Cautious Regret Minimization: Online Optimization with Long-Term Budget Constraints

Nikolaos Liakopoulos [1 2] **Apostolos Destounis** [1] **Georgios Paschos** [1] **Thrasyvoulos Spyropoulos** [2]
Panayotis Mertikopoulos [3]

Abstract

We study a class of online convex optimization problems with long-term budget constraints that arise naturally as reliability guarantees or total consumption constraints. In this general setting, prior work by Mannor et al. (2009) has shown that achieving no regret is impossible if the functions defining the agent's budget are chosen by an adversary. To overcome this obstacle, we refine the agent's regret metric by introducing the notion of a "K-benchmark", i.e., a comparator which meets the problem's allotted budget over any window of length K. The impossibility analysis of Mannor et al. (2009) is recovered when $K = T$; however, for $K = o(T)$, we show that it *is* possible to minimize regret while still meeting the problem's long-term budget constraints. We achieve this via an online learning algorithm based on *cautious online Lagrangian descent* (COLD) for which we derive explicit bounds, in terms of both the incurred regret and the residual budget violations.

1. Introduction

Consider the following bare-bones model of an online portfolio management problem: At each stage $t = 1, 2, \dots$, an investor places an investment over d diverse goods. This investment is modeled as a vector of unit bid prices $x_t = (x_t^1, \dots, x_t^d)$, with each x_t^i representing the amount of money the investor is willing to pay for a unit of the i-th good – for instance, for an advertiser requesting ad space from different publishers, x_t^i would denote the cost-per-click (CPC). Based on the performance of each individual asset (e.g., the number of clicks), the investor pays a total cost as a function of x_t^i and a performance parameter p_t^i of the i-th asset, and concurrently collects a corresponding reward w_t from their investment portfolio.

In our running example of online ad placement, the agent's utility $u_t(x_t; p_t)$ would be typically assumed concave in the agent's investment vector (to model diminishing returns), but otherwise stage-dependent, reflecting the variability of the performance parameter p_t^i of each investment. As such, utility maximization in this setting leads to an *online optimization problem* with the goal of maximizing the total reward $\sum_{t=1}^T u_t(x_t)$ accrued over T stages.

A considerable complication arises in this problem when the agent also needs to balance their total investment against an allotted budget (daily, monthly, or otherwise). In more detail, assume that the agent must meet a long-term budget constraint of the form $\sum_{t=1}^T c_t \leq b_T$, where $c_t = \langle p_t, x_t \rangle$ denotes the total expenditure of the consumer at time t. Since the performance parameters p_t are not known ahead of time (nor can they be assumed to follow a stationary probability law), techniques based on dynamic programming and optimal control cannot be applied in this context.

More generally, long-term budget constraints of this type can be formulated as $\sum_{t=1}^T g_t(x_t) \leq 0$ (1), where g_t is a convex function representing the impact to the budget at time t. For instance, in our previous example, we have $g_t(x_t) = \langle p_t, x_t \rangle - b_T/T$, so it simply represents the target $\sum_{t=1}^T \langle p_t, x_t \rangle \leq b_T$. Importantly, these constraint functions are not only a priori unknown, but their evolution could even be adversarial: for instance, in online ad markets, competitors may click on ads to deplete their rivals' advertising budget, fraudulent publishers may attempt to manufacture revenue by increasing the click-through-rate (CTR) without legitimate buying intent, as shown in The Economist (2005).

In this way, we obtain the following archetype of an *online optimization problem with long-term budget constraints:*

$$\begin{aligned} &\text{minimize} && \textstyle\sum_{t=1}^T f_t(x_t), \\ &\text{subject to} && \textstyle\sum_{t=1}^T g_t(x_t) \leq 0. \end{aligned} \tag{2}$$

In the above, the problem's loss and constraint functions (f_t and g_t respectively) are assumed convex and Lipschitz

[1]Paris Research Center, Huawei Technologies, Paris, France [2]EURECOM, Sophia-Antipolis, France [3]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, Grenoble, France. Correspondence to: Nikolaos Liakopoulos <liakopoulosnp@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Paper	Constraint	Benchmark window	Regret	Residual[(a)]	Parameter	Assumption
Yuan & Lamperski (2018)	Fixed	T	$\mathcal{O}(\sqrt{T}+T/V)$	$\mathcal{O}(\sqrt{VT})$	$V \in (1,T)$	-
Yu et al. (2017)	Stochastic	T	$\mathcal{O}(\sqrt{T})$	$\mathcal{O}(\sqrt{T})$		St. Slater[(b)]
Neely & Yu (2017)	Adversarial	1	$\mathcal{O}(\sqrt{T})$	$\mathcal{O}(\sqrt{T})$		Slater[(c)]
Sun et al. (2017)	Adversarial	1	$\mathcal{O}(\sqrt{T})$	$\mathcal{O}(T^{3/4})$		-
Mannor et al. (2009)	Adversarial	T	$\Omega(T)$	$o(T)$		-
This paper	Adversarial	K	$\mathcal{O}(\sqrt{T}+KT/V)$	$\mathcal{O}(\sqrt{VT})$	$V \in (K,T)$	-

Table 1. State of the art results in OCO with long-term budget constraints. All papers assume f_t, g_t are convex and Lipschitz continuous. (a) Residual refers to the long-term budget constraint violation. (b) Stochastic Slater assumes there exists an action $x_* \in \mathcal{X}$ such that $\mathbb{E}\left[g_t(x_*)\right] < 0$ for all t. (c) Slater assumes there exists an action $x_* \in \mathcal{X}$ such that $g_t(x_*) < 0$ for all t.

on their definition domain, but are otherwise arbitrary. Our aim in the rest of this paper will thus be to *a*) quantify the trade-offs between regret minimization and budget violations in this setting; and *b*) propose online algorithms capable of achieving the problem's minimization objective while exceeding the allotted budget by a minimal amount.

1.1. Related work

When the agent's action are constrained to lie on a fixed convex set $\mathcal{X}$ – i.e., in the absence of long-term constraints – standard methods based on online gradient/mirror descent enjoy an $\mathcal{O}(\sqrt{T})$ bound on the incurred regret, as in Zinkevich (2003); Shalev-Shwartz (2012), which is well-known to be optimal in this setting, see Abernethy et al. (2008).

Beyond this classic regret minimization framework, the first work to examine online optimization problems with long-term budget constraints is Mahdavi et al. (2012), where the constraint functions are the same for all t, i.e., the constraint is of the form $\sum_t g(x_t) \leq 0$. For deterministic, non-adversarial constraints of this form, Mahdavi et al. (2012) achieved $\mathcal{O}(\sqrt{T})$ regret and an $\mathcal{O}(T^{2/3})$ constraint residual (defined here as $\sum_{t=1}^T g(x_t)$). These bounds were subsequently improved by Jenatton et al. (2016) to $\mathcal{O}(T^{\max[\beta,1-\beta]})$ and $\mathcal{O}(T^{1-\beta/2})$ respectively, with $\beta \in (0,1)$ a free parameter, using varying stepsizes and regularization parameters. Finally, Yuan & Lamperski (2018) generalized these works by proving the same regret and $\mathcal{O}(T^{1-\beta})$ constraint residual for the tighter constraints $\sum_{t=1}^T \left([g(x_t)]^+\right)^2$. All above works use roughly the same algorithm: at each round the dual variable of the long-term budget constraint is updated with g, and the algorithm takes a step along the (online) subgradient of the instantaneous augmented Lagrangian.

An alternative model to capture budget constraints (constraints on the resources consumed over time) is the framework of multi-armed bandits with concave rewards and convex knapsacks, presented in Agrawal & Devanur (2014). In this framework, the decision space is a discrete set of arms and pulling an arm i_t at round t generates a vector v_t according to some fixed distribution. The objective is to satisfy a constraint on the average outcome vector $\bar{v}_t = \frac{1}{T}\sum_{t=1}^T v_t$ while maximizing a reward function $f(\bar{v}_T)$. They extend the setting of Badanidiyuru et al. (2013), where pulling an arm consumes some resource, stopping the process if the resource has been depleted, and prove near-optimal bounds for regret and constraint residual.

Moving on to time-varying constraint functions g_t, Yu et al. (2017) examined the case where the losses f_t are adversarial but g_t are stochastic (non-adversarial), drawn from some unknown (but otherwise *stationary*) distribution. They define the regret with respect to the best action that satisfies $\mathbb{E}\left[g_t(x_*)\right] < 0$ at each round. Assuming such an action exists, a combination of OGD with a virtual queue playing the role of a dual relaxation variable guarantees a bound $\mathcal{O}(\sqrt{T})$ on both regret and constraint residual.

In a concurrent line of work by Chen et al. (2017) and Cao & Liu (2018), the performance of an online optimization algorithm is compared to that of an instantaneous minimizer of f_t subject to $g_t(x) \leq 0$. As expected, regret guarantees against this dynamic comparator require very strong assumptions – for instance, that the variation of two consecutive constraint functions is bounded by the slack achieved by a static action over all constraint functions (the existence of such an action is already difficult to guarantee). [1]

Our aim in this paper is to study online convex optimization problems with time-varying (and possibly *adversarial*) long-term budget constraints. In this general setting, prior work by Mannor et al. (2009) provided a simple counterexample showing that the regret of any causal algorithm is lower bounded as $\Omega(T)$; as such, achieving no regret is impossible if the functions defining the agent's budget are chosen by an adversary. More recently, Neely & Yu (2017) proved

[1] We should mention here that our work can also be extended in this direction using the work of Besbes et al. (2015). However, because we want to focus on regret minimization with minimal assumptions, we only use static comparators throughout.

that a combination of OGD with a virtual queue can indeed provide no regret, compared to a static action that is strictly feasible *for all* functions g_t, i.e., x_* must satisfy $g_t(x_*) < 0$ for all $t = \{1, \dots, T\}$. Sun et al. (2017) further proved that mirror descent on a properly augmented Lagrangian can achieve no regret even if the static action is just feasible. However, especially in an adversarial setting, this assumption might not be easy to achieve because, even a small degree of residual constraint violation injected by the adversary could disqualify *any* comparator action. More importantly, since x_* is artificially constrained in this way, the obtained regret guarantee can be fairly loose.

A summary of the state of the art can be found in Table 1.

1.2. Our contributions

Our overarching objective is to examine the various tradeoffs between regret minimization and long-term budget constraint violations. Our first contribution in this direction is the introduction of a refined regret metric which compares the agent's incurred losses to those of a "K-benchmark", i.e., a comparator which meets the problem's allotted budget over any window of length K. In other words, a K-benchmark comparator satisfies:

$$\sum_{\tau=t}^{t+K-1} g_\tau(x) \leq 0, \quad \forall t \in \{1, \dots, T-K+1\} \tag{3}$$

and the "regret over a K-benchmark" compares the loss accrued by an online algorithm to that of the best K-benchmark in hindsight.

