

# Online Learning for Approximately-Convex Functions with Long-term Adversarial Constraints

Dhruv Sarkar  
IIT Kharagpur, India

Samrat Mukhopadhyay  
IIT (ISM) Dhanbad, India

Abhishek Sinha  
TIFR Mumbai, India

## Abstract

We study an online learning problem with long-term budget constraints in the adversarial setting. In this problem, at each round  $t$ , the learner selects an action from a convex decision set, after which the adversary reveals a cost function  $f_t$  and a resource consumption function  $g_t$ . The cost and consumption functions are assumed to be  $\alpha$ -approximately convex — a broad class that generalizes convexity and encompasses many common non-convex optimization problems, including DR-submodular maximization, Online Vertex Cover, and Regularized Phase Retrieval. The goal is to design an online algorithm that minimizes cumulative cost over a horizon of length  $T$  while approximately satisfying a long-term budget constraint of  $B_T$ . We propose an efficient first-order online algorithm that guarantees  $O(\sqrt{T})$   $\alpha$ -regret against the optimal fixed feasible benchmark while consuming at most  $O(B_T \log T) + \tilde{O}(\sqrt{T})$  resources in both full-information and bandit feedback settings. In the bandit feedback setting, our approach yields an efficient solution for the **Adversarial Bandits with Knapsacks** problem with improved guarantees. We also prove matching lower bounds, demonstrating the tightness of our results. Finally, we characterize the class of  $\alpha$ -approximately convex functions and show that our results apply to a broad family of problems.

## 1 Introduction

Online Convex Optimization (OCO) with long-term constraints has emerged as a fundamental framework for modeling sequential decision-making while satisfying time-varying and uncertain constraints [Mannor et al., 2009, Guo et al., 2022, Immorlica et al., 2022, Sinha and Vaze, 2024]. This problem can be formulated as a sequential game played between a learner and an adversary over a horizon of length  $T$ . Specifically, a learner with a long-term resource budget  $B_T$  selects an action  $x_t$  at round  $t \in [T]$  from a convex decision set  $\mathcal{X}$ . Consequently, at the  $t^{\text{th}}$  round, the learner incurs a cost of  $f_t(x_t)$  and consumes  $g_t(x_t) \geq 0$  amount of resources. The cost and consumption functions may be non-convex and can be chosen adversarially. The goal is to minimize the cumulative cost over the horizon, *i.e.*,  $\sum_{t=1}^T f_t(x_t)$ , while approximately satisfying the given budget constraint  $\sum_{t=1}^T g_t(x_t) \leq B_T$ .

Constrained learning problems arise in a variety of practical settings, including online resource allocation, dynamic pricing [Besbes and Zeevi, 2009], online ad markets [Slivkins, 2013, Liakopoulos et al., 2019], learning with fairness and safety constraints [Sun et al., 2017, Sinha, 2023], and Bandits with Knapsacks [Immorlica et al., 2022], and online packing problems [Mehta et al., 2013, Agrawal and Devanur, 2014]. Significant progress has been made in the setting where the cost and consumption functions are convex and the static comparator satisfies per-round constraints [Sinha and Vaze, 2024, Guo et al., 2022, Mahdavi et al., 2012]. However, the *non-convex* setting with long-term budget constraints presents substantial computational and algorithmic challenges. Please refer to Section 2 for the state-of-the-art results on this problem. In this paper, we simultaneously address the challenges of the non-convexity of the cost and constraints, along with satisfying the long-term budget constraints in an adversarial setup. Our contributions are twofold:

1. **Characterizing  $\alpha$ -approximately Convex functions:** We focus on the setting where both

the cost and constraint functions belong to the class of  $\alpha$ -approximately convex functions [Pedramfar and Aggarwal, 2025]. In Section 3.1, we motivate our study by showing that the objective function for some foundational problems, such as **Online Vertex Cover**, belongs to the category. In Theorem 3, we provide several necessary and sufficient characterizations of this class using convex conjugate theory. Using this result, in Appendix A.3 and A.4, we demonstrate that **DR-submodular** functions (which arises as a continuous extension of submodular functions) and the objective function of the **Regularized Phase Retrieval** problem belongs to this class.

2. **Learning with Long-term Budget Constraints:** We propose an efficient first-order online learning algorithm, which guarantees a sublinear  $\alpha$ -regret while approximately satisfying a given long-term budget constraint of  $B_T$ , when the cost and constraint functions are  $\alpha$ -approximately convex. These results improve the state-of-the-art even in the convex setting where  $\alpha = 1$ . The regret of the proposed algorithm is measured against the best fixed offline action in hindsight that satisfies the budget constraint over the entire horizon of length  $T$ . This must be contrasted with a weaker benchmark, commonly used in the literature, which is required to remain feasible in *every* round [Sinha and Vaze, 2024, Guo et al., 2022, Yi et al., 2021]. These results are obtained by reducing the constrained learning problem to an instance of standard Online Linear Optimization (OLO) problem, and showing that the proposed algorithm achieves  $O(\sqrt{T})$   $\alpha$ -regret while consuming at most  $O(\log T)B_T + O(\sqrt{T})$  resources. We do not make any assumptions on Slater’s condition. In Section 6, we establish converse results that show that the above performance bounds are tight. Our proposed algorithms generalize almost immediately to the bandit feedback setting, described in Appendix D.

This paper is organized as follows. Section 2 reviews the related work to contextualize our contribution within the existing literature. Section 3 formally defines the problem and motivates it via the **Online Vertex Cover** problem. Section 4 introduces the class of  $\alpha$ -approximately convex functions and provides several equivalent characterizations. Our main algorithm is presented in Section 5, followed by lower bounds in Section 6. We conclude the paper in Section 7.

## 2 Related Work

### Online Learning with Per-Round Constraints:

Online learning with long-term constraints was first studied by Mannor et al. [2009]. In the context of two-player infinite-horizon stochastic games, they established a fundamental impossibility result: it is not possible to simultaneously achieve sublinear bounds for both regret as well as cumulative constraint violation (CCV) against the best fixed offline action that satisfies the long-term constraint over the entire horizon. This negative result motivated subsequent works to consider a weaker benchmark, in which the benchmark satisfies the budget constraints at *every* round [Mahdavi et al., 2012, Neely and Yu, 2017, Guo et al., 2022]. The goal in this line of work is to obtain the tightest regret and CCV bounds. For time-invariant constraints, Mahdavi et al. [2012] were the first to use Online Gradient Descent (OGD) and mirror prox-based online policies to obtain sublinear regret and sublinear CCV guarantees. Later, Castiglioni et al. [2022b] proposed a unified meta-algorithm that achieves  $O(T^{3/4})$  bounds for both approximate regret and CCV in the non-convex setting with long-term constraints. However, their results rely on Slater’s condition and the assumption that the constraint functions vary no faster than  $O(T^{-1/4})$ . Guo et al. [2022] studied the setting with adversarial constraints, without assuming Slater’s condition, and achieved  $O(\sqrt{T})$  regret and  $O(T^{3/4})$  CCV. The closest related work is Sinha and Vaze [2024], which achieves optimal  $O(\sqrt{T})$  regret and  $\tilde{O}(\sqrt{T})$  CCV guarantees in the convex setting by reducing the problem to standard online convex optimization. However, all of the above works compare against a weaker fixed offline benchmark that is required to remain feasible at *every* round. In addition to being too restrictive, it is possible that there does not exist any fixed action that satisfies all  $T$  (potentially adversarially chosen) constraints simultaneously, leading to vacuous guarantees.

### Online Learning with Long-term Constraints:

Online learning with long-term constraints have also been considered in the literature, where the goal is to establish tight competitive ratio guarantees [Immorlica et al., 2022, Slivkins, 2013, Badanidiyuru et al., 2018, Rivera Cardoso et al., 2025]. In this problem, known as Bandits with Knapsacks (BwK), we have  $K$  arms and  $d$  resources. Corresponding to each pull of the arm, the algorithm incurs some loss and some of the resources get consumed where the total consumption of each resource is limited by a budget  $B$ . The problem is to find the best possible regret guarantee within the budget constraint. Naturally, the algorithm stops at time horizon  $T$ , or when the total consumption of some resource exceeds its budget.

Closely related to our work is the paper by Immorlica et al. [2022]. The authors studied the BwK prob-

lem in the *adversarial* setting where both the cost and consumption vectors are chosen by an oblivious adversary. The authors showed that it is impossible to achieve sublinear regret guarantees for this problem. Thus, they instead try to get meaningful guarantees on the *competitive ratio*. In this setting, they designed a primal-dual based online algorithm that achieves  $O(\log T)$  competitive ratio. However, in addition to assume existence of a *NULL* arm (defined to be an arm with zero cost and consumption at all rounds) and running two different regret minimizers (for each of the primal and dual problems), their multi-phase algorithm needs to guess the value of the optimal benchmark on an exponential scale, which further adds to its complexity. Furthermore, a major limitation of their theoretical result is that their regret bound becomes vacuous unless the budget  $B_T$  is at least  $\tilde{\Omega}(\sqrt{T})$ . On the contrary, our first-order and efficient algorithm yields near-optimal guarantees for any budget, including  $B_T = 0$ , without any further assumptions (see Appendix D).

The work of Immorlica et al. [2022] was further extended by Castiglioni et al. [2022a] where they designed a primal-dual-based algorithm for both stochastic and adversarial setting in the regime where the long-term budget scales linearly with the time-horizon, *i.e.*,  $B_T = \Omega(T)$ . They showed that, their algorithm achieves a constant competitive ratio in this regime. Stradi et al. [2025] considers the setting where the learner is equipped with a *spending plan*, which is essentially a budget allocation profile that prescribes how much of each resource may be consumed at each round. They design a primal-dual-based algorithm that achieve sublinear regret of  $\tilde{O}((\rho_{\min})^{-1} \cdot \sqrt{T})$ , where  $\rho_{\min}$  is the Slater constant. In the worst case, they design a meta-algorithm that guarantees  $\tilde{O}(T^{3/4})$  regret, even when the spending plan is highly imbalanced. Raut et al. [2021] considered the online DR-submodular maximization problem with long-term constraints where the constraints are assumed to be linear and stochastic. A common limitation of primal-dual-based algorithms, *e.g.*, the ones proposed by Castiglioni et al. [2022a] and Stradi et al. [2025], is that they assume Slater’s condition. This assumption results in bounds which depends inversely on the Slater’s constant, resulting in vacuous bounds when the Slater’s constant is arbitrarily small. In contrast, we do not make any assumptions on the allocated budget  $B_T$  or the spending plan, or assume Slater’s condition.

**Online Non-Convex Optimization:** Online non-convex optimization has garnered increasing attention in recent years. Agarwal et al. [2019] showed that online learning with adversarial losses is computation-

ally intractable in the absence of structural assumptions. Setting aside computational challenges, Sugala and Netrapalli [2020] showed that the Follow-the-Perturbed-Leader policy (FTPL) can attain sublinear regret for non-convex losses when equipped with an optimization oracle. However, their approach relies on a strong assumption: the existence of an offline oracle capable of approximately solving non-convex optimization problems. The online non-convex learning problem has also been studied under structural assumptions, such as the Polyak–Łojasiewicz condition by Mulvaney-Kemp et al. [2023] and the weak pseudo-convexity condition by Gao et al. [2018]. Closely related to our work, the paper by Pedramfar and Aggarwal [2025] introduced the notion of upper-linearizable functions and proved  $\alpha$ -regret guarantees for the same. The class of  $\alpha$ -convex functions introduced in this paper coincides with the negative of upper-linearizable functions and extends the class of convex functions. We allow both the cost and constraint functions to be  $\alpha$ -approximately convex functions, which strictly generalize convex functions and encompass a wide range of practically relevant objectives such as phase retrieval and DR-submodular optimization.

