Multi-Armed Bandits with Guaranteed Revenue per Arm

Dorian Baudry[1,4] **Nadav Merlis**[1] **Hugo Richard**[2]
Mathieu Molina[1,3] **Vianney Perchet**[1,2]

[1]ENSAE, FAIRPLAY joint team, CREST, Palaiseau, France
[2]Criteo AI Lab, FAIRPLAY joint team, Paris, France
[3]Inria, FAIRPLAY joint team, Palaiseau, France
[4]Institut Polytechnique de Paris, Palaiseau, France

Abstract

We consider a Multi-Armed Bandit problem with *covering* constraints, where the primary goal is to ensure that each arm receives a minimum expected reward while maximizing the total cumulative reward. In this scenario, the optimal policy then belongs to some unknown *feasible set*. Unlike much of the existing literature, we do not assume the presence of a safe policy or a feasibility margin, which hinders the exclusive use of conservative approaches. Consequently, we propose and analyze an algorithm that switches between pessimism and optimism in the face of uncertainty. We prove both precise problem-dependent and problem-independent bounds, demonstrating that our algorithm achieves the best of the two approaches – depending on the presence or absence of a feasibility margin – in terms of constraint violation guarantees. Furthermore, our results indicate that playing greedily on the constraints actually outperforms pessimism when considering *long-term* violations rather than violations on a *per-round* basis.

1 INTRODUCTION

1.1 Preliminaries

The Multi-Armed Bandit (MAB) is a classical model for sequential decision-making in the face of uncertainty (Lattimore and Szepesvári, 2020). In the standard formulation of the problem, the objective of the learner is to maximize the total sum of rewards by efficiently balancing exploration (testing different arms to learn their rewards) and exploitation (choosing arms with the highest expected rewards based on available information). While the classic MAB framework is applicable in various domains, in many real-world problems, additional constraints and considerations come into play.

Motivation Consider, for instance, an online content recommendation system (Zhou and Brunskill, 2016; Zong et al., 2016), aiming to maximize user engagement and satisfaction. Its success is intricately bound to the diversity of content available on the platform. To sustain this essential diversity, platforms must ensure sufficient exposure for all content creators, guaranteeing a sufficient *revenue* for their continued activity. This revenue could be, for instance, the number of times a content was played, or revenues from ads displayed with the content, which cannot be reduced to the number of times the content is suggested by the platform. In this context, a *constrained* MAB problem (see e.g. Slivkins et al. (2022); Sinha (2023)) emerges naturally. In this setting, the primary goal of the learner is to guarantee a fixed (known) minimum expected revenue to each arm (e.g. content recommendation) and consider the maximization of the total cumulative reward as a complementary objective.

Setting and notation We consider a K-armed bandit $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{F}^K$, where $\mathcal{F}$ is a family of distributions. We denote by μ_k, the expected reward of arm k, and by $\Delta_k = \mu_1 - \mu_k$, the sub-optimality gap of arm k, assuming w.l.o.g. that arm 1 is optimal. At each time step t, the decision-maker selects an arm A_t and receives a reward $r_t \sim \nu_{A_t}$ drawn independently at random. At time t, the policy π that chooses the actions can rely on past observations $\mathcal{H}_{t-1} = (A_1, r_1, \dots, A_{t-1}, r_{t-1})$ and internal randomization. For each arm k, we denote by

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

$N_k(t) = \sum_{s=1}^{t} \mathbb{I}(A_s = k)$, the number of times it was selected up to time t. We denote $(\cdot)_+ = \max(\cdot, 0)$, $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$.

The goal of the decision maker is to maximize its rewards $\mathbb{E}_{\nu,\pi}\left[\sum_{t=T}^{K} r_t\right]$ under the constraint that the expected revenue of each arm scales linearly in T, namely,

$$\forall k \in [K] : \ \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} r_t \mathbb{I}(A_t = k)\right] \geqslant \lambda_k T \ , \qquad (1)$$

where the scaling parameters $(\lambda_k)_{k\in[K]} \in (\mathbb{R}^+)^K$ are known a priori. If $\mathbb{E}_\nu[r_t | A_t = k] = \mu_k > 0$, satisfying the constraint of arm k is equivalent to ensuring that

$$\frac{1}{T}\mathbb{E}_{\nu,\pi}[N_k(T)] \geqslant p_k^\star \coloneqq \frac{\lambda_k}{\mu_k}.$$

Conveniently, the previous constrained optimization problem can be reduced to the linear program

$$\min_{p\in\triangle_K} \quad \sum_{k=1}^{K} \Delta_k(p_k^\star - p_k) \text{ s.t.} \quad \forall k \in [K], p_k \geqslant p_k^\star \ ,$$

where $\triangle_K$ denotes the K-dimensional simplex. Clearly, an optimal solution consists of playing each arm with probability $p_k^\star$ (if possible) and allocating the remaining probability over the optimal arms. Unfortunately, a non-anticipative policy cannot compute $(p_k^\star)_{k\in[K]}$ or know the optimal arms because the means are unknown. Moreover, the feasibility of this linear program depends on the problem parameters, as there exists a solution if and only if $\forall k,\ \mu_k \geqslant \lambda_k$ and $\sum_{k=1}^{K} p_k^\star \leqslant 1$.

Definition 1 (Feasibility gap)**.** *The feasibility gap of a problem with parameters $(\lambda, \mu) \in (\mathbb{R}^{+K})^2$ is*

$$\rho_\lambda(\mu) = 1 - \sum_{k=1}^{K} \frac{\lambda_k}{\mu_k} \ . \qquad (2)$$

*A problem instance (λ, μ) is **feasible** if $\rho_\lambda(\mu) \geqslant 0$.*

In the following, we simply denote the feasibility gap by ρ_λ when the context is clear.

Evaluation We do not assume any prior knowledge on $p_k^\star$, so relevant metrics must tolerate constraint violation to a certain level. Inspired by the literature on *safe bandits* (see next section for details), we consider two criteria to evaluate a policy on a given problem: the *excess regret* (for the regret-minimization objective), and the *constraint violation*. Denoting by $p_{k,t} = \mathbb{E}[\mathbb{I}(A_t = k)|\mathcal{H}_{t-1}]$ the *sampling probability* of arm k, we consider the following metrics.

Definition 2. *The total per-round excess-regret and constraint violation are respectively defined by*

$$\mathcal{R}_T^\pi(\nu, \lambda) = \sum_{k=1}^{K} \Delta_k \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} (p_{k,t} - p_k^\star)_+\right], \text{ and}$$

$$\mathcal{V}_T^\pi(\nu, \lambda) = \sum_{k=1}^{K} \mu_k \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} (p_k^\star - p_{k,t})_+\right] .$$

When the context is clear we omit (π, ν, λ) in the notation for simplicity. Intuitively, the per-round metrics encourage policies that smoothly converge to an optimal stationary policy. This is a desirable feature in real systems, which motivates providing policies with strong guarantees under these metrics.

1.2 Comparison with the Literature

Due to the numerous possible applications, the general problem of online learning with constraints covers several active research areas. In the literature, constraints typically originate from safety, fairness, or budget considerations to name but a few.

The generic problem that we consider is part of the literature on *Bandits with Linear Constraints*, notably including knapsacks (or packing) and covering constraints. Bandits with knapsacks have been extensively studied in the stochastic setting with finitely many arms (Badanidiyuru et al., 2018) as well as in the contextual (Agrawal and Devanur, 2016) and adversarial (Immorlica et al., 2022) setting. Logarithmic problem-dependent bounds have also surfaced in Sankararaman and Slivkins (2020); Li et al. (2021); Kumar and Kleinberg (2022). Generally, positive costs are incurred and the algorithm runs until some positive threshold is violated. On the contrary, covering constraints necessitates managing negative budgets and costs. A line of works considers deterministic covering (Claure et al., 2020; Patil et al., 2021; Wang et al., 2021; Chen et al., 2019), ensuring that each arm is pulled at least at a minimal known frequency, or Liu et al. (2022) with deterministic linear constraints. The core setting of this paper, which is a covering problem, is more challenging because the constraints are stochastic. Some works tackle this case (Agrawal and Devanur, 2019; Slivkins et al., 2022), and obtain $\mathcal{O}(\sqrt{T})$ constraint violation that holds for the setting that we consider. However in Slivkins et al. (2022); Chzhen et al. (2023), this guarantee holds only with at least one strong assumption: knowledge of an initially *safe* policy[1] or a feasibility margin ($\rho_\lambda \geqslant \delta$ for some $\delta > 0$). Finally, Sinha (2023) studies the same revenue guarantees as described in (1), and propose the `BanditQ` algorithm, that implements the natural idea of sampling the arm that is the

[1] In the current setting, this would consist in knowing for any arm k an allocation $\widetilde{p}_k \geqslant p_k^\star$ satisfying $\sum_k \widetilde{p}_k \leqslant 1$

most "late" w.r.t. its revenue constraint at the current time step, up to additional mechanisms to simultaneously minimize the regret. They obtain bounds of order $\mathcal{O}(T^{3/4})$ for a long-term evaluation of constraint violation and regret, weaker than Definition 2.

Other areas of research are also closely related. For instance, in safe bandits (Amani et al., 2019; Moradipari et al., 2021; Pacchiano et al., 2021; Liu et al., 2021; Zhou and Ji, 2022; Chen et al., 2022; Hutchinson et al., 2023) the goal is to only play actions belonging to an unknown feasibility set, with the objective of guaranteeing no violation of this constraint with high probability. The per-round evaluation metrics (def. 2) are inspired by some of these works. Their common approach is *pessimism-optimism* (PO): the algorithm plays the action maximizing reward (optimism) into a set included w.h.p. into the feasible set (pessimism). However, a safe-action and a known feasibility gap are again instrumental to design these algorithms. In contrast, Chen et al. (2023); Agrawal and Devanur (2019) obtained $\mathcal{O}(\sqrt{T})$ per-round safety violation without these assumptions, with a *doubly-optimistic* (DO) approach, considering instead an extended feasible set at each round. This good performance motivate the use of (DO) as a "worst-case" policy in the switching policies presented in Section 2. Furthermore, the algorithm DOC is the K-armed instance of (DO), for which we present a tighter analysis tailored for MAB.

Finally, we mention additional related fields. In online convex optimization with long term constraints (Mannor et al., 2009; Jenatton et al., 2015; Yu et al., 2017; Castiglioni et al., 2022) the learner receives full feedback of rewards and constraints. The question of unknown constraints is also considered in Chaudhary and Kalathil (2021); Liang et al. (2023), but as done in the safe bandits literature some safe action is assumed known. This pre-existing safe policy assumption is also often made in safe Reinforcement Learning (Bura et al., 2021; Ding et al., 2020; Xu et al., 2020; Efroni et al., 2020), which additionally mainly study problem-independent guarantees. Repeated auctions with ROI constraints (Castiglioni et al., 2023; Deng et al., 2023) is also similar in spirit, but again indirectly assumes some null action. Finally we mention Carlsson et al. (2023) that considers a setting similar to ours but with a best-policy identification objective.

In this work, we study problem dependent-bounds for stochastic bandits with unknown specific covering constraints, when no feasible actions is known beforehand and when per-round constraint violations are measured, and propose algorithms that switch between optimism and pessimism to minimize constraint violation for any possible problem.

1.3 Outline and Contributions

We propose several algorithms to solve the MAB problem with revenue guarantees presented in Section 1.1, which we frame as *Revenue-Guaranteeing Bandits* (RGB). RGB decouples the two objectives (Definition 2) by using a *target allocation* to satisfy the constraints, and a standard *optimistic* bandit algorithm for regret minimization. It is hence inspired by doubly-optimistic (DO) and pessimistic-optimistic (PO) approaches in the literature.

Typically, to obtain strong guarantees for the per-round constraints violations, the first natural idea is to take inspiration from (PO) methods (Pacchiano et al., 2021; Li et al., 2021) even though there are no initial feasible actions (according to the (PO) formulation). However, for problems with a small feasibility gap, this produces poorly performing or unfeasible algorithms.

Hence we first propose and analyze DOC (Algorithm 2), based on (DO), which achieves $\mathcal{O}(\sqrt{T})$ constraint violation for all problems We refine this result with novel problem-dependent bounds (Theorem 1) and further prove *constant* excess regret $\mathcal{O}(\sum_k (p_k^\star \Delta_k^2)^{-1})$ if $\min_k \lambda_k > 0$.

In order to get even better problem-dependent bounds for $\mathcal{V}_T$, we introduce a new algorithm, named SPOC (Algorithm 3), built on a hybrid combination of (PO) and (DO). This policy also achieves $\mathcal{O}(\sqrt{T})$ violation for all problems, and it even gets *constant* constraints violations $\mathcal{O}(\rho_\lambda^{-1})$ (Theorem 2) on strictly feasible problem instances, where ρ_λ is the *feasibility gap* (Definition 1). SPOC thus achieves the **best of the two approaches in terms of constraint violations**. This is illustrated in Table 1, summarizing problem-dependent results. We only include the scaling of first-order terms and omit logarithmic factors for clarity.

We additionally prove a lower bound (Theorem 3) which implies that the upper bound derived for $\mathcal{V}_T$ (resp. $\mathcal{R}_T$) for DOC (resp. SPOC) cannot be improved by more than logarithmic factors.

Finally, our experiments (Section 4.1) suggest investigating the long-term properties of a *greedy* algorithm, named SGOC – which is thus in-between optimism and pessimism. For SGOC, we prove that the cumulative (and not the average !) *long-term* constraints violation converges to 0 (Theorem 4) for a phase-based version of this approach.

2 ALGORITHMS

In this section, we detail the *revenue-guaranteeing bandit* (RGB) framework, as well as specific implementations. As detailed in Section 1.2, our inspiration comes from

Table 1: Problem-dependent bounds (Section 3), contribution of arm k

ALG.	$\mathcal{V}_T$	$\mathcal{R}_T$
DOC	$\sqrt{p_k^\star T}$	$\frac{1}{p_k^\star \Delta_k} \wedge \frac{\log(T)}{\Delta_k}$
SPOC	$\rho_\lambda^{-1} \wedge \sqrt{p_k^\star T}$	$\frac{1}{\mu_k}\sqrt{p_k^\star T}$
SGOC	$\sqrt{p_k^\star T}$	$\frac{1}{\mu_k}\sqrt{p_k^\star T}$

(PO) and (DO) approaches that have been proposed in the literature. In the K-armed problem considered, an RGB policy implements these principles as follows: at each time step, a *target allocation* routine proposes a feasible target allocation $\widehat{p}_{k,t}$, where $\widehat{p}_{k,t}$ aims at estimating the allocation $p_k^\star$, while a standard *bandit algorithm* (that we call *base bandit*) chooses (independently) one arm to allocate the remaining probability $1-\sum_{j\in[K]}\widehat{p}_{j,t}$. An arm is then chosen at random from the mixture. We detail RGB in Algorithm 1 below.

Algorithm 1 Revenue-Guaranteeing Bandits (RGB)

Input: $\lambda_1,\dots,\lambda_K$ (constraint levels),

algorithms TargetAlloc and BaseBandit.

Init: $\mathcal{H}_0 = \{\}$ (history of observations)

for $t \geqslant 1$ **do**

$(\widehat{p}_{1,t},\dots,\widehat{p}_{K,t}) = \texttt{TargetAlloc}(\mathcal{H}_{t-1})$

$k_t = \texttt{BaseBandit}(\mathcal{H}_{t-1})$

Set $p_{k,t} = \left(\widehat{p}_{k,t} + \mathbb{I}(k_t = k)\left(1 - \sum_{j=1}^{K}\widehat{p}_{j,t}\right)\right)$

Draw $A_t \sim \text{Mult}\,(p_{1,t},\dots,p_{K,t})$, collect $r_t \sim \nu_{A_t}$

Update $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{(A_t, r_t)\}$

We now explore possible choices for the base bandit algorithm and for the target allocation.

Base bandit As the revenue constraints are already handled by the target allocation, the goal of the base bandit is simply to try playing rewarding arms. This can be handled by a classic bandit algorithm, which could be chosen among any standard policy, like UCB (Auer et al., 2002), ETC (Perchet et al., 2016) TS (Thompson, 1933; Agrawal and Goyal, 2012), KL-UCB (Cappé et al., 2013) or MED (Honda and Takemura, 2011; Baudry et al., 2023). As detailed in the next section, choosing an optimistic algorithm is convenient for the analysis of RGB policies, so in the rest of the paper we assume that the base bandit is UCB (Auer et al., 2002) or KL-UCB (Cappé et al., 2013), instantiated to achieve logarithmic regret on the family of distributions $\mathcal{F}$. In the rest of the paper, we denote this algorithm by $\overline{\texttt{UCB}}$ to avoid any ambiguity with the target allocation.