By varying K, this refined regret metric provides sufficient flexibility to study the difficult question of adversarial long-term budget constraints. Specifically, letting $K = T^\kappa$ for some $\kappa \in [0, 1]$, we recover the result of Mannor et al. (2009) for $\kappa = 1$ (i.e., every causal algorithm is regretful in the long run). At the other end of the spectrum, for $\kappa = 0$ – i.e., $K = \Theta(1)$ – we recover the framework of Neely & Yu (2017), where no regret *is* achievable. In this way, we are led to the following fundamental questions: (*i*) What is the largest κ for which "no regret over K-benchmark" can be achieved? and (*ii*) For a given $\kappa \in (0, 1)$, what is the regret guarantee for a given tolerance on the residual constraint violation?

Building on prior work by Jenatton et al. (2016), Yuan & Lamperski (2018), Yu et al. (2017) and Neely & Yu (2017), we attack the first question by means of an online optimization policy which we call *cautious online Lagrangian descent* (COLD). As we show in the sequel, COLD achieves no regret over any benchmark of length $K = T^\kappa$, for all $\kappa \in [0, 1)$. Since no regret is impossible without further assumptions for $\kappa = 1$, our result closes the gap with respect to achieving no regret with adversarial long-term budget constraints. Finally, regarding the second question above, we show that the COLD algorithm can simultaneously achieve the tradeoffs

$$\underbrace{\mathcal{O}(KT/V + \sqrt{T})}_{\text{regret over } K\text{-benchmark}} \quad \text{and} \quad \underbrace{\mathcal{O}(\sqrt{VT})}_{\text{constraint residual}}$$

for any choice of $V \in [K, T)$. The "cautiousness parameter" V can be tuned by the optimizer and, in so doing, we derive the region of COLD relative to the trade-off between regret minimization and long-term residual constraint violation.

Our theoretical findings are also validated by a series of numerical experiments which suggest that increasing K – that is, enlarging the window over which the budget must be balanced – makes the K-benchmark guarantee tighter. Hence, proving no regret for large K results in tighter performance guarantees – an observation which is not a priori obvious in a bona fide adversarial setting.

2. OCO with long-term budget constraints

To motivate the formal setup of the problem under study, we begin by discussing in more detail the online ad placement problem presented in Section 1.

Specifically, assume that, at every round, an advertiser chooses an investment vector of bid prices $x_t = (x_t^1, \dots, x_t^d)$ over d different websites, with x_t^i denoting the cost-per-click (CPC) for the i-th website. It is implied that each website offers different deals for ad display with varying position prominence and frequency of display, and accordingly arranged prices per click. The ultimate cost of an investment in dollars is determined when the number of click(s) the ad receives (denoted here by p_t^i and measuring the performance of the corresponding investment) is revealed, and is equal to $\langle p_t, x_t \rangle$. In this setting, the values of p_t^i fluctuate in an unpredictable manner – for instance, following website popularity, viewer interest, and/or possible attacks by competitors who click on an ad without a legitimate intent to buy, but only to increase the cost to the advertiser. As a result, satisfying the monthly budget given by $\sum_t \langle p_t, x_t \rangle \leq b_T$ is an adversarial long-term budget constraint of the form Eq.(1).

The goal of the customer in this framework is to invest the available budget wisely. Specifically, the reward from ad display at site i also fluctuates unpredictably according to the website's popularity and the relation of the users to the advertised product (in some websites the value of a user click is higher since the user is more probable to become a customer). With these considerations in mind, the collected utility is given by an unknown concave function[2]

[2]The DR-submodular reward functions (Chen et al., 2018) could be an interesting generalization capturing diminishing returns, but incorporating constraints is still an open problem.

$u_t(x_t) = \sum_i u_t^i(x_t^i)$. The concavity of u_t reflects the diminishing returns for a fixed website characteristic (e.g. making the ad more visible will attract proportionally less extra viewers). Needless to say, this task is very challenging because of the unpredictable fluctuations of price and reward, but also because an early aggressive choice might consume the budget resulting in missing out on opportunities towards the end of the horizon.

2.1. Problem formulation and assumptions

To state the above in a more formal framework, we will focus on the online optimization problem Eq.(2) over a play horizon of $t = 1, \ldots, T$ rounds. In round t, the action $x_t \in \mathcal{X}$ incurs $f_t(x_t)$ loss and impacts the budget by the amount $g_t(x_t)$. Here, functions f_t and g_t are not required to be differentiable and $f_t'(x_t), g_t'(x_t)$ denote subgradients at x_t. To analyze this problem we require the following basic assumptions.

(A1) The set $\mathcal{X}$ is convex and compact with diameter D.

(A2) For all $t = 1, \ldots, T$ functions $f_t, g_t : \mathcal{X} \to \mathbb{R}$ are convex and Lipschitz, with $\|f_t'\|_2 \leq G$ and $\|g_t'\|_2 \leq G$.

(A3) For a given $K \leq T$, consider the set of all actions that maintain a balanced budget within all windows of K rounds:

$$\mathcal{X}_K = \left\{ x \in \mathcal{X} : \sum_{\tau=t}^{t+K-1} g_\tau(x) \leq 0,\ 1 \leq t \leq T-K+1 \right\}$$

We assume that $\mathcal{X}_K$ is non-empty.

Since $\mathcal{X}$ is compact and f_t, g_t Lipschitz, it follows that they are bounded, i.e., $|f_t(x)| \leq F$, $|g_t(x)| \leq F$, for all $x \in \mathcal{X}$.

Assumptions A.1–A.2 are the blanket assumptions of all aforementioned OCO papers. A.3 is essential in order to define the regret metric that we use, and is significantly less stringent than the nonempty interior Slater assumption $\cap_t \{x : g_t(x) < 0\}$ of Neely & Yu (2017). For example, if g_t are nonnegative the Slater assumption cannot hold. This case appears in problems where we want to ensure that a rate of failures does not cross a threshold, as in Yuan & Lamperski (2018).

2.2. Performance metric

We classify algorithms based on how they fare with respect to the constraint and the aggregate loss.

2.2.1. Feasibility

Regarding the constraint Eq.(1), we take the common approach in the OCO literature, that of a relaxed notion of feasibility.

Definition 1 (Asymptotic feasibility)**.** *An algorithm is asymptotically feasible if it satisfies:*

$$Ctr(T) \triangleq \sum_{t=1}^{T} g_t(x_t) = o(T).$$

An asymptotically feasible algorithm has the desirable property that it learns to produce a vanishingly small *constraint residual* $\sum_{t=1}^{T} g_t(x_t)$ over a large horizon T, i.e., we have $\sum_{t=1}^{T} g_t(x_t)/T \to 0$ as $T \to \infty$, and hence produces an asymptotically feasible solution of Eq.(2).

2.2.2. Regret over K-benchmark

The algorithmic efficiency is measured in OCO with the *static regret*: the aggregate loss difference between the algorithm and that of a benchmark action with hindsight. In our work, however, we encounter a further complication that does not appear in the literature. The appropriate regret definition must clarify how the benchmark action will behave with respect to the long-term budget constraint. To this end, we introduce the following family of benchmarks.

Definition 2 (K-benchmark)**.** *Fix a $K \in \{1, \ldots, T\}$. The K-benchmark x_*^K is an action that satisfies:*

$$x_*^K \in \operatorname*{argmin}_{x \in \mathcal{X}_K} \sum_{t=1}^{T} f_t(x), \tag{4}$$

where $\mathcal{X}_K$ is defined in A.3.

This allows us to extend the definition of regret in the following manner.

Definition 3 (Regret of x_t over x_*^K)**.** *Fix $K \in \{1, \ldots, T\}$, and suppose x_*^K is a K-benchmark. The regret of x_t over x_*^K is defined to be:*

$$R_K(T) \triangleq \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_*^K).$$

Remark 1: If an asymptotically feasible online algorithm has *no regret* over x_*^K, it follows that the average losses $R_K(T)/T \to 0$ as $T \to \infty$, which implies that our policy approximates the benchmark x_*^K under any sequence of functions, while additionally asymptotically satisfying the long-term budget constraint.

Remark 2: The actual guarantee provided by the novel regret criterion depends on K. As K increases, actions in $\mathcal{X}_K$ must balance the budget in longer periods, and therefore become more aggressive. In fact, it can be checked that $\mathcal{X}_1 \subseteq \mathcal{X}_K$, hence $x_*^1 \leq x_*^K$ for all K. Consequently, a no regret guarantee over x_*^K is tighter than a no regret guarantee over x_*^1; section 5 illustrates this with a numerical example. We are, therefore, motivated to prove no regret for as large K as possible.

Interestingly, however, for $K > 1$ the K-benchmarks are not necessarily monotonic in K as the next example shows.

Example 1 (Non-monotonicity of K-benchmark)**.** Consider the ad display example with only one website. The K-benchmark is the largest action x constrained to the use of Kb_T/T budget within every K-round window. Given hindsight and prices $p_1, \ldots, p_T$, and assuming non-increasing f_t (i.e. more investment means larger utility and smaller loss), we have

$$x_*^K = \frac{Kb_T}{T \max_{t=0,\ldots,T-K-1} \sum_{\tau=t+1}^{t+K} p_\tau}.$$

Consider the instance where $T = 3$, $p = (10, 0, 8)$, and $b_T = 30$. We have $x_*^1 = 1, x_*^2 = 2, x_*^3 = 5/3$, and surprisingly $x_*^2 > x_*^3$. The latter is due to that the window of size 3 has higher values than the mean at the two extremes.

The lack of monotonicity is indicative of a powerful adversary who can tweak functions g_t to disturb the agent's algorithm in non-trivial ways. In spite of this complication, we present next a general algorithm that establishes no regret over x_*^K for any $K = o(T)$. Additionally, although a strong monotonicity result can not be established, in the numerical section we observe an improvement trend with increasing K, verifying the intuition that a larger window allows the agent to handle its budget in a better manner.

3. The algorithm

The main idea behind handling the long-term budget constraints is to weigh their importance against the loss in a Lagrangian fashion. Specifically, consider a *regularized instantaneous Lagrangian* for problem Eq.(2) that takes the following form in round t:

$$L_t(x, Q(t)) = V f_t(x) + Q(t) g_t(x) + \alpha \|x - x_{t-1}\|^2, \quad (5)$$

where *a*) V is a configurable *cautiousness parameter*; *b*) $Q(t)$ is a virtual queue that plays the role of the Lagrangian multiplier; *c*) the term $\|x - x_{t-1}\|^2$ is a L_2 regularizer that smoothens the differences between consecutive actions; and *d*) α is the strength of the regularization.

The main difference between Eq.(5) and traditional Lagrangian relaxations is that the cautiousness parameter V can be used to control the tradeoff between regret and constraint residual (smaller V makes the algorithm more cautious); likewise, the regularization parameter α can be tuned to enhance the algorithm's robustness to fluctuations.