### 3 Problem Formulation

Consider the following repeated game between a learner and an adversary played for  $T$  rounds. At each round  $t$ , the learner chooses an action  $x_t$  from an admissible set  $\mathcal{X} \subseteq \mathbb{R}^d$  for some  $d \geq 1$ . The set  $\mathcal{X}$  is assumed to be non-empty, closed, and convex with a finite Euclidean diameter of  $D$ . Upon observing the action  $x_t$ , the adversary chooses two non-negative functions - a *cost* function  $f_t : \mathcal{X} \rightarrow \mathbb{R}_+$  and a *resource-consumption* function (*a.k.a.* constraint function)  $g_t : \mathcal{X} \rightarrow \mathbb{R}_+$ . The cost and constraint functions are assumed to belong to the class of  $\alpha$ -approximately convex functions, which will be defined and characterized in Section 4. The consumption  $g_t(x_t)$  denotes the amount of resources consumed on round  $t$  due to the action  $x_t$ . The allotted resource consumption budget for the entire horizon of length  $T$  is specified to be  $B_T$ . The performance of any online algorithm is characterized by comparing its cumulative cost against that of a fixed feasible action  $x^* \in \mathcal{X}$  which satisfies the long-term budget constraint  $\sum_{t=1}^T g_t(x^*) \leq B_T$ . Since we allow the cost and consumption functions to be non-convex, we use the  $\alpha$ -regret as the performance metric [Chen et al., 2018], which generalizes the notion of the static regret [Hazan, 2022]. In the  $\alpha$ -regret metric, we use a potentially weaker benchmark by scaling up its cumulative cost by a factor of  $\alpha \geq 1$ . Specifically, let  $\mathcal{X}^* \subseteq \mathcal{X}$  be the (non-empty) subset of all fixed actions

satisfying the long-term constraint, *i.e.*,

$$\mathcal{X}^* = \{x \in \mathcal{X} : \sum_{t=1}^T g_t(x) \leq B_T\}. \quad (1)$$

Assuming the feasible set to be non-empty, we define the  $\alpha$ -Regret and the Cumulative Consumption (CC) of any algorithm as follows:

$$\text{Regret}_T(\alpha) = \sup_{x^* \in \mathcal{X}^*} \sum_{t=1}^T (f_t(x_t) - \alpha f_t(x^*)), \quad (2)$$

$$\text{CC}_T = \sum_{t=1}^T g_t(x_t). \quad (3)$$

Our objective is to simultaneously upper bound the  $\alpha$ -Regret (2) and the CC (3) for a suitably chosen small value of  $\alpha$ . For  $\alpha = 1$ , the  $\alpha$ -regret reduces to the standard (static) regret.

**Remarks:** As mentioned in Section 2, previous works on the COCO problem considered *round wise feasibility* with a restricted benchmark which is required to incur zero constraint violation in every round, *i.e.*,  $g_t(x^*) = 0, \forall 1 \leq t \leq T$ . Apart from being severely restrictive, a more serious issue with this assumption is that the feasible set may be empty, *i.e.*,  $\cap_{t=1}^T \{x^* : g_t(x^*) = 0\} = \emptyset$ , resulting in vacuous bounds. In this paper, we avoid this restrictive assumption by requiring the offline benchmark to satisfy the budget constraint only over the entire horizon of length  $T$ . Note that by setting  $B_T = 0$ , and using the non-negativity of the consumption function, we recover the instantaneous feasibility condition as above. In this setting, the cumulative consumption (CC) metric is known as Cumulative Constraint Violation (CCV) [Sinha and Vaze, 2024, Guo et al., 2022].

Secondly, unlike Immorlica et al. [2022], which assumes the existence of a NULL action with zero cost and zero consumption, we only assume the feasible set  $\mathcal{X}^*$  to be non-empty, which is necessary to make the problem well-defined. Although the above formulation considers only a single resource, extension of our algorithm to multiple resources is straightforward; see Appendix C. In the following, we describe the **Online Vertex Cover** problem, which concretely illustrates various components of the above problem.

### 3.1 A Motivating Example: The Online Vertex Cover Problem

Consider a sequence of graphs defined on a fixed set of  $n$  vertices  $V$ , with time-varying vertex prices  $\{c_t\}_{t \geq 1}$  and time-varying edges  $\{E_t\}_{t \geq 1}$ . A learner and an adversary play the following repeated game on this sequence of graphs. At each round, the learner selects

a subset of vertices, and, at the same time, the vertex prices and the current edges of the graph are chosen by the adversary. The goal of the learner is to select a subset of vertices on each round to maximize the total number of edges covered over a horizon of length  $T$  with a given long-term budget constraint  $B_T$ . Specifically, in every round  $t \geq 1$ , assume that the learner randomly selects a subset of vertices, denoted by the indicator variables  $\mathbf{X}_t \in \{0, 1\}^n$ . Simultaneously, the adversary reveals the current set of edges  $E_t$  and the current prices for the vertices  $c_t : V \rightarrow \mathbb{R}_+^n$ . Hence, on round  $t$ , the learner pays an expected price of

$$C_t = \mathbb{E} \sum_{i \in V} c_{t,i} X_{t,i}, \quad (4)$$

and receives an expected reward equal to the expected number of edges covered, *i.e.*,

$$\begin{aligned} R_t &= \mathbb{E} \sum_{(i,j) \in E_t} \max(X_{t,i}, X_{t,j}) \\ &= \sum_{(i,j) \in E_t} \mathbb{P}(X_{t,i} = 1 \vee X_{t,j} = 1), \end{aligned} \quad (5)$$

where we note that an edge is covered if either of its end-points are selected by the learner. In the above, the expectations are taken with respect to the randomness of the policy. We emphasize the fact that, unlike the classical Minimum Vertex Cover problem, in the online version, the learner selects the vertices on round  $t$  *without* observing the current edges  $E_t$  or the current prices  $c_t$ . Since the classical offline variant of the vertex cover problem, where the graph is revealed *a priori*, is well-known to be **NP-hard** [Garey and Johnson, 2002], we instead seek approximate solutions in the online setting.

Towards this end, we consider a class of randomized policies where, on round  $t$ , vertex  $i$  is independently selected with probability  $x_{t,i}, i \in V$ . Thus the decision set  $\mathcal{X}$  is given by the hypercube  $[0, 1]^n$ . The goal is to design a sequence of inclusion probability vectors  $\{x_t\}_{t \geq 1}$  to maximize the cumulative rewards subject to the long-term budget constraints.

For any randomized policy, the probability that an edge  $(i, j)$  is covered on round  $t$  is given by:

$$\begin{aligned} \mathbb{P}(X_{t,i} = 1 \vee X_{t,j} = 1) &= 1 - \mathbb{P}(X_{t,i} = 0 \wedge X_{t,j} = 0) \\ &= 1 - (1 - x_{t,i})(1 - x_{t,j}) \\ &= x_{t,i} + x_{t,j} - x_{t,i}x_{t,j} \\ &\geq \frac{1}{2}(x_{t,i} + x_{t,j}), \end{aligned} \quad (6)$$

where in the last inequality, we have used the fact that  $\frac{1}{2}(x_{t,i} + x_{t,j}) \stackrel{(a)}{\geq} \frac{1}{2}(x_{t,i}^2 + x_{t,j}^2) \stackrel{(\text{AM-GM})}{\geq} x_{t,i}x_{t,j}$ , where the



equality (a) holds because  $0 \leq x_{t,i} \leq 1, \forall t, i$ . Furthermore, using the union bound, we have

$$\begin{aligned} \mathbb{P}(X_{t,i} = 1 \vee X_{t,j} = 1) &\leq \mathbb{P}(X_{t,i} = 1) + \mathbb{P}(X_{t,j} = 1) \\ &= x_{t,i} + x_{t,j}. \end{aligned} \quad (7)$$

Clearly, the cost (4) and the reward (5) are functions of the inclusion probability vector  $x_t$ . Using the linearity of expectation, while the cost  $C_t = \sum_i c_{t,i} x_{t,i}$  is linear, the reward function  $R_t(x_t) = \sum_{(i,j) \in E_t} \mathbb{P}(X_{t,i} = 1 \vee X_{t,j} = 1) = \sum_{(i,j) \in E_t} (x_{t,i} + x_{t,j} - x_{t,i} x_{t,j})$  is non-linear and non-concave in the decision variable  $x_t$ . Nevertheless, from Eqns. (6) and (7), it follows that the function  $R_t(x_t)$  satisfies the following inequality for any  $x_t, u_t \in \mathcal{X}$ , which generalizes the first-order condition for concavity:

$$\begin{aligned} R_t(x_t) - \frac{1}{2} R_t(u_t) &\geq \\ \frac{1}{2} \sum_{(i,j) \in E_t} \{ (x_{t,i} + x_{t,j}) - (u_{t,i} + u_{t,j}) \} & \\ = \langle \text{Deg}_t, x_t - u_t \rangle, & \end{aligned} \quad (8)$$

where  $\text{Deg}_{t,i}$  denotes the degree of vertex  $i$  on round  $t$ . Inequality (8) motivates our definition of the class of approximately convex (equivalently, concave) functions given in the following section.

## 4 The Class of $\alpha$ -Approximately Convex Functions

**Definition 1.** *The class of  $\alpha$ -approximately convex functions ( $\alpha \geq 1$ ), denoted by  $\mathcal{L}_\alpha$ , is defined to be the family of non-negative real-valued functions defined on a convex domain  $\mathcal{X}$  such that for any point  $x \in \mathcal{X}$ , there exists a vector  $H(x)$ , called a generalized sub-gradient at  $x$ , so that the following inequality, which we call  $\alpha$ -APPROXIMATE CONVEXITY, holds uniformly for any  $x, u \in \mathcal{X}$ :*

$$f(x) \leq \alpha f(u) + \langle H(x), x - u \rangle, \quad \forall f \in \mathcal{L}_\alpha. \quad (9)$$

We analogously define  $\alpha$ -approximately concave functions with  $0 \leq \alpha \leq 1$ , where the direction of the inequality (9) is reversed (see, e.g., Eqn. (8)). Clearly, with  $\alpha = 1$ , the class  $\mathcal{L}_\alpha$  includes the class of all non-negative convex functions where  $H(x)$  can be taken to be a sub-gradient at  $x$ . Under standard assumptions, we can also bound the norm of the generalized sub-gradients (please refer to Lemma 8 in the Appendix).

As we show in Appendix A.3, the class  $\mathcal{L}_\alpha$  appears in several common non-convex optimization problems, including weakly DR-submodular maximization, regularized phase retrieval, and online vertex cover. The class of  $\alpha$ -approximately convex functions was first

introduced in an equivalent form by Pedramfar and Aggarwal [2025], who called it the class of *upper-linearizable* functions.

**Remarks:** It is interesting to note that, in sharp contrast with convex functions, even if an  $\alpha$ -approximately convex (concave) function is differentiable, its gradient *need not* correspond to a generalized sub-gradient. For example, in the online vertex cover example in Section 3.1, the  $t^{\text{th}}$  reward function  $R_t(x) \equiv \sum_{(i,j) \in E_t} (x_i + x_j - x_i x_j)$  is differentiable and  $1/2$ -approximately concave. Yet its gradient does not correspond to a generalized sub-gradient.

The following is an immediate consequence of Definition (9).

**Proposition 2.** *The class  $\mathcal{L}_\alpha$  is closed under non-negative linear combinations.*

See Appendix A.1 for the proof.

Recall that the Fenchel conjugate  $f^* : \mathbb{R}^n \rightarrow \mathbb{R}$  of a function  $f : \mathcal{X} \mapsto \mathbb{R}$  is defined as

$$f^*(y) = \sup_{x \in \mathcal{X}} (\langle y, x \rangle - f(x)).$$

Being a pointwise supremum of a family of affine functions, the function  $f^*$  is convex [Boyd and Vandenberghe, 2004]. The biconjugate of  $f$  is defined to be the Fenchel conjugate of the function  $f^*$ . The following theorem gives equivalent characterizations for the class of  $\alpha$ -approximately convex functions.

**Theorem 3.** *Let  $f : \mathcal{X} \rightarrow \mathbb{R}_+$  be a non-negative function and  $\alpha \geq 1$ . Then the following statements are equivalent:*

1.  $f$  is  $\alpha$ -approximately convex.
2. The biconjugate of the function  $f$  satisfies,  $f(x) \leq \alpha f^{**}(x)$ ,  $\forall x \in \mathcal{X}$ . Since for any function  $f^{**}(x) \leq f(x)$ , the function  $f$  is sandwiched between  $f^*$  and  $\alpha f^*$  pointwise, i.e.,

$$f^{**}(x) \leq f(x) \leq \alpha f^{**}(x), \quad \forall x \in \mathcal{X}.$$

3. There exists a non-negative convex function  $g : \mathcal{X} \rightarrow \mathbb{R}_+$  such that  $g(x) \leq f(x) \leq \alpha g(x)$  for all  $x \in \mathcal{X}$ .
4. (APPROXIMATE JENSEN'S INEQUALITY) For any set of  $N$  points  $\{x_i\}_{i=1}^N$ , all from the set  $\mathcal{X}$ , and any probability distribution  $p$  on these  $N$  points, the following approximate version of the Jensen's inequality holds:

$$f\left(\sum_i p_i x_i\right) \leq \alpha \sum_i p_i f(x_i).$$

The proof of Theorem 3 is given in Appendix A.2. Theorem 3 is useful for establishing  $\alpha$ -approximate

convexity for many useful non-convex functions. See Appendix A.4 for an example involving the phase retrieval problem.

## 5 Online Learning with Budget Constraints

In this Section, we propose an online policy for the constrained learning problem introduced in Section 3 with  $\alpha$ -approximately convex cost and constraint functions (the case of  $\alpha$ -approximately concave functions can be treated similarly). As stated earlier, we benchmark our online policy against the best fixed action in hindsight satisfying the long-term budget constraint (Eqn. (1)).