We further remark that if $\max_k \lambda_k = 0$ (no revenue guarantee) then our algorithm simply follows the recommendation of $\overline{\texttt{UCB}}$, so RGB naturally interpolates standard MABs and revenue-guaranteeing bandits.

Target allocation All the target allocations considered in this paper are of the form

$$\widehat{p}^{\,\pi}_{k,t} = \frac{\lambda_k}{\widehat{\mu}^{\pi}_{k,t}}\,, \tag{3}$$

where for all $k \in [K]$, $\widehat{\mu}_{k,t}$ is an estimate of the mean μ_k computed with the $N_k(t-1)$ observations obtained up to time $t-1$. With a slight abuse, we say that $(\widehat{p}^{\,\pi}_{k,t})_{k\in[K]}$ is a *feasible* (resp. unfeasible) allocation if $\sum_k \widehat{p}^{\,\pi}_{k,t} \leqslant 1$ (resp. $\geqslant 1$). We consider $(\widehat{\mu}_{k,t})_{k\in[K],t\in[T]}$ that are either empirical means or confidence bounds. To design the latter, we use the following standard assumption.

Assumption 1 (Sub-Gaussian rewards)**.** *$\mathcal{F}$ is the family of 1-sub-Gaussian distributions.*

Under the sub-Gaussian model, for some parameter $c > 0$ we can use the following mean estimates:

- $\widehat{\mu}^{\text{Greedy}}_{k,t} = \bar{\mu}_{k,t} \coloneqq \frac{1}{N_k(t-1)}\sum_{s=1}^{t-1} r_s \mathbb{I}(A_s = k)$.
- $\widehat{\mu}^{\texttt{LCB}}_{k,t} = \texttt{LCB}_{k,t} \coloneqq \bar{\mu}_{k,t} - \sqrt{\frac{6(1+c)\log(t)}{N_k(t-1)}}$.
- $\widehat{\mu}^{\texttt{UCB}}_{k,t} = \texttt{UCB}_{k,t} \coloneqq \bar{\mu}_{k,t} + \sqrt{\frac{6(1+c)\log(t)}{N_k(t-1)}}$.

so that $\mathbb{P}\left(\mu_k \in [\texttt{LCB}_{k,t}, \texttt{UCB}_{k,t}]\right) \geqslant 1 - 2t^{-2(1+c)}$ (using a simple union bound on $N_k(t-1)$). As detailed in the next section, this confidence level (with $2(1+c) > 2$) is crucial for the theoretical analysis of RGB policies. Note however that $\overline{\texttt{UCB}}$ can use a different confidence bound than $\texttt{UCB}_{k,t}$ (e.g., with lower confidence). In the following, we use the shorthand formulation "LCB allocation" (resp. UCB or greedy) to refer to the target allocation corresponding to this estimate.

Algorithms We now detail *Doubly-Optimistic Covering* (DOC), *Safe Pessimistic-Optimistic Covering* (SPOC) and *Safe Greedy-Optimistic Covering* (SGOC).

It is known (Agrawal and Devanur, 2019; Chen et al., 2023) that (DO) can provide surprisingly better candidates than (PO) to satisfy the revenue guarantees when the problem has low feasibility gap. This stems from the following property:

If the problem is feasible, then the UCB allocation is feasible with high probability.

On the contrary, LCB allocations may take a long time to become feasible when ρ_λ is small (and may never be if $\rho_\lambda = 0$), which causes the failure of (PO) for low feasibility gaps. This motivates the idea of using the UCB allocation as a backup policy when the LCB one is unfeasible, which we later exploit with SPOC. Before that, we detail DOC in Algorithm 2. For completeness, DOC needs to be able to provide a target allocation also when the UCB one is unfeasible, even if this situation is unlikely. In the following implementation, we simply assume that a routine UnfeasAlloc is chosen beforehand to tackle that case. In our code we simply choose to normalize the UCB allocation if it is unfeasible, $\texttt{UnfeasAlloc}(\mathcal{H}_{t-1}) \propto \widehat{p}^{\texttt{UCB}}_{k,t}$, because dividing all revenues by the same factor seems to be a fair way to tackle unfeasibility. We discuss other choices at the end of this section.

Algorithm 2 Doubly-Optimistic Covering (DOC)

Input: $\lambda = (\lambda_1, \dots, \lambda_K)$, UnfeasAlloc

Play $\texttt{RGB}(\lambda, \texttt{UCB-Alloc}, \overline{\texttt{UCB}})$ (Alg. 1), with

UCB-Alloc:

$$\mathcal{H}_{t-1} \to \begin{cases} \left(\frac{\lambda_k}{\texttt{UCB}_{k,t}}\right)_{k\in[K]} & \text{if feasible,} \\ \texttt{UnfeasAlloc}(\mathcal{H}_{t-1}) & \text{otherwise.} \end{cases}$$

We then detail the implementation of SPOC in Algorithm 3 below. The idea is to play the LCB allocation whenever it is feasible, and to switch to the UCB allocation otherwise. We build SGOC on the same design as SPOC, replacing the LCB allocation with the greedy one. We report its implementation in Algorithm 4. Interestingly, we obtain from Equation (3) that for each $k \in [K], t \in [T]$, given the same observations,the sampling probabilities of the three algorithms satisfy

$$\widehat{p}^{\texttt{DOC}}_{k,t} \leqslant \widehat{p}^{\texttt{SGOC}}_{k,t} \quad \text{and} \quad \widehat{p}^{\texttt{DOC}}_{k,t} \leqslant \widehat{p}^{\texttt{SPOC}}_{k,t}\,, \tag{4}$$

so SGOC and SPOC are expected to serve the constraint at least as well as DOC in any situation (omitting the role of $\overline{\texttt{UCB}}$ for simplicity). This is the motivation for qualifying SPOC (resp. SGOC) as a "safe" way to implement a pessimistic (resp. greedy) approach in the RGB framework.

Remark 1 (Individual switches)**.** *In practice, we can implement a variant of SPOC that switches as few arms as possible to the UCB allocation, in order to guarantee $\widehat{p}_{k,t} \geqslant p^\star_k$ w.h.p. for as many arms as possible. We chose the implementation of Algorithm 3 to simplify the presentation, but the theoretical guarantees derived for SPOC in Theorem 2 (next section) trivially hold for any more subtle implementation of switches.*

Algorithm 3 Safe Pessimistic-Optimistic Covering (SPOC)

Input: $\lambda = (\lambda_1, \dots, \lambda_K)$, UnfeasAlloc

Play $\texttt{RGB}(\lambda, \texttt{SPOC-Alloc}, \overline{\texttt{UCB}})$ (Alg. 1), with

SPOC-Alloc:

$$\mathcal{H}_{t-1} \to \begin{cases} \left(\frac{\lambda_k}{\texttt{LCB}_{k,t}}\right)_{k\in[K]} & \text{if feasible, else} \\ \left(\frac{\lambda_k}{\texttt{UCB}_{k,t}}\right)_{k\in[K]} & \text{if feasible,} \\ \texttt{UnfeasAlloc}(\mathcal{H}_{t-1}) & \text{otherwise.} \end{cases}$$

Policy for the unfeasible case In practice, the decision-maker should decide in advance what strategy to adopt if the initial problem appears to be unfeasible (which is true w.h.p. if the UCB allocation is unfeasible). For instance, for recommendation systems, there may be no way to certify in advance that some content may work "well enough" or not. However, there is no unique way to define this new goal, i.e., a new target allocation $(\widetilde{p}^\star_k)_{k\in[K]}$. This depends on the exact context of the problem, and UnfeasAlloc should be tailored to reach the chosen objective. One, that we choose in our implementations, is to avoid discriminating between arms (e.g., for fairness reasons) by defining $\widetilde{p}^\star_k \propto p^\star_k$: every arm receives the same fraction of their initial guaranteed revenue. Another is to define an implicit ranking of the arms $(i_1, \dots, i_K)$ that can be learned (by knowing a rule set depending on problem parameters) and to serve the constraints of the better-ranked arms in priority. For instance, ranking the arms by decreasing expectation minimizes the total constraint violation, while ranking the arms by increasing values of $p^\star_k$ maximizes the number of constraints that are satisfied.

Algorithm 4 Safe Greedy-Optimistic Covering (SGOC)

Input: $\lambda = (\lambda_1, \dots, \lambda_K)$, UnfeasAlloc

Play $\texttt{RGB}(\lambda, \texttt{SGOC-Alloc}, \overline{\texttt{UCB}})$ (Alg. 1), with

SGOC-Alloc:

$$\mathcal{H}_{t-1} \to \begin{cases} \left(\frac{\lambda_k}{\bar{\mu}_{k,t}}\right)_{k\in[K]} & \text{if feasible, else} \\ \left(\frac{\lambda_k}{\texttt{UCB}_{k,t}}\right)_{k\in[K]} & \text{if feasible,} \\ \texttt{UnfeasAlloc}(\mathcal{H}_{t-1}) & \text{otherwise.} \end{cases}$$

3 THEORETICAL RESULTS

In this section, we provide upper bounds on $\mathcal{R}_T$ and $\mathcal{V}_T$ for DOC, SPOC and SGOC, and provide lower bounds that exhibit the trade-off between the two metrics. We assume that **all problems considered are feasible** (Definition 1), which is essential for the interpretation of the results.

3.1 Auxiliary Results

Before presenting the main theorems, we define the key quantities used in their statement.

Regret due to UCB For arms with positive revenue guarantees, it is clear that UCB benefits from the plays of sub-optimal arms caused by the target allocation. In Lemma 1 (Appendix A) we prove that UCB only selects such arms a finite number of times. More precisely, we show that under Assumption 1 there exists a constant multiplicative factor α such that the number of selection of any arm $k \in [K]$ by UCB is upper bounded by

$$\overline{N}_k(T) \coloneqq \overline{N}_k^\star \wedge \alpha \frac{\log(T)}{\Delta_k^2} \wedge T \ , \tag{5}$$

with a constant $\overline{N}_k^\star \in \mathbb{R} \cup \{+\infty\}$ satisfying

$$\overline{N}_k^\star = \mathcal{O}\left(\frac{\log\left(3 \vee (p_k^\star \Delta_k^2)^{-1}\right)}{p_k^\star \Delta_k^2} \right) \ . \tag{6}$$

Equation (5) further exhibits different regimes according to the time horizon: if $p_k^\star \leqslant (\log(T))^{-1}$ we can use the standard logarithmic bound, which is intuitive.

Sufficient sampling We now introduce a crucial result for the derivation of the problem-dependent bounds presented in the next section. It formalizes the intuition that, when playing a revenue-guaranteeing algorithm, at each step t any arm k should satisfy $N_k(t) = \Omega(p_k^\star t)$ with high probability. More specifically, we show in Lemma 2 (Appendix A) that there exists a constant Γ_k such that if a policy π satisfies $p_{k,t}^\pi \geqslant p_{k,t}^{\texttt{DOC}}$ (which is the case for SPOC and SGOC), it holds that

$$\sum_{t=1}^{+\infty} \mathbb{P}_\pi \left(N_k(t) \leqslant \frac{p_k^\star}{8} t \right) \leqslant \Gamma_k \ , \text{ with} \tag{7}$$

$$\Gamma_k = \mathcal{O}\left(\frac{\log\left((p_k^\star \mu_k^3)^{-1}\right)^{\frac{3}{2}}}{p_k^\star \mu_k^3} \vee \frac{-\log(p_k^\star)}{p_k^\star} \vee \frac{1}{c} \right) . \tag{8}$$

The proof of this result is non-trivial and relies on a variant of Freedman's inequality (Theorem 1 from Beygelzimer et al. (2011)). It is also noteworthy that the factor c^{-1} in (8) justifies the confidence level adopted in Algorithms 2 and 3: a smaller level may not guarantee that the arms are sufficiently sampled with high probability.

3.2 Upper Bounds on $\mathcal{V}_T$ and $\mathcal{R}_T$

We can now formalize our main results. When unspecified, the guarantees are problem-dependent, while problem-independent results will be explicitly stated as such. We start with DOC, which serves as a basis for the other algorithms.

Theorem 1 (Upper bounds for DOC)**.** *Under Assumption 1, the excess-regret of DOC satisfies*

$$\mathcal{R}_T^{\texttt{DOC}} \leqslant \sum_{k=1}^K \left(\Delta_k (\overline{N}_k^\star + \Gamma_k) \right) \wedge \frac{\alpha \log(T)}{\Delta_k} + \frac{K \max_k \Delta_k}{1+c} \ ,$$

where α, $\overline{N}_k^\star$ and Γ_k are respectively defined in Equations (5), (6) and (7). If $\max_k \Delta_k \leqslant \Delta^+$ for a fixed $\Delta^+ \in \mathbb{R}$ it furthermore holds that $\mathcal{R}_T^{\texttt{DOC}} = \mathcal{O}(\sqrt{KT\log(T)})$ (pb. independent bound).

Moreover, there exists an absolute constant C_0 such that the constraint violation DOC satisfies

$$\mathcal{V}_T^{\texttt{DOC}} \leqslant C_0 \sum_{k=1}^K \sqrt{p_k^\star T \log(T)} + \sum_{k=1}^K \lambda_k \Gamma_k + \frac{K \max_k \mu_k}{1+c} \ ,$$

and $\mathcal{V}_T^{\texttt{DOC}} = \mathcal{O}(K\sqrt{T\log(T)})$ (pb. independent).

The details of the proof can be found in Appendix B.1. Theorem 1 first establishes that both $\mathcal{R}_T^{\texttt{DOC}}$ and $\mathcal{V}_T^{\texttt{DOC}}$ admit a problem-independent bound scaling in $\mathcal{O}(\sqrt{T})$, which is on par with the best results obtained in the literature for (DO) approaches (see e.g. Chen et al. (2023)). These results are refined with novel problem-dependent bounds: we obtain a constant for $\mathcal{R}_T^{\texttt{DOC}}$, and a bound for $\mathcal{V}_T^{\texttt{DOC}}$ that improves the scaling of the first-order term. For instance, if $\mu_k \gg \lambda_k$ then $\sqrt{p_k^\star T \log(T)}$ and $\lambda_k \Gamma_k = \mathcal{O}(\mu_k^{-2})$ can both be much smaller than $\sqrt{T}$. It is also noteworthy that we employed a different proof scheme to derive the two results for $\mathcal{V}_T^{\texttt{DOC}}$.

We now present upper bounds for SPOC, that we prove in Appendix B.2.

Theorem 2 (Upper bounds for SPOC)**.** *Under Assumption 1, SPOC satisfies $\mathcal{V}_T^{\texttt{SPOC}} \leqslant \mathcal{V}_T^{\texttt{DOC}}$ as well as*

$$\mathcal{V}_T^{\texttt{SPOC}} = \mathcal{O}\left(\frac{\sqrt{\log\left(\rho_\lambda^{-2} \vee e\right)}}{\rho_\lambda} \sqrt{K D_{\lambda,\mu}} \right) \ ,$$

where $D_{\lambda,\mu} = \max_{j \in [K]: \lambda_j > 0} \frac{\log\left(e \vee (\lambda_j \mu_j)^{-1}\right)}{\lambda_j \mu_j}$. Moreover, there exists an absolute constant $C_1 > 0$ such that

$$\mathcal{R}_T^{\texttt{SPOC}} \leqslant \sum_{k=1}^K \Delta_k \left(C_1 \frac{\sqrt{p_k^\star T \log(T)}}{\mu_k} + \overline{N}_k(T) + 2\Gamma_k \right) \ ,$$

where $\overline{N}_k(T)$ and Γ_k are resp. defined in (5) and (7).

The main result is that $\mathcal{V}_T^{\texttt{SPOC}}$ admits a constant upper bound as soon as $\rho_\lambda > 0$, while simultaneously guaranteeing no more constraints violation than DOC for any horizon T. This justifies calling SPOC a "safe" implementation of pessimism-optimism for the revenue-guaranteeing problem. However, we note that the constant bound may be vacuous if one of the revenue parameters is very small (high constant $D_{\lambda,\mu}$). This effect may be reduced by implementing more subtle switches (see Remark 1), but improving $D_{\lambda,\mu}$ seems quite intricate in general.