To estimate the value of the (otherwise unknown) functions f_t and g_t, we will employ their linear surrogates:

$$\hat{f}_t(x) \triangleq f_{t-1}(x_{t-1}) + \langle f'_{t-1}(x_{t-1}), x - x_{t-1} \rangle, \quad (6)$$

$$\hat{g}_t(x) \triangleq g_{t-1}(x_{t-1}) + \langle g'_{t-1}(x_{t-1}), x - x_{t-1} \rangle. \quad (7)$$

Let $L'_t(x, Q)$ denote the subgradient of $L_t(x, Q(t))$ at x. Plugging the above surrogate terms in Eq.(5) and using basic subgradient algebra, we get:

$$L'_t(x, Q) = V f'_{t-1}(x_{t-1}) + Q(t) g'_{t-1}(x_{t-1}) + 2\alpha(x - x_{t-1})$$

Our algorithm is designed to compute a stationary point of L_t in round t, and then project it on $\mathcal{X}$, while updating queue $Q(t+1)$ with the surrogate $\hat{g}_t(x_t)$. The latter is in fact the subgradient of $L_t(x, Q)$ with respect to Q.

Cautious Online Lagrangian Descent (COLD)

For $t = 1, \ldots, T$:

$$x_t = \Pi_{\mathcal{X}} \left[x_{t-1} - \frac{V f'_{t-1}(x_{t-1}) + Q(t) g'_{t-1}(x_{t-1})}{2\alpha} \right] \quad (8)$$

$$Q(t+1) = [Q(t) + \hat{g}_t(x_t)]^+. \quad (9)$$

with initialization $Q(1) = 0$, $x_0 \in \mathcal{X}$, and where:

- $\Pi_{\mathcal{X}}[.]$ denotes the Euclidean projection on set $\mathcal{X}$,
- V is the configurable cautiousness parameter,
- f'_{t-1}, g'_{t-1} are the subgradient vectors in round $t-1$,
- $Q(t)$ is a virtual queue that is updated according to Eq.(9), and it is called the *predictor queue*,
- α is the configurable regularization strength parameter,
- $\hat{g}_t(x_t)$ is the surrogate of $g_t(x_t)$ from Eq.(7),
- $[.]^+$ is $\max\{., 0\}$,

We remark that if we fix $Q(t) = 0, \forall t$, COLD reduces to the OGD of Zinkevich (2003) with stepsize $V/2\alpha$, which however would fail to address the long-term budget constraints. In what follows, we provide the logical steps that are used in the design of COLD, as well as to prove its performance guarantees. The same steps can be used to derive variations of COLD for different problems.

3.1. Regularized drift plus loss framework

The framework is inspired by the unification of two theories, namely the theory of stochastic network optimization, explained in Neely (2010a), that handles time-average constraints by stabilizing virtual queues; and the standard framework of OCO, described in Zinkevich (2003); Shalev-Shwartz (2012); Belmega et al. (2018).

Our mathematical analysis is based on an instantaneous metric called *drift plus loss plus smoothness* (DPLPS), which at each round measures the quality of an action by weighing three competing factors: (i) the *quadratic Lyapunov drift* of the predictor queue (which reflects the urgency of the constraint), (ii) the predicted instantaneous loss, and (iii) the L_2 regularizer to smoothen the changes in the sequence of actions. The values of all three above depend on the action x_t, and we will show that our algorithm arises as the action that minimizes an upper bound of the DPLPS. The remaining of this subsection provides further detail.

We first define the quadratic *Lyapunov drift* as the change in

the quadratic predictor queue length after action x_t is taken:

$$\Delta(x_t) \triangleq \frac{1}{2}[Q^2(t+1) - Q^2(t)]. \quad (10)$$

Since x_t determines $\hat{g}_t(x_t)$, it implicitly affects $\Delta(x_t)$ via $Q(t+1)$, see Eq.(9). By taking actions to minimize the drift $\Delta(x_t)$ we may keep the queue length small at the end of the horizon. We will then show that a bound on $Q(T)$ can be manipulated into proving asymptotic feasibility. In summary, the minimization of $\Delta(x_t)$ at each round pushes towards actions that satisfy the long-term budget constraint.

From the literature of stochastic network optimization (Neely (2010a)), the drift can be combined with the instantaneous loss (here adapted to its surrogate from Eq.(6)): $\Delta(x_t) + V\hat{f}_t(x_t)$. Minimizing this weighted metric pushes towards actions that simultaneously satisfy the long-term budget constraints and achieve low aggregate loss in the stochastic setting, see Neely (2010b). However, such algorithms have a bang-bang behavior, oscillating between extreme actions, which is inappropriate for many real applications. Therefore, in this work we further add the L_2 regularizer (following the methodology of Neely & Yu (2017)), which brings us to the DPLPS metric:

$$DPLPS(x_t) \triangleq \Delta(x_t) + V\hat{f}_t(x_t) + \alpha||x_t - x_{t-1}||_2^2. \quad (11)$$

With some work on the definition of the Lyapunov drift (see also Lemma 4.2 in Georgiadis et al. (2006)), we have:

$$\Delta(x_t) \leq B + Q(t)\hat{g}_t(x),$$

where $B \triangleq (F+GD)^2/2$ is a constant. Therefore, we get

$$DPLPS(x) \leq \underbrace{B + V\hat{f}_t(x) + Q(t)\hat{g}_t(x) + \alpha||x - x_{t-1}||_2^2}_{r_t(x)}.$$

Observe that the term $r_t(x)$ equals Eq.(5), plus a constant B. The following Lemma proves formally that $r_t(x)$ is minimized by COLD at each round.

Lemma 1 (DPLPS bound minimizer)**.** *COLD minimizes $r_t(x)$ at each round.*

Proof. First, observe that by definition our policy is:

$$x_t \triangleq \Pi_{\mathcal{X}}\left[x_{t-1} + \frac{H_t}{2\alpha}\right],$$

where H_t is the following constant:

$$H_t \triangleq Vf'_{t-1}(x_{t-1}) + Q(t)g'_{t-1}(x_{t-1}).$$

It suffices to show that this projection actually returns a minimizer of $r_t(x)$.

We have that, $\langle H_t, x_t - x_{t-1}\rangle = Q(t)\langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle + V\langle f'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle$. Hence:

$$\begin{aligned}
x_t &= \operatorname*{argmin}_{x\in\mathcal{X}} \{r_t(x)\} \\
&\overset{(a)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \{\langle H_t, x - x_{t-1}\rangle + \alpha||x - x_{t-1}||_2^2\} \\
&\overset{(b)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\langle H_t, x - x_{t-1}\rangle + \alpha||x - x_{t-1}||_2^2 + \frac{H_t^2}{4\alpha}\right\} \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \left|\left|\frac{H_t}{2\sqrt{\alpha}} + \sqrt{\alpha}(x - x_{t-1})\right|\right|_2^2 \\
&= \operatorname*{argmin}_{x\in\mathcal{X}} \left|\left|x - \left(x_{t-1} - \frac{H_t}{2\alpha}\right)\right|\right|_2^2 \overset{(c)}{=} \Pi_{\mathcal{X}}\left[x_{t-1} - \frac{H_t}{2\alpha}\right].
\end{aligned}$$

In (a) we discard the terms that do not depend on x. In (b) we add a constant term that completes the square norm but does not change the minimizer. Finally, (c) is the definition of Euclidean projection on set $\mathcal{X}$ completing the proof. □

This methodology constitutes a powerful framework, where for a new problem we may define appropriate virtual queues for the constraints, determine the corresponding DPLPS metric, and then extract asymptotically feasible online algorithms by minimizing a bound on the DPLPS. In the technical proofs, we show how this bound minimization can be used to derive the performance guarantees of our algorithm.

4. Performance analysis

In this section we provide our main theoretical results, which characterize the performance of COLD algorithm. Recall that T is the horizon, V, α are configurable parameters, F, G are universal constants (see section 2.1), D is the diameter of set $\mathcal{X}$, and $B = (F+GD)^2/2$.

Proposition 1 (COLD performance)**.** *If the assumptions in Sec. (2.1) are satisfied, and actions are taken according to the COLD algorithm, the constraint residual is bounded by:*

$$\begin{aligned}
Ctr(T) \triangleq \sum_{t=1}^{T} g_t(x_t) \leq \Bigg\{2BKT + 4FVT + \frac{V^2G^2T}{\alpha} + \\
2\alpha D^2 + 2BK^2 + 2(T+1)BK\Bigg\}^{\frac{1}{2}} + \frac{GVT}{2\alpha} + \\
\frac{G^2\left[\sqrt{2BK} + \sqrt{4FV} + \sqrt{\frac{V^2G^2}{\alpha}} + \sqrt{2BK}\right]\frac{T^{\frac{3}{2}}+T}{\sqrt{2}}}{2\alpha} + \\
\frac{G^2\left[\sqrt{2\alpha D^2} + \sqrt{2BK^2} + \sqrt{2BK}\right]T}{2\alpha}
\end{aligned} \quad (12)$$

and the regret over the K-benchmark is bounded by:

$$\begin{aligned}
R_K(T) \triangleq \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_*^K) \leq \frac{BKT}{V} + \frac{G^2VT}{2\alpha} + \\
+ B\frac{(K+1)(2K+1)}{6V} + \frac{D^2\alpha}{V} + 2F(K-1).
\end{aligned} \quad (13)$$

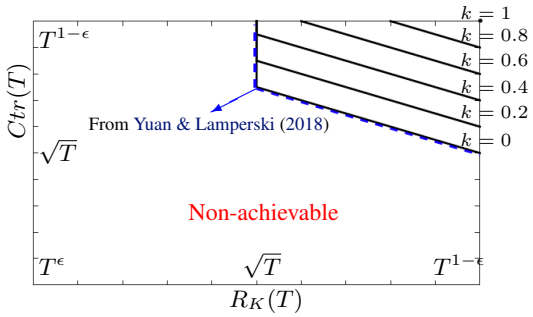


Figure 1. Achievable bounds for $K = T^k, k = 0 : 1 : 0.2$.

Based on the above fundamental bounds, we optimize the parameters V, α to get a region of achievable asymptotic laws, such that both the constraint residual and the regret are $o(T)$. In general, we may choose the value of parameter V in the range (K, T) and different choices provide different tradeoffs. For instance, choosing V close to K makes the algorithm cautious and provides the best guarantees on the constraint residual, while choosing it close to T makes the algorithm to aggressively pursue the best regret.

