilde{\mathbf{x}}_t^{I_t})$ ,  $a =$  and  $b =$ ;

(b) follows from the definition that  $\ell_t(\tilde{\mathbf{x}}) = \nabla_t^\top (\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_t)$  is a linear function in  $\tilde{\mathbf{x}}$ ; (c) follows from Eq.(5).

Hence, telescoping the above we get

$$\log \left( \frac{W_{t+1}}{W_1} \right) \leq -\epsilon \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau) + \frac{t\epsilon^2}{8} \tag{47}$$

On the other hand, we have

$$\begin{aligned}
\log\left(\frac{W_{t+1}}{W_1}\right) &= \log(W_{t+1}) - \log(W_1) \\
&\geq \log(\max_{i \in [N]} w_{i,t}) - \log(N) \\
&= \max_{i \in [N]} \log(w_{i,t}) - \log(N) \\
&\stackrel{(a)}{=} \max_{i \in [N]} \log\left(w_{i,1} \exp\left(-\epsilon \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i)\right)\right) - \log(N) \\
&= -\epsilon \min_{i \in [N]} \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i) - \log(N)
\end{aligned} \tag{48}$$

Hence, combining Eqs.(47) and (48), and dividing both sides by  $\epsilon > 0$  we get

$$\begin{aligned}
-\sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau) + \frac{t\epsilon}{8} &\geq -\min_{i \in [N]} \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i) - \frac{\log(N)}{\epsilon} \\
\Rightarrow \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau) - \min_{i \in [N]} \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i) &\leq \frac{t\epsilon}{8} + \frac{\log(N)}{\epsilon} \\
\Rightarrow \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau) - \sum_{\tau \in [t]} \ell_\tau(\tilde{\mathbf{x}}_\tau^i) &\leq \frac{t\epsilon}{8} + \frac{\log(N)}{\epsilon}, \quad \forall i \in [N]
\end{aligned} \tag{49}$$

□

## Appendix B: Supplementary lemmas

**Lemma B.1 (Hoeffding's lemma)** *Let  $X$  be some random variable such that  $a \leq X \leq b$  almost surely for some  $a, b \in \mathbb{R}$ . Then for any  $\epsilon \in \mathbb{R}$ , we have  $\mathbb{E}[\exp(-\epsilon X)] \leq \exp\left(-\epsilon \mathbb{E}[X] + \frac{\epsilon^2(b-a)^2}{8}\right)$ .*

**Lemma B.2 ([26] Lemma 2.1)** *Let  $h : \mathcal{X} \rightarrow \mathbb{R}$  be some convex function (not necessarily differentiable). Then for any  $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$  and  $\delta > 0$  we have*

$$\nabla_{\mathbf{x}} \mathbb{E}_{\mathbf{v} \sim U(\mathbb{B})} [h(\mathbf{x} + \delta \mathbf{v})] = \frac{d}{\delta} \cdot \mathbb{E}_{\mathbf{u} \sim U(\mathbb{S})} [h(\mathbf{x} + \delta \mathbf{u}) \cdot \mathbf{u}] \tag{50}$$