Symmetrically to $\mathcal{V}_T^{\texttt{DOC}}$, the dominant term of the excess-regret upper bound scales as $\mathcal{O}(\sqrt{T})$. However, the factor μ_k^{-1} does not permit to obtain problem-independent results: the upper bound is vacuous for $\mu_k \leqslant T^{-1/2}$. Still, by assuming that the problem is feasible we know that $\mu_k^{-1} \leqslant \lambda_k^{-1}$, which in turn provides an upper bound on $\mathcal{R}_T^{\texttt{SPOC}}$ that is known by the decision-maker.

We recall that the assumption that the problem is feasible makes the UCB allocation feasible w.h.p., which explains why the bounds of the two theorems hold for any choice of UnfeasAlloc.

Finally, theoretical guarantees for SGOC can be easily derived from the bounds on $\mathcal{R}_T^{\texttt{SPOC}}$ and $\mathcal{V}_T^{\texttt{DOC}}$. We detail this in Section 4.2.

Remark 2. *A minor modification of SPOC leads to problem-independent bounds on $\mathcal{R}_T^{\texttt{SPOC}}$: choose thresholds $(\tau_t)_{t\in[T]}$ such that SPOC plays the UCB allocation for arm k if $\texttt{LCB}_{k,t} \leqslant \tau_t$. With this mechanism, we can get $\mathcal{R}_T^{\texttt{SPOC}} = \mathcal{O}\left(\tau_T^{-1}\sqrt{KT} \vee \tau_T^{-3}\right)$ (up to logarithms). The drawback is that it is necessary to wait that $\tau_T \leqslant \mu_k$ to play LCB, which degrades the guarantees on $\mathcal{V}_T$ from a non-asymptotic perspective.*

3.3 Lower Bounds

In this section we prove lower bounds that show that the problem-dependent bounds obtained in previous section for $\mathcal{V}_T^{\texttt{DOC}}$ and $\mathcal{R}_T^{\texttt{SPOC}}$ cannot be improved by more than logarithmic factors, that can depend on T but not on the problem constants. We define a revenue-guaranteeing problem by $(\lambda, \nu) \in \mathbb{R}^K \times \mathcal{F}^K$, and use the notation $\mathcal{C}$ to denote the set of feasible problems, and by $\mathcal{C}^0 \subset \mathcal{C}$ its interior (problems with positive feasibility gaps).

Definition 3 (Admissible policies)**.** *A policy π belongs to the set of admissible policies Π if*

$$\forall (\nu,\lambda) \in \mathcal{C}: \liminf_{T\to\infty} \inf_{k\in[K]} \frac{\mathbb{E}_{\nu,\pi}[N_k(T)]}{p_k^\star(\nu,\lambda)T} \geqslant 1,$$

where for $k \in [K]$, $p_k^\star(\nu,\lambda) = \frac{\lambda_k}{\mathbb{E}_{X\sim\nu_k}[X]}$.

In other words, π is admissible if all revenue guarantees are satisfied asymptotically. We now consider more precisely two subsets of admissible policies.,

Definition 4. *A policy $\pi \in \Pi$ can be*

$$\mathcal{R}\text{-targeting: if } \forall (\nu,\lambda) \in \mathcal{C}: \limsup T^{-\frac{1}{2}}\mathcal{R}_{T,\nu}^\pi = 0,$$
$$\mathcal{V}\text{-targeting: if } \forall (\nu,\lambda) \in \mathcal{C}^0: \limsup T^{-\frac{1}{2}}\mathcal{V}_{T,\nu}^\pi = 0\,.$$

We denote by Π_R (resp. Π_V) the set of $\mathcal{R}$-targeting (resp. $\mathcal{V}$-targeting) policies.

It is clear from Theorem 1 and 2 that DOC is $\mathcal{R}$-targeting while SPOC is $\mathcal{V}$-targeting. We now state our main result for this part.

Theorem 3 (Lower bounds)**.** *Consider $\lambda \in \mathbb{R}^K$ and a bandit $\nu \in \mathcal{F}^K$ with means $(\mu_k)_{k\in[K]}$. For any policy $\pi \in \Pi_R$, it holds that*

$$\forall (\nu,\lambda) \in \mathcal{C},\ \limsup_{T\to\infty} \frac{\mathcal{V}_T^\pi(\nu,\lambda)}{\sqrt{Tp_k^\star(\nu,\lambda)}} \geqslant \frac{1}{2\sqrt{e}}\,,$$

and, for any policy $\pi \in \Pi_V$ it holds that

$$\forall (\nu,\lambda) \in \mathcal{C}^0,\ \limsup_{T\to\infty} \frac{\mathcal{R}_T^\pi(\nu,\lambda)}{\frac{1}{\mu_k}\sqrt{Tp_k^\star(\nu,\lambda)}} \geqslant \frac{1}{2\sqrt{e}}\,.$$

The proof details can be found in Appendix B.3, and rely on standard change-of-measure arguments. Theorem 3 indicates that a targeting policy must pay at least $\sqrt{T}$ for the "non-targeted" objective. We furthermore observe that the upper bounds on $\mathcal{V}_T^{\texttt{DOC}}$ and $\mathcal{R}_T^{\texttt{SPOC}}$ obtained in Section 3.2 match the lower bounds of Theorem 3 up to logarithmic factors that do not depend on (λ,ν). Finally, Theorem 3 also confirms that the factors μ_k^{-1} in the upper bounds on $\mathcal{R}_T$ are unavoidable for $\mathcal{V}$-targeting policies.

4 PRACTICAL RESULTS

4.1 Experiments

The experiments can be reproduced using the code, available online[2]. As highlighted in Section 1.2, only a small fraction of the literature is directly applicable to our setting. Thus, we benchmark DOC, SPOC and SGOC in terms of excess-regret and constraint violation with BanditQ (Sinha, 2023) and a primal-dual algorithm by Slivkins et al. (2022). However, we present our results for the latter only in Appendix C, since we did not obtain good performance with this algorithm. Additionally, for fair comparison with BanditQ we also

[2] https://github.com/DBaudry/Revenue_guaranteeing_bandits

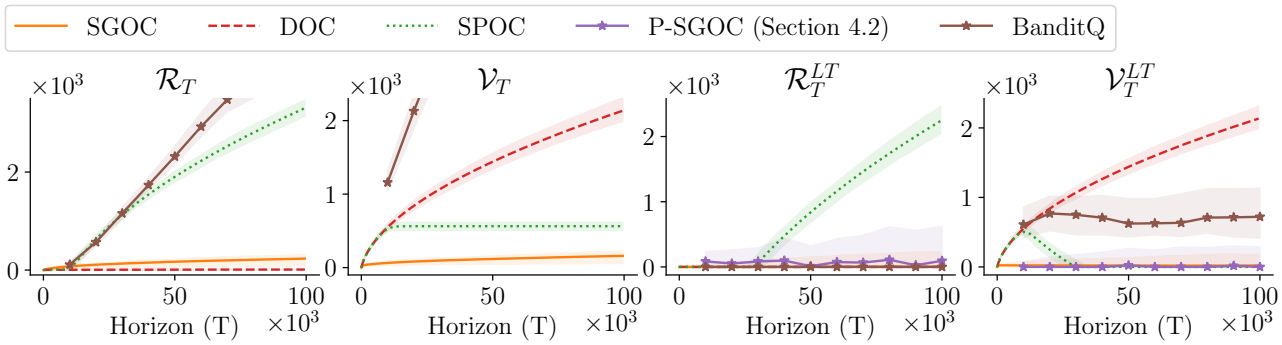

Figure 1: Reproducing the simulation setup from Sinha (2023)

consider the following long term metrics:

$$\mathcal{R}^{\mathrm{LT}}_{\pi,T}(\nu,\lambda) = \sum_{k=1}^{K} \Delta_k \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(p_{k,t}-p_k^\star)\right]_+ \text{, and}$$

$$\mathcal{V}^{\mathrm{LT}}_{\pi,T}(\nu,\lambda) = \sum_{k=1}^{K} \mu_k \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(p_k^\star-p_{k,t})\right]_+ .$$

We replicate the experiment presented in Sinha (2023)[3], using 200 seeds and with T varying in $[10^2, 10^5]$, reporting 10 values for non-anytime algorithms. We compute the excess regret and constraints violation as well as their long-term counterparts and display the result in Figure 1, averaging the values across seeds. Error bars represent the first and the last decile.

As predicted by our analysis, `DOC` has small excess regret and square root violation while `SPOC` exhibits constant violation and square root excess regret. `SGOC` exhibits $\mathcal{O}(\sqrt{T})$ excess regret and violation (see Figure 3 for better resolution) but still achieves lower excess regret than `SPOC` and lower constraints violation than `DOC`. In this example, we also observe the transition of `SPOC` from optimism to pessimism, making the long-term violation converge to 0. In Appendix C, we further study the impact of ρ_λ on the performance of the algorithms, confirming at the same time our previous observations.

If we consider more specifically the long-term metrics, `BanditQ` seems to converge to 0 regret and to constant violation. However, simulations with different problem parameters available in Appendix C show that `BanditQ` exhibits positive regret on some instances, and positive violation on others, contrasting with the predictable behaviour of `SPOC` and `DOC`.

With long term metrics, `SGOC` seems to be the go-to approach, reaching both very small long-term regret and violation. It is not clear that these quantities still scales in $\sqrt{T}$. This observation motivates a closer investigation of the performance of `SGOC` w.r.t. the long term metrics. Unfortunately, providing a tight analysis for `SGOC` may be intricate, because the mean estimates $(\bar{\mu}_k(t))_{k\in[K]}$ are not independent of the trajectory. For this reason, we introduce a phase based algorithm called `P-SGOC`, in order to mimic the long-term behavior of `SGOC`. We describe and analyze this algorithm in the next section. Observe that `P-SGOC` seems to follow closely `SGOC` in the presented experiment.

4.2 Greedy Algorithms

We start by elaborating on the theoretical performance of `SGOC`. It is clear that, given the same sequence of observations, it holds that $p_{k,t}^{\texttt{DOC}} \leqslant p_{k,t}^{\texttt{SGOC}} \leqslant p_{k,t}^{\texttt{SPOC}}$. Hence, by following closely the proofs we can show that the upper bounds obtained for $\mathcal{V}_T^{\texttt{DOC}}$ (Theorem 1) hold for $\mathcal{V}_T^{\texttt{SGOC}}$, and similarly that the upper bound obtained for $\mathcal{R}_T^{\texttt{SPOC}}$ (Theorem 2) hold for $\mathcal{R}_T^{\texttt{SGOC}}$. This leads to the results presented in Table 1. Although the scaling in T and the problem parameters $(\lambda_k, \mu_k)_{k\in[K]}$ is the same, smaller multiplicative constants than C_0 and C_1 presented in the two theorems can be obtained. This is simply because the proof can involve tighter confidence intervals (since the greedy estimates do not use confidence bounds). This is on par with the simulations presented in the previous section. We now focus on the long-term metrics, providing a refined analsis for an algorithm inspired by `SGOC`.

`SGOC` with phases In this part, we assume that the horizon T is known, and that the means are bounded by 1 for simplicity[4]. `P-SGOC` proceeds in four phases. The first two phases are used to build a target allocation $\widehat{p} = (\widehat{p}_k)_{k\in[K]}$. The first phase, of length $\sum \frac{\lambda_k}{4} T$, provides a rough estimate of $(p_k^\star)$, that is then used to calibrate the length of the second estimation phase. In the third phase, `P-SGOC` serves the constraints by

[3] $K = 5$, $\boldsymbol{\mu} = (0.335, 0.203, 0.241, 0.781, 0.617)$ and $\lambda = (0.167, 0.067, 0, 0, 0)$

[4] otherwise, a short preliminary phase can provide a crude upper bound on each mean.

ensuring that $\widehat{p}_k T$ samples are collected from each arm k. Finally, UCB plays for the remaining time steps. We provide a detailed pseudo-code in Appendix B.4 (Algorithm 5). In the exact implementation of P-SGOC, we carefully tune the length of phase 1 and 2 to ensure that the algorithm goes to phase 3 with high probability. Furthermore, using two estimation phases allows P-SGOC to obtain (again, w.h.p.) an uncertainty on the estimate of $p_k^\star$ on par with SGOC (for which the errors depends on $N_k(t) = \Omega(p_k^\star t)$ w.h.p.). For these reasons, we believe that P-SGOC is a good proxy for SGOC.

We now present the long-term guarantees of P-SGOC, assuming for simplicity a positive feasibility gap and only positive revenue guarantee.

Theorem 4 (Long-term excess-regret and constraint violation of P-SGOC)**.** *Assume that* $\min_{k\in[K]} \lambda_k > 0$*, that* $\rho_\lambda > 0$*, and that* $\max_{k\in[K]} \mu_k \leqslant 1$*. If the distributions are* σ*-sub-Gaussian then P-SGOC satisfies*

$$\limsup_{T\to\infty} \mathcal{R}_T^{LT} \leqslant 24 \sum_{k=1}^{K} \frac{\sigma^2}{\mu_k^2} \Delta_k \ , \ \textit{and} \quad \limsup_{T\to\infty} \mathcal{V}_T^{LT} \leqslant 0 \ .$$

The results are stated in an asymptotic formulation to simplify their interpretation. The detailed proof, with explicit bounds, is available in Appendix B.4.

By Theorem 4, P-SGOC asymptotically satisfies all the long-term constraints in expectation, and achieves constant excess-regret. This result is of course much stronger than what we obtained for the per-round metrics with SGOC, proving that a "greedy-optimistic" have merits if long-term goals are also considered.

CONCLUSION

In this paper we tackle a Multi-Armed Bandit problem with guaranteed per-arm revenue. Setting the *per-round* satisfaction of the revenue constraint as the main goal encourages the design of policies that switch between *pessimism* and *optimism* for the constraint estimation. This approach achieves strong theoretical guarantees, even for difficult problems with small *feasibility gap*. Numerical experiments support these findings, and further reveals the strong long-term performance of a *greedy* approach. Further theoretical results indicate that greedy outperforms pessimism under new metrics defined for *long-term* satisfaction of the constraints, by achieving constant regret and no violation.

In future works, extensions of MAB with revenue guarantees could be considered. The contextual setting (Slivkins et al., 2022) is a natural extension but challenging as it is currently unknown whether a (DO) algorithm can get both $\mathcal{O}(\sqrt{T})$ excess regret and constraint violation. Another promising direction would be to extend the hybrid approach presented in this paper to handle more complicated constraint structures. For instance, a legal contract may specify the policies that should be targeted, whether some constraints are feasible or not, or depending on some problem parameters. We could then design algorithms with multiple switches depending on their own evaluation of their location on the decision tree.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034255.

Nadav Merlis is partially supported by the Viterbi Fellowship, Technion.

Dorian Baudry thanks the support of ANR-19-CHIA-02 SCAI.

References

S. Agrawal and N. R. Devanur. Linear contextual bandits with knapsacks. In *Neural Information Processing Systems*, 2016.

S. Agrawal and N. R. Devanur. Bandits with global convex constraints and objective. *Operation Research*, 2019.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 2012.

S. Amani, M. Alizadeh, and C. Thrampoulidis. Linear stochastic bandits under safety constraints. In *Neural Information Processing Systems*, 2019.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2002.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *Journal of the ACM*, 2018.

D. Baudry, K. Suzuki, and J. Honda. A general recipe for the analysis of randomized multi-armed bandit algorithms. *ArXiv : 2303.06058*, 2023.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 2013.

A. Bura, A. HasanzadeZonuzy, D. M. Kalathil, S. Shakkottai, and J.-F. Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe

reinforcement learning. In *Neural Information Processing Systems*, 2021.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 2013.

E. Carlsson, D. Basu, F. D. Johansson, and D. P. Dubhashi. Pure exploration in bandits with linear constraints. *ArXiv : 2306.12774*, 2023.

M. Castiglioni, A. Celli, A. Marchesi, G. Romano, and N. Gatti. A unifying framework for online optimization with long-term constraints. In *Advances in Neural Information Processing Systems*, 2022.

M. Castiglioni, A. Celli, and C. Kroer. Online learning under budget and roi constraints and applications to bidding in non-truthful auctions. *ArXiv : 2302.01203*, 2023.

S. Chaudhary and D. M. Kalathil. Safe online convex optimization with unknown linear safety constraints. In *AAAI Conference on Artificial Intelligence*, 2021.