Theorem 1 (Achievable tradeoffs)**.** *Fix $K \geq 1$ such that $K = o(T)$ (higher K makes the K-benchmark tighter). Choose some $V \in (K, T)$, and $\alpha = \max\{T, V\sqrt{T}\}$. Then, Eq.(12)-(13) simplify to:*

$$\underbrace{\mathcal{O}(KT/V + \sqrt{T})}_{\text{regret over } K\text{-benchmark}} \quad \text{and} \quad \underbrace{\mathcal{O}(\sqrt{VT})}_{\text{constraint residual}}$$

Furthermore, suppose $K = T^{1-\epsilon}$ for some small $\epsilon > 0$ and choose $V = T^{1-\frac{\epsilon}{2}}$, and $\alpha = V\sqrt{T}$. Then, Eq.(12)-(13) simplify to:

$$\underbrace{\mathcal{O}(T^{1-\frac{\epsilon}{2}})}_{\text{regret over } T^{1-\epsilon}\text{-benchmark}} \quad \text{and} \quad \underbrace{\mathcal{O}(T^{1-\frac{\epsilon}{4}})}_{\text{constraint residual}}$$

In Fig.(1) we illustrate the results of Th.(1): the black curves indicate the Pareto frontier for different values of $K = T^k$ so all the values north-east of the frontier are achievable. As $K \to T$, the Pareto frontier vanishes to the north-east corner point $\mathcal{O}(T^{1-\epsilon/2}), \mathcal{O}(T^{1-\epsilon/4})$. The blue dotted line shows the tradeoffs achieved by Yuan & Lamperski (2018), which interestingly coincide with our case of $K = 1$, though we mention that they address fixed (non-adversarial) constraints. Finally, note that for $\epsilon = 1$, we retrieve the result of Sun et al. (2017). On the other hand, the point $\mathcal{O}(\sqrt{T}), \mathcal{O}(\sqrt{T})$ achieved by Neely & Yu (2017) for $K = 1$ is not part of our achievable guarantees; we attribute this gap to the stricter Slater assumption studied by Neely & Yu (2017).

4.1. Outline of the proofs

The technical proofs of Prop.(1) and Th.(1) are deferred to the appendix (in supplemental material) due to space limitations. Here, we provide a brief outline.

The COLD algorithm is designed to minimize the DPLPS, hence one can directly compare it to the 1-benchmark, as for example in Neely & Yu (2017). The novel element in our analysis is that we compare K steps of DPLPS of COLD to the DPLPS of the K-benchmark (Def.(2)). This comparison forms the basis of our analysis, and it is given in Lem.(8) and Cor.(2) in the appendix (in supplemental material).

First, based on the feasibility of the K-benchmark in windows of size K, we establish a bound on $Q(t)$ of COLD for any $t \in \{1, \dots, T+1\}$. The bound builds on the above-explained comparison between COLD and the K-benchmark. Then, the bound on $Q(t)$ is manipulated into proving the upper bound on the constraint residual Eq.(12). We note that an important part of the proof is to obtain a good upper bound of $\sum_{t=1}^{T} Q(t)$. A similar strategy is used in comparing the losses of COLD to those of the K-benchmark and proving the regret bound Eq.(13).

We mention that the bound on $Q(t)$ can be strengthened if we make the Slater assumption, i.e., assume the existence of a vector x_* such that $g_t(x_*) < -\eta, \ \ \forall t$. In particular, this is the approach taken by Neely & Yu (2017) achieving the point $(\mathcal{O}(\sqrt{T}), \mathcal{O}(\sqrt{T}))$ for $K = 1$. In our case, we would assume the existence of actions satisfying $\sum_{\tau=t}^{t+K-1} g_\tau(x) < -\eta, \ \ \forall t$, with which we could improve further the bounds on $Q(t)$, $\sum_{t=1}^{T} Q(t)$ and eventually our Proposition. In this work, we chose to present the most general bounds, and left this direction for future research.

Finally, regarding the proofs of Th.(1), our strategy is to restrict progressively the values of K, α, and V, such that both $Ctr(T)$ and $R_K(T)$ are $o(T)$. An optimization over parameters α, V provides the presented trade-offs.

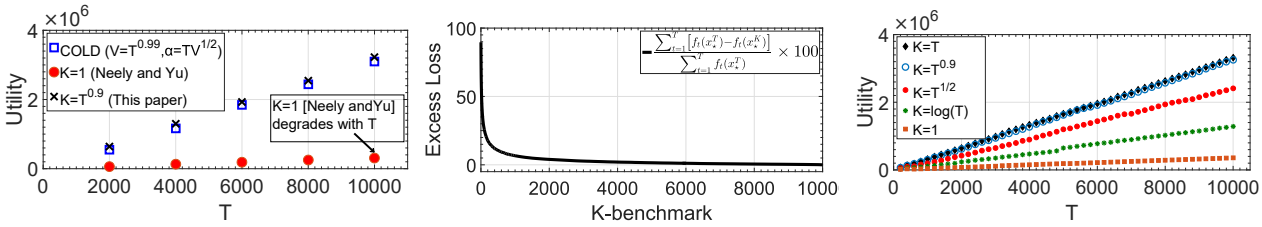
5. Numerical results

In this section we test COLD and the performance guarantees given by K-benchmarks on an instance of our example application of online ad placement. We also showcase the effect of the cautiousness parameter V on COLD's utility and constraint residual.

5.1. Accuracy of performance guarantee

We simulate a scenario with one website, where $x_t \in [0, \infty)$, $f_t(x_t) = -w_t x_t$, and $g_t(x_t) = p_t x_t - b_T/T$, where w_t, p_t are generated by exponential distributions $w_t \sim Exp(11)$ and $p_t \sim Exp(10)$. We run the experiment for different horizons $T = \{2000, 4000, \dots, 10000\}$, budget $b_T = 300T$, and parameters set to $\alpha = \max\{T, V\sqrt{T}\}$ and $V = T^{0.99}$ for each of the experiments. Fig.(2(a)) shows the utility (minus the loss) for the 1-benchmark (Neely & Yu (2017)), the $T^{0.9}$-benchmark (an instance of this paper) and finally the utility achieved by the COLD algorithm.

In Fig.(2(a)) we observe that COLD's utility is approximated much more accurately by the $T^{0.9}$-benchmark than by 1-benchmark. In fact, the approximation by the 1-



(a) COLD vs $K = 1, K = T^{0.9}$ (b) K-benchmark vs T-benchmark (c) Utility for different scaling of K

Figure 2. Simulations of the example of online ad placement. (a) COLD utility comparison versus 1-benchmark and $T^{0.9}$-benchmark. (b) Relative excess loss of K-benchmark compared to T-benchmark. (c) Utilities of K-benchmark for different values of K.

benchmark becomes even worse as T increases. Since proving no regret over K-benchmark essentially bounds the algorithm's losses, we conclude that the importance of the regret guarantee lies with how large K is.

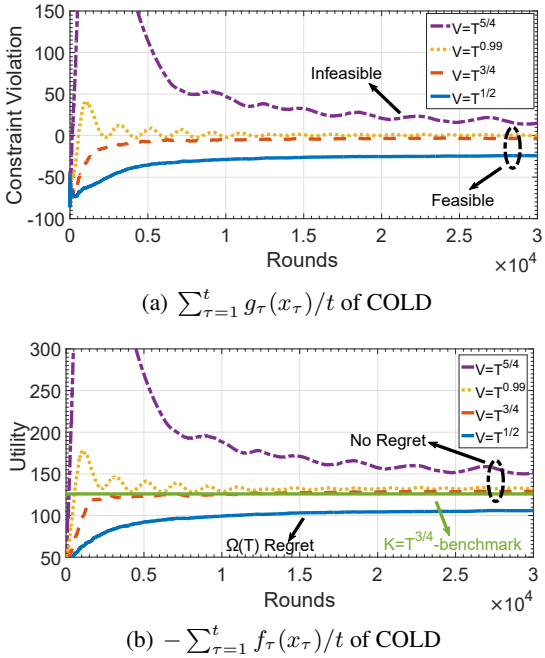
In a similar experiment, Fig.(2(b)) presents the relative excess loss of the K-benchmark with respect to the T-benchmark, for all values of $K = 1, \dots, T$. This relative excess loss is depictive of the approximation error of a K-benchmark with respect to the T-benchmark. We remind the reader that the excess loss is due to constraining the benchmark action to be feasible on a window that is $K < T$, and that the regret guarantees can be established for $K = o(T)$ only. The points are averages over 150 sample paths. Due to averaging over sample paths we observe a monotonous decrease of excess loss as K increases and more importantly that the 1-benchmark can have as much as 85% excess loss.

On Fig.(2(c)), we observe that a K-benchmark that scales sublinearly with the horizon ($K = o(T)$) has a much better approximation compared to a constant K, like $K = 1$, which is very pessimistic even for moderate values of T. All the above experiments support that intuition that taking K large produces a tighter regret guarantee.

5.2. Impact of the Cautiousness Parameter V

In this subsection we explore the effect of the cautiousness parameter V on the performance of COLD algorithm. On the same example as before, we choose $K = T^{3/4}$ and $V = \{T^{1/2}, T^{3/4}, T^{0.99}, T^{5/4}\}$. In our theoretical analysis, we have proven that V has to be greater than K to achieve no regret. On Fig.(3(b)) the running average utility for $V = T^{1/2}$ indeed fails to reach the benchmark. Furthermore, we can deduce from Fig.(3(a)) that, for $V > T$, the constraint residual is not sub-linear to the horizon T, which is in accordance to our bounds.

On the other hand when $K < V < T$, indeed the constraint residuals are sub-linear and the average utility approximates the utility of the K-benchmark. Increasing V up to T, one can observe in Fig.(3) the different tradeoffs between the regret and the constraint residual.



(a) $\sum_{\tau=1}^{t} g_\tau(x_\tau)/t$ of COLD

(b) $-\sum_{\tau=1}^{t} f_\tau(x_\tau)/t$ of COLD

Figure 3. Running averages of constraint residual and utility performance of COLD for different values of parameter V.

6. Conclusion

In this paper we study OCO with long-term budget constraints. In particular, we deal with the case where the constraints are adversarial, which captures the scenario where the long-term budget constraint must be addressed in the presence of poor prediction quality. We introduce the K-benchmark, which allows us to refine the regret metric used to provide performance guarantees for online algorithms. Although for $K = T$ prior work has established that no algorithm can provide no regret, we prove that the COLD algorithm achieves no regret for any $K = o(T)$. Our numerical results suggest that a K-benchmark with K large can provide a more accurate performance guarantee than the previous state of the art $K = 1$. Finally, we provide a new region of regret-constraint residual tradeoffs that characterize the performance of COLD in the general setting.

References

Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. Optimal Strategies and Minimax Lower Bounds for Online Convex Games. 2008.

Agrawal, S. and Devanur, N. R. Bandits with Concave Rewards and Convex Knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, 2014.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with Knapsacks. In *IEEE 54th Annual Symposium on Foundations of Computer Science*, 2013.