**The Regret Decomposition Inequality:** Let  $Q(t)$  be the amount of resources consumed up to round  $t$ , *i.e.*,

$$Q(t) = Q(t-1) + g_t(x_t), \quad Q(0) = 0. \quad (10)$$

Let  $\Phi(\cdot)$  be a non-decreasing and convex Lyapunov function. The increase of the value of the Lyapunov function from round  $t-1$  to  $t$  can be upper bounded as follows:

$$\begin{aligned} \Phi(Q(t)) - \Phi(Q(t-1)) &\stackrel{(a)}{\leq} \Phi'(Q(t))(Q(t) - Q(t-1)) \\ &\stackrel{(b)}{=} \Phi'(Q(t))g_t(x_t), \end{aligned}$$

where in step (a) we have used the convexity of the function  $\Phi(\cdot)$  and in (b), we have used Eqn. (10). Let  $x^* \in \mathcal{X}^*$  be any fixed action from the feasible set (1). Adding the term  $V(f_t(x_t) - \alpha f_t(x^*))$  to both sides of the above inequality, we obtain:

$$\begin{aligned} &\Phi(Q(t)) - \Phi(Q(t-1)) + V(f_t(x_t) - \alpha f_t(x^*)) \\ &\leq (Vf_t(x_t) + \Phi'(Q(t))g_t(x_t)) - \alpha(Vf_t(x^*) + \Phi'(Q(t))g_t(x^*)) + \alpha\Phi'(Q(t))g_t(x^*), \end{aligned} \quad (11)$$

where we have added and subtracted the term  $\alpha\Phi'(Q(t))g_t(x^*)$ . Define the surrogate cost function  $\hat{f}_t : \mathcal{X} \mapsto \mathbb{R}_+$  for round  $t$  as follows:

$$\hat{f}_t = Vf_t + \Phi'(Q(t))g_t, \quad t \geq 1, \quad (12)$$

Since both  $f_t$  and  $g_t$  are  $\alpha$ -approximately convex and the Lyapunov function  $\Phi(\cdot)$  is non-decreasing, from Proposition 2, it follows that the surrogate cost function  $\hat{f}_t$  is also  $\alpha$ -approximately convex. Summing up the inequalities (11) for  $1 \leq t \leq T$ , we obtain the following Regret Decomposition inequality

$$\begin{aligned} &\Phi(Q(T)) - \Phi(Q(0)) + V\text{Regret}_T(\alpha) \\ &\stackrel{(a)}{\leq} \text{Regret}'_T(\alpha) + \alpha\Phi'(Q(T)) \sum_{t=1}^T g_t(x^*), \end{aligned}$$

$$\stackrel{(b)}{\leq} \text{Regret}'_T(\alpha) + \alpha\Phi'(Q(T))B_T, \quad (13)$$

where  $\text{Regret}_T(\alpha)$  and  $\text{Regret}'_T(\alpha)$  respectively denote the  $\alpha$ -regrets for learning the original cost functions  $\{f_t\}_{t \geq 1}$  and the surrogate cost functions  $\{\hat{f}_t\}_{t \geq 1}$  w.r.t. the feasible action  $x^*$  (see Eqn. (2) for the definition of  $\alpha$ -regret). In step (a) above, we have used the monotonicity of the sequence  $\{Q(t)\}_{t \geq 1}$  and the convexity of the Lyapunov function  $\Phi(\cdot)$ , and in step (b), we have used the fact that the offline benchmark  $x^*$  satisfies the long-term budget constraint of  $B_T$ .

### 5.1 Algorithm Design and Analysis

As mentioned above, the surrogate cost function  $\hat{f}_t$  (12), is non-negative and  $\alpha$ -approximately convex. Let  $H_{f_t}(x)$  and  $H_{g_t}(x)$  be generalized subgradients at  $x$  for  $f_t$  and  $g_t$  respectively. Then, as in the proof of Proposition 2, the vector  $H_{\hat{f}_t}(x)$  defined as

$$H_{\hat{f}_t}(x) \equiv VH_{f_t}(x) + \Phi'(Q(t))H_{g_t}(x) \quad (14)$$

is a generalized subgradient for the surrogate cost function  $\hat{f}_t$ , *i.e.*, we have

$$\hat{f}_t(x_t) - \alpha\hat{f}_t(x^*) \leq \langle H_{\hat{f}_t}(x_t), x_t \rangle - \langle H_{\hat{f}_t}(x_t), x^* \rangle. \quad (15)$$

Summing up inequalities (15) for  $1 \leq t \leq T$ , we conclude that the  $\alpha$ -regret for the surrogate costs is upper bounded as:

$$\text{Regret}'_T(\alpha) \leq \text{Regret}''_T, \quad (16)$$

where  $\text{Regret}''_T$  is the standard regret ( $\alpha = 1$ ) of the surrogate Online Linear Optimization (OLO) problem where the cost function  $\tilde{f}_t : \mathcal{X} \mapsto \mathbb{R}_+$  on round  $t$  is defined to be:

$$\tilde{f}_t(x) = \langle H_{\hat{f}_t}(x_t), x \rangle, \quad 1 \leq t \leq T. \quad (17)$$

Combining Eqns. (13) and (47), we obtain the following inequality, which constitutes the key to the subsequent analysis:

$$\begin{aligned} &\Phi(Q(T)) - \Phi(Q(0)) + V\text{Regret}_T(\alpha) \\ &\leq \text{Regret}''_T + \alpha\Phi'(Q(T))B_T. \end{aligned} \quad (18)$$

Eqn. (18) suggests that in order to control both  $Q(T)$  and  $\text{Regret}_T(\alpha)$ , which appear on the LHS of (18), we can minimize the regret of the corresponding OLO problem, which appear in the upper bound in the inequality (18).

Standard online policies for the OLO problem, such as Online Gradient Descent [Hazan, 2022], require a uniform upper bound on the norms of the cost function gradients. Since the norm of the gradient of the surrogate OLO problem  $\|H_{\hat{f}_t}(x)\|$  scale with  $Q(t)$  (an

algorithm-dependent variable), it can not be upper bounded at the beginning of the game. Thus we use an adaptive learning policy, such as ADAGRAD, which does not need us to specify the scale of the gradients, yet achieves near-optimal bounds. Algorithm 1 describes our proposed online learning policy. Theorem

---

**Algorithm 1** Online policy for  $\alpha$ -approximately convex functions with constraints

---

- 1: **Inputs:** Convex decision set  $\mathcal{X}$  with a finite Euclidean diameter  $D$ , Euclidean projection operator  $\text{PROJ}_{\mathcal{X}}(\cdot)$  on the set  $\mathcal{X}$ , sequence of  $\alpha$ -approximately convex cost functions  $\{f_t\}_{t \geq 1}$ , and consumption functions  $\{g_t\}_{t \geq 1}$ , Budget  $B_T$ , Parameters  $V, \lambda$
- 2: Initialize  $x_1 \in \mathcal{X}$  arbitrarily
- 3: **for**  $t = 1 : T$  **do**
- 4:   Play  $x_t$ ; compute  $H_{f_t}(x_t)$ , and  $H_{g_t}(x_t)$ .
- 5:   Compute  $H_{\hat{f}_t}(x_t)$  as follows:

$$H_{\hat{f}_t}(x_t) \equiv V H_{f_t}(x_t) + \Phi'(Q(t)) H_{g_t}(x_t)$$

- 6:   Use the ADAGRAD step sizes:

$$\eta_t \leftarrow \frac{\sqrt{2}D}{2\sqrt{\sum_{\tau=1}^t \|H_{\hat{f}_\tau}(x_\tau)\|^2}}.$$

- 7:   Compute the next action  $x_{t+1}$  using Online Gradient Descent with step size  $\eta_t$ :

$$x_{t+1} \leftarrow \text{PROJ}_{\mathcal{X}}(x_t - \eta_t H_{\hat{f}_t}(x_t)).$$

- 8: **end for**
- 

4 constitutes the main result of this paper.

**Theorem 4.** *Consider the constrained online learning problem described in Section 3 with a sequence of  $\alpha$ -approximately convex cost and constraint functions and a long-term budget of  $B_T$ . Assume that the generalized subgradients of all cost and constraint functions are upper bounded by  $\alpha G$  for some  $G > 0$ . Then, Algorithm 1, with  $\Phi(x) = \exp(\lambda x)$ ,  $\lambda = \frac{1}{2}(\alpha G D \sqrt{2T} + \alpha B_T)^{-1}$ ,  $V = (\alpha G D)^{-1}$ , achieves near-optimal  $\alpha$ -regret while consuming close to the allocated budget. Specifically:*

$$\text{Regret}_T(\alpha) = O(\alpha \sqrt{T}), \text{CC}_T = \tilde{O}(\alpha B_T + G D \sqrt{T}).$$

The proof of Theorem 4 is given in Section 5.2.

**Remarks:** In case of bandit feedback (*a.k.a.* the Adversarial Bandits with Knapsacks (BwK) problem in the literature), we replace the full-information-based ADAGRAD sub-routine with an adaptive bandit algorithm and use a power-law Lyapunov function for

technical reasons. Due to space constraints, the details are deferred to Appendix D. Note that, the competitive ratio bound for the BwK problem, given by Immorlica et al. [2022, Theorem 5.1], becomes vacuous unless the budget is at least  $\tilde{\Omega}(\sqrt{T})$  [Immorlica et al., 2022, Remark 5.2]. On the other hand, Theorem 4 and Theorem 10 give non-trivial regret and cumulative consumption bounds for *any* arbitrary budget  $B_T \geq 0$  in the full-information and bandit feedback settings respectively. Furthermore, compared to Immorlica et al. [2022], Castiglioni et al. [2022a], which run two different regret-minimizers - one for the primal and the other for the dual, our primal-only algorithm with a single regret minimizer is computationally efficient. Finally, we do not make any assumption on the Slater condition [Castiglioni et al., 2022a] or the existence of a NULL arm [Immorlica et al., 2022].

## 5.2 Proof of Theorem 4

The norm of the gradients of the surrogate OLO cost functions (17) can be upper bounded as follows:

$$\begin{aligned} \|H_{\hat{f}_t}(x_t)\|_2 &\stackrel{(a)}{\leq} V \|H_{f_t}(x_t)\|_2 + \Phi'(Q(t)) \|H_{g_t}(x_t)\|_2 \\ &\stackrel{(b)}{\leq} \alpha G (V + \Phi'(Q(T))). \end{aligned} \quad (19)$$

where (a) follows from using the triangle inequality in Eqn. (14), and (b) follows from the assumption that the norm of the generalized sub-gradients are uniformly bounded by  $\alpha G$  for some  $G > 0$ . Using the adaptive regret bound of the ADAGRAD sub-routine, given by Theorem 9 in Appendix A.6, we have the following upper bound for the standard regret of the surrogate OLO problem:

$$\text{Regret}_T'' \leq \sqrt{2} G D \alpha (V + \Phi'(Q(T))) \sqrt{T}. \quad (20)$$

Hence Eqn. (18) yields:

$$\begin{aligned} \Phi(Q(T)) + V \text{Regret}_T(\alpha) &\leq \Phi(Q(0)) + \\ &\alpha V G D \sqrt{2T} + \alpha \Phi'(Q(T)) (G D \sqrt{2T} + B_T). \end{aligned}$$

We now choose  $\Phi(\cdot)$  to be the exponential Lyapunov function  $\Phi(x) = \exp(\lambda x)$ , where the parameter  $\lambda$  will be fixed below. With this choice for  $\Phi(\cdot)$ , we have

$$\begin{aligned} \exp(\lambda Q(T)) + V \text{Regret}_T(\alpha) &\leq 1 + \alpha V G D \sqrt{2T} + \\ &\lambda \alpha \exp(\lambda Q(T)) (G D \sqrt{2T} + B_T). \end{aligned} \quad (21)$$

We now choose the free parameters to be  $\lambda = \frac{1}{2\alpha} (G D \sqrt{2T} + B_T)^{-1}$  and  $V = (\alpha G D)^{-1}$ . Hence, the inequality above simplifies to:

$$\frac{1}{2} \exp(\lambda Q(T)) + (\alpha G D)^{-1} \text{Regret}_T(\alpha) \leq 1 + \sqrt{2T}. \quad (22)$$

The regret and CC bounds follow upon solving the above inequality.