T. Chen, A. Gangrade, and V. Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, 2022.

T. Chen, A. Gangrade, and V. Saligrama. Doubly-optimistic play for safe linear bandits. *ArXiv : 2209.13694*, 2023.

Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 2019.

E. Chzhen, C. Giraud, Z. Li, and G. Stoltz. Small total-cost constraints in contextual bandits with knapsacks, with application to fairness. *ArXiv : 2305.15807*, 2023.

H. Claure, Y. Chen, J. Modi, M. F. Jung, and S. Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. *IEEE International Conference on Human-Robot Interaction*, 2020.

Y. Deng, N. Golrezaei, P. Jaillet, J. C. N. Liang, and V. Mirrokni. Multi-channel autobidding with budget and ROI constraints. In *International Conference on Machine Learning*, 2023.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanovi'c. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained mdps. *ArXiv : 2003.02189*, 2020.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 2011.

S. Hutchinson, B. Turan, and M. Alizadeh. Exploiting problem geometry in safe linear bandits. *ArXiv : 2308.15006*, 2023.

N. Immorlica, K. Sankararaman, R. Schapire, and A. Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 2022.

R. Jenatton, J. C. Huang, and C. Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, 2015.

R. Kumar and R. Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. In *Advances in Neural Information Processing Systems*, 2022.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

X. Li, C. Sun, and Y. Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.

J. C. N. Liang, H. Lu, and B. Zhou. Online ad procurement in non-stationary autobidding worlds. *ArXiv : 2307.05698*, 2023.

Q. Liu, W. Xu, S.-Y. Wang, and Z. Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. In *Neural Information Processing Systems*, 2022.

X. Liu, B. Li, P. Shi, and L. Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Neural Information Processing Systems*, 2021.

S. Mannor, J. N. Tsitsiklis, and J. Y. Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 2009.

A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis. Safe linear bandits. In *Annual Conference on Information Sciences and Systems*, 2021.

A. Pacchiano, M. Ghavamzadeh, P. L. Bartlett, and H. Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, 2021.

V. Patil, G. Ghalme, V. J. Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 2021.

V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg. Batched bandit problems. 2016.

K. A. Sankararaman and A. Slivkins. Bandits with knapsacks beyond the worst case. In *Neural Information Processing Systems*, 2020.

A. Sinha. Banditq: Fair multi-armed bandits with guaranteed rewards per arm. In *Arxiv:2304.05219*, 2023.

A. Slivkins, K. A. Sankararaman, and D. J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *Annual Conference Computational Learning Theory*, 2022.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

L. Wang, Y. Bai, W. Sun, and T. Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, 2021.

T. Xu, Y. Liang, and G. Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, 2020.

H. Yu, M. J. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Neural Information Processing Systems*, 2017.

L. Zhou and E. Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. In *International Joint Conference on Artificial Intelligence*, 2016.

X. Zhou and B. Ji. On kernelized multi-armed bandits with constraints. In *Advances in Neural Information Processing Systems*, 2022.

S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton. Cascading bandits for large-scale recommendation problems. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. See Section 1 and Algorithms 2 to 5.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. See theorems 1, 2 and 4, and it is clear that the computational complexity of the algorithm is $\mathcal{O}(K)$ at each step.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes code will be provided in the supplementary material.]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes. See the Appendix.]
   (c) Clear explanations of any assumptions. [Yes.]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, the code will be given in the supplementary materials.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, see section 4.1]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, see Appendix.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes.]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A TECHNICAL LEMMAS

In this section we formalize and prove the results presented at the beginning of Section 3 on the regret caused by the base bandit (Lemma 1) and the sufficient sampling of each arm (Lemma 2). We also provide in Lemma 4 a simple result used to derive the constants presented in the two previously introduced lemmas.

In order to assess the generality of the approaches presented in this paper, we prove the aforementioned result under the following Assumption 2, more general than Assumption 1.

Assumption 2 (Confidence sets)**.** *For any $c > 0$ and collected data $\mathcal{H}_{t-1}$, the target allocation can use a confidence interval $[\mathit{LCB}_{k,t}, \mathit{UCB}_{k,t}]$ satisfying*

$$\mathbb{P}\left(\mu_k \in [\mathit{LCB}_{k,t}, \mathit{UCB}_{k,t}]\right) \geqslant 1 - \frac{1}{t^{2(1+c)}} ,$$

Furthermore, there exists a constant $C > 0$ such that

$$\mathit{UCB}_{k,t} - \mathit{LCB}_{k,t} \leqslant C\sqrt{\frac{\log(t)}{N_k(t-1)}} . \tag{9}$$

Indeed, Assumption 2 is not only satisfied by sub-Gaussian distributions, but by more general exponential families of distributions, and can also be applied to some families of heavy-tail distributions by building the confidence intervals with appropriate robust estimators (see e.g. Bubeck et al. (2013)).

Lemma 1 (Excess-regret caused by $\overline{\mathtt{UCB}}$)**.** *We assume that the confidence bound used by $\overline{\mathit{UCB}}$ satisfy Assumption 2, and use the notation $\bar{p} = (\bar{p}_k)_{k\in[K]}$ for some arbitrary $\bar{p}_k \geqslant 0$. Then, for any time step t we denote by $\overline{\mathit{UCB}}_{j,t}$ the upper confidence bound used for arm j, and define $\mathcal{B}_t = \{\forall j \in [K] : \mu_j \leqslant \overline{\mathit{UCB}}_{j,t}\}$ (optimism), and $\mathcal{N}_{k,t} = \{N_k(t) \geqslant \bar{p}_k t\}$ (sufficient sampling). Then, the number of pulls of any sub-optimal arm k by $\overline{\mathit{UCB}}$ under $\mathcal{B}_t$ and $\mathcal{N}_{k,t-1}$ satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(A_{t+1} = k_t = k, \mathcal{B}_t, \mathcal{N}_{k,t-1})\right] \leqslant \overline{N}_k(T) := \frac{3C^2 \log\left(3 \vee \frac{C^2}{\bar{p}_k \Delta_k^2}\right)}{\bar{p}_k \Delta_k^2} \wedge C^2 \frac{\log(T)}{\Delta_k^2} \wedge T .$$

Furthermore, for large enough time horizons this bound becomes $\overline{N}_k^\star = \frac{3C^2 \log\left(3 \vee \frac{C^2}{\bar{p}_k \Delta_k^2}\right)}{\bar{p}_k \Delta_k^2}$ if $\bar{p}_k > 0$, and $C^2 \frac{\log(T)}{\Delta_k^2}$ otherwise.

Proof. Let us start by fixing a sub-optimal arm k and considering $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(k_t = k, \mathcal{B}_t, \mathcal{D}_t, \mathcal{N}_{k,t-1})\right]$, with $\mathcal{D}_t = \{A_{t+1} = k_t\}$. First, the upper bound by T is trivial. Then, because the base bandit is based on the UCB principle we have that if $k_t = k$ then arm k has the largest upper confidence bound among all arms. In particular, if $\mathcal{B}_t$ holds then $\overline{\mathtt{UCB}}_{k,t} \geqslant \overline{\mathtt{UCB}}_{1,t} \geqslant \mu_1$. We thus obtain that

$$\begin{aligned}
\{k_t = k, \overline{\mathtt{UCB}}_{k,t} \geqslant \max_j \overline{\mathtt{UCB}}_{j,t}, \mathcal{D}_t, \mathcal{B}_t, \mathcal{N}_{k,t-1}\} &\subset \{k_t = k, \mathcal{D}_t, \overline{\mathtt{UCB}}_{k,t} \geqslant \overline{\mathtt{UCB}}_{1,t}, \mathcal{B}_t, \mathcal{N}_{k,t-1}\} \\
&\subset \{k_t = k, \mathcal{D}_t, \overline{\mathtt{UCB}}_{k,t} \geqslant \mu_1, \mathcal{B}_t, \mathcal{N}_{k,t-1}\} \\
&\subset \{k_t = k, \mathcal{D}_t, \mu_k + C\sqrt{\frac{\log(t)}{N_k(t-1)}} \geqslant \mu_1, \mathcal{B}_t, \mathcal{N}_{k,t-1}\} \\
&= \left\{k_t = k, \mathcal{D}_t, C\sqrt{\frac{\log(t)}{N_k(t-1)}} \geqslant \Delta_k, \mathcal{B}_t, \mathcal{N}_{k,t-1}\right\} ,
\end{aligned}$$

where we used Assumption 2 to upper bound $\overline{\mathtt{UCB}}_{k,t} - \mu_k$. It is clear that the final event cannot happen if $N_k(t-1) \geqslant C^2 \frac{\log(t)}{\Delta_k^2}$. Hence, $C^2 \frac{\log(T)}{\Delta_k^2}$, provides the second upper bound on the number of pulls due to the base bandit, independently of the value of $\bar{p}_k$. This result is standard for the analysis of UCB algorithms. Under $\mathcal{N}_{k,t-1}$, we can further prove that the base bandit will not cause pulls arm k after a large enough time. More precisely,

$$t > t_k(\bar{p}_k) := \sup\left\{t \in \mathbb{N} : \bar{p}_k t \leqslant \frac{C^2}{\Delta_k^2}\log(t)\right\} \Rightarrow \mathbb{I}(k_t = k, \mathcal{B}_t, \mathcal{D}_t, \mathcal{N}_{k,t-1}) = 0 .$$

Furthermore, using Lemma 4 we obtain that $t_k(\bar{p}_k) \leqslant \frac{3C^2 \log\left(3 \vee \frac{C^2}{\bar{p}_k \Delta_k^2}\right)}{\bar{p}_k \Delta_k^2}$, giving the first part of Lemma 1. This concludes the proof. □

Lemma 2 (Sufficient sampling). *Under Assumption 2, for any arm $k \in [K]$ there exists a problem-dependent constant Γ_k such that if a policy π satisfies $p^{\pi}_{k,t} \geqslant p^{\texttt{DOC}}_{k,t}$ at all time it holds that*

$$\sum_{t=1}^{+\infty} \mathbb{P}_{\pi}\left(N_k(t) \leqslant \frac{p_k^\star}{8} t\right) \leqslant \Gamma_k = \mathcal{O}\left(\frac{\log\left(\frac{1}{p_k^\star \mu_k^3}\right)^{\frac{3}{2}}}{p_k^\star \mu_k^3} \vee \frac{-\log(p_k^\star)}{p_k^\star} \vee \frac{1}{c}\right).$$

Proof. The proof is based on the following concentration result, that proves that the sample size of each arm k is "close" to the sum of sampling probabilities with high probability.

Lemma 3 (Application of Freedman's inequality). *For any $\delta_t > 0$ and $\eta \in (0, 1]$ it holds that*

$$N_k(t) \geqslant (1-\eta) \sum_{s=1}^{t} p_{k,s} - \frac{1}{\eta} \log(1/\delta_t) \quad \text{with probability at least } 1 - \delta_t \,. \tag{10}$$

Proof. We apply Theorem 1 from Beygelzimer et al. (2011) with the martingale difference $(X_s)_{s \leqslant t}$ defined by $\forall s \leqslant t$, $X_s = \mathbb{I}(A_s = k) - p_{k,s}$, using that $\mathbb{E}[X_s^2 | \mathcal{F}_{s-1}] = p_{k,s}(1 - p_{k,s}) \leqslant p_{k,s} \leqslant 1$. □

Using Lemma 3 with $\delta_t = t^{-(1+p_k^\star)}$ and a parameter $\eta \in (0, 1)$ (defined later) we first obtain that

$$\begin{aligned}
\mathbb{P}\left(N_k(t) \leqslant \frac{p_k^\star t}{8}\right) &\leqslant \mathbb{P}\left(N_k(t) \leqslant (1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1}{\eta}\log(1/\delta_t)\right) + \mathbb{P}\left((1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1}{\eta}\log(1/\delta_t) \leqslant \frac{p_k^\star}{8} t\right) \\
&\leqslant \frac{1}{t^{1+p_k^\star}} + \underbrace{\mathbb{P}\left((1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1+p_k^\star}{\eta}\log(t) \leqslant \frac{p_k^\star}{8} t\right)}_{P_t} .
\end{aligned}$$

We then upper bound P_t by analyzing different scenarios for the sequence $(p_{k,s})_{s \leqslant t}$. We consider two events: $\mathcal{E}_t = \{\cap_{s \in [t/4, t]} \mathcal{B}_s\}$ (μ_k belongs to the confidence intervals for all rounds between $t/4$ and t), and the event $\{N_k(t/2) \geqslant n_t := \frac{C^2 \log(t)}{\mu_k^2}\}$. We then obtain that

$$\begin{aligned}
P_t \leqslant &\underbrace{\mathbb{P}\left((1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1+p_k^\star}{\eta}\log(t) \leqslant \frac{p_k^\star}{8} t, N_k\left(\frac{t}{2}\right) \geqslant n_t, \mathcal{E}_t\right)}_{P'_t} \\
&+ \mathbb{P}\left(N_k\left(\frac{t}{2}\right) \leqslant n_t\right) + \mathbb{P}\left(\bar{\mathcal{E}}_t\right)
\end{aligned}$$

We first upper bound $\sum_{t=1}^{t} \mathbb{P}(\bar{\mathcal{E}}_t)$ thanks to our assumptions on $\mathbb{P}(\bar{\mathcal{B}}_s)$,

$$\sum_{t=1}^{T} \mathbb{P}(\bar{\mathcal{E}}_t) \leqslant \sum_{t=1}^{T} \sum_{s=t/4}^{t} \mathbb{P}\left(\bar{\mathcal{B}}_s\right) \leqslant 12 \sum_{t=1}^{+\infty} \frac{1}{t^{1+c}} \leqslant \frac{12}{c} \,, \tag{11}$$

We then upper bound P'_t. We first remark that

$$\mathcal{E}_t \cap \{N_k(t/2) \geqslant n_t\} \Rightarrow \forall s \in [t/2, t],\ p_{k,s} \geqslant \frac{\lambda_k}{\mu_k + C\sqrt{\frac{\log(s)}{N_k(s)}}} \geqslant \frac{p_k^\star}{2} \,. \tag{12}$$

Using this result, we obtain the following deterministic upper bound

$$\sum_{t=1}^{T} P'_t \leqslant \sum_{t=1}^{T} \mathbb{I}\left(\frac{p_k^\star t}{4}(1-\eta) - \frac{1+p_k^\star}{\eta}\log(t) \leqslant \frac{p_k^\star t}{8}\right)$$
$$\leqslant \sum_{t=1}^{T} \mathbb{I}\left(\frac{p_k^\star t}{8} - \left(\frac{p_k^\star t}{4}\eta + \frac{1+p_k^\star}{\eta}\log(t)\right) \leqslant 0\right)$$

We now optimize η to make this quantity as small as possible, obtaining $\eta = 2\sqrt{\frac{(1+p_k^\star)\log(t)}{p_k^\star t}}$. We can use this value only if $\eta < 1$, but remark that this check is redundant with the indicator being 0. Indeed, we obtain that

$$\sum_{t=1}^{T} P'_t \leqslant \sum_{t=1}^{T} \mathbb{I}\left(\frac{p_k^\star t}{8} \leqslant \sqrt{p_k^\star(1+p_k^\star)t\log(t)} \cup 2\sqrt{\frac{(1+p_k^\star)\log(t)}{p_k^\star t}} > 1\right)$$
$$\leqslant \sum_{t=1}^{T} \mathbb{I}\left(\frac{t}{\log(t)} \leqslant 64\left(1+\frac{1}{p_k^\star}\right)\right) := t_k^0 \,.$$

Using Lemma 4 we further obtain that $t_k^0 = \mathcal{O}\left(\frac{-\log(p_k^\star)}{p_k^\star}\right)$.