Belmega, E. V., Mertikopoulos, P., Negrel, R., and Sanguinetti, L. Online Convex Optimization and No-Regret Learning: Algorithms, Guarantees and Applications. 2018. URL http://arxiv.org/abs/1804.04529.

Besbes, O., Gur, Y., and Zeevi, A. Non-Stationary Stochastic Optimization. *Operations Research*, 63(5):1227–1244, 2015.

Cao, X. and Liu, K. J. R. Online Convex Optimization with Time-Varying Constraints and Bandit Feedback. *IEEE Transactions on Automatic Control*, 2018.

Chen, L., Harshaw, C., Hassani, H., and Karbasi, A. Projection-Free Online Optimization with Stochastic Gradient: From Convexity to Submodularity. In *ICML*, July 2018.

Chen, T., Ling, Q., and Giannakis, G. B. An Online Convex Optimization Approach to Proactive Network Resource Allocation. *IEEE Transactions on Signal Processing*, 65 (24):6350–6364, 2017.

Georgiadis, L., Neely, M. J., and Tassiulas, L. Resource Allocation and Cross-Layer Control in Wireless Networks. *Foundations and Trends® in Networking*, 2006.

Jenatton, R., Huang, J. C., and Archambeau, C. Adaptive Algorithms for Online Convex Optimization with Long-Term Constraints. ICML, pp. 402–411, 2016.

Mahdavi, M., Jin, R., and Yang, T. Trading Regret for Efficiency: Online Convex Optimization with Long Term Constraints. *Journal of Machine Learning Research*, 2012.

Mannor, S., Tsitsiklis, J. N., and Yu, J. Y. Online Learning with Sample Path Constraints. *Journal of Machine Learning Research*, 2009.

Neely, M. J. Stochastic Network Optimization with Application to Communication and Queueing Systems. *Synthesis Lectures on Communication Networks*, 2010a.

Neely, M. J. Universal Scheduling for Networks with Arbitrary Traffic, Channels, and Mobility. In *IEEE Decision and Control (CDC)*, pp. 1822–1829, 2010b.

Neely, M. J. and Yu, H. Online Convex Optimization with Time-Varying Constraints. *arXiv preprint arXiv:1702.04783*, 2017.

Shalev-Shwartz, S. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 2012.

Sun, W., Dey, D., and Kapoor, A. Safety-Aware Algorithms for Adversarial Contextual Bandit. In *ICML*, pp. 3280–3288, Sydney, Australia, August 2017.

The Economist. Truth in Advertising: "Click Fraud Poses a Threat to the Boom in Internet Advertising". 2005.

Yu, H., Neely, M. J., and Wei, X. Online Convex Optimization with Stochastic Constraints. NeurIPS, 2017.

Yuan, J. and Lamperski, A. Online Convex Optimization for Cumulative Constraints. NeurIPS, 2018.

Zinkevich, M. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. ICML, 2003.

7. Appendix with proofs

7.1. Proof of Proposition 1

In this subsection we prove that the COLD algorithm is a feasible policy with no regret against any K-benchmark with $K = o(T)$. We first prove an upper bound on the predictor queue $Q(t)$ at the beginning of round t. Then we use this to upper bound the constraint residual, which will, in turn be used to prove feasibility; picking x_t according to Eq.(8) will give sub-linear long-term constraint residual in a time horizon T (see 7.1.1). After having proved feasibility, we compare our policy with the K-benchmark and prove, in 7.1.2, that it achieves no regret for any $K = o(T)$. For proving these an intermediate technical result regarding the behavior of the sums of $Q(t)\hat{g}_t(x_t)$ over any window of K consecutive rounds for any sequence of actions $\{x_t\}$ is necessary; the statement and its proof are deferred to Lem.(8) and Cor.(2) at the end of this subsection.

7.1.1. Proof of the bound on the Constraint Residual

In the first part of the proof, we will prove the bound (12) on the constraint residual. First, we consider the quadratic Lyapunov Drift (defined as $\Delta(t) \triangleq \frac{1}{2}Q(t+1)^2 - \frac{1}{2}Q(t)^2$); this measures the impact of our policy on the predictor queue. We bound the Lyapunov drift by the following lemma:

Lemma 2 (Drift Upper Bound)**.** *The predictor queue drift is upper bounded by:*

$$\Delta(t) \leq B + Q(t)\hat{g}_t(x_t),\ \forall t \in \{0,\dots,T\}, \forall x \in \mathcal{X}. \tag{14}$$

where $B \triangleq \frac{(F+GD)^2}{2}$*, is a constant. Reminder:*

$$\hat{g}_t(x_t) \triangleq g_{t-1}(x_{t-1}) + \langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle.$$

Proof. Squaring the queue update function Eq.(9), since $\max\{0,x\}^2 \leq x^2$ we get:

$$\begin{aligned} Q(t+1)^2 &= \left([Q(t) + \hat{g}_t(x_t)]^+\right)^2 \\ &\leq \left(Q(t) + \hat{g}_t(x_t)\right)^2 \\ &\leq Q(t)^2 + 2Q(t)\hat{g}_t(x_t) + \hat{g}_t(x_t)^2. \end{aligned}$$

To continue we bound $\hat{g}_t(x_t)$:

$$\begin{aligned} |\hat{g}_t(x_t)| &\leq |g_{t-1}(x_{t-1}) + \langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle| \\ &\leq |g_{t-1}(x_{t-1})| + |\langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle| \\ &\leq F + GD. \end{aligned}$$

Using the above we get that:

$$\begin{aligned} Q(t+1)^2 &\leq Q(t)^2 + 2Q(t)\hat{g}_t(x_t) + \hat{g}_t(x_t)^2 \\ &\leq Q(t)^2 + 2Q(t)\hat{g}_t(x_t) + (F+GD)^2. \end{aligned}$$

Rearranging the terms, dividing by 2 and using the definitions of B and $\Delta(t)$ completes the proof. □

Corollary 1. *Plugging Lem.(2) into DPLPS definition, Eq.(11) we get the resulting upper bound:*

$$\begin{aligned} &\Delta(t) + V\hat{f}_{t-1}(x_{t-1}) + \alpha||x_t - x_{t-1}||_2^2 \leq B + \\ &Q(t)\hat{g}_t(x_t) + V\hat{f}_{t-1}(x_{t-1}) + \alpha||x_t - x_{t-1}||_2^2. \end{aligned}$$

We define $r_t(x)$*:*

$$r_t(x) \triangleq B + Q(t)\hat{g}_t(x) + V\hat{f}_{t-1}(x_{t-1}) + \alpha||x - x_{t-1}||_2^2.$$

Hence, $DPLPS(x_t) \leq r_t(x_t)$.

Lemma 3. *Our policy* x_t *minimizes* $r_t(x)$

$$x_t \in \operatorname*{argmin}_{x\in\mathcal{X}} \{r_t(x)\}.$$

Proof. First, observe that by definition our policy is:

$$x_t \triangleq \Pi_{\mathcal{X}}\left[x_{t-1} + \frac{H_t}{2\alpha}\right],$$

where H_t is the constant from past observations:

$$H_t \triangleq Vf'_{t-1}(x_{t-1}) + Q(t)g'_{t-1}(x_{t-1}).$$

It suffices to show that this projection actually returns a minimizer of $r_t(x)$. We have that, $\langle H_t, x_t - x_{t-1}\rangle = Q(t)\langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle + V\langle f'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle$.

$$\begin{aligned} x_t &= \operatorname*{argmin}_{x\in\mathcal{X}} \{r_t(x)\} \\ &\overset{(a)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \{\langle H_t, x - x_{t-1}\rangle + \alpha||x - x_{t-1}||_2^2\} \\ &\overset{(b)}{=} \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\langle H_t, x - x_{t-1}\rangle + \alpha||x - x_{t-1}||_2^2 + \frac{||H_t||^2}{4\alpha}\right\} \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \left|\left|\frac{H_t}{2\sqrt{\alpha}} + \sqrt{\alpha}(x - x_{t-1})\right|\right|_2^2 \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \left|\left|x - \left(x_{t-1} - \frac{H_t}{2\alpha}\right)\right|\right|_2^2 \\ &\overset{(c)}{=} \Pi_{\mathcal{X}}\left[x_{t-1} - \frac{H_t}{2\alpha}\right]. \end{aligned}$$

Here, in the (a) we discard the constant terms $Q(t)g_{t-1}(x_{t-1}) + Vf_{t-1}(x_{t-1})$ that do not depend on variable x. The (b) equality does not change the minimizer since we add a constant term that completes the square norm. Finally, (c) is the definition of euclidean projection on set $\mathcal{X}$, hence the last equality follows and completes the proof. □

Using the fact that our policy minimizes the DPLPS, by Lem.(3) and the bound on the drift by Lem.(2), we will

prove that the predictor queue is bounded; this result is an intermediate step towards proving that the constraint residuals are bounded. The idea is to compare our policy's DPLPS, with the upper bound for a policy that satisfies the constraint at a rolling window of K rounds, namely the K-benchmark policy x_*^K. The properties of the K -benchmark policy will help us bound the performance of our policy, that minimizes DPLPS.

Lemma 4 (Queue Bound)**.** *For any $t \in \{0, 1, \dots, T\}$, the queue is bounded as:*

$$Q(t+1) \leq \Big\{ 2BKt + 4FVt + \frac{V^2G^2t}{\alpha} + 2\alpha D^2 + 2BK^2 + 2BK(t+1) \Big\}^{\frac{1}{2}}$$

Proof. First, by Lem.(3), x_t is the minimizer of $r_t(x)$ at round t. Furthermore, $r_t(x)$ is a 2α-strongly convex function, hence $r_t(x_t) \leq r_t(y) - \alpha||y - x_t||_2^2$ for any $y \in \mathcal{X}$. We have that $DPLPS(x_t) \leq r_t(x_t) \leq r_t(y) - \alpha||y - x_t||_2^2$. We expand and re-arrange terms:

$$\Delta(t) \leq B + \underbrace{Q(t)\hat{g}_t(y)}_{\text{(a)}} + V \underbrace{\langle f'_{t-1}(x_{t-1}), (y - x_{t-1})\rangle}_{\text{(b)}} \underbrace{-V\langle f'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle - \alpha||x_t - x_{t-1}||_2^2}_{\text{(c)}} + \underbrace{\alpha||y - x_{t-1}||_2^2 - \alpha||y - x_t||_2^2}_{\text{(d)}}. \tag{15}$$

To continue with the proof we need to work on the terms (a),(b),(c) and (d) appearing on the right hand side. Since $f_t(x)$ is convex, (b) term is upper bounded by:

$$\begin{aligned} f_{t-1}(y) &\geq f_{t-1}(x_{t-1}) + \langle f'_{t-1}(x_{t-1}), (y - x_{t-1})\rangle \\ \langle f'_{t-1}(x_{t-1}), (y - x_{t-1})\rangle &\leq f_{t-1}(y) - f_{t-1}(x_{t-1}) \\ \langle f'_{t-1}(x_{t-1}), (y - x_{t-1})\rangle &\leq 2F. \end{aligned}$$