**Regret Bound:** Using the fact that  $\exp(\lambda Q(T)) \geq \exp(\lambda Q(0)) \geq 1$ , Eqn. (22) yields

$$\text{Regret}_T(\alpha) \leq \alpha GD\sqrt{2T} + \frac{\alpha}{2}GD, \quad T \geq 1.$$

**CC Bound:** Let  $F$  denote the maximum value of  $f_t$  over the decision set, i.e.,  $F = \max_{1 \leq t \leq T} \max_{x \in \mathcal{X}} f_t(x)$ . Then, using the non-negativity of the cost functions, we have

$$f_t(x_t) - \alpha f_t(x^*) \geq 0 - \alpha F.$$

This implies that  $\text{Regret}_T(\alpha) \geq -\alpha FT$ . Hence, from Eqn. (22), we have for any  $T \geq 1$ :

$$\exp(\lambda Q(T)) \leq 2(1 + FT/GD + \sqrt{2T}).$$

Hence, the total resource consumption over the horizon is bounded as:

$$Q(T) \leq \lambda^{-1}O(\log T) = (\alpha B_T + GD\sqrt{T})O(\log T).$$

## 6 Lower bounds

Recall from Theorem 4 that our proposed online policy achieves a cumulative consumption (CC) bound of  $O(\log T)B_T + \tilde{O}(\sqrt{T})$ . First, setting  $B_T = 0$  recovers the notion of *round-wise feasibility*, in which any feasible offline benchmark ( $x^*$ ) incurs zero consumption in every round. In this setting, it was previously established by Sinha and Vaze [2024, Theorem 3] that the additive  $\tilde{O}(\sqrt{T})$  factor in the CC bound cannot be improved. In this Section, we further show that the  $O(\log T)$  multiplicative factor in front of  $B_T$  (equivalently, competitive ratio against any fixed action in hindsight) in the above expression for CC cannot be improved while maintaining a sublinear regret guarantee for the cumulative costs. In particular, we demonstrate that this impossibility result holds even when both the cost and constraint functions are linear, i.e.,  $\alpha = 1$ . For notational simplicity, we state the results in terms of rewards instead of costs.

**Theorem 5** (Lower bound for the competitive ratio). *Consider the above constrained learning problem with linear cost and linear consumption functions and a long-term budget of  $B_T$  for a horizon of length  $T$ . Let  $\pi$  be any online policy and  $\pi^*$  be a fixed offline optimal policy in the hindsight that samples actions from a fixed distribution in every round and consumes at most  $B_T$  resources in expectation, thus satisfying the budget constraint. Let  $\text{REW}_T(\pi)$  and  $\text{OPT}_T$  be the cumulative rewards accumulated by  $\pi$  and  $\pi^*$  respectively up to round  $T$ . Furthermore, let  $\text{CC}_T$  be the cumulative amount of resources consumed by the online policy  $\pi$  up to round  $T$ . Assume that for some constant  $\kappa > 0$ ,*

*and any  $T \geq 1$  the online policy  $\pi$  enjoys the following guarantee:*

$$\text{OPT}_T - \text{REW}_T(\pi) \leq h(T), \quad \text{CC}_T(\pi) - \kappa B_T \leq s(T),$$

*where  $s(T)$  and  $h(T)$  are some non-negative sublinear functions of the horizon length  $T$ , which do not depend on budget  $B_T$ . Then we must have  $\kappa \geq \Omega(\log T)$ .*

**Proof outline:** Our proof adapts the construction from the lower bound on the competitive ratio for the adversarial Bandits with Knapsacks (BwK) problem [Immorlica et al., 2022, Construction 8.7]. At a high-level, the key difference between the adversarial BwK and our settings is that while in the BwK problem, we stop playing as soon as the budget is exhausted (resulting in zero constraint violations), in our problem, we continue playing throughout the entire horizon at the expense of violating the prescribed resource constraints. In the following, we show that by appropriately rescaling the budget, a lower bound for the parameter  $\kappa$  in Theorem 5 can be obtained from the existing lower bound of the competitive ratio for the adversarial BwK problem.

## 7 Conclusion

In this paper, we introduced a framework for online non-convex optimization with long-term adversarial budget constraints for  $\alpha$ -approximately convex functions. We proposed an efficient first-order online policy that guarantees  $O(\sqrt{T})$   $\alpha$ -regret while exceeding the budget only by a factor of at most  $O(\log T)$  in both full-information and bandit settings. We also show that our performance bounds are tight. In the future, it will be interesting to extend the algorithm to more general class of non-convex functions.

## 8 Acknowledgement

AS was supported in part by the Department of Atomic Energy, Government of India, under project no. RTI4001 and in part by a Google India faculty Research Award.



## References

- Naman Agarwal, Alon Gonen, and Elad Hazan. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, pages 18–29. PMLR, 2019.
- Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1405–1424. SIAM, 2014.
- Francis Bach et al. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in machine learning*, 6(2-3): 145–373, 2013.
- Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.
- Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.
- Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations research*, 57(6): 1407–1420, 2009.
- Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022a.
- Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35:33589–33602, 2022b.
- Lin Chen, Hamed Hassani, and Amin Karbasi. Online continuous submodular maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 1896–1905. PMLR, 2018.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- Xiand Gao, Xiaobo Li, and Shuzhong Zhang. Online learning with non-convex losses and non-stationary regret. In *International Conference on Artificial Intelligence and Statistics*, pages 235–243. PMLR, 2018.
- Michael R Garey and David S Johnson. *Computers and intractability*, volume 29. wh freeman New York, 2002.
- Hengquan Guo, Xin Liu, Honghao Wei, and Lei Ying. Online convex optimization with hard constraints: Towards the best of two worlds and beyond. *Advances in Neural Information Processing Systems*, 35:36426–36439, 2022.
- Elad Hazan. *Introduction to online convex optimization*. MIT Press, 2022.
- Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6): 1–47, 2022.
- Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *Optical compressive imaging*, pages 279–312, 2016.
- Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3944–3952. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/liakopoulos19a.html>.
- 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.
- Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.
- Aranyak Mehta et al. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science*, 8(4):265–368, 2013.
- Julie Mulvaney-Kemp, SangWoo Park, Ming Jin, and Javad Lavaei. Dynamic regret bounds for constrained online nonconvex optimization based on polyak-lojasiewicz regions. *IEEE Transactions on Control of Network Systems*, 10(2):599–611, 2023.
- Michael J Neely and Hao Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.
- Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
- Mohammad Pedramfar and Vaneet Aggarwal. From linear to linearizable optimization: a novel framework with applications to stationary and non-stationary dr-submodular optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS ’24, Red

- Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.
- Sudeep Raja Putta and Shipra Agrawal. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 910–930. PMLR, 2022.
- Prasanna Raut, Omid Sadeghi, and Maryam Fazel. Online dr-submodular maximization: Minimizing regret and constraint violation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9395–9402, 2021.
- Adrian Rivera Cardoso, He Wang, and Huan Xu. The online saddle point problem and online convex optimization with knapsacks. *Mathematics of Operations Research*, 50(1):1–39, 2025.
- Abhishek Sinha. BanditQ - Fair Multi-Armed Bandits with Guaranteed Rewards per Arm. *arXiv preprint arXiv:2304.05219*, 2023.
- Abhishek Sinha and Rahul Vaze. Optimal algorithms for online convex optimization with adversarial constraints. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=TxffvJMnBy>.
- Aleksandrs Slivkins. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*, 2013.
- Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Christian Kroer. No-regret learning under adversarial resource constraints: A spending plan is all you need! *arXiv preprint arXiv:2506.13244*, 2025.
- Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861, 2020.
- Wen Sun, Debadeepta Dey, and Ashish Kapoor. Safety-aware algorithms for adversarial contextual bandit. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning Research*, volume 70 of *Proceedings of Machine Learning Research*, pages 3280–3288. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/sun17a.html>.
- Gang Wang, Georgios B Giannakis, Yousef Saad, and Jie Chen. Phase retrieval via reweighted amplitude flow. *IEEE Trans. Signal Process.*, 66(11):2818–2833, 2018.
- Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Johansson. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. In *International Conference on Machine Learning*, pages 11998–12008. PMLR, 2021.
- Qixin Zhang, Zengde Deng, Zaiyi Chen, Haoyuan Hu, and Yu Yang. Stochastic continuous submodular maximization: Boosting via non-oblivious function. In *International Conference on Machine Learning*, pages 26116–26134. PMLR, 2022.

## A Appendix

### A.1 Proof of Proposition 2

Suppose we have  $f \in \mathcal{L}_\alpha$  and  $g \in \mathcal{L}_\alpha$ . Then by definition, we have for all  $x, u \in \mathcal{X}$ :

$$f(x) - \alpha f(u) \leq \langle H_f(x), x - u \rangle \quad (23)$$

and

$$g(x) - \alpha g(u) \leq \langle H_g(x), x - u \rangle \quad (24)$$

Let  $h$  be a non-negative linear combination of  $f$  and  $g$ , i.e.,  $h = c_1 f + c_2 g$ , where  $c_1, c_2 \geq 0$ . Then for  $H_h = c_1 H_f + c_2 H_g$  the following holds

$$h(x) - \alpha h(u) \leq \langle H_h(x), x - u \rangle \quad (25)$$

Thus  $h \in \mathcal{L}_\alpha$ .

### A.2 Proof of Theorem 3

(1)  $\implies$  (2): Since  $f$  is  $\alpha$ -approximately convex, for a given  $x \in \mathcal{X}$ ,

$$\begin{aligned} & \exists g' \in \mathbb{R}^n \text{ s.t. } f(x) \leq \alpha f(u) + \langle g', x - u \rangle, \forall u \in \mathcal{X} \\ \Leftrightarrow & \exists g \in \mathbb{R}^n \text{ s.t. } \frac{f(x)}{\alpha} \leq f(u) + \langle g, x - u \rangle, \forall u \in \mathcal{X} \\ \Leftrightarrow & \frac{f(x)}{\alpha} \leq \sup_{g \in \mathbb{R}^n} \inf_{u \in \mathcal{X}} (f(u) + \langle g, x - u \rangle) \\ \Leftrightarrow & \frac{f(x)}{\alpha} \stackrel{(a)}{\leq} \sup_{g \in \mathbb{R}^n} (\langle g, x \rangle - f^*(g)) \\ \Leftrightarrow & f(x) \stackrel{(b)}{\leq} \alpha f^{**}(x), \end{aligned} \quad (26)$$

where in steps (a) and (b), we have used the definition of Fenchel conjugate.

(2)  $\implies$  (3): This holds since  $f^{**}$  is a convex function that satisfies  $f^{**}(x) \leq f(x), \forall x \in \mathcal{X}$  [Bertsekas et al., 2003, Proposition 7.1.1].

(3)  $\implies$  (1): We have

$$g(x) \leq f(x) \leq \alpha g(x), \forall x \in \mathcal{X}. \quad (27)$$

Since  $g$  is convex, for any arbitrary  $u \in \mathcal{X}$ , we have

$$g(u) \geq g(x) + \langle G(x), u - x \rangle,$$

where  $G(x)$  is a subgradient of the convex function  $g$  at the point  $x$ . Hence, from the given condition, we have

$$f(u) \geq g(u) \geq g(x) + \langle G(x), u - x \rangle \geq \frac{f(x)}{\alpha} + \langle G(x), u - x \rangle.$$

This implies that for all  $x, u$ , we have

$$f(x) - \alpha f(u) \leq \langle H(x), x - u \rangle,$$

which shows that the function  $f$  is  $\alpha$ -approximate convex. In the above equation, we have defined  $H(x) \equiv \alpha G(x)$ .

(3)  $\implies$  (4): If (3) holds then there is a convex function  $g$  such that  $g(x) \leq f(x) \leq \alpha g(x), \forall x \in \mathcal{X}$ . Then,

$$f\left(\sum_i p_i x_i\right) \leq \alpha g\left(\sum_i p_i x_i\right) \leq \alpha \sum_i p_i g(x_i), \quad (28)$$

where the last inequality follows from Jensen's inequality.

(4)  $\implies$  (1): Let, for any  $N \geq 1$ ,  $p \in \Delta_N$  and  $x_i \in \mathcal{X}, \forall i \in [N]$ . Then, note that  $(\sum_i p_i x_i, \sum_i p_i f(x_i)) \in \text{co}(\text{epi}(f))$ . Also, by the condition, as  $f(\sum_i p_i x_i) \leq \alpha \sum_i p_i f(x_i)$ , we have that if  $w \geq \sum_i p_i f(x_i)$ , then,  $\alpha w \geq f(\sum_i p_i x_i)$ . Consequently,  $\text{co}(\text{epi}(f)) \subset \text{epi}_\alpha(f)$ , where we have defined  $\text{epi}_\alpha(f) = \{(x, w) : w \geq f(x)/\alpha\}$ . Since  $\text{epi}(f^{**}) = \text{cl}(\text{co}(\text{epi}(f)))$ , we have,  $\text{epi}(f^{**}) \subset \text{cl}(\text{epi}_\alpha(f))$ . Therefore, for any  $x$ ,  $(x, f^{**}(x)) \in \text{epi}(f^{**}) \implies (x, f^{**}(x)) \in \text{cl}(\text{epi}_\alpha(f))$ . Therefore, there is a sequence  $\{(x_k, w_k)\} \in \text{epi}_\alpha(f)$  such that  $x_k \rightarrow x$  and  $w_k \rightarrow f^{**}(x)$ . Since  $(x_k, w_k) \in \text{epi}_\alpha(f)$ ,  $f(x_k) \leq \alpha w_k$ . Assuming  $f$  to be a closed function, we have,  $f(x) = \lim_k f(x_k) \leq \alpha f^{**}(x)$ . Then, by (2),  $f$  is  $\alpha$ -approximately convex.