It remains to upper bound $\sum_{t=1}^{T}\mathbb{P}\left(N_k(t/2) \leqslant n_t\right)$. We can use the exact same scheme as before, with the significant advantage that n_t scales in $\log(t)$ instead of t, so we can use a cruder lower bound on $\sum_{s=1}^{t/2} p_{k,s}$ to conclude the proof. We now define $\varepsilon'_t = \{\cap_{s\in[t/4,t/2]}\mathcal{B}_s\}$, which provides

$$\varepsilon'_t \Rightarrow \sum_{s=1}^{t/2} p_{k,s} \geqslant \frac{t}{4} \times \frac{\lambda_k}{\mu_k + C\sqrt{\log(t)}} \,.$$

Using this result along with Lemma 3 for $\delta'_t = \frac{1}{t^{1+\lambda_k}}$ we now obtain that for any $\eta \in (0,1)$,

$$\mathbb{P}\left(N_k\left(\frac{t}{2}\right) \leqslant n_t\right) \leqslant \delta'_t + \mathbb{P}\left((1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1+\lambda_k}{\eta}\log(t) \leqslant n_t, \mathcal{E}'_t\right) + \mathbb{P}(\bar{\mathcal{E}}'_t) \,.$$

Similarly as for $\sum\mathbb{P}(\bar{\mathcal{E}}_t)$ we obtain that $\sum_{t=1}^{T}\mathbb{P}(\bar{\mathcal{E}}'_t) \leqslant \frac{8}{c}$, and we also obtain by choosing $\eta = \sqrt{8\frac{(1+\lambda_k)\log(t)}{\lambda_k t}(\mu_k + C\sqrt{\log(t)})}$ in Lemma 3 that

$$\sum_{t=1}^{T}\mathbb{P}\left((1-\eta)\sum_{s=1}^{t} p_{k,s} - \frac{1+\lambda_k}{\eta}\log(t) \leqslant n_t, \mathcal{E}'_t\right) \leqslant \sum_{t=1}^{T}\mathbb{I}\left((1-\eta)\frac{\lambda_k}{\mu_k + C\sqrt{\log(t)}}\frac{t}{8} - \frac{1+\lambda_k}{\eta}\log(t) \leqslant n_t\right)$$
$$= \sum_{t=1}^{T}\mathbb{I}\left(\frac{\lambda_k}{\mu_k + C\sqrt{\log(t)}}\frac{t}{8} - \sqrt{\frac{\lambda_k(1+\lambda_k)t\log(t)}{2(\mu_k + C\sqrt{\log(t)})}} \leqslant n_t\right)$$
$$\leqslant \sum_{t=1}^{T}\mathbb{I}\left(\frac{\lambda_k}{\mu_k + C\sqrt{\log(t)}}\frac{t}{8} \leqslant 2n_t,\right) \vee \sum_{t=1}^{T}\mathbb{I}\left(\frac{\lambda_k}{\mu_k + C\sqrt{\log(t)}}\frac{t}{8} \leqslant 2\sqrt{\frac{\lambda_k(1+\lambda_k)t\log(t)}{2(\mu_k + C\sqrt{\log(t)})}}\right)$$
$$:= t_k^1 \vee t_k^2 \,,$$

that are both problem-dependent constants. Similarly to t_k^0, an upper bound on t_k^1 and t_k^2 can be derived explicitly thanks to Lemma 4, and again we used that the value of η that we choose is valid for $t \geqslant t_k^1 \vee t_k^2$. We thus easily obtain that $t_k^1 = \mathcal{O}\left(\frac{1}{\lambda_k\mu_k^2}\log\left(\frac{1}{\lambda_k\mu_k^2}\right)^{3/2}\right)$ and $t_k^2 = \mathcal{O}\left(\frac{1}{\lambda_k}\log\left(\frac{1}{\lambda_k}\right)^{3/2}\right)$. A summary of all the results obtained so far finally leads to

$$\sum_{t=1}^{+\infty}\mathbb{P}\left(N_k(t) \leqslant \frac{p_k^* t}{8}\right) \leqslant \Gamma_k := t_k^0 \vee t_k^1 \vee t_k^2 + \underbrace{\frac{1}{p_k^\star} + \frac{1}{\lambda_k}}_{\sum_{t=1}^{T}(\delta_t + \delta'_t)} + \frac{20}{c} \,. \tag{13}$$

If c is not unreasonably small, the "characteristic times" $(t_k^i)_{i\in[3]}$ dominate this bound $p_k^\star$ and/or μ_k are small. Thanks to Lemma 4, we obtain the scaling

$$\Gamma_k = \mathcal{O}\left(\frac{1}{p_k^\star\mu_k^3}\log\left(\frac{1}{p_k^\star\mu_k^3}\right)^{3/2} \vee \frac{1}{p_k^\star\mu_k}\log\left(\frac{1}{p_k^\star\mu_k}\right)^{3/2} \vee \frac{-\log(p_k^\star)}{p_k^\star} \vee \frac{1}{c}\right) ,$$

□

and remark that the second term can be removed without changing the result.

Lemma 4. *For any $\alpha \geqslant 1$, the mapping $f_\alpha : x \in [(\alpha+2)^\alpha \vee 3, \infty) \mapsto \sup\left\{t \in \mathbb{N} : \frac{t}{\log(t)^\alpha} \leqslant x\right\}$ satisfies*

$$f_\alpha(x) \leqslant (\alpha+2)^\alpha \times \log(x)^\alpha x.$$

Proof. We start by remarking that the function $g(x) = \frac{x}{\log(x)^\alpha}$ is strictly increasing for all $x \geq e^\alpha$. Now, consider a value $s = Ax\log(x)^\alpha$ for some $A > 0$, such that $s \geqslant 3 \vee e^\alpha$. By the monotonicity of $\frac{t}{(\log t)^\alpha}$, we have that

$$t > s \Rightarrow \frac{t}{(\log(t)^\alpha} > \frac{s}{(\log(s)^\alpha} = x \times \frac{A\log(x)^\alpha}{(\log(A) + \log(x) + \alpha\log(\log(x)))^\alpha} .$$

Then, for $x \geqslant A \geqslant 3$, it holds that $\log(A) + \log(x) + \alpha\log(\log(x)) \leqslant (\alpha+2)\log(x)$, so we can simply choose $A = (\alpha+2)^\alpha$ to obtain the result.

All that is left is to verify that for this choice, $s = (\alpha+2)^\alpha \times \log(x)^\alpha x \geqslant 3 \vee e^\alpha$, but this clearly holds for all $x \geq 3$ and $\alpha > 0$. □

B PROOF OF THEOREMS 1–4

In this section we prove all the main results presented in this paper.

B.1 Proof of Theorem 1

We recall the theorem before detailing the proof.

Theorem 1 (Upper bounds for DOC)**.** *Under Assumption 1, the excess-regret of DOC satisfies*

$$\mathcal{R}_T^{\texttt{DOC}} \leqslant \sum_{k=1}^K \left(\Delta_k(\overline{N}_k^\star + \Gamma_k)\right) \wedge \frac{\alpha\log(T)}{\Delta_k} + \frac{K\max_k\Delta_k}{1+c} ,$$

where α, $\overline{N}_k^\star$ and Γ_k are respectively defined in Equations (5), (6) and (7). If $\max_k\Delta_k \leqslant \Delta^+$ for a fixed $\Delta^+ \in \mathbb{R}$ it furthermore holds that $\mathcal{R}_T^{\texttt{DOC}} = \mathcal{O}(\sqrt{KT\log(T)})$ (pb. independent bound).

Moreover, there exists an absolute constant C_0 such that the constraint violation DOC satisfies

$$\mathcal{V}_T^{\texttt{DOC}} \leqslant C_0\sum_{k=1}^K\sqrt{p_k^\star T\log(T)} + \sum_{k=1}^K\lambda_k\Gamma_k + \frac{K\max_k\mu_k}{1+c} ,$$

and $\mathcal{V}_T^{\texttt{DOC}} = \mathcal{O}(K\sqrt{T\log(T)})$ (pb. independent).

Proof. We divide the proof between the upper bounds on the excess-regret and constraint violation, starting with the excess-regret. As for the technical results of Appendix A, we write the proof under the more general Assumption 2, and instantiate the constants for Assumption 1 only when the final results are derived.

Upper bound on the excess-regret We consider the events $\mathcal{N}_{k,t} = \{N_k(t) \geqslant \bar{p}_k t\}$ for some $(\bar{p}_k)_{k\in[K]}$, $\mathcal{B}_t = \{\forall j \in [K] : \mu_j \in [\mathsf{LCB}_{j,t}, \mathsf{UCB}_{j,t}]\}$, and $\mathcal{D}_t = \{A_{t+1} = k_t\}$. Let us fix a sub-optimal arm k. For any time step t, we use that $p^{\mathsf{DOC}}_{k,t} \leqslant \frac{\lambda_k}{\mathsf{UCB}_{k,t}}$ to obtain that

$$\begin{aligned}
\mathbb{E}[(p^{\mathsf{DOC}}_{k,t} - p^\star_k)_+] &\leqslant \mathbb{E}[(p^{\mathsf{DOC}}_{k,t} - p^\star_k)_+ \mathbb{I}(\mathcal{N}_{k,t-1}, \mathcal{B}_t)] + \mathbb{P}(\bar{\mathcal{N}}_{k,t-1}) + \mathbb{E}\left[p^{\mathsf{DOC}}_{k,t}\mathbb{I}(\bar{\mathcal{B}}_t)\right] \\
&\leqslant \mathbb{E}\left[\left(\frac{\lambda_k}{\mathsf{UCB}_{k,t}} - p^\star_k\right)_+ \mathbb{I}(\mathcal{N}_{k,t-1}, \mathcal{B}_t)\right] + \mathbb{E}\left[\mathbb{I}(k_t = k, \mathcal{N}_{k,t-1}, \mathcal{B}_t, \mathcal{D}_t)\right] + \mathbb{P}(\bar{\mathcal{N}}_{k,t-1}) + \mathbb{E}\left[p^{\mathsf{DOC}}_{k,t}\mathbb{I}(\bar{\mathcal{B}}_t)\right] \\
&\leqslant 0 + \mathbb{E}\left[\mathbb{I}(k_t = k, \mathcal{N}_{k,t-1}, \mathcal{B}_t, \mathcal{D}_t)\right] + \mathbb{P}(\bar{\mathcal{N}}_{k,t-1}) + \mathbb{E}\left[p^{\mathsf{DOC}}_{k,t}\mathbb{I}(\bar{\mathcal{B}}_t)\right] \ .
\end{aligned}$$

We emphasize that due to the optimism under $\mathcal{B}_t$, this equality holds for any choice of $(\bar{p}_k)_{k\in[K]}$.

Now, we first obtain the term $\frac{K\max_k \Delta_k}{1+c}$ by upper bounding the following term,

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \Delta_k \mathbb{E}\left[p_{k,t}\mathbb{I}(\bar{\mathcal{B}}_t)\right] \leqslant \sum_{t=1}^{T} \max_k \Delta_k \mathbb{P}(\bar{\mathcal{B}}_t) \tag{14}$$

$$\leqslant \left(\sum_{t=1}^{T} \delta_t\right) K \max_k \Delta_k = \frac{K\max_k \Delta_k}{1+c} \ . \tag{15}$$

We then obtain the first-order term of the result by using Lemma 1 and Lemma 2. We now consider two different choices for $(\bar{p}_k)_{k\in[K]}$ to bound the two remaining terms.

Case I: $\bar{p}_k = \frac{p^\star_k}{8}$. We can use both lemmas and obtain that $\forall k \in [K]$,

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\mathbb{I}(k_t = k, \mathcal{N}_{k,t-1}, \mathcal{B}_t, \mathcal{D}_t)\right] + \mathbb{P}(\bar{\mathcal{N}}_{k,t-1})\right) \leqslant \overline{N}^\star_k + \Gamma_k \ ,$$

where $\overline{N}^\star_k$ and Γ_k are respectively defined in the statement of Lemma 1 and of Lemma 2.

Case II: $\bar{p}_k = 0$. For this choice, $(\mathcal{N}_{k,t})$ always holds ($\mathbb{P}(\bar{\mathcal{N}}_{k,t}) = 0$), and we complete the result by using Lemma 1 to obtain that, at the same time, this quantity is also bounded by $\alpha\frac{\log(T)}{\Delta_k^2}$, for some $\alpha = C^2$ (with C defined in Assumption 2).

Moreover, this second upper bound by $\alpha\frac{\log(T)}{\Delta_k}$ also guarantees the standard $\mathcal{O}\left(\sqrt{KT\log(T)}\right)$ problem-independent bound, which is directly obtained by taking the maximum between the logarithmic bound and the trivial bound by T. We remark that the upper bound Δ^+ on the gap is necessary to upper bound the term $\frac{K\max_k \Delta_k}{1+c}$.

Constraint violation To upper bound $\mathcal{V}^{\mathsf{DOC}}_T$, we consider any arm $k \in [K]$ for which $\lambda_k > 0$. We again use the events $\mathcal{N}_{k,t}$ and $\mathcal{B}_t$ that we used to upper bound $\mathcal{R}^{\mathsf{DOC}}_T$, so that under $\mathcal{B}_t$ the UCB allocation is feasible. We first write that

$$\sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p^\star_k - p^{\mathsf{DOC}}_{k,t})_+\right] \leqslant \sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p^\star_k - p^{\mathsf{DOC}}_{k,t})_+ \mathbb{I}(\mathcal{N}_{k,t-1}, \mathcal{B}_t)\right] + \lambda_k \sum_{t=1}^{T} \mathbb{P}(\bar{\mathcal{N}}_{k,t-1}) + \mu_k \sum_{t=1}^{T} \mathbb{E}\left[p^*_k \mathbb{I}(\bar{\mathcal{B}}_t)\right] \ .$$

We upper bound the second order terms using Assumption 2 and Lemma 2, obtaining (similar to the proof for the excess regret) $\sum_{k=1}^{K} \lambda_k \Gamma_k + \frac{K\max_k \mu_k}{1+c}$ when summing over the k arms. For the remaining term, we use Assumption 2 to write that

$$\begin{aligned}
\leqslant \sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p^\star_k - p^{\mathsf{DOC}}_{k,t})_+ \mathbb{I}(\mathcal{N}_{k,t-1}, \mathcal{B}_t)\right] &\leqslant \sum_{t=1}^{T} \mu_k \lambda_k \mathbb{E}\left[\left(\frac{\mathsf{UCB}_{k,t} - \mu_k}{\mathsf{UCB}_{k,t} \times \mu_k}\right)_+ \mathbb{I}(\mathcal{N}_{k,t-1}, \mathcal{B}_t)\right] \\
&\leqslant \sum_{t=1}^{T} p^\star_k \times C\sqrt{8\frac{\log(t)}{p^\star_k t}} \ ,
\end{aligned}$$

which concludes the proof for the problem-dependent bound, with $C_0 = 2\sqrt{8}C$. Under Assumption 2 we can define $C = 2\sqrt{6(1+c)}$, which further provides $C_0 = 16\sqrt{3(1+c)}$.

For the problem-independent bound, we take another path that do not use the events $\mathcal{N}_{k,t}$, since the scaling of Γ_k provided by Lemma 2 does not allow us to recover the desired bound. We use that the value of $\mathtt{UCB}_{k,t}$ is determined by $(\widehat{\mu}_{k,n})_{n\in\mathbb{N}}$ (empirical average with sample size $N_k(t-1)=n$), t and the confidence level δ_t. Since the confidence level is increasing with t and δ_t, we can claim that there exists an absolute constant D such that

$$\forall t \in T\,, \mathtt{UCB}_{k,t} \leqslant \widetilde{\mathtt{UCB}}(N_k(t-1), T) \coloneqq \mu_k + D\sqrt{\frac{\log(T)}{N_k(t-1)}} \text{ with probability at least } 1 - \frac{1}{T}\,.$$

We denote by $\mathcal{B}$ the corresponding good event and $\widetilde{p}_{k,t} = \frac{\lambda_k}{\widetilde{\mathtt{UCB}}(N_k(t-1),T)}$, and write that

$$\begin{aligned}
\sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t}^{\mathtt{DOC}})_+\right] &\leqslant \sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p_k^\star - \widetilde{p}_{k,t}^{\mathtt{DOC}})_+\right] \\
&\sum_{t=1}^{T} \mu_k \mathbb{E}\left[(p_k^\star - \widetilde{p}_{k,t}^{\mathtt{DOC}})_+ \mathbb{I}(\mathcal{B})\right] + T\mu_k \mathbb{P}(\bar{\mathcal{B}}) \\
&\leqslant \sum_{t=1}^{T} \mu_k \lambda_k \mathbb{E}\left[\left(\frac{\widetilde{\mathtt{UCB}}_k(N_k(t-1),T) - \mu_k}{\widetilde{\mathtt{UCB}}(N_k(t-1),T) \times \mu_k}\right)_+ \mathbb{I}(\mathcal{B})\right] + \mu_k \\
&\leqslant D p_k^\star \times \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \frac{\mu_k}{\widetilde{\mathtt{UCB}}(N_k(t-1),T)}\sqrt{\frac{\log(T)}{N_k(t-1)}}\mathbb{I}(\mathcal{B})\right]}_{Z_T} + \mu_k\,.
\end{aligned}$$

We further upper bound Z_T using a union bound on $N_k(t-1)$,

$$\begin{aligned}
Z_T &\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=1}^{T} \frac{\mu_k}{\widetilde{\mathtt{UCB}}(n,T)}\sqrt{\frac{\log(T)}{n}}\mathbb{I}(N_k(t-1)=n,\mathcal{B})\right] \\
&\sum_{n=1}^{T} \frac{\mu_k}{\widetilde{\mathtt{UCB}}(n,T)}\sqrt{\frac{\log(T)}{n}}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(N_k(t-1)=n,\mathcal{B})\right]\,.
\end{aligned}$$

Without loss of generality, we assume that $n \geq 1$; the case of $n=0$ is taken care of by sampling each arm once at the beginning, which would not change any of the results.