For (c) term we prove a technical Lemma:

Lemma 5. *Neely & Yu (2017)*

$$-\langle V f'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle - \alpha||x_t - x_{t-1}||_2^2 \leq \frac{V^2G^2}{2\alpha}.$$

Proof. Since $||\alpha||_2^2 + ||\beta||_2^2 + 2\langle\alpha, \beta\rangle = ||\alpha + \beta||_2^2$:

$$\begin{aligned} &- \langle V f'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle - \alpha||x_t - x_{t-1}||_2^2 \leq \\ &- \left|\left|\frac{V f'_{t-1}(x_{t-1})}{\sqrt{2\alpha}} + \frac{\sqrt{\alpha}(x_t - x_{t-1})}{\sqrt{2}}\right|\right|_2^2 + \frac{V||f'_{t-1}(x_{t-1})||_2^2}{2\alpha} \\ &\leq \frac{V^2G^2}{2\alpha}. \end{aligned}$$

□

Now we pick $y = x_\star^K$, according to Def.(2), and we sum the inequality for K consecutive rounds.

$$\sum_{\tau=0}^{K-1} \Delta(t+\tau) \leq BK + \underbrace{\sum_{\tau=0}^{K-1} Q(t+\tau)\hat{g}_{t+\tau}(y)}_{\text{(a)}} + 2FVK + \underbrace{\frac{V^2G^2K}{2\alpha} + \alpha \sum_{\tau=0}^{K-1} \left\{||y - x_{t+\tau-1}||_2^2 - ||y - x_{t+\tau}||_2^2\right\}}_{\text{(d)}}.$$

For the term (a) we use Cor.(2) (see at the end of this subsection for a proof),

$$\sum_{\tau=0}^{K-1} Q(t+\tau)\hat{g}_{t+\tau}(y) \leq BK(K-1).$$

We get to:

$$\sum_{\tau=0}^{K-1} \Delta(t+\tau) \leq BK^2 + 2FVK + \frac{V^2G^2K}{2\alpha} + (d).$$

We take the telescopic sum for $t = \{1, \dots, T-K\}$:

$$\begin{aligned} \sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K} \Delta(t+\tau) &\leq BK^2(T-K) + 2FVK(T-K) + \frac{V^2G^2K(T-K)}{2\alpha} + \sum_{t=1}^{T-K}(d) \\ &\leq BK^2T + 2FVKT + \frac{V^2G^2KT}{2\alpha} + \sum_{t=1}^{T-K}(d). \end{aligned}$$

For the term (d) we have that:

$$\begin{aligned} \alpha \sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K} \left(||y - x_{t+\tau-1}||_2^2 - ||y - x_{t+\tau}||_2^2\right) &= \\ \alpha \sum_{\tau=0}^{K-1} \left(||y - x_\tau||_2^2 - ||y - x_{T-K+\tau}||_2^2\right) &\leq \\ \alpha \sum_{\tau=0}^{K-1} ||y - x_\tau||_2^2 \leq \alpha K D^2. \end{aligned} \tag{16}$$

The intermediate drift terms in the telescopic sum are canceled out, hence:

$$\underbrace{\sum_{\tau=0}^{K-1} \frac{Q(T-K+\tau+1)^2}{2}}_{\text{e}} - \underbrace{\sum_{\tau=0}^{K-1} \frac{Q(\tau+1)^2}{2}}_{\text{f}} \leq BK^2T + 2FVKT + \frac{V^2G^2KT}{2\alpha} + \alpha K D^2.$$

Since $\hat{g}_t(x_t) \leq \sqrt{2B}$, the following holds for all t:

$$\begin{aligned} Q(t+1) &= [Q(t) + b_t(x_t)]^+ \leq |Q(t) + \hat{g}_t(x_t)| \\ &\leq |Q(t)| + |\hat{g}_t(x_t)| \leq Q(0) + \sum_{\tau=0}^{T} |\hat{g}_t(x_t)| \\ &\leq \sum_{\tau=0}^{t} \sqrt{2B} \leq (t+1)\sqrt{2B} \end{aligned} \tag{17}$$

We use Eq.(17) to upperbound (f):

$$(f) \leq \frac{1}{2} \sum_{\tau=0}^{K-1} ((\tau+1)\sqrt{2B})^2 \leq B\frac{K(K+1)(2K+1)}{6},$$

giving us the new bound:

$$\underbrace{\sum_{\tau=0}^{K-1} \frac{Q(T-K+\tau+1)^2}{2}}_{\text{(e)}} \leq BK^2T + 2FVKT + \frac{V^2G^2KT}{2\alpha} + \alpha KD^2 + B\frac{K(K+1)(2K+1)}{6}. \tag{18}$$

To continue we need to come up with a lower bound for term (e) based on $Q(T+1)$. We will use Eq.(26):

$$Q(T-K+\tau+1) \geq Q(T+1) - \sum_{n=T-K+\tau+1}^{T} |\hat{g}_n(x_n)| \overset{(i)}{\geq} 0. \tag{19}$$

here (i) is true if $Q(T+1) \geq K\sqrt{2B} \geq (\tau+1)\sqrt{2B}$[3]. We take the square of each side of Eq.(19):

$$\begin{aligned} Q(T-K+\tau+1)^2 &\geq \left(Q(T+1) - \sum_{n=T-K+\tau+1}^{T} |\hat{g}_n(x_n)|\right)^2 \\ &\geq Q(T+1)^2 - 2Q(T+1)\sum_{n=T-K+\tau+1}^{T} |\hat{g}_n(x_n)|. \end{aligned}$$

Now, we sum the resulting inequality for K consecutive slots, arriving at:

$$\begin{aligned} &\sum_{\tau=0}^{K-1} Q(T-K+\tau+1)^2 \geq \\ &\geq KQ(T+1)^2 - \sum_{\tau=0}^{K-1} 2Q(T+1)\sum_{n=T-K+\tau+1}^{T} |\hat{g}_n(x_n)| \\ &\overset{(i)}{\geq} KQ(T+1)^2 - \sum_{\tau=0}^{K-1} 2(T+1)\sqrt{2B}\sum_{n=T-K+\tau+1}^{T} \sqrt{2B} \\ &\geq KQ(T+1)^2 - 4(T+1)B\sum_{\tau=0}^{K-1}(K-\tau-1) \\ &\geq KQ(T+1)^2 - 4(T+1)B\frac{K(K-1)}{2}, \end{aligned}$$

where (i) follows from Eq.(17) and $\hat{g}_t(x_t) \leq \sqrt{2B}$. We rearrange and use Eq.(18):

$$\begin{aligned} &KQ(T+1)^2 \leq \\ &\leq \sum_{\tau=0}^{K-1} Q(T-K+\tau+1)^2 + 4(T+1)B\frac{K(K-1)}{2} \\ &\overset{(i)}{\leq} 2BK^2T + 4FVKT + \frac{V^2G^2KT}{\alpha} + 2\alpha KD^2 + \\ &\quad B\frac{K(K+1)(2K+1)}{3} + 2(T+1)BK^2, \end{aligned}$$

where (i) is true due to Eq.(18). Dividing by K and taking the square root, we get

$$\begin{aligned} Q(T+1) &\leq \left\{2BKT + 4FVT + \frac{V^2G^2T}{\alpha} + 2\alpha D^2 + \right. \\ &\quad \left. + B\frac{(K+1)(2K+1)}{3} + 2(T+1)B(K+1)\right\}^{\frac{1}{2}} \\ &\overset{(i)}{\leq} \left\{2BKT + 4FVT + \frac{V^2G^2T}{\alpha} + 2\alpha D^2 + \right. \\ &\quad \left. 2BK^2 + 2(T+1)BK\right\}^{\frac{1}{2}}, \end{aligned} \tag{20}$$

where in (i) $B\frac{(K+1)(K+2)}{3} \leq 2BK^2$, for $K \geq 1$. We note here that we proved Eq.(20) for $Q(T+1) \geq K\sqrt{2B}$, for $Q(T+1) < K\sqrt{2B}$ obviously Eq.(20) is an upper bound. This result can be easily generalized for any $T' \in \{K, K+1, \ldots, T\}$. For $T' \in \{1, \ldots, K-1\}$, we have from Eq.(17) that $Q(T') \leq T'\sqrt{2B} \leq K\sqrt{2B}$, which is again upper bounded by Eq.(20). This finishes the proof. □

We may now use the queue bound from Lem.(4) to prove that our policy guarantees sublinear growth of constraint residual.

Lemma 6 (Constraint Residual Bound)**.** *The Constraint Residual is bounded as:*

$$\begin{aligned} \sum_{t=1}^{T} g_t(x_t) \leq &\left\{2BKT + 4FVT + \frac{V^2G^2T}{\alpha} + \right. \\ &\left. 2\alpha D^2 + 2BK^2 + 2(T+1)BK\right\}^{\frac{1}{2}} + \frac{GVT}{2\alpha} + \\ &\frac{G^2\left[\sqrt{2BK} + \sqrt{4FV} + \sqrt{\frac{V^2G^2}{\alpha} + \sqrt{2BK}}\right]\frac{T^{\frac{3}{2}}+T}{\sqrt{2}}}{2\alpha} + \\ &\frac{G^2\left[\sqrt{2\alpha D^2} + \sqrt{2BK^2} + \sqrt{2BK}\right]T}{2\alpha}. \end{aligned}$$

Proof. We start with the queue update from Eq.(9):

$$\begin{aligned} Q(t+1) &= [Q(t) + \hat{g}_t(x_t)]^+ \geq Q(t) + \hat{g}_t(x_t) \\ &\geq Q(t) + g_{t-1}(x_{t-1}) + \langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle. \end{aligned}$$

[3] We will later prove our bound holds even if this assumption is not true.