### A.3 Approximate Convexity of Weakly DR-submodular functions

**Weakly DR-submodular functions:** Consider a product-form decision set  $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$  where each  $\mathcal{X}_i$  is a compact subset of non-negative reals  $\mathbb{R}_+$ . For any  $(x, y) \in \mathcal{X} \times \mathcal{X}$  we define the partial order  $\leq$  such that  $x \leq y$  iff  $x_i \leq y_i, \forall i$ . We say a differentiable function  $F(\cdot)$  is weakly DR (diminishing return) submodular with parameter  $\gamma$  if we have:

$$\nabla F(x) \geq \gamma \nabla F(y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{X}, \text{ s.t. } x \leq y,$$

This class of functions generalizes the class of differentiable DR submodular functions which have  $\gamma = 1$  [Raut et al., 2021].

**Example:** An important example of a DR-submodular function is the multilinear extension  $F : [0, 1]^n \mapsto \mathbb{R}$  of a submodular function  $f : 2^V \mapsto \mathbb{R}$  defined on the subsets of a ground set  $V$  as below:

$$F(x) = \sum_{S \subseteq V} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) f(S).$$

In other words,  $F(x)$  is the expectation of the set function  $f(S)$  when the element  $i$  is included in the subset  $S$  independently w.p.  $x_i, \forall i$ . Some examples of submodular set functions include: the Cut function in a graph, Rank function of a Matroid, Coverage function, Log-determinant function of a positive semidefinite matrix etc. See Bach et al. [2013] for the definition of these functions and an excellent treatment of submodular optimization.

Theorem 6 below shows that the function  $F$  is  $\frac{1}{1-e^{-\gamma}}$ -approximately concave.

**Theorem 6.** Let  $F : \mathcal{X} \rightarrow \mathbb{R}_+$  be a weakly DR-submodular and monotone function with parameter  $\gamma > 0$ . Then for any two vectors  $x, y \in \mathcal{X}$ , we have

$$F(x) - (1 - e^{-\gamma})F(y) \geq \langle \nabla \tilde{F}(x), x - y \rangle,$$

where  $\tilde{F} : \mathcal{X} \rightarrow \mathbb{R}_+$  is the non-oblivious function corresponding to  $F$  defined as:

$$\nabla \tilde{F}(x) = \int_0^1 e^{\gamma(z-1)} \nabla F(z) dz. \quad (29)$$

*Proof.* Lemma 2 of Zhang et al. [2022] states that, for any two vectors  $x, y \in \mathcal{X}$ , we have

$$\langle y - x, \nabla \tilde{F}(x) \rangle \geq \gamma \left( \int_0^1 w(z) dz \right) (F(y) - \theta(w)F(x)), \quad (30)$$

where the expressions  $w(z)$  and  $\theta(w)$  have been defined in Zhang et al. [2022]. Using expressions of  $w(z)$  and  $\theta(w)$  from Theorem 1 of Zhang et al. [2022], we obtain the desired result.  $\square$

### A.4 Approximate Convexity of Regularized Phase Retrieval

**The Phase Retrieval Problem:** Let us consider the problem of  $l_2$ -regularized Phase Retrieval (PR), where the problem is estimate an unknown signed vector from the absolute (unsigned) values of its linear measurements [Jaganathan et al., 2016]. The standard approach for the PR problem solves the following optimization problem:

$$\min_{x \in \mathcal{X}} f(x) = \min_{x \in \mathcal{X}} \frac{1}{2} \|y - |\Phi x|\|^2 + \frac{\lambda}{2} \|x\|^2, \quad (31)$$



where  $\lambda > 0$  is a regularization parameter,  $\Phi \in \mathbb{R}^{m \times n}$  is the measurement matrix and  $y$  is the measurement vector with non-negative co-ordinates and  $\mathcal{X}$  is the constraint set. The objective function  $f(x)$  is known to be non-convex [Wang et al., 2018]. We prove the following:

**Theorem 7.** Let  $\|\Phi\|_{2 \rightarrow 2}$  denote the operator norm of the measurement matrix  $\Phi$ , and let  $\lambda > 0$  in (31). Then  $f$ , defined in (31) is  $(1 + 1/\gamma)$ -approximately convex, where  $\gamma = \frac{\lambda}{\|\Phi\|_{2 \rightarrow 2}^2}$ .

*Proof.* In the following, we denote  $\gamma = \frac{\lambda}{\|\Phi\|_{2 \rightarrow 2}^2}$ , where  $\|\Phi\|_{2 \rightarrow 2}$  is the operator norm of  $\Phi$ .

We now define the following candidate function  $g$  which appears in part 3 of Theorem 3:

$$g(x) := f(x) - \frac{(1 + \gamma)}{2} \left\| \left( \frac{y}{1 + \gamma} - |\Phi x| \right)_+ \right\|^2, \quad (32)$$

where for any vector  $v$ , we define  $(v)_+$  as the vector with  $[(v)_+]_i = \max\{0, v_i\}$ . Clearly  $g(x) \leq f(x)$ . We will now prove that  $g$  is convex.

To see this, note that we can re-express  $f$  as below:

$$\begin{aligned} f(x) &= \frac{\|y\|^2}{2} - y^\top |\Phi x| + \frac{\|\Phi x\|^2}{2} \left( 1 + \frac{\lambda}{\|\Phi\|_{2 \rightarrow 2}^2} \right) + \frac{\lambda}{2} \left( \|x\|^2 - \frac{\|\Phi x\|^2}{\|\Phi\|_{2 \rightarrow 2}^2} \right) \\ &= \frac{\|y\|^2}{2} - y^\top |\Phi x| + \frac{\|\Phi x\|^2}{2} (1 + \gamma) + \frac{\lambda}{2} \left( \|x\|^2 - \frac{\|\Phi x\|^2}{\|\Phi\|_{2 \rightarrow 2}^2} \right) \\ &= \frac{\|y\|^2}{2} \left( 1 - \frac{1}{1 + \gamma} \right) + \frac{\|y\|^2}{2(1 + \gamma)} - y^\top |\Phi x| + \frac{\|\Phi x\|^2}{2} (1 + \gamma) + \frac{\lambda}{2} \left( \|x\|^2 - \frac{\|\Phi x\|^2}{\|\Phi\|_{2 \rightarrow 2}^2} \right) \\ &= \frac{\|y\|^2 \gamma}{2(1 + \gamma)} + \frac{\lambda}{2} x^\top \left( I - \frac{\Phi^\top \Phi}{\|\Phi\|_{2 \rightarrow 2}^2} \right) x + \frac{(1 + \gamma)}{2} \left\| \frac{y}{1 + \gamma} - |\Phi x| \right\|^2. \end{aligned} \quad (33)$$

Consequently, we obtain from the definition of  $g$ ,

$$g(x) = \underbrace{\frac{\|y\|^2 \gamma}{2(1 + \gamma)}}_{T_1} + \underbrace{\frac{\lambda}{2} x^\top \left( I - \frac{\Phi^\top \Phi}{\|\Phi\|_{2 \rightarrow 2}^2} \right) x}_{T_2} + \underbrace{\frac{(1 + \gamma)}{2} \left\| \left( |\Phi x| - \frac{y}{1 + \gamma} \right)_+ \right\|^2}_{T_3}. \quad (34)$$

The term  $T_1$  is a constant,  $T_2$  is convex as the Hessian is  $\lambda \left( I - \frac{\Phi^\top \Phi}{\|\Phi\|_{2 \rightarrow 2}^2} \right)$ , which is positive semi-definite. The function in  $T_3$  can be shown to be convex as below:

$$\frac{2T_3}{1 + \gamma} = \sum_{j=1}^m \left( |\phi_j^\top x| - \frac{y_j}{1 + \gamma} \right)_+^2 = \sum_{j=1}^m h_j(\phi_j^\top x), \quad (35)$$

where  $h_j(u) = \left( |u| - \frac{y_j}{1 + \gamma} \right)_+^2$ ,  $1 \leq j \leq m$ . Since the squared ReLU function is convex,  $h_j$ 's are convex, making  $T_3$  convex. Consequently,  $g$  is convex.

To find  $\alpha > 1$  such that  $f(x) \leq \alpha g(x)$ , using the expressions of  $f, g$  it therefore suffices to find  $\alpha$  such that

$$\begin{aligned} f(x) &\leq \alpha \left( f(x) - \frac{(1 + \gamma)}{2} \left\| \left( \frac{y}{1 + \gamma} - |\Phi x| \right)_+ \right\|^2 \right) \\ \Leftrightarrow \frac{(1 + \gamma)\alpha}{2(\alpha - 1)} &\leq \frac{f(x)}{\left\| \left( \frac{y}{1 + \gamma} - |\Phi x| \right)_+ \right\|^2}, \forall x. \end{aligned} \quad (36)$$

Since the RHS have to be minimized, let us focus on the polyhedron  $\mathcal{C} = \{x : |\Phi x| \leq \frac{y}{1+\gamma}\}$ . Then, such an  $\alpha$  can be found if it satisfies the following:

$$\frac{(1+\gamma)\alpha}{2(\alpha-1)} \leq \frac{\frac{\|y\|^2\gamma}{2(1+\gamma)} + \frac{\lambda}{2}x^\top \left(I - \frac{\Phi^\top\Phi}{\|\Phi\|_{2 \rightarrow 2}^2}\right)x + \frac{(1+\gamma)}{2} \left\| |\Phi x| - \frac{y}{1+\gamma} \right\|^2}{\left\| \frac{y}{1+\gamma} - |\Phi x| \right\|^2}, \quad \forall x \in \mathcal{C}, \quad (37)$$

which in turn is satisfied if

$$\frac{(1+\gamma)\alpha}{2(\alpha-1)} \leq \frac{1+\gamma}{2} + \frac{\frac{\|y\|^2\gamma}{2(1+\gamma)}}{\sum_{j=1}^m \frac{y_j^2}{(1+\gamma)^2} (1-t_j)^2}, \quad t_j \in [0, 1], \quad j = 1, 2, \dots, m, \quad (38)$$

where we define  $t_j = \frac{(1+\gamma)|\phi_j^\top x|}{y_j} \in [0, 1]$  whenever  $x \in \mathcal{C}$ . The above is satisfied if

$$\frac{(1+\gamma)\alpha}{2(\alpha-1)} \leq \frac{1+\gamma}{2} + \frac{\frac{\|y\|^2\gamma}{2(1+\gamma)}}{\sum_{j=1}^m \frac{y_j^2}{(1+\gamma)^2}} = \frac{(1+\gamma)^2}{2}. \quad (39)$$

The choice  $\alpha = 1 + \frac{1}{\gamma}$  satisfies the above. □

### A.5 On Bounding the Norms of Generalized Subgradients

**Lemma 8.** *Let  $f$  be an  $\alpha$ -approximately convex function with domain  $\mathcal{X}$ . Then from Theorem 3, part 3, there exists a convex function  $g$  such that  $g(x) \leq f(x) \leq \alpha g(x), \forall x \in \mathcal{X}$ . If  $h(x)$  is a sub-gradient of  $g$  at the point  $x \in \mathcal{X}$  then  $\alpha h(x)$  is a generalized sub-gradient of  $f$  at  $x \in \mathcal{X}$ .*

*As a corollary, if  $\|h(x)\|_2 \leq G, \forall x \in \mathcal{X}$ , then the  $\ell_2$ -norms of the generalized subgradients of  $f$  as constructed above can be uniformly upper bounded by  $\alpha G$ .*

*Proof.* Note that  $h(x)$  always exists since  $g$  is convex. Therefore, for any  $u \in \mathcal{X}$ , we obtain,

$$\begin{aligned} \alpha f(u) + \langle \alpha h(x), x - u \rangle &= \alpha(f(u) + \langle h(x), x - u \rangle) \\ &\stackrel{(a)}{\geq} \alpha(g(u) + \langle h(x), x - u \rangle) \\ &\stackrel{(b)}{\geq} \alpha g(x) \\ &\stackrel{(c)}{\geq} f(x), \end{aligned}$$

where step (a) follows from the assumption that  $f(u) \geq g(u)$ , step (b) follows from the fact that  $h(x)$  is a sub-gradient of  $g$  at the point  $x$ , and (c) follows from the assumption that  $\alpha g(x) \geq f(x)$ . The final inequality shows that  $h(x)$  is a generalized sub-gradient of  $f$  at  $x \in \mathcal{X}$ . □

### A.6 Adaptive regret bounds for OCO

In this Section, we briefly recall the first-order methods (*a.k.a.* Projected Online Gradient Descent (OGD)) for the standard OCO problem [Orabona, 2019, Algorithm 2.1] [Hazan, 2022]. These methods differ among each other in the way the step sizes are chosen. For a sequence of convex cost functions  $\{\hat{f}_t\}_{t \geq 1}$ , a projected OGD algorithm selects the successive actions as:

$$x_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \eta_t \nabla_t), \quad \forall t \geq 1, \quad (40)$$

where  $\nabla_t \equiv \nabla \hat{f}_t(x_t)$  is a subgradient of the function  $\hat{f}_t$  at  $x_t$ ,  $\mathcal{P}_{\mathcal{X}}(\cdot)$  is the Euclidean projection operator on the set  $\mathcal{X}$  and  $\{\eta_t\}_{t \geq 1}$  is a specified step size schedule. The (diagonal version of the) AdaGrad policy adaptively chooses the step size sequence as a function of the previous subgradients as  $\eta_t = \frac{\sqrt{2D}}{2\sqrt{\sum_{\tau=1}^t G_\tau^2}}$ , where  $G_t = \|\nabla_t\|_2, t \geq 1$  [Duchi et al., 2011].<sup>1</sup> This algorithm enjoys the following adaptive regret bound.