Notice that the sum in the expectation is the number of rounds since we get to $N_k(t-1)=n$ until we play arm k another time and move to $N_k(t-1)=n+1$, under $\mathcal{B}$. We use again that under $\mathcal{B}$, for all time steps t the sampling probability of k is larger than $\widetilde{p}_k(t) = \frac{\lambda_k}{\widetilde{\mathtt{UCB}}_k(n,T)}$ when $N_k(t-1)=n$. Thus, $\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}(N_k(t-1)=n,\mathcal{B})\right]$ is smaller than the expectation of a geometric random variable with probability $\frac{\lambda_k}{\widetilde{\mathtt{UCB}}_k(n,T)}$. We hence obtain that

$$Z_T \leqslant \sum_{n=1}^{T} \frac{\mu_k}{\widetilde{\mathtt{UCB}}(n,T)}\sqrt{\frac{\log(T)}{n}} \times \frac{\widetilde{\mathtt{UCB}}(n,T)}{\lambda_k} = \sum_{n=1}^{T}\frac{1}{p_k^\star}\sqrt{\frac{\log(T)}{n}}$$

Multiplying Z_T by $Dp_k^\star$, we then conclude that

$$\sum_{t=1}^{T} \mu_k \mathbb{E}\left[p_k^\star - p_{k,t}^{\mathtt{DOC}}\right] \leqslant 2D\sqrt{\log(T)T}\,,$$

which gives the problem-independent bound of $\mathcal{O}(K\sqrt{T\log(T)}$ when summing over the K constraints. □

B.2 Proof of Theorem 2

We recall the theorem below.

Theorem 2 (Upper bounds for SPOC). *Under Assumption 1, SPOC satisfies* $\mathcal{V}_T^{\texttt{SPOC}} \leqslant \mathcal{V}_T^{\texttt{DOC}}$ *as well as*

$$\mathcal{V}_T^{\texttt{SPOC}} = \mathcal{O}\left(\frac{\sqrt{\log\left(\rho_\lambda^{-2} \vee e\right)}}{\rho_\lambda}\sqrt{K D_{\lambda,\mu}}\right) ,$$

where $D_{\lambda,\mu} = \max_{j\in[K]:\lambda_j>0} \frac{\log\left(e\vee(\lambda_j\mu_j)^{-1}\right)}{\lambda_j\mu_j}$. *Moreover, there exists an absolute constant* $C_1 > 0$ *such that*

$$\mathcal{R}_T^{\texttt{SPOC}} \leqslant \sum_{k=1}^K \Delta_k \left(C_1 \frac{\sqrt{p_k^\star T \log(T)}}{\mu_k} + \overline{N}_k(T) + 2\Gamma_k\right) ,$$

where $\overline{N}_k(T)$ *and* Γ_k *are resp. defined in* (5) *and* (7).

Proof. We decompose the proof in two parts, starting with the per-round constraint violation. We again write the proof under the more general Assumption 2, and instantiate the constants for Assumption 1 only when the final results are derived.

Constraint violation: By design, SPOC uses the UCB target allocation if the LCB one is unfeasible. Hence, it directly holds that

$$\forall k \in [K] : \; p_{k,t}^{\texttt{SPOC}} \geqslant p_{k,t}^{\texttt{DOC}} \Rightarrow \mathcal{V}_T^{\texttt{SPOC}} \leqslant \mathcal{V}_T^{\texttt{DOC}} .$$

This gives a first part of the result. We now consider the refined upper bound for problems with positive feasibility gap. Assume that $\rho_\lambda > 0$, and denote by $\mathcal{G}_t = \left\{\sum_{j=1}^K \frac{\lambda_j}{\texttt{LCB}_{j,t}} \leqslant 1\right\}$ the event that the LCB allocation is feasible, and by $\mathcal{B}_t = \{\forall j \in [K] : \; \mu_j \in [\texttt{LCB}_{j,t}, \texttt{UCB}_{j,t}]\}$ the event that all means are well concentrated. We further define $\mathcal{N}_t = \cap_j \mathcal{N}_{j,t} \coloneqq \cap_j \{N_j(t) \geqslant p_j^\star \frac{t}{8}\}$, and first write that for any arm k with $\lambda_k > 0$ it holds that

$$\sum_{t=1}^T \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t})_+\right] \leqslant \sum_{t=1}^T \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t})_+ \mathbb{I}(\mathcal{N}_{t-1}, \mathcal{B}_t)\right] + \lambda_k \sum_{t=1}^T \mathbb{P}(\bar{\mathcal{N}}_t) + \mu_k \sum_{t=1}^T \mathbb{E}[p_k^\star \mathbb{I}(\bar{\mathcal{B}}_t)]$$

We can upper bound the last two terms similarly as in the proof of Theorem 1. Using Lemma 2 and Assumption 2 we obtain that

$$\sum_{k=1}^K \left(\lambda_k \sum_{t=1}^T \mathbb{P}(\bar{\mathcal{N}}_t) + \mu_k \sum_{t=1}^T \mathbb{E}[p_k^\star \mathbb{I}(\bar{\mathcal{B}}_t)]\right) \leqslant \left(\sum_{j=1}^K \lambda_j\right) \sum_{k=1}^K \Gamma_k + \frac{K \max_k \mu_k}{1+c} , \tag{16}$$

where we used a union bound to derive the first term. We now consider the first-order term of the result. Again, we use that by design $p_{k,t}^{\texttt{SPOC}} \geqslant p_{k,t}^{\texttt{DOC}}$ for any k, t. The result comes from using the same proof as for DOC up to a characteristic time t_ρ depending on ρ_λ and other problem parameters. We start by writing that

$$\begin{aligned} V_T^k \coloneqq \sum_{t=1}^T \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t})_+ \mathbb{I}(\mathcal{N}_{t-1}, \mathcal{B}_t)\right] &\leqslant \sum_{t=1}^T \mu_k \mathbb{E}\left[(p_k^\star - p_k^\star \mathbb{I}(\mathcal{G}_t) + p_{k,t}^{\texttt{DOC}} \mathbb{I}(\bar{\mathcal{G}}_t))_+ \mathbb{I}(\mathcal{N}_{t-1}, \mathcal{B}_t)\right] \\ &= \sum_{t=1}^T \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t}^{\texttt{DOC}})_+ \mathbb{I}(\mathcal{N}_{t-1}, \mathcal{B}_t, \bar{\mathcal{G}}_t)\right] . \end{aligned}$$

Then, we use a simple property to introduce ρ_λ in the analysis: if $\forall j \in [K] : , \bar{\mu}_j^{\texttt{LCB}}(t) \geqslant (1-\rho_\lambda)\mu_j$, then the LCB allocation is feasible and $\mathcal{G}_t$ holds. Combining this property with $\mathcal{B}_t$ and $\mathcal{N}_{t-1}$, we obtain that the events considered cannot hold simultaneously when t is larger than a problem-dependent constant, formally

$$t > t_\rho \coloneqq \sup\left\{t \in \mathbb{N} : \; \exists j \in [K] : \; C\sqrt{8\frac{\log(t)}{p_j^\star t}} \geqslant \rho_\lambda \mu_j\right\} . \tag{17}$$

By Lemma 4, we furthermore obtain that $t_\rho \leqslant \max_{j\in[K]} \frac{24C^2}{p_j^\star(\rho_\lambda \mu_j)^2}$. We can thus write that

$$\begin{aligned} V_T^k &\leqslant \sum_{t=1}^{t_\rho} \mu_k \mathbb{E}\left[(p_k^\star - p_{k,t}^{\texttt{DOC}})_+ \mathbb{I}(\mathcal{N}_{t-1}, \mathcal{B}_t)\right] \\ &\leqslant C \sum_{t=1}^{t_\rho} \sqrt{\frac{8 p_k^\star \log(t)}{t}} && \text{(as in Theorem 1)} \\ &\leqslant 2C\sqrt{8 p_k^\star t_\rho \log(t_\rho)} \\ &\leqslant \frac{56C^2}{\rho_\lambda} \max_{j\in[K]} \sqrt{\frac{p_k^\star}{\lambda_j \mu_j} \log\left(\frac{112 C^2}{\lambda_j \mu_j \rho_\lambda^2}\right)} \,. \end{aligned}$$

Remarking that for $(x, y) > 0$ it holds that $x + y \leqslant ((y \vee 1)(1 + x))$, we can further write that

$$V_T^k \leqslant 56C^2 \times \frac{1 \vee \sqrt{\log\left(\frac{1}{\rho_\lambda^2}\right)}}{\rho_\lambda} \max_{j\in[K]} \sqrt{\frac{p_k^\star}{\lambda_j \mu_j}\left(1 + \log\left(\frac{112 C^2}{\lambda_j \mu_j}\right)\right)}, \tag{18}$$

and finally using that $\sum_k \sqrt{p_k^\star} \leqslant \sqrt{K}$ we get

$$\sum_{k=1}^K V_T^k \leqslant \frac{\sqrt{\log\left(\frac{1}{\rho_\lambda^2} \vee e\right)}}{\rho_\lambda} \sqrt{K D_{\lambda,\mu}} \,,$$

where $D_{\lambda,\mu}$ is explicitly defined in Equation (18). This completes the proof regarding the constraint violation.

Excess regret: Let us consider again the events $\mathcal{B}_t = \{\forall j \in [K] \mu_j \in [\texttt{LCB}_{j,t}, \texttt{UCB}_{j,t}]\}$ and $\mathcal{N}_{k,t} = \{N_k(t) \geqslant \frac{p_k^\star}{8} t\}$, $\mathcal{D}_t = \{A_{t+1} = k_t\}$, and $\mathcal{G}_t = \left\{\sum_{k=1} p_{k,t}^{\texttt{LCB}} \leqslant 1\right\}$. We re-use our results for the LCB allocation, remarking that

$$\mathcal{R}_T^{\texttt{SPOC}} \leqslant \underbrace{\sum_{k=1}^K \Delta_k \mathbb{E}\left[(p_{k,t}^{\texttt{SPOC}} - p_k^\star)_+ \mathbb{I}(\mathcal{B}_t, \mathcal{N}_{k,t-1}, \mathcal{G}_t)\right]}_{\widetilde{\mathcal{R}}_T^{\texttt{LCB}}} + \underbrace{\sum_{k=1}^K \Delta_k \mathbb{E}\left[(p_{k,t}^{\texttt{DOC}} - p_k^\star)_+ \left(\mathbb{I}(\mathcal{B}_t, \mathcal{N}_{k,t-1}, \mathcal{D}_t) + \mathbb{I}(\overline{\mathcal{B}_t, \mathcal{N}_{k,t-1}})\right)\right]}_{\widetilde{\mathcal{R}}_T^{\texttt{DOC}}} \,.$$

With the same arguments as for the proof of Theorem 1, we first obtain that

$$\widetilde{\mathcal{R}}_T^{\texttt{DOC}} \leqslant \sum_{k=1}^K \Delta_k \left(\overline{N}_k(T) + \Gamma_k\right) \,.$$

It remains to upper bound the term $\widetilde{\mathcal{R}}_T^{\texttt{LCB}}$. We first remark that if $\rho_\lambda = 0$ then the LCB allocation is unfeasible under $\mathcal{B}_t$ so $\mathcal{R}_T^{\texttt{LCB}} = 0$. Let us now consider the case $\rho_\lambda > 0$. Since lower bounding the number of rounds for which the LCB allocation is feasible is intricate, we simply drop the event $\mathcal{G}_t$ in the rest of the proof after using that

$$(p_{k,t}^{\texttt{SPOC}} - p_k^\star)_+ \mathbb{I}(\mathcal{G}_t) \leqslant p_k^\star \left(\frac{\mu_k - \texttt{LCB}_{k,t}}{\texttt{LCB}_{k,t}}\right)_+ \vee 1 \,.$$

so that we can now work with the confidence intervals. Under $\mathcal{N}_{k,t-1}$ and $\mathcal{B}_t$, we are sure that $\texttt{LCB}_{k,t} \geqslant \frac{\mu_k}{2}$ if

$$t > t_k(1/2) \coloneqq \sup\left\{s \in \mathbb{N} : \; C\sqrt{\frac{8\log(t)}{p_k^\star t}} \geqslant \frac{\mu_k}{2}\right\} \,.$$

We remark from the proof of Lemma 2 that this term is one of the component of Γ_k, so for simplicity we use $t_k(1/2) \leqslant \Gamma_k$ in the statement of Theorem 2. We hence obtain that

$$\widetilde{\mathcal{R}}_T^{\texttt{LCB}} \leqslant \sum_{k=1}^K \Delta_k \Gamma_k + \underbrace{\sum_{k=1}^K \Delta_k \sum_{t=t_k(1/2)}^T 2\mathbb{E}\left[p_k^\star \left(\frac{\mu_k - \texttt{LCB}_{k,t}}{\mu_k}\right) \mathbb{I}(\mathcal{B}_t, \mathcal{N}_{k,t-1})\right]}_{R_T} \,.$$

Finally, using Assumption 2 we obtain that

$$R_T \leqslant \sum_{k=1}^{K} \Delta_k \sum_{t=t_k(1/2)}^{T} 2C \frac{p_k^\star}{\mu_k} \sqrt{\frac{8\log(t)}{p_k^\star t}}$$
$$\leqslant 4C \sum_{k=1}^{K} \frac{\Delta_k}{\mu_k} \sqrt{8 p_k^\star T \log(T)} \, ,$$

which concludes the proof, obtaining that $C_1 = 4C\sqrt{8}$. Under Assumption 1, since $C = 2\sqrt{6(1+c)}$ we obtain that $C_1 = 32\sqrt{3(1+c)}$. □

B.3 Proof of Theorem 3

We recall from Definition 4 that Π_R denotes the set of $\mathcal{R}$-targeting policies ($\mathcal{R}_T = o(\sqrt{T})$) and Π_V denotes the set of $\mathcal{V}$-targeting policies ($\mathcal{V}_T = o(\sqrt{T})$). We prove the following result.

Theorem 3 (Lower bounds)**.** *Consider $\lambda \in \mathbb{R}^K$ and a bandit $\nu \in \mathcal{F}^K$ with means $(\mu_k)_{k\in[K]}$. For any policy $\pi \in \Pi_R$, it holds that*

$$\forall (\nu, \lambda) \in \mathcal{C}, \ \limsup_{T\to\infty} \frac{\mathcal{V}_T^\pi(\nu,\lambda)}{\sqrt{T p_k^\star(\nu,\lambda)}} \geqslant \frac{1}{2\sqrt{e}} \, ,$$

and, for any policy $\pi \in \Pi_V$ it holds that

$$\forall (\nu, \lambda) \in \mathcal{C}^0, \ \limsup_{T\to\infty} \frac{\mathcal{R}_T^\pi(\nu,\lambda)}{\frac{1}{\mu_k}\sqrt{T p_k^\star(\nu,\lambda)}} \geqslant \frac{1}{2\sqrt{e}} \, .$$

Proof. Let us fix a set of parameters $(\lambda_1, \dots, \lambda_k)_{\in(\mathbb{R}^+)^K}$. We start by proving the first statement, assuming that the bandit regret under the policy $\pi \in \Pi$ is dominated by $\sqrt{T}$ asymptotically for all problems. We fix a bandit instance $\nu \in \mathcal{F}$ and, for simplicity, let us consider an arbitrary arm with constraint parameter $\lambda > 0$. In the following, we assume w.l.o.g. that the selected arm is arm 1 (up to re-indexing the arms and constraints).

Then, consider another bandit instance $\nu' \in \mathcal{F}^K$, where the distributions of the arms in (ν, ν') are the same except for arm 1. To keep simple notation, let us now denote by ν and ν' the distribution of arm 1 (only) under the two models.