Re-arranging terms in the above, we have:

$$g_{t-1}(x_{t-1}) \leq Q(t+1) - Q(t) - \langle g'_{t-1}(x_{t-1}), (x_t - x_{t-1})\rangle \overset{(a)}{\leq} Q(t+1) - Q(t) + ||g'_{t-1}(x_{t-1})||_2 ||(x_t - x_{t-1})||_2., \tag{21}$$

where in (a) we use the Cauchy-Schwarz inequality. Let us now denote:

$$y \triangleq x_{t-1} - \frac{1}{2\alpha}\left(V f'_{t-1}(x_{t-1}) + Q(t) g'_{t-1}(x_{t-1})\right),$$

as in Eq.(8). We have:

$$||x_t - x_{t-1}||_2 = ||\Pi_{\mathcal{X}}(y) - x_{t-1}||_2 \overset{(a)}{\leq} ||y - x_{t-1}||_2 \overset{(b)}{\leq} ||y - x_{t-1}||_1,$$

where (a) is true due to the non-expansiveness of euclidean projection on a convex set, while for (b) we used the norm inequality. Combing this finding, Eq.(21) and the fact that $||g'_t(x)||_2 \leq G$ we get:

$$g_{t-1}(x_{t-1}) \leq Q(t+1) - Q(t) + G||y - x_{t-1}||_1.$$

We now expand y as

$$g_{t-1}(x_{t-1}) \leq Q(t+1) - Q(t) + G\frac{||V f'_{t-1}(x_{t-1}) + Q(t) g'_{t-1}(x_{t-1})||_1}{2\alpha} \leq Q(t+1) - Q(t) + \frac{VG^2}{2\alpha} + \frac{G^2 Q(t)}{2\alpha}.$$

We take the sum for T rounds and we drop $Q(1) = 0$:

$$\sum_{t=1}^{T} g_{t-1}(x_{t-1}) \leq Q(T+1) + \frac{G^2 VT}{2\alpha} + \frac{G^2 \sum_{t=1}^{T} Q(t)}{2\alpha}. \tag{22}$$

We now use the queue bound (Lem.(4)), from which it follows that

$$\sum_{t=1}^{T} Q(t) \leq \sum_{t=1}^{T} \left\{ 2BKt + 4FVt + \frac{V^2 G^2 t}{\alpha} + 2\alpha D^2 + 2BK^2 + 2(t+1)BK \right\}^{\frac{1}{2}} \leq \sum_{t=1}^{T} \left[\sqrt{2BKt} + \sqrt{4FVt} + \sqrt{\frac{V^2 G^2 t}{\alpha}} + \sqrt{2\alpha D^2} + \sqrt{2BK^2} + \sqrt{2BK(T+1)} \right] \leq \left[\sqrt{2BK} + \sqrt{4FV} + \sqrt{\frac{V^2 G^2}{\alpha}} + \sqrt{2BK} \right] \sum_{t=1}^{T} \sqrt{t} + \left[\sqrt{2\alpha D^2} + \sqrt{2BK^2} + \sqrt{2BK} \right] T.$$

To proceed futher, we use the norm inequality (Cauchy-Schwarz) for the vector $[\sqrt{1}, \sqrt{2}, \sqrt{3}, ..., \sqrt{t}, ..., \sqrt{T}]$, from which we get

$$\sqrt{\sum_{t=1}^{T} t} \leq \sum_{t=1}^{T} \sqrt{t} \leq \sqrt{T \sum_{t=1}^{T} t} \leq \frac{T^{\frac{3}{2}} + T}{\sqrt{2}},$$

hence:

$$\sum_{t=1}^{T} Q(t) \leq \left[\sqrt{2BK} + \sqrt{4FV} + \sqrt{\frac{V^2 G^2}{\alpha}} + \sqrt{2BK} \right] \times \frac{T^{\frac{3}{2}} + T}{\sqrt{2}} + \left[\sqrt{2\alpha D^2} + \sqrt{2BK^2} + \sqrt{2BK} \right] T.$$

By using the bounds for $Q(t)$ and $\sum_t Q(t)$ in inequality Eq.(22) we complete the proof.

□

7.1.2. Proof of the Regret Bound

In the second part of the proof, we will prove the regret bound (13) against any K-benchmark. Since our policy x_t minimizes the DPLPS by Lem.(3) and $x_\star^K$ by Def.(2) has has minimum cost subject to no constraint residuals over any window of K rounds, comparing $x_\star^K$ with our policy will produce the result.

Lemma 7 (Regret Bound)**.** *The upper bound of the regret of our policy against an $x_\star^K$ (Eq.(2)) benchmark is:*

$$\sum_{t=0}^{T-1} f_t(x_t) - \sum_{t=0}^{T-1} f_t(y) \leq \frac{BKT}{V} + \frac{VG^2 T}{2\alpha} + B\frac{(K+1)(2K+1)}{6V} + \frac{\alpha D^2}{V} + 2F(K-1).$$

Proof. We begin by Eq.(15), where we add on both sides the term $V f_{t-1}(x_{t-1})$, with $y = x_\star^K$. We remind that:

$$\hat{f}_t(x_t) \triangleq f_{t-1}(x_{t-1}) + f'_{t-1}(x_{t-1})(x_t - x_{t-1}).$$

Using Lem.(5) on Eq.(15)(c) and summing the resulting inequality for $K = \{t, \ldots, t+K-1\}$ consequent rounds, we get:

$$\sum_{\tau=0}^{K-1} \Delta(t+\tau) + \sum_{\tau=0}^{K-1} V f_{t+\tau-1}(x_{t+\tau-1}) \leq BK + \underbrace{\sum_{\tau=0}^{K-1} Q(t+\tau)\hat{g}_{t+\tau}(y)}_{(a)} + V \sum_{\tau=0}^{K-1} \underbrace{\hat{f}_{t+\tau}(y)}_{(b)} + \frac{V^2 G^2 K}{2\alpha} + \alpha \sum_{\tau=0}^{K-1} ||y - x_{t+\tau-1}||_2^2 - \alpha \sum_{\tau=0}^{K-1} ||y - x_{t+\tau}||_2^2. \tag{23}$$

Table 2. Black terms already exist, blue and red terms need to be added and subtracted to complete regret

$$f_0(x_0)+f_1(x_1)+f_2(x_2)+\ldots+f_{T-K-1}(x_{T-K-1})+f_{T-K}(x_{T-K})+f_{T-K+1}(x_{T-K+1})+\ldots+f_{T-1}(x_{T-1})$$
$$f_0(x_0)+f_1(x_1)+f_2(x_2)+\ldots+f_{T-K-1}(x_{T-K-1})+f_{T-K}(x_{T-K})+f_{T-K+1}(x_{T-K+1})+\ldots+f_{T-1}(x_{T-1})$$

$$f_0(x_0)+f_1(x_1)+f_2(x_2)+\cdots+f_{T-K-1}(x_{T-K-1})+f_{T-K}(x_{T-K})+f_{T-K+1}(x_{T-K+1})+\ldots+f_{T-1}(x_{T-1})$$

$$\vdots$$
$$f_0(x_0)+\ldots+f_{K-1}(x_{K-1})+\cdots+f_{T-K-1}(x_{T-K-1})+f_{T-K}(x_{T-K})+f_{T-K+1}(x_{T-K+1})+\ldots+f_{T-1}(x_{T-1})$$

We now use Lem.8 (see at the end of this subsection for a proof) in order to bound term (a) in a way so we can use the sample path property of the static policy $x_\star^K$, and convexity of $f_t(x)$, hence $\hat{f}_t(x) \leq f_{t-1}(x)$, in order to bound term (b).

Taking Eq.(23) and using the Cor.(2) for term (a) and convexity of $f_t(x)$ for term (b), we get the following:

$$\sum_{\tau=0}^{K-1}\Delta(t+\tau)+V\sum_{\tau=0}^{K-1}f_{t+\tau-1}(x_{t+\tau-1}) \leq BK+ \underbrace{BK(K-1)}_{(a)}+V\sum_{\tau=0}^{K-1}\underbrace{f_{t+\tau-1}(y)}_{(b)}+\frac{V^2G^2K}{2\alpha}+ \alpha\sum_{\tau=0}^{K-1}\left[||y-x_{t+\tau-1}||_2^2-||y-x_{t+\tau}||_2^2\right].$$

We sum the inequality from $t=\{1,\ldots,T-K\}$:

$$\underbrace{\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}\Delta(t+\tau)}_{(d)}+V\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}f_{t+\tau-1}(x_{t+\tau-1}) \overset{(i)}{\leq} BK^2T+V\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}f_{t+\tau-1}(y)+\frac{V^2G^2KT}{2\alpha}+ \underbrace{\alpha\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}||y-x_{t+\tau-1}||_2^2-\alpha\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}||y-x_{t+\tau}||_2^2}_{(e)}. \tag{24}$$

where in (i), we use where needed that $T > T-K$. To continue with the proof we need to lower bound the negative terms of the drift expression (d):

$$\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}\Delta(t+\tau)= \frac{1}{2}\sum_{\tau=0}^{K-1}Q(T-K+\tau+1)^2-\frac{1}{2}\sum_{\tau=0}^{K-1}Q(\tau+1)^2 \geq -\frac{1}{2}\sum_{\tau=0}^{K-1}Q(\tau+1)^2 \overset{Eq.(17)}{\geq} -\frac{1}{2}\sum_{\tau=0}^{K-1}((\tau+1)\sqrt{2B})^2 \geq -B\frac{K(K+1)(2K+1)}{6}.$$

Furthermore, term (e) is upper bounded by $(e) \leq \alpha KD^2$, as shown by Eq.(16). Realigning Eq.(24) with the new bounds we get:

$$V\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}f_{t+\tau-1}(x_{t+\tau-1})-V\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}f_{t+\tau-1}(y) \leq BK^2T+\frac{V^2G^2KT}{2\alpha}+B\frac{K(K+1)(2K+1)}{6}+\alpha KD^2.$$

Here, this expression already compares the cost of our policy against the $x_\star^K$ policy. An illustration is given in Tab.(2), which shows how to manipulate the sums on the left hand side. To continue the proof we add and subtract the blue and red terms of Tab.(2) in order to make Regret ($R_K(T) =$

$\sum_{t=0}^{T-1}[f_t(x_t) - f_t(y)]$) appear K times:

$$\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}(f_{t+\tau-1}(x_{t+\tau-1}) - f_{t+\tau-1}(y)) =$$
$$\sum_{\tau=0}^{K-1}\sum_{t=1}^{T-K}(f_{t+\tau-1}(x_{t+\tau-1}) - f_{t+\tau-1}(y)) +$$
$$\sum_{\tau=0}^{K-1}\sum_{t=0}^{\tau-1}(f_t(x_t) - f_t(y)) - \sum_{\tau=0}^{K-1}\sum_{t=0}^{\tau-1}(f_t(x_t) - f_t(y)) +$$
$$\sum_{\tau=0}^{K-1}\sum_{t=T-K+\tau+1}^{T-1}\left[f_t(x_t) - f_t(y) - f_t(x_t) - f_t(y)\right] =$$
$$K\sum_{t=1}^{T}(f_{t-1}(x_{t-1}) - f_{t-1}(y)) -$$
$$\sum_{\tau=0}^{K-1}\sum_{t=0}^{\tau-1}(f_t(x_t) - f_t(y)) - \sum_{\tau=0}^{K-1}\sum_{t=T-K+\tau+1}^{T-1}(f_t(x_t) - f_t(y))$$

Then, we upper bound the red and blue terms:

$$\sum_{\tau=0}^{K-1}\sum_{t=0}^{\tau-1}(f_t(x_t) - f_t(y)) - \sum_{\tau=0}^{K-1}\sum_{t=T-K+\tau+1}^{T-1}(f_t(x_t) - f_t(y)) \leq 2FK(K-1)$$

By dividing both sides with V, K and using the above inequalities we can complete the proof:

$$\sum_{t=0}^{T-1}f_t(x_t) - \sum_{t=0}^{T-1}f_t(y) \leq \frac{BKT}{V} + \frac{VG^2T}{2\alpha} + B\frac{(K+1)(2K+1)}{6V} + \frac{\alpha D^2}{V} + 2F(K-1).$$

□

Finally, we prove an important technical result regarding upper bounding the quantity $\sum_{\tau=0}^{K-1}Q(t+\tau)\hat{g}_{t+\tau}(x_{t+\tau})$ for any sequence of actions x_t, and then specifically for the action selected by the K-benchmark.