<sup>1</sup>We set  $\eta_t = 0$  if  $G_t = 0$ .

**Theorem 9.** [Orabona, 2019, Theorem 4.14] The AdaGrad policy, with the above step size sequence, achieves the following regret bound for the standard OCO problem:

$$\text{Regret}_T \leq \sqrt{2}D \sqrt{\sum_{t=1}^T G_t^2}. \quad (41)$$

## B Proof of Theorem 5

Consider an ensemble of constrained learning problems defined in Section 3 with linear rewards, where each instance consists of two arms  $\mathcal{A}_0, \mathcal{A}_1$  and a budget of  $B_T = \sqrt{2} \max(\sqrt{Th(T)}, s(T))$ . Note that this implies that  $B_T = \Omega(\sqrt{T})$ . An online randomized policy selects one of these two arms in every round. In line with our deterministic formulation, we convexify the decision set and work with the expected rewards and consumptions. In particular, the decision set  $\mathcal{X}$  in this problem is taken to be the closed interval  $[0, 1]$ , which denotes the probability of pulling the arm  $\mathcal{A}_1$ .

We partition the time horizon into  $T/B_T$  phases of duration  $B_T$  each<sup>2</sup>. Next, we define  $T/B_T$  problem instances: for instance  $I_\tau, \tau \in [\frac{T}{B}]$ , arm  $\mathcal{A}_1$  has positive rewards up to and including phase  $\tau$ ; rewards for all subsequent phases are zero. In phase  $\sigma \in [\tau]$ , arm  $\mathcal{A}_1$  has reward  $\sigma B_T/T$  in each round. Arm  $\mathcal{A}_1$  consumes unit resource in each round. On the other hand, arm  $\mathcal{A}_0$  has zero rewards and zero consumptions on all rounds for all instances. In every round  $t \geq 1$ , let the randomized policy pulls arm  $\mathcal{A}_1$  with probability  $x_t$ , and pulls arm  $\mathcal{A}_0$  with the complementary probability  $1 - x_t$ . Hence, the expected reward and consumption functions for round  $t$  for the instance  $I_\tau$  are given to be:

$$\begin{aligned} f_t^{I_\tau}(x_t) &= \begin{cases} \sigma B x_t / T; & t \in [\sigma B/T, (\sigma + 1)B/T], \sigma \in [\tau] \\ 0 & \end{cases} \\ g_t^{I_\tau}(x_t) &= x_t. \end{aligned}$$

It should be noted that, for each instance, the cost and constraint functions are *linear* in the action variable  $x_t$ .

**Analysis:** We call a policy *feasible* if it satisfies the long-term budget constraint, *i.e.*, does not violate the budget constraint. Fix some problem instance  $I_\tau, \tau \in [B/T]$ . Let  $\text{OPT}_T$  be the reward obtained by the best fixed feasible randomized policy for this instance. Consider any feasible randomized online policy  $\pi'$ . From Immorlica et al. [2022, Theorem 8.1, part (b) and Lemma 8.6], it follows that for any feasible policy, there exists a problem instance  $\mathcal{I}_\tau$  s.t.:

$$\text{OPT}_T / \text{Rew}_T(\pi') \geq \Omega(\log T). \quad (42)$$

Now consider an online policy  $\pi$  with budget constraint  $B_T$ , which pulls arm  $\mathcal{A}_1$  with probability  $x_t$  in every round  $t \geq 1$ . Note that the policy  $\pi$  is not necessarily feasible as its cumulative consumption after  $T$  rounds may exceed the budget  $B_T$ . We now modify the policy  $\pi$  to obtain a new online policy  $\pi'$  which is *feasible*. The modified policy  $\pi'$  pulls arm  $\mathcal{A}_1$  with probability  $x_t/\eta(T)$  and arm  $\mathcal{A}_0$  with probability  $1 - x_t/\eta(T)$  on round  $t$ , where  $\eta(T) = (\kappa + \frac{s(T)}{B_T})$ . Due to the linearity of the rewards and consumptions with respect to the variable  $x_t$ , we have:

$$\text{REW}_T(\pi') = \frac{1}{\eta(T)} \text{REW}_T(\pi), \quad \text{CC}_T(\pi') = \frac{1}{\eta(T)} \text{CC}_T(\pi).$$

Finally, using the cumulative consumption bound for the policy  $\pi$ , we can write

$$\text{CC}_T(\pi') \leq \frac{\kappa B_T + s(T)}{\eta(T)} = B_T.$$

This shows that the modified policy  $\pi'$  is indeed feasible. Furthermore, using the regret guarantee for the policy  $\pi$ , we have

$$\text{REW}_T(\pi') \geq \frac{\text{OPT}_T - h(T)}{\eta(T)} = \frac{\text{OPT}_T - h(T)}{\kappa + \frac{s(T)}{B_T}}. \quad (43)$$

<sup>2</sup>Without any loss of generality, we assume that  $B_T$  divides  $T$ .

From Immorlica et al. [2022, Lemma 8.9], we have that for any of the constructed instances, we have  $\text{OPT}_T \geq \frac{B_T^2}{T}$ . Thus

$$\text{OPT}_T - h(T) = \text{OPT}_T \left(1 - \frac{h(T)}{\text{OPT}_T}\right) \geq \text{OPT}_T \left(1 - \frac{Th(T)}{B_T^2}\right) \stackrel{(a)}{\geq} \text{OPT}_T \left(1 - \frac{Th(T)}{2Th(T)}\right) \geq \text{OPT}_T/2.$$

where in (a), we have used the fact that  $B_T \geq \sqrt{2Th(T)}$ . Thus, from Eqn. (43), we have

$$\begin{aligned} \text{REW}_T(\pi') &\geq \frac{\text{OPT}_T/2}{\kappa + \frac{s(T)}{B_T}} \\ \text{i.e., } \kappa + \frac{s(T)}{B_T} &\geq \frac{\text{OPT}_T}{2\text{REW}_T(\pi')}. \end{aligned}$$

Since  $B_T \geq \sqrt{2}s(T)$ , using the lower bound from Eqn. (42), it follows that there exists a problem instance  $I_T$  with  $\kappa \geq \Omega(\log T)$ .

## C Extension to Multiple Resources

Instead of a single resource as described in the main paper, we now assume that there are  $k \geq 1$  separate resources such that each resource has a separate budget constraint of  $B_T$ .<sup>3</sup> Note that since the we have a separate long-term budget constraint for each resource, unlike Sinha and Vaze [2024, Section 2.1], we can not reduce multiple resources into a single effective resource by taking the pointwise supremum of the consumption functions.

This is because Sinha and Vaze [2024] assumed per-round feasibility for all constraints and the same would hold for pointwise supremum. Formally, if the following holds for each resource

$$f_{t,i}(x^*) \leq 0 \quad \forall t$$

then

$$\max_i f_{t,i}(x^*) \leq 0 \quad \forall t$$

However, in our setting, just because the sum of consumptions for each resource satisfies the budget individually, it does not extend to their pointwise supremum. As a toy example consider the setting where we have 2 resources and the horizon is of length  $T$ . The first resource has cumulative consumption  $B_T$  in the first  $T/2$  rounds and 0 in the rest. The second resource has 0 cumulative consumption in the first  $T/2$  rounds and  $B_T$  in the rest. It is clear that the cumulative consumption of both of these resources would individually be  $B_T$ . However, when we take their pointwise supremum, the resulting effective resource would have cumulative consumption  $B_T$  in both halves of the horizon. Overall, the cumulative consumption would be  $2B_T$  which would violate the budget.

We now extend our previous analysis to handle this general case.

Let  $Q_i(t)$  be the cumulative consumption of the  $i^{\text{th}}$  resource, which evolves as follows:

$$Q_i(t) = Q_i(t-1) + g_{t,i}(x_t), i \in [k].$$

Let  $\Phi(\cdot)$  be a non-decreasing and convex Lyapunov function. We compute the drift for the  $i^{\text{th}}$  resource as

$$\Phi(Q_i(t)) - \Phi(Q_i(t-1)) \leq \Phi'(Q_i(t))(Q_i(t) - Q_i(t-1)) = \Phi'(Q_i(t))g_{t,i}(x_t).$$

Summing both sides of the inequality over all resources and then adding  $V(f_t(x_t) - \alpha f_t(x^*))$  to both sides, we obtain

$$\begin{aligned} &V(f_t(x_t) - \alpha f_t(x^*)) + \sum_i (\Phi(Q_i(t)) - \Phi(Q_i(t-1))) \\ &\leq (Vf_t(x_t) + \sum_i \Phi'(Q_i(t))g_{t,i}(x_t)) - \alpha(Vf_t(x^*) + \sum_i \Phi'(Q_i(t))g_{t,i}(x^*)) + \alpha \sum_i \Phi'(Q_i(T))g_{t,i}(x^*) \end{aligned}$$

<sup>3</sup>The case where the budget constraint for each of the resources could be different can be handled by scaling the  $i^{\text{th}}$  consumption function by  $B_T/B_{T,i}, i \in [k]$ .



where, in the last step, we have used the facts that  $\Phi'(\cdot)$  is monotone (since  $\Phi(\cdot)$  is convex),  $Q(t)$  is non-decreasing, and  $g_t \geq 0$ . Summing up the above inequality over  $1 \leq t \leq T$ , we have the following regret decomposition inequality

$$\sum_i (\Phi(Q_i(T)) - \Phi(Q_i(0))) + V\text{Regret}_T(\alpha) \leq \text{Regret}'_T(\alpha) + \sum_i \Phi'(Q_i(T)) \sum_{t=1}^T g_{t,i}(x^*), \quad (44)$$

$$\leq \text{Regret}'_T(\alpha) + \alpha \sum_i \Phi'(Q_i(T))B \quad (45)$$

where we have used the fact that  $\sum_t g_{t,i}(x^*) \leq B$  and  $\text{Regret}'_T(\alpha)$  is defined as the regret for learning the surrogate cost function sequence

$$\hat{f}_t = Vf_t + \sum_i \Phi'(Q_i(t))g_{t,i}, t \geq 1, \quad (46)$$

with the comparator taken to be  $\hat{f}_t(x^*)$ , where  $x^*$  is a feasible action belonging to the set  $\mathcal{X}^*$ . Note that the regret decomposition inequality (45) holds for any cost and non-negative constraint functions.