Let $\mu = \mathbb{E}_{X\sim\nu}[X]$ be the expectation of this arm under ν. Then, choose $\nu' \in \mathcal{F}$ to be absolutely continuous w.r.t. ν and such that $\mathbb{E}_{X\sim\nu'}[X] = \mu + \varepsilon$ for some $\varepsilon > 0$. Assume that $(\lambda_1, \dots, \lambda_K)$ are such that the problem is feasible under ν (by extension, it is feasible under ν'). We further denote by p_t the sampling probability of the selected arm at time step t for a trajectory under π, and use the shorthand notation $p_\nu^\star = \frac{\lambda}{\mu}$ and $p_{\nu'}^\star = \frac{\lambda}{\mu+\varepsilon}$. We consider the event

$$\mathcal{E} = \left\{ \sum_{t=1}^{T} p_t \leqslant \frac{\lambda}{\mu + \frac{\varepsilon}{2}} T \right\} \, .$$

The result follows from a standard change of measure argument. We first remark that

$$\mathcal{E} \text{ holds under } \nu \Rightarrow \sum_{t=1}^{T} (p_\nu^\star - p_t)_+ \geqslant \sum_{t=1}^{T} (p_\nu^\star - p_t) \geqslant T\lambda \frac{\varepsilon}{2\mu\left(\mu + \frac{\varepsilon}{2}\right)} \, ,$$

and similarly that

$$\bar{\mathcal{E}} \text{ holds under } \nu' \Rightarrow \sum_{t=1}^{T} (p_t - p_{\nu'}^\star)_+ \geqslant \sum_{t=1}^{T} (p_t - p_{\nu'}^\star) \geqslant T\lambda \frac{\varepsilon}{2(\mu+\varepsilon)\left(\mu + \frac{\varepsilon}{2}\right)} \, .$$

We use the Bretagnolle-Huber inequality (see e.g. Theorem 14.2 of Lattimore and Szepesvári (2020)),

$$\begin{aligned}\mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(p_\nu^\star - p_t)_+\right] + \mathbb{E}_{\nu',\pi}\left[\sum_{t=1}^{T}(p_t - p_{\nu'}^\star)_+\right] &\geqslant \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(p_\nu^\star - p_t)_+\mathbb{I}(\mathcal{E})\right] + \mathbb{E}_{\nu',\pi}\left[\sum_{t=1}^{T}(p_t - p_{\nu'}^\star)_+\mathbb{I}(\bar{\mathcal{E}})\right] \\ &\geqslant T\lambda\frac{\varepsilon}{2(\mu+\varepsilon)\left(\mu+\frac{\varepsilon}{2}\right)}\left(\mathbb{P}_{\nu,\pi}(\mathcal{E}) + \mathbb{P}_{\nu',\pi}(\bar{\mathcal{E}})\right) \\ &\geqslant T\lambda\frac{\varepsilon}{2(\mu+\varepsilon)\left(\mu+\frac{\varepsilon}{2}\right)}\exp\left(-\mathbb{E}_{\nu,\pi}[N_1(T)]D(\nu,\nu')\right) \ .\end{aligned}$$

By Assumption 1, as ν and ν' are sub-Gaussian it holds that $D(\nu,\nu') \geqslant \frac{\varepsilon^2}{2}$

$$\mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(p_\nu^\star - p_t)_+\right] + \mathbb{E}_{\nu',\pi}\left[\sum_{t=1}^{T}(p_t - p_{\nu'}^\star)_+\right] \geqslant \frac{\lambda}{2\mu^2}\varepsilon T \times \underbrace{\frac{\exp\left(-\mathbb{E}_{\nu,\pi}[N_1(T)]\frac{\varepsilon^2}{2}\right)}{\left(1+\frac{\varepsilon}{\mu}\right)^2}}_{B(T,\varepsilon)}$$

We now consider the properties of $B(T,\varepsilon)$ for small values of ε and large values of T. Since π is an admissible policy (Definition 3), it must hold that $\liminf \mathbb{E}_{\nu,\pi}[N_1(T)] \geqslant p_\nu^\star T$. Hence, for $T \to \infty$ and $\varepsilon \to 0$ we have $\varepsilon B(T,\varepsilon) \sim \varepsilon e^{-p_\nu^\star T\frac{\varepsilon^2}{2}}$, which is maximized by choosing $\varepsilon = (p_\nu^\star T)^{-\frac{1}{2}}$. This choice provides $B(T,\varepsilon) \sim e^{-1/2}$. To complete the proof of the first lower bound, we use that π is $\mathcal{R}$-targeting so $\frac{\mathbb{E}_{\nu',\pi}\left[\sum_{t=1}^{T}(p_t - p_\nu^\star)_+\right]}{\sqrt{T}} \to 0$. Hence, the scaling in $\sqrt{T}$ must come from the contribution of arm 1 to $\mathcal{V}_T^\pi(\nu,\lambda)$. Furthermore, we obtain an asymptotic rate of $\frac{1}{2\sqrt{e}}\sqrt{Tp_\nu^\star}$.

We omit the proof of the second statement of the theorem, since it consists in the exact same steps. The only subtlety is that we now need to assume that ν satisfies $\rho_\lambda(\nu) > 0$ (as indicated in the statement), so that for $\varepsilon > 0$ small enough the relevant alternative problem ν' (such that $\mathbb{E}_{X\sim\nu_1'} = \mu_1 - \varepsilon$) can be feasible for ε small enough. Furthermore, the event of interest for this part of the proof becomes $\mathcal{E} = \left\{\sum_{t=1}^{T} p_t \geqslant \frac{\lambda}{\mu-\frac{\varepsilon}{2}}T\right\}$, so that we can lower bound the excess regret suffered by arm 1 under the assumption that π is $\mathcal{V}$-targeting. □

B.4 Proof of Theorem 4

In order to make the presentation of the proof of Theorem 4 clearer, we detail in Algorithm 5 the implementation of `P-SGOC`.

We then recall the theorem, before proving it.

Theorem 4 (Long-term excess-regret and constraint violation of `P-SGOC`)**.** *Assume that* $\min_{k\in[K]}\lambda_k > 0$*, that* $\rho_\lambda > 0$*, and that* $\max_{k\in[K]}\mu_k \leqslant 1$*. If the distributions are* σ*-sub-Gaussian then* `P-SGOC` *satisfies*

$$\limsup_{T\to\infty}\mathcal{R}_T^{LT} \leqslant 24\sum_{k=1}^{K}\frac{\sigma^2}{\mu_k^2}\Delta_k \ , \ \textit{and} \quad \limsup_{T\to\infty}\mathcal{V}_T^{LT} \leqslant 0 \ .$$

Proof. We start by defining the favorable high-probability events that guarantee the performance of the algorithm.

Success events We consider events that ensure that the algorithm goes to phase 3 with a good estimate $\widehat{\mu_k^2}$. First, we use that if for an arm j the estimate $\widehat{\mu_j^1}$ is well concentrated it is possible to collect $N_j^2 = \Omega(p_j^\star T)$ samples from this arm. More precisely, we have that

$$\mathcal{G}_1 := \left\{\forall j\in[K] : \ \widehat{\mu_j^1} \in \left[\frac{2\mu_j}{3}, 2\mu_j\right]\right\} \Rightarrow \forall j\in[K] : \ N_j^2 \in \left[\frac{p_j^\star}{12}T, \frac{p_j^\star}{2}T\right] \ .$$

Algorithm 5 Phased Safe Greedy-Optimistic Covering (P-SGOC)

Input: $\lambda = (\lambda_1, \dots, \lambda_K)$, time horizon T

Set $\mathcal{S}_K = \{k \in [K] : \lambda_k > 0\}$

Phase 1: (Initial estimation)

for $k \in \mathcal{S}_K$ **do**

Collect $N_k^1 = \frac{\lambda_k}{4} T$ samples $(r_{k,1}, \dots r_{k,N_k^1})$; ▷ Total phase duration smaller than $\frac{T}{4}$ if (λ, μ) feasible.

Compute $\widehat{\mu_k^1} = \frac{1}{N_k^1} \sum_{i=1}^{N_k^1} r_{k,i}$; ▷ First "crude" mean estimate with $N_k^1 = \frac{\lambda_k}{4} T$ samples.

Store the data collected during the phase in $\mathcal{H}_1$.

Phase 2: (Refined estimation)

for $k \in \mathcal{S}_K$ **do**

if $\sum_{j=1}^K \left(\frac{\lambda_j}{6\widehat{\mu_j^1}} \vee \frac{\lambda_j}{2} \right) T \leqslant \frac{T}{2}$; ▷ Ensure a phase duration smaller than $\frac{T}{2}$.

then

$N_k^2 = \left(\frac{\lambda_j}{6\widehat{\mu_j^1}} \vee \frac{\lambda_j}{2} \right) T$; ▷ For T large enough, $N_k^2 = \frac{\lambda_k}{6\widehat{\mu_k^1}} T$ w.h.p. ...

else

$N_k^2 = \frac{\lambda_k}{2}$; ▷ ...But it is guaranteed that $N_k^2 \geqslant \frac{\lambda_k}{2}$ in any case.

Collect N_k^2 samples $(r_{k,1,2}, \dots r_{k,N_k^2,2})$ from arm k

Compute $\widehat{\mu_k^2} = \frac{1}{N_k^2} \sum_{i=1}^{N_k^2} r_{k,i,2}.$; ▷ Refined mean estimate with $N_k^2 = \Omega(p_k^\star T)$ samples w.h.p.

Store the data collected during the phase in $\mathcal{H}_2$

Phase 3: (Target Allocation)

for $k \in \mathcal{S}_K$ **do**

if $\sum_{j=1}^K \frac{\lambda_j}{\widehat{\mu_j^2}} \leqslant 1$ **then**

$N_k^3 = \left(\frac{\lambda_k}{\widehat{\mu_k^2}} - \frac{N_k^1 + N_k^2}{T} \right)_+ T$; ▷ Target a total number of samples $\frac{\lambda_k}{\widehat{\mu_k^2}} T$

else

$\forall j \in [K]$, compute $\widehat{\mathtt{UCB}_j^2}$ such that $\mathbb{P}\left(\widehat{\mathtt{UCB}_j^2} \leqslant \mu_j \right) \leqslant \frac{1}{T}$

Set $N_k^3 = \left(\frac{\lambda_k}{\widehat{\mathtt{UCB}_k^2}} - \frac{N_k^1 + N_k^2}{T} \right)_+ T.$; ▷ Switch to optimistic allocation if unfeasible.

Collect N_k^3 samples from each arm $k \in \mathcal{S}_K$ using round-robin, stop if horizon T is reached.

Store the data collected during the phase in $\mathcal{H}_3$

Phase 4: (Regret minimization)

Play $\overline{\mathtt{UCB}}$ using $\mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$ until horizon T is reached. ; ▷ Last phase with the base bandit.

In particular, under $\mathcal{G}_1$ the algorithm goes to phase 3 with a number of samples for each arm that is a fraction of their optimal allocation. Furthermore, the number of pulls of arm k in the third phase is positive if $\widehat{\mu}_k^2 \leqslant 2\mu_k$. Considering that we also would not want the sampling probability in the third phase to be too high, we naturally consider the second type of good events,

$$\forall k : \mathcal{G}_k^2 \coloneqq \left\{ \widehat{\mu}_k^2 \in \left[\frac{1}{2}\mu_k, 2\mu_k \right] \right\} ,$$

so that under $\mathcal{G} = \mathcal{G}^1 \cap \mathcal{G}_k^2$ the algorithm collects $\frac{\lambda_k}{\widehat{\mu}_k^2} T$ samples for arm k, and $N_k^2 \geqslant \frac{p_k^\star}{12} T$.

Before upper bounding the two metrics, we first provide an auxiliary results that will be used to prove both statements. We upper bound the probability of the bad event, as follows

$$\begin{aligned}
\mathbb{P}(\bar{\mathcal{G}}) &\leqslant \sum_{j=1}^{K} \left\{ \mathbb{P}\left(\widehat{\mu}_j^1 \notin \left[\frac{2\mu_j}{3}, 2\mu_j \right] \right) + \mathbb{P}\left(\widehat{\mu}_k^2 \in \left[\frac{1}{2}\mu_k, 2\mu_k \right], \widehat{\mu}_j^1 \in \left[\frac{2\mu_j}{3}, 2\mu_j \right] \right) \right\} \\
&\leqslant \sum_{j=1}^{K} \left\{ \mathbb{P}\left(\widehat{\mu}_j^1 \notin \left[\frac{2\mu_j}{3}, 2\mu_j \right] \right) + \mathbb{P}\left(\widehat{\mu}_k^2 \in \left[\frac{1}{2}\mu_k, 2\mu_k \right], N_j^2(T) \geqslant \frac{p_j^\star}{12} T \right) \right\} .
\end{aligned}$$

All the terms can then be upper bounded using Hoeffding's inequality, so we obtain that

$$\mathbb{P}(\bar{\mathcal{G}}) \leqslant \sum_j \left\{ e^{-\frac{\lambda_j T}{2\sigma_j^2}\left(\frac{2\mu_j}{3}\right)^2} + e^{-\frac{\lambda_j T}{2\sigma_j^2}(2\mu_j)^2} + e^{-\frac{p_j^\star}{24\sigma_j^2} T\left(\frac{1}{2}\mu_j\right)^2} + e^{-\frac{p_j^\star}{24\sigma_j^2} T(2\mu_j)^2} \right\} . \tag{19}$$

Then, before proving the statements we can also remark that (1) for the long-term metrics the quantity of interest for the analysis is $\mathbb{E}\left[\sum_{t=1}^T p_{k,t}^{\texttt{P-SGOC}}\right]$, and (2) that for deterministic algorithm we can consider the expected number of pulls of each arm instead of the sampling probabilities. Indeed, we can write a phase decomposition as follows

$$\mathbb{E}\left[\sum_{t=1}^{T} p_{k,t}^{\texttt{P-SGOC}} \right] = \mathbb{E}[N_k(T)] = \sum_{s=1}^{4} \mathbb{E}[N_k^i]$$

where N_k^i denotes the number of pulls of arm k in phase i, for $i \in [4]$ (set to 0 if P-SGOC does not reach this phase).

Upper bounding $\mathcal{R}_T^{\mathbf{LT}}$ Under $\mathcal{G}$, the algorithm goes to phase 3 and plays the greedy allocation obtained at the end of phase 2. Further using that $\mathcal{G} \subset \mathcal{G}' \coloneqq \{N_k^2 \geqslant \frac{p_k^\star}{12} T, \widehat{\mu}_k^2 \geqslant \mu_k/2\}$ (as explained above), we can then upper bound $\sum_{s=1}^4 \mathbb{E}[N_k^i]$ as follows,

$$\begin{aligned}
\sum_{s=1}^{4} \mathbb{E}[N_k^i(T)] &\leqslant T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu}_k^2} \times \mathbb{I}\left(\mathcal{G}'\right) \right] + T\mathbb{P}(\bar{\mathcal{G}}') + \mathbb{E}[N_k^4(T)\mathbb{I}(\mathcal{G}')] \\
&= T p_k^\star \underbrace{\mathbb{E}\left[\frac{\mu_k}{\widehat{\mu}_k^2} \times \mathbb{I}\left(\mathcal{G}'\right) \right]}_{A_T} + T\mathbb{P}(\bar{\mathcal{G}}) + \mathbb{E}[N_k^4(T)\mathbb{I}(\mathcal{G}')] .
\end{aligned}$$

The upper bound of Equation (19) indicates that $\lim_{T\to\infty} T\mathbb{P}(\bar{\mathcal{G}}) = 0$. Then, by a direct adaptation of the proof of Lemma 1 we can also obtain that, since $\lambda_k > 0$, $\lim_{T\to\infty} \mathbb{E}[N_k^4(T)\mathbb{I}(\mathcal{G})] = 0$. Hence, for a large enough horizon T excess-regret can only be caused by the term A_T.