Lemma 8. *For any sequence of actions* $\{x_t\}$*:*

$$\sum_{\tau=0}^{K-1}Q(t+\tau)\hat{g}_{t+\tau}(x_{t+\tau}) \leq Q(t)\sum_{\tau=0}^{K-1}\hat{g}_{t+\tau}(x_{t+\tau}) + BK(K-1).$$

Proof. We start by the queue update equation Eq.(9) and give a lower bound for $Q(t+K)$ for any $K \geq 1$:

$$\begin{aligned}Q(t+K) &= [Q(t+K-1) + \hat{g}_{t+K-1}(x_{t+K-1})]^+ \\ &\geq Q(t+K-1) + \hat{g}_{t+K-1}(x_{t+K-1}) \\ &\geq Q(t) + \sum_{\tau=0}^{K-1}\hat{g}_{t+\tau}(x_{t+\tau}). \end{aligned} \tag{25}$$

and an upper bound:

$$\begin{aligned}Q(t+K) &= [Q(t+K-1) + \hat{g}_{t+K-1}(x_{t+K-1})]^+ \\ &\leq Q(t+K-1) + |\hat{g}_{t+K-1}(x_{t+K-1})| \\ &\leq Q(t) + \sum_{\tau=0}^{K-1}|\hat{g}_{t+\tau}(x_{t+\tau})|. \end{aligned} \tag{26}$$

Next, we use these bounds so that $Q(t)$ appears as a common term in the sum. We denote $\max\{0, f(\cdot)\} \triangleq [f(\cdot)]^+$ and $\min\{0, f(\cdot)\} \triangleq [f(\cdot)]^-$:

$$\sum_{\tau=0}^{K-1}Q(t+\tau)\hat{g}_{t+\tau}(x_{t+\tau}) =$$
$$\sum_{\tau=0}^{K-1}Q(t+\tau)\left\{[\hat{g}_{t+\tau}(x_{t+\tau})]^+ + [\hat{g}_{t+\tau}(x_{t+\tau})]^-\right\} \overset{(a)}{\leq}$$
$$\sum_{\tau=0}^{K-1}\left\{Q(t) + \sum_{i=0}^{\tau-1}|\hat{g}_{t+i}(x_{t+i})|\right\}[\hat{g}_{t+\tau}(x_{t+\tau})]^+ +$$
$$\sum_{\tau=0}^{K-1}\left\{Q(t) + \sum_{i=0}^{\tau-1}\hat{g}_{t+i}(x_{t+i})\right\}[\hat{g}_{t+\tau}(x_{t+\tau})]^- \overset{(b)}{\leq}$$
$$Q(t)\sum_{\tau=0}^{K-1}\{[\hat{g}_{t+\tau}(x_{t+\tau})]^+ + [\hat{g}_{t+\tau}(x_{t+\tau})]^-\} +$$
$$\sum_{\tau=0}^{K-1}\left\{\sum_{i=0}^{\tau-1}|\hat{g}_{t+i}(x_{t+i})|\right\}[\hat{g}_{t+\tau}(x_{t+\tau})]^+ +$$
$$\sum_{\tau=0}^{K-1}\left\{\sum_{i=0}^{\tau-1}\hat{g}_{t+i}(x_{t+i})\right\}[\hat{g}_{t+\tau}(x_{t+\tau})]^- \overset{(c)}{\leq}$$
$$Q(t)\sum_{\tau=0}^{K-1}\hat{g}_{t+\tau}(x_{t+\tau}) +$$
$$\sum_{\tau=0}^{K-1}\left\{\sum_{i=0}^{\tau-1}|\hat{g}_{t+i}(x_{t+i})|\right\}|\hat{g}_{t+\tau}(x_{t+\tau})| \overset{(d)}{\leq}$$
$$Q(t)\sum_{\tau=0}^{K-1}\hat{g}_{t+\tau}(x_{t+\tau}) + 2B\sum_{\tau=0}^{K-1}\sum_{i=0}^{\tau-1}1 \leq$$
$$Q(t)\sum_{\tau=0}^{K-1}\hat{g}_{t+\tau}(x_{t+\tau}) + BK(K-1).$$

For (a) we take the upper bound for queue on the positive terms (Eq.(26)) and the lower bound on the queue for the negative terms (Eq.(25)), this gives an upper bound on the

total. In (b) we rewrite the equation by bringing in the front the common $Q(t)$ terms. Next, at (c) we upper bound the non-$Q(t)$ terms with their norm and by passing the norm to every element the inequality follows. Finally, (d) follows by taking the upper bound $|\hat{g}_t(x)| \leq \sqrt{2B}$. □

Corollary 2. *For policy $y = x_\star^K$:*

$$\sum_{\tau=0}^{K-1} Q(t+\tau)\hat{g}_{t+\tau}(y) \leq BK(K-1).$$

Proof. Using the above Lem.(8), replacing $x_t + \tau$ for all τ with y, since y has the property that $\sum_{\tau=0}^{K-1} g_{t+\tau-1}(y) \leq 0,\ \forall t$ and $\hat{g}_t(x) \leq g_{t-1}(x)$ due to convexity, the proof follows. □

7.2. Proof of Theorem 1

Keeping in mind that K is a free variable, we restrict our analysis to choices of α, V such that both $Ctr(T), R_K(T)$ are $o(T)$, and the bounds are optimized. These restrictions allow us to drop some terms in (12)-(13) which are of smaller order than the rest terms.

First, we note that $1 \leq K \leq T$. Hence, we may eliminate from (12) the following dominated terms: $2BK^2 = \mathcal{O}(2BKT)$, $\sqrt{K^2/T} = \mathcal{O}(\sqrt{K})$.

Second, we observe that $K < V < T$. This is because: if $V \leq K$, then the term BKT/V will become $\Omega(T)$, and if $V \geq T$, then the term $\sqrt{4FVT}$ will become $\Omega(T)$. Accordingly, we have the following dominated terms: $2BKT = \mathcal{O}(4FVT)$, $\sqrt{K} = \mathcal{O}(\sqrt{4FV})$, $B(K-1)(2K-1)/(6V) = \mathcal{O}(2F(K-1))$, $2F(K-1) = \mathcal{O}(BKT/V)$.

Dropping constants, and applying the above relations, we have arrived at the following simplified bounds:

$$Ctr(T) = \mathcal{O}\left(\frac{VT}{\alpha} + \frac{T}{\sqrt{\alpha}} + \sqrt{VT + \alpha + \frac{V^2T}{\alpha}} + \right. \\ \left. + \frac{T^{\frac{3}{2}}}{\alpha}\left(\sqrt{V} + \frac{V}{\sqrt{\alpha}}\right)\right), \tag{27}$$

$$R_K(T) = \mathcal{O}\left(\frac{KT}{V} + \frac{VT}{\alpha} + \frac{\alpha}{V}\right). \tag{28}$$

Next, we work to simplify (27). From the terms $\frac{VT}{\alpha}, \frac{\alpha}{V}$ in the regret expression, we may infer that: $V = o(\alpha)$, $\alpha = o(VT)$.

Hence, we have the following dominated terms: $V\frac{T}{\alpha} = \mathcal{O}(\sqrt{VT}\frac{T}{\alpha})$, $\frac{T}{\sqrt{\alpha}} = \mathcal{O}(\frac{\sqrt{VT}}{\sqrt{\alpha}}\frac{T}{\sqrt{\alpha}})$, $\alpha = \mathcal{O}(\sqrt{VT})$, $\frac{V\sqrt{T}}{\sqrt{\alpha}} = \sqrt{VT}\frac{\sqrt{V}}{\sqrt{\alpha}} = \mathcal{O}(\sqrt{VT})$, $\frac{V}{\sqrt{a}} = \mathcal{O}(\sqrt{V})$. Therefore (27) simplifies to:

$$Ctr(T) = \mathcal{O}\left(\sqrt{VT} + \sqrt{VT}\frac{T}{\alpha}\right).$$

Now comparing the above with (28), we observe that the only term that benefits from setting $\alpha < T$ is the term $\frac{\alpha}{V}$, and all other terms improve when α increases. However, for $\alpha < T$ the term $\frac{\alpha}{V}$ is dominated by $\frac{KT}{V}$. Hence, we may restrict $\alpha \geq T$ with no loss of optimality. In the sequel we further restrict $\alpha \geq T$. This brings us to the following simplified bounds:

$$Ctr(T) = \mathcal{O}(\sqrt{VT}),$$
$$R_K(T) = \mathcal{O}\left(\frac{KT}{V} + \frac{VT}{\alpha} + \frac{\alpha}{V}\right).$$

These bounds contain all cases where both constraint residuals and regret are sublinear to T. Next, consider two cases:

Case 1: $K < V < \sqrt{T}$. This case exists only when $K < \sqrt{T}$. In this case, we obtain no benefit by increasing α beyond T (since the benefiting term VT/α is already less than $\sqrt{T}$), hence the best choice is $\alpha = T$. Then, notice that $\frac{T}{V} = \mathcal{O}(\frac{KT}{V})$, and $V < \sqrt{T} = \mathcal{O}(\frac{KT}{V})$. Thus we get:

$$Ctr(T) = \mathcal{O}(\sqrt{VT}), \quad R_K(T) = \mathcal{O}\left(\frac{KT}{V}\right).$$

Case 2: $\sqrt{T} \leq V < T$. In this case, there are admissible values for α to make the two terms (containing α) equal, hence we select $\alpha = V\sqrt{T}$. This yields:

$$Ctr(T) = \mathcal{O}(\sqrt{VT}), \quad R_K(T) = \mathcal{O}\left(\frac{KT}{V} + \sqrt{T}\right).$$

Last, we observe that in the first case, we also have $\frac{KT}{V} = \Omega(\sqrt{T})$ due to the restricted values of V, hence we can express both cases with the expression:

$$Ctr(T) = \mathcal{O}(\sqrt{VT}), \quad R_K(T) = \mathcal{O}\left(\frac{KT}{V} + \sqrt{T}\right).$$

This concludes the proof.