We now make the assumption that the cost and constraint functions are  $\alpha$ -approximately convex and  $G$ -Lipschitz and we can bound the surrogate  $\alpha$ -regret by the regret incurred by passing  $\tilde{f}_t(x) = \langle H_{\hat{f}_t}(x_t), x \rangle$  to an OLO algorithm. Using the analysis of section A.6, we have the following upper bound on the surrogate regret:

$$\text{Regret}'_T(\alpha) \leq \text{Regret}'(\text{OLO}), \quad (47)$$

$$\|H_{\hat{f}_t}(x_t)\|_2 \leq V\|H_{f_t}(x_t)\|_2 + \sum_i \Phi'(Q_i(t))\|H_{g_t}(x_t)\|_2 \leq \alpha G(V + \sum_i \Phi'(Q_i(T))). \quad (48)$$

$$\text{Regret}'_T(\text{OLO}) \leq \sqrt{2}GD\alpha(V + \sum_i \Phi'(Q_i(T)))\sqrt{T}. \quad (49)$$

Hence (45) yields:

$$\sum_i \Phi(Q_i(T)) + V\text{Regret}_T(x^*) \leq \sum_i \Phi(Q_i(0)) + \alpha VGD\sqrt{2T} + \alpha \sum_i \Phi'(Q_i(T))(GD\sqrt{2T} + B). \quad (50)$$

Consider the exponential Lyapunov function:  $\Phi(x) = \exp(\lambda x)$ , where the value of  $\lambda$  will be fixed later. With this, inequality (50) yields

$$\sum_i \exp(\lambda Q_i(T)) + V\text{Regret}_T(x^*) \leq k + \alpha VGD\sqrt{2T} + \lambda \alpha \sum_i \exp(\lambda Q_i(T))(GD\sqrt{2T} + B).$$

Now we set  $\lambda = \frac{1}{2}(\alpha GD\sqrt{2T} + \alpha B)^{-1}$  and  $V = (\alpha GD)^{-1}$ . With this choice for the parameters, the above inequality yields:

$$\frac{1}{2} \sum_i \exp(\lambda Q_i(T)) + V\text{Regret}_T(x^*) \leq 1 + \sqrt{2T}. \quad (51)$$

**Regret Bound:** Using the fact that  $\exp(\lambda Q_i(T)) \geq 1$ , Eqn. (51) yields

$$\text{Regret}_T(x^*) \leq \alpha GD\sqrt{2T} + \alpha GD\frac{k}{2},$$

where  $k$  is the numbers of resources.

**Bounding the CC:** Since  $\text{Regret}_T(x^*) \geq -\alpha FT$ , where  $F$  is a uniform upper bound for the losses, Eqn. (51) yields for  $T \geq 1$ :

$$\sum_i \exp(\lambda Q_i(T)) \leq 2(1 + FT/GD + \sqrt{2T})$$

Hence,

$$Q_i(T) \leq (\alpha B_T + GD\sqrt{T})O(\log T) \quad \forall i.$$

## D Adversarial Bandits with Knapsacks

In this Section, we demonstrate how the proposed online algorithm (Algorithm 1) and its analysis can be extended to the setting where the learner receives bandit feedback, *i.e.*, only the losses and consumption of the selected actions are revealed to the learner. The setting we consider here is the same as the Bandits with Knapsacks (BwK) problem, considered by Immorlica et al. [2022], with the key difference that, in our case, the interaction between the learner and the adversary continues over the entire time-horizon and, consequently, we allow the constraints to be violated. Compared to the primal-dual-based algorithm proposed in Immorlica et al. [2022], which needs to guess the value of the optimal offline algorithm, our algorithm is simpler and does not need any such guesses and offers improved guarantees as shown next.

Specifically, we consider a Multi-Armed Bandit (MAB) with  $K$  arms and a single resource with the budget constraint  $B_T$ . When arm  $a_t \in [K]$  is pulled on round  $t$ , it incurs a loss of  $l_t(a_t) \in [0, 1]$  and consumes  $c_t(a_t) \in [0, 1]$  amount of resource. The objective is to minimize the total loss while consuming close to the allocated budget over a horizon of length  $T$ .

Because of the bandit feedback, the learner is informed of these two scalars only at round  $t$ . We make the standard assumption that the loss and consumption sequences, *i.e.*,  $\{l_t, c_t\}_{t=1}^T$ , are generated in an oblivious fashion, *i.e.*, they are fixed before the game begins. Hence, any randomized policy, that samples an arm  $a_t$  from a distribution  $x_t \in \Delta_K$ <sup>4</sup> on round  $t$ , incurs an expected cost of  $f_t(x_t)$  and consumes an expected  $g_t(x_t)$  amount of resource as given below:

$$f_t(x_t) = \langle l_t, x_t \rangle, \quad g_t(x_t) = \langle c_t, x_t \rangle.$$

The regret and cumulative consumptions are defined as in Eqn. (2) and (3) as before. Our proposed constrained bandits algorithm, described in Algorithm 2, simply runs an adaptive bandit algorithm on a sequence of surrogate cost functions defined similar to Eqn. (12) as in the full-information case. However, for technical reasons which will be clear in the analysis, we use a power-law Lyapunov function rather than the exponential Lyapunov function as in the full-information case.

---

### Algorithm 2 Algorithm for Adversarial Bandits with Knapsacks

---

- 1: **Inputs:** The set of arms  $[K]$ , horizon length  $T$ , sequence of losses  $\{l_t\}_{t=1}^T$  and consumptions  $\{c_t\}_{t=1}^T$ , Budget  $B_T$ .
  - 2: **Parameters:**  $V := \frac{(e(18K\sqrt{T}(\log T)^3 + B_T \log T))^{\log T}}{36K\sqrt{T}(\log T)^2}$ , Lyapunov function  $\Phi(x) = x^{\log T}$
  - 3: Initialize a uniform sampling distribution  $x_1 \leftarrow (1/K, \dots, 1/K)$  and set  $Q(0) \leftarrow \log T$ .
  - 4: **for**  $t = 1 : T$  **do**
  - 5:   Sample an arm  $a_t$  from the probability distribution  $x_t$
  - 6:   Observe  $l_t(a_t)$  and  $c_t(a_t)$
  - 7:   Compute surrogate loss  $\hat{l}_t(a_t) = V l_t(a_t) + e\Phi'(Q(t-1))c_t(a_t)$
  - 8:   Update the cumulative consumed resource  $Q(t) = Q(t-1) + c_t(a_t)$
  - 9:   Pass the observed surrogate loss  $\hat{l}_t(a_t)$  to the adaptive MAB algorithm (Algorithm 3), which returns the next sampling distribution  $x_{t+1} \in \Delta_K$ .
  - 10: **end for**
- 

**Theorem 10.** *Algorithm 2 achieves a regret bound of  $\tilde{O}(K\sqrt{T})$  while consuming at most  $\tilde{O}(K\sqrt{T}) + O(B_T \log T)$  amount of resources in expectation.*

<sup>4</sup> $\Delta_K$  denotes the standard probability simplex on  $K$  atoms (arms). All logarithms are taken w.r.t. the natural base.

*Proof.* We start with slightly modifying the derivation of the regret decomposition inequality in the full information setting from Section 5. The cumulative consumption evolves as

$$Q(t) = Q(t-1) + c_t(a_t). \quad (52)$$

As before, let  $\Phi(\cdot)$  be a non-decreasing and convex Lyapunov function, which will be fixed later. We can bound the change in the Lyapunov function at round  $t$  as follows:

$$\Phi(Q(t)) - \Phi(Q(t-1)) \stackrel{(a)}{\leq} \Phi'(Q(t))(Q(t) - Q(t-1)) \stackrel{(b)}{=} \Phi'(Q(t))c_t(a_t) \stackrel{(c)}{\leq} \Phi'(Q(t-1)+1)c_t(a_t) \stackrel{(d)}{\leq} e\Phi'(Q(t-1))c_t(a_t)$$

where (a) follows from the convexity of  $\Phi(\cdot)$ , (b) follows from Eqn. (52), (c) follows from the fact that  $c_t(a_t) \leq 1$ , and (d) holds for our particular choice of the Lyapunov function with a proper initialization for  $Q(0)$  as shown in Lemma 12. Adding  $V(l_t(a_t) - l_t(a^*))$  to both sides of the above inequality, we obtain:

$$\begin{aligned} & \Phi(Q(t)) - \Phi(Q(t-1)) + V(l_t(a_t) - l_t(a^*)) \\ & \leq (Vl_t(a_t) + e\Phi'(Q(t-1))c_t(a_t)) - (Vf_t(x^*) + e\Phi'(Q(t-1))c_t(a^*)) + e\Phi'(Q(t-1))c_t(a^*), \end{aligned} \quad (53)$$

where the comparator  $a^*$  is taken to be a fixed randomized benchmark action that minimizes the expected cumulative costs subject to that it satisfies the budget constraint in expectation, *i.e.*,  $a^* \sim D^*$  where the distribution  $D^*$  solves the following optimization problem

$$\min_{D \in \Delta_K} \mathbb{E}_{a^* \sim D} \sum_{t=1}^T l_t(a^*), \text{ s.t. } \mathbb{E}_{a^* \sim D} \sum_{t=1}^T c_t(a^*) \leq B_T. \quad (54)$$

Clearly, the distribution of the benchmark action  $a^*$  may depend on the entire sequence of loss and consumption vectors but not on the actions of the online policy. Similar to Eqn. (12), we now define the surrogate losses for the arms at round  $t$  as follows:

$$\hat{l}_t(a) = Vl_t(a) + e\Phi'(Q(t-1))c_t(a), \quad \forall a \in [K]. \quad (55)$$

The main difference between the definitions of surrogate loss in the full information setting (12) and the bandit setting (55) is that the quantity  $\Phi'(Q(t))$  in the former is replaced with  $\Phi'(Q(t-1))$  in the latter. Thus, in the above definition, the surrogate loss on round  $t$  *does not* depend on the action  $a_t$  of the algorithm at the same round. This is an essential requirement in the bandit feedback setting as, unlike the full-information algorithms, standard adversarial MAB algorithms randomize their actions to estimate the unseen loss components (*e.g.*, using the inverse propensity score). This estimation process fails if the losses at round  $t$  also depend on the action of the policy at the same round. To summarize, Eqn. (55) implies that the surrogate loss  $\hat{l}_t$  is  $\mathcal{F}_{t-1}$  measurable, where  $\{\mathcal{F}_\tau\}_{\tau \geq 1}$  is the standard filtration.

Summing up inequalities (53) for  $1 \leq t \leq T$  and telescoping, we conclude that

$$\Phi(Q(T)) - \Phi(Q(0)) + V\text{Regret}_T \leq \text{Regret}'_T + e\Phi'(Q(T)) \sum_{t=1}^T c_t(a^*), \quad (56)$$

where, as before, we have used the non-decreasing property of  $\Phi'(\cdot)$  and the non-negativity of the consumption functions. As before,  $\text{Regret}_T$  and  $\text{Regret}'_T$  correspond to the regrets w.r.t. the original and surrogate losses, both of which are computed against the fixed randomized action  $a^*$ . Taking expectations of both sides of (56) w.r.t. the randomness of the policy and the offline benchmark  $a^*$ , we have:

$$\begin{aligned} \mathbb{E}\Phi(Q(T)) - \mathbb{E}\Phi(Q(0)) + V\mathbb{E}\text{Regret}_T & \stackrel{(a)}{\leq} \mathbb{E}\text{Regret}'_T + e\mathbb{E}\Phi'(Q(T))\mathbb{E}\left(\sum_{t=1}^T c_t(a^*)\right) \\ & \stackrel{(b)}{\leq} \mathbb{E}\text{Regret}'_T + e\mathbb{E}\Phi'(Q(T))B_T, \end{aligned} \quad (57)$$

where in step (a), we have used the fact that the benchmark  $a^*$  is independent of the online policy, and hence, is independent of  $Q(T)$ , and in step (b), we have used the fact that  $a^*$  satisfies the budget constraint in expectation (Eqn. (54)). Eqn. (57) is analogous to the regret decomposition inequality (13) in the full-information setting.

However, instead of using a full-information online learning policy, we now must use an adversarial MAB policy for learning the surrogate losses (55). Note that due to the factor  $\Phi'(Q(t-1))$  in the surrogate loss, a tight upper bound to the surrogate losses can not be obtained *a priori* as the evolution of the sequence  $\{Q(t)\}_{t \geq 1}$  depends on the online policy.

Because of the above reasons, we use an adaptive MAB policy, proposed by Putta and Agrawal [2022], which does not need any *a priori* upper bound on the magnitude of the losses, and yields a scale-free regret bound. On a high-level, the MAB algorithm proposed by Putta and Agrawal [2022] uses the FTRL sub-routine with a time-varying adaptive learning rate with the standard inverse-propensity score (IPS) estimator to estimate the unseen losses. For completeness, we give the pseudocode of the policy in Algorithm 3.