We then rewrite A_T in the form of a bias and variance term as follows,

$$A_T = \mathbb{E}\left[\frac{1}{1+\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}} \times \mathbb{I}\left(\mathcal{G}'\right)\right]$$
$$= \mathbb{E}\left[\left(1 - \frac{\widehat{\mu_k^2}-\mu_k}{\mu_k} + \frac{\left(\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}\right)^2}{1+\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}}\right)\mathbb{I}\left(\mathcal{G}'\right)\right], \text{ since } \frac{1}{1+x} = 1 - x + \frac{x^2}{1+x} \text{ for } x > -1.$$

Then, we use that for any sample size n it holds that $\mathbb{E}\left[\widehat{\mu_k^2}\mathbb{I}(\widehat{\mu_k^2} \geqslant \mu_k/2, N_k^2 = n)\right] \geqslant \mu_k$ (the estimate is positively biased by $\mathcal{G}'$), so the bias term is negative. We can thus further upper bound A_T by

$$A_T \leqslant 1 + \mathbb{E}\left[\frac{\left(\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}\right)^2}{1+\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}}\mathbb{I}\left(\mathcal{G}'\right)\right]$$
$$\leqslant 1 + 2\mathbb{E}\left[\left(\frac{\widehat{\mu_k^2}-\mu_k}{\mu_k}\right)^2\mathbb{I}\left(\mathcal{G}'\right)\right], \text{ since } \widehat{\mu_k^2} \geqslant \frac{\mu_k}{2}$$
$$\leqslant 1 + \frac{2}{\mu_k^2} \times \mathbb{E}\left[\left(\widehat{\mu_k^2}-\mu_k\right)^2\mathbb{I}\left(N_k^2 \geqslant \frac{p_k^\star}{12}T\right)\right].$$

Then, for the simplicity of notation we denote by $\bar{\mu}_{k,n}$ the mean estimate corresponding to $N_k^2 = n$. Using that the sample size and that the variance of the mean estimate are independent, we obtain that

$$\mathbb{E}\left[\left(\widehat{\mu_k^2}-\mu_k\right)^2\mathbb{I}\left(N_k^2 \geqslant \frac{p_k^\star}{12}T\right)\right] = \sum_{n=\frac{p_k^\star}{12}T}^{T} \mathbb{P}(N_k^2 = n)\mathbb{E}\left[\left(\bar{\mu}_{k,n}-\mu_k\right)^2\right]$$
$$\leqslant \max_{n \geqslant \frac{p_k^\star}{12}T} \mathbb{E}\left[\left(\bar{\mu}_{k,n}-\mu_k\right)^2\right]$$
$$\leqslant 12\frac{\sigma^2}{p_k^\star T}.$$

We then obtain the desired constant by multiplying this upper bound by $2\frac{p_k^\star T}{\mu_k^2}$, which proves the upper bound provided on $\mathcal{R}_T^{\mathrm{LT}}$: `P-SGOC` suffers constant regret when $\min_k \lambda_k > 0$.

Lower Bounding $\mathcal{V}_T^{\mathbf{LT}}$ We now lower bound $\mathbb{E}[N_k(T)]$, and consider the successful event $\mathcal{G}_1$. Under this event, we are sure that the duration of phase 2 is no more than $T/2$ rounds, that the third phase occurs and that the mean estimate of each arm j used in phase 3 is computed with at least $\frac{p_k^\star}{12}T$ samples. We furthermore define $\mathcal{J} = \left\{\sum_{j=1}^K \frac{\lambda_j}{\widehat{\mu_j^2}} \leqslant 1\right\}$ the event that the greedy allocation proposed at the end of phase 2 is feasible. In this part of the analysis, we omit phase 4 for simplicity, as well as the case when $\bar{\mathcal{J}}$ holds. Indeed, since we assume that $\rho_\lambda > 0$ and we consider the asymptotic problem-dependent bound, we can avoid considering the cases where the UCB allocation is played. Hence, we simply consider the following lower bound,

$$\mathbb{E}[N_k(T)] \geqslant \mathbb{E}\left[\left(N_k^1 + N_k^2 + N_k^3\right)\mathbb{I}(\mathcal{G}, \mathcal{J})\right]$$
$$\geqslant T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu_k^2}}\mathbb{I}\left(\mathcal{J}, \mathcal{G}_1\right)\right].$$

In general, there might be a problem of definition for $\widehat{\mu_k^2}^{-1}$ (that can technically be negative or infinite, even though μ_k is assumed to be positive). However under $\mathcal{J}$ the greedy allocation is feasible, which is possible only if $\widehat{\mu_k^2} \geqslant \lambda_k \geqslant \frac{\lambda_k}{2}$. We can thus use that

$$\frac{\lambda_k}{\widehat{\mu_k^2}}\mathbb{I}\left(\mathcal{J}, \mathcal{G}_1\right) = \frac{\lambda_k}{\widehat{\mu_k^2} + \left(\frac{\lambda_k}{2} - \widehat{\mu_k^2}\right)_+}\mathbb{I}\left(\mathcal{J}, \mathcal{G}_1\right) \ ,$$

which is a convenient re-writing because the right-hand term is now well-defined even when ignoring the events $\mathcal{J}$ and $\mathcal{G}_1$. Using also that under $\mathcal{G}_1$ it holds that $N_k^2 \geqslant \frac{p_k^\star}{12}T$, we obtain that

$$\begin{aligned}
\mathbb{E}[N_k(T)] &\geqslant T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu_k^2} + \left(\frac{\lambda_k}{2} - \widehat{\mu_k^2}\right)_+}\mathbb{I}\left(\mathcal{J}, \mathcal{G}_1\right)\right] \\
&= T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu_k^2} + \left(\frac{\lambda_k}{2} - \widehat{\mu_k^2}\right)_+}\mathbb{I}\left(\mathcal{J}, \mathcal{G}_1, N_k^2 \geqslant \frac{p_k^\star}{12}T\right)\right] \\
&= T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu_k^2} + \left(\frac{\lambda_k}{2} - \widehat{\mu_k^2}\right)_+}\mathbb{I}\left(N_k^2 \geqslant \frac{p_k^\star}{12}T\right)\left(1 - \mathbb{I}\left(\overline{\mathcal{J}, \mathcal{G}_1}\right)\right)\right] \\
&\geqslant \underbrace{T\mathbb{E}\left[\frac{\lambda_k}{\widehat{\mu_k^2} + \left(\frac{\lambda_k}{2} - \widehat{\mu_k^2}\right)_+}\mathbb{I}\left(N_k^2 \geqslant \frac{p_k^\star}{12}T\right)\right]}_{B_T} - 2T\left(\mathbb{P}(\bar{\mathcal{J}}, \mathcal{G}_1) + \mathbb{P}(\bar{\mathcal{G}}_1)\right) \ .
\end{aligned}$$

We then prove that the probabilities corresponding to $\bar{\mathcal{J}}$ and $\bar{\mathcal{G}}_1$ are negligible asymptotically. First, we can upper bound $T\mathbb{P}\left(\bar{\mathcal{G}}_1\right)$ by (19) (only the first two terms of the r.h.s. are necessary). Then, we similarly upper bound

$$T\mathbb{P}(\bar{\mathcal{J}}, \mathcal{G}_1) \leqslant T\mathbb{P}\left(\exists j : \ \lambda_j > 0 \text{ and } \widehat{\mu_j^2} \leqslant (1-\rho_\lambda)\mu_j, \mathcal{G}_1\right) \leqslant T\sum_{j=1}^K e^{-\frac{p_j^\star}{24\sigma^2}T(\rho_\lambda\mu_j)^2} \to 0 \ ,$$

which is again negligible asymptotically, because $\rho_\lambda > 0$. We can thus focus on lower bounding B_T, and use the independence of N_k^2 and $\widehat{\mu}_k^2$ to write (using the same notation $\bar{\mu}_{k,n}$ that we used when upper bounding $\mathcal{R}_T^{\mathrm{LT}}$) that

$$\begin{aligned}
B_T &= p_k^\star T \sum_{n=\frac{p_k^\star}{12}T} \mathbb{P}(N_k^2 = n)\mathbb{E}\left[\frac{\mu_k}{\bar{\mu}_{k,n} + \left(\frac{\lambda_k}{2} - \bar{\mu}_{k,n}\right)_+}\right] \\
&\geqslant p_k^\star T \sum_{n=\frac{p_k^\star}{12}T} \mathbb{P}(N_k^2 = n)\left(1 - \frac{1}{\mu_k}\mathbb{E}\left[\bar{\mu}_{k,n} - \mu_k\right] - \frac{1}{\mu_k}\mathbb{E}\left[\left(\frac{\lambda_k}{2} - \bar{\mu}_{k,n}\right)_+\right]\right) \\
&= p_k^\star T \sum_{n=\frac{p_k^\star}{12}T} \mathbb{P}(N_k^2 = n)\left(1 - \frac{1}{\mu_k}\mathbb{E}\left[\left(\frac{\lambda_k}{2} - \bar{\mu}_{k,n}\right)_+\right]\right) \\
&\geqslant p_k^\star T\left(1 - \mathbb{P}\left(N_k^2 \leqslant \frac{p_k^\star}{12}T\right)\right)\left(1 - \frac{p_k^\star}{2}e^{-\frac{p_k^\star}{96\sigma^2}T\lambda_k^2)}\right) \\
&\geqslant p_k^\star T\left(1 - \mathbb{P}\left(N_k^2 \leqslant \frac{p_k^\star}{12}T\right) - \frac{p_k^\star}{2}e^{-\frac{p_k^\star}{96\sigma^2}T\lambda_k^2)}\right) \ .
\end{aligned}$$

We finally upper bound $\mathbb{P}\left(N_k^2 \leqslant \frac{p_k^\star}{12}T\right) \leqslant \mathbb{P}(\bar{\mathcal{G}}_1)$, that we can again upper bound thanks to (19). We finally obtain that $B_T \sim p_k^\star T$ when $T \to +\infty$. Combining all the results developed in this part, we finally obtain the second statement of the theorem: when $\rho_\lambda > 0$, `P-SGOC` suffers no constraint violation asymptotically. □

C ADDITIONAL EXPERIMENTS

In this section, we provide additional experiments to further support the results presented in Section 4.1. We first consider different value for the feasibility gap ρ_λ on a synthetic experiment with fixed distributions. Then, we redraw the Figure 1 presented in the main text but look only at the performance of SGOC to illustrate its behaviour with respect to the horizon T with better resolution. Lastly, we plot the results we obtain with the approach of Slivkins et al. (2022). All code is written in python. Computations were run on a cluster with 10 cpus and 100 GB of RAM.

Impact of the feasability gap ρ_λ

In Figure 2, we study the impact of the feasibility gap ρ_λ on the performance of SGOC, DOC, SPOC, BanditQ in terms of violation $\mathcal{V}_T$ and excess regret $\mathcal{R}_T$ and the performance of SGOC, DOC, SPOC, BanditQ, SGOC, P-SGOC in terms of long term violation $\mathcal{V}_T$ and long term excess regret $\mathcal{R}_T^{LT}$. We take $K = 3$ arms, $\boldsymbol{\mu} = (0.8, 0.9, 0.7)$ with the feasibility ρ_λ varying in $\{0, 0.1, 0.5, 0.9\}$ and the horizon varying in $[10^2, 10^5]$. We set $\lambda = \boldsymbol{\mu}(1 - \rho_\lambda)/K$. We take 200 seeds and report the mean value across seeds. Error bars represent the first and last decile.

When $\rho_\lambda = 0$, SPOC, DOC and SGOC behave like DOC and therefore have low regret but a constraint violation in $\sqrt{T}$. Looking at long term metrics, we see that all algorithms have low regret and positive fairness violation. BanditQ seems to get the best trade-off in this setting.

As the feasibility increases, SGOC, DOC, SPOC, and SGOC behave similarly as in the experiment in Figure 1 in the main text. In particular, we observe when $\rho_\lambda = 0.1$, the transition from optimism to pessimism of SPOC yielding a $\sqrt{T}$ regret but constant fairness violation. SGOC still has excellent performance with respect to long term metrics. The difference is the behavior of BanditQ. When $\rho_\lambda > 0$, BanditQ seems to have high long term excess regret but low long term violation. This is the opposite of what was observed in the experiment in Figure 1 in the main text. Such behaviour contrasts with the predictability of approaches like SPOC or DOC.

The regret and violation in $\sqrt{T}$ of SGOC

We redraw in Figure 3 the plot in Figure 1 but displaying only the performance of SGOC. We observe that the regret and constraints violation of SGOC evolve as $\sqrt{T}$ as expected from the analysis.

Results with LagrangeBwK (Slivkins et al., 2022)

In Figure 4 we consider the same simulation setup as the one in Figure 1, including the LagrangeBwK algorithm from Slivkins et al. (2022). We observe a linear constraint violation, showing that this implementation of the algorithm does not seem to converge to an optimal allocation for the problem and time horizon considered. We used the learning rates suggested in Slivkins et al. (2022), and for completeness we provide the implementation that we used in the supplementary material.

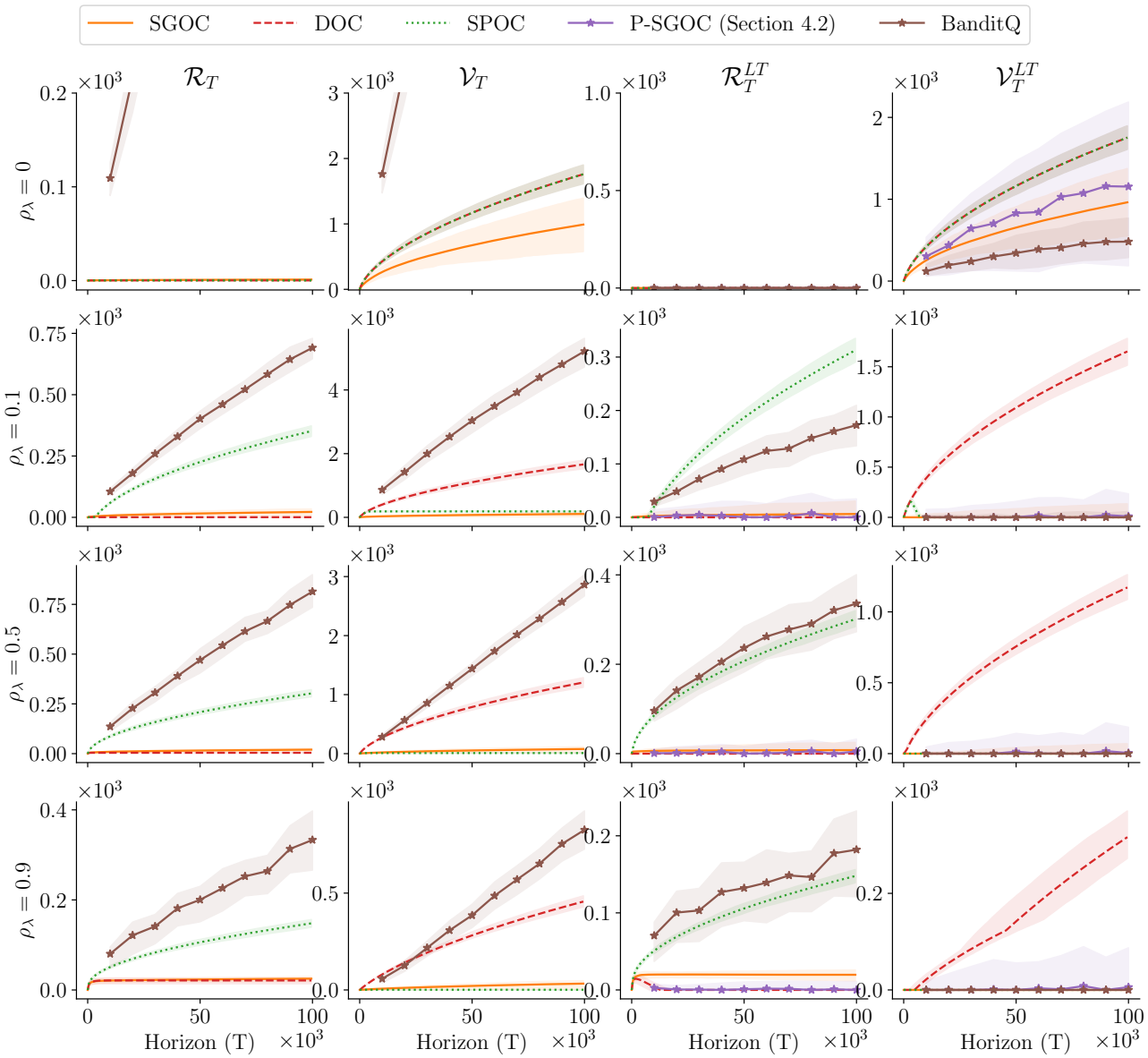

Figure 2: Simulations with increasing feasibility gap ρ_λ. The plots in each row uses a different value of ρ_λ, the plots in each column represents a different metric.