---

**Algorithm 3** Scale-Free Multi Armed Bandit
 

---

- 1: **Parameter initialization:**  $\eta_0 = K$ ,  $\gamma_0 = 1/2$
- 2: **Regularizer:**  $F(q) = \sum_{i=1}^K (f(q(i)) - f(1/K))$ , where  $f(x) = -\log(x)$
- 3: **Initialization:**  $p_1 = (1/K, \dots, 1/K)$
- 4: **for**  $t = 1$  **to**  $T$  **do**
- 5:   **Sampling Scheme:**  $p'_t = (1 - \gamma_{t-1})p_t + \frac{\gamma_{t-1}}{K}$
- 6:   Sample arm  $i_t \sim p'_t$  and see loss  $\tilde{\ell}_t(i_t)$ .
- 7:   **Estimation Scheme:**  $\tilde{\ell}_t(i) = \frac{\tilde{\ell}_t(i_t)}{p'_t(i_t)} \mathbf{1}(i_t = i), \forall i$ .
- 8:   Compute  $\gamma_t$  for next step:  $\gamma_t = \min(1/2, \sqrt{K/t})$
- 9:   Compute  $\eta_t = \frac{K}{1 + \sum_{s=1}^t M_s(\eta_{s-1})}$ , where

$$M_t(\eta) = \sup_{q \in \Delta_K} \left[ \tilde{\ell}_t^\top (p_t - q) - \frac{1}{\eta} \text{Breg}_F(q \| p_t) \right]$$

- 10:   **Find the next sampling distribution using FTRL:**

$$p_{t+1} = \arg \min_{q \in \Delta_K} \left[ F(q) + \eta_t \sum_{s=1}^t q^\top \tilde{\ell}_s \right]$$

- 11: **end for**
- 

**Theorem 11** (Putta and Agrawal [2022]). *The MAB policy described in Algorithm 3, when run with the sequence of loss vectors  $\{\hat{\mathbf{l}}_t\}_{t=1}^T$ , enjoys the following scale-free regret bound:*

$$\mathbb{E} \text{Regret}_T \leq 2 \left( 1 + \sqrt{K \sum_{t=1}^T \|\hat{\mathbf{l}}_t\|_2^2 + \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \sqrt{KT}} \right) \left( 2 + \log(1 + \|\sum_{t=1}^T \hat{\mathbf{l}}_t\|_\infty) \right), \quad (58)$$

where the expectation is taken over the randomness of the algorithm.

**Remarks:** Note that although the surrogate loss (55) on round  $t$  depends on the actions of the online algorithm up to round  $t-1$ , the benchmark  $a^*$ , as discussed above, is oblivious to the action of the algorithms and can be decided at the start of the play by an offline oracle. Since the action of the algorithm at round  $t$  is independent of the losses at round  $t$ , it is clear that we can upper bound the regret of the surrogate problem against the benchmark  $a^*$  using any oblivious MAB regret bound.

We now return to the proof of our main results. Using the adaptive regret bound from Eqn. (58) to the regret decomposition inequality in Eqn. (57) with the surrogate loss  $\hat{\mathbf{l}}_t$  defined in Eqn. (55), we obtain

$$\begin{aligned} & \mathbb{E} \Phi(Q(T)) - \mathbb{E} \Phi(Q(0)) + V \mathbb{E} \text{Regret}_T \\ & \leq 2 \mathbb{E} \left[ 1 + \sqrt{N \sum_{t=1}^T \|\hat{\mathbf{l}}_t\|_2^2 + \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \sqrt{KT}} \right] \left( 2 + \log(1 + T \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty) \right) + e \mathbb{E} \Phi'(Q(T)) B_T. \end{aligned} \quad (59)$$



Using the fact that  $\|\hat{\mathbf{l}}_t\|_2 \leq \sqrt{K}\|\hat{\mathbf{l}}_t\|_\infty$ , and  $\|\sum_t \hat{\mathbf{l}}_t\|_\infty \leq T \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty$ , we have the following bound

$$1 + \sqrt{N \sum_{t=1}^T \|\hat{\mathbf{l}}_t\|_2^2} + \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \sqrt{KT} \leq 1 + \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty K\sqrt{T} + \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \sqrt{KT} \leq 3 \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty K\sqrt{T}.$$

Further, in Lemma 13, we show that for our choice of the Lyapunov function  $\Phi(\cdot)$ , the parameter  $V$ , and using the trivial bound  $Q(T) \leq T$ , we have  $\max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \leq (e^2 T \log T)^{\log T}$ . Hence, we can upper bound the logarithmic pre-factor as follows:

$$2 + \log(1 + T \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty) \leq 2 + \log(2T \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty) \leq 2 \log(T \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty) \leq 3(\log T)^2.$$

Plugging in the above bounds in the regret decomposition inequality (59), we obtain

$$\mathbb{E}\Phi(Q(T)) - \mathbb{E}\Phi(Q(0)) + V\mathbb{E}\text{Regret}_T \leq 18 \max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty K\sqrt{T}(\log T)^2 + e\mathbb{E}\Phi'(Q(T))B_T. \quad (60)$$

Now, note that

$$|\hat{l}_t(a)| \leq V l_t(a) + e\Phi'(Q(t-1))c_t(a) \leq V + e\Phi'(Q(T)), \quad \forall t, a,$$

where, in the above, we have used the fact that  $l_t(a) \in [0, 1], c_t(a) \in [0, 1], \forall t, a$ , and the monotonicity of the function  $\Phi'(\cdot)$ . This yields

$$\max_{t \in [T]} \|\hat{\mathbf{l}}_t\|_\infty \leq V + e\Phi'(Q(T)).$$

Using the above bounds, Eqn. (60) simplifies to

$$\begin{aligned} \mathbb{E}\Phi(Q(T)) - \mathbb{E}\Phi(Q(0)) + V\mathbb{E}\text{Regret}_T &\leq 18(V + e\mathbb{E}\Phi'(Q(T)))K\sqrt{T}(\log T)^2 + e\mathbb{E}\Phi'(Q(T))B_T \\ &= 18VK\sqrt{T}(\log T)^2 + e(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}\Phi'(Q(T)). \end{aligned} \quad (61)$$

Finally, we choose a power-law Lyapunov function  $\Phi(x) = x^m$  with the exponent  $m = \log T$ , and initialize  $Q(0) = \log T$ . With these choices, inequality (61) yields

$$\mathbb{E}Q^m(T) + V\mathbb{E}\text{Regret}_T \leq 18VK\sqrt{T}(\log T)^2 + me(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}Q^{m-1}(T) + (\log T)^m. \quad (62)$$

We now analyze the above inequality for bounding both regret and the cumulative consumptions (CC).

**Bounding the Cumulative Consumption (CC):** Since the losses on each round are bounded by one, we trivially have  $\mathbb{E}\text{Regret}_T \geq -T$ . Plugging this in (62) yields

$$\begin{aligned} \mathbb{E}Q^m(T) &\leq 18VK\sqrt{T}(\log T)^2 + VT + me(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}Q^{m-1}(T) + (\log T)^m \\ &\leq 2 \max \left( 18VK\sqrt{T}(\log T)^2 + VT + (\log T)^m, me(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}Q^{m-1}(T) \right). \end{aligned} \quad (63)$$

If the first term is dominant in the above  $\max(\cdot)$  operator in Eqn. (63), we have

$$\begin{aligned} \mathbb{E}Q^m(T) &\leq 36VK\sqrt{T}(\log T)^2 + 2VT + 2(\log T)^m \\ &\stackrel{(a)}{\implies} (\mathbb{E}Q(T))^m \leq 36VK\sqrt{T}(\log T)^2 + 2VT + 2(\log T)^m \\ &\stackrel{(b)}{\implies} \mathbb{E}Q(T) \leq (36VK\sqrt{T}(\log T)^2)^{\frac{1}{m}} + (2VT)^{\frac{1}{m}} + 2\log T. \end{aligned} \quad (64)$$

where (a) follows from the convexity of the mapping  $x \mapsto x^m$  and applying Jensen's inequality and (b) follows from the fact that  $(a+b)^{1/m} \leq a^{1/m} + b^{1/m}$  for any  $m \geq 1, a \geq 0, b \geq 0$ .

Similarly, if the second term within the  $\max(\cdot)$  operator in (63) is dominant, we have

$$\mathbb{E}Q^m(T) \leq 2me(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}Q^{m-1}(T)$$

$$\begin{aligned}
 &\stackrel{(a)}{\implies} \mathbb{E}[Q^{m-1}(T)]^{\frac{m}{m-1}} \leq 2me(18K\sqrt{T}(\log T)^2 + B_T)\mathbb{E}Q^{m-1}(T) \\
 &\implies \mathbb{E}(Q^{m-1}(T))^{\frac{1}{m-1}} \leq 2me(18K\sqrt{T}(\log T)^2 + B_T) \\
 &\stackrel{(b)}{\implies} \mathbb{E}Q(T) \leq 2me(18K\sqrt{T}(\log T)^2 + B_T),
 \end{aligned} \tag{65}$$

where (a) follows from applying Jensen's inequality on the LHS to the convex map  $x \mapsto x^{\frac{m}{m-1}}$ , resulting in

$$\mathbb{E}(Q^m(T)) = \mathbb{E}[(Q^{m-1}(T))^{\frac{m}{m-1}}] \geq \mathbb{E}[Q^{m-1}(T)]^{\frac{m}{m-1}},$$

and (b) follows from the convexity of the map  $x \mapsto x^{m-1}$  and using Jensen's inequality.

Combining (64) and (65), we conclude that the cumulative consumption of the proposed policy is upper bounded as:

$$\mathbb{E}Q(T) \leq \max((36VK\sqrt{T}(\log T)^2)^{\frac{1}{m}} + (2VT)^{\frac{1}{m}} + 2\log T, 2me(18K\sqrt{T}(\log T)^2 + B_T)). \tag{66}$$

**Bounding the Regret:** Next, we start from (62) to bound the regret as follows. Transposing the term  $\mathbb{E}Q^m(T)$  to the right, we have

$$\mathbb{E}\text{Regret}_T \leq 18K\sqrt{T}(\log T)^2 + \frac{1}{V} \underbrace{\mathbb{E}[(me(18K\sqrt{T}(\log T)^2 + B_T) - Q(T))Q^{m-1}(T)]}_{\leq (me(18K\sqrt{T}(\log T)^2 + B_T))^m}.$$

The upper bound on the last term is obtained by considering two possible cases:

**Case I:**  $Q(T) > me(18K\sqrt{T}(\log T)^2 + B_T)$  : In this case, the last term is non-positive.

**Case II:**  $0 \leq Q(T) \leq me(18K\sqrt{T}(\log T)^2 + B_T)$  : In this case, we simply use the upper bound  $Q(T) \leq me(18K\sqrt{T}(\log T)^2 + B_T)$  to bound  $Q^{m-1}(T)$ .

Finally, choosing the parameter  $V = \frac{(me(18K\sqrt{T}(\log T)^2 + B_T))^m}{36K\sqrt{T}(\log T)^2}$ , and  $m = \log T$ , the regret can be bounded as follows:

$$\mathbb{E}\text{Regret}_T \leq 54K\sqrt{T}(\log T)^2.$$

Substituting the above parameter choices in (66), the cumulative consumption can be bounded as follows:

$$\mathbb{E}Q(T) \leq e^2(18K\sqrt{T}(\log T)^3 + B_T \log T).$$

This concludes our analysis of the proposed algorithm in the bandit setting.  $\square$

### Supporting Lemmas:

**Lemma 12.** For the power law potential  $\Phi(x) = x^m$  with  $m = \log T$  and  $Q(0) = \log T$  the following holds

$$\Phi'(Q(t-1) + 1) \leq e\Phi'(Q(t-1)).$$

*Proof.* We have

$$\begin{aligned}
 \Phi'(Q(t-1) + 1) &= m(Q(t-1) + 1)^{m-1} \\
 &= mQ^{m-1}(t-1)\left(1 + \frac{1}{Q(t-1)}\right)^{m-1} \\
 &\stackrel{(a)}{\leq} mQ^{m-1}(t-1)\left(1 + \frac{1}{Q(0)}\right)^{m-1} \\
 &\stackrel{(b)}{\leq} e\Phi'(Q(t-1)).
 \end{aligned}$$

where (a) follows because  $Q(0) \leq Q(t-1)$  and (b) follows because  $Q(0) = \log T = m$  and  $(1 + \frac{1}{m})^{m-1} < (1 + \frac{1}{m})^m \leq e$ .  $\square$

**Lemma 13.** *The magnitude of the surrogate loss  $\hat{l}_t(a) = Vl_t(a) + e\Phi'(Q(t-1))c_t(a)$  can be uniformly bounded as follows:*

$$\max_{t \in [T]} \|\hat{l}_t\|_\infty \leq (e^2 T \log T)^{\log T}$$

where  $V = \frac{(me(18K\sqrt{T}(\log T)^2 + B_T))^m}{36K\sqrt{T}(\log T)^2}$  and  $\Phi(x) = x^m$  with  $m = \log T$ .

*Proof.*

$$\max_{t \in [T]} \|\hat{l}_t\|_\infty \leq V + e\Phi'(Q(T)) \leq (me(18K\sqrt{T}(\log T)^2 + B_T))^m + em(T + \log T)^m \leq (e^2 T \log T)^{\log T}$$

where the second last inequality follows because  $V < (me(18K\sqrt{T}(\log T)^2 + B_T))^m$ ,  $B_T \leq T$  (the sum of constraint violations is bounded by  $T$ ) and  $Q(T) \leq T + \log T$ .  $\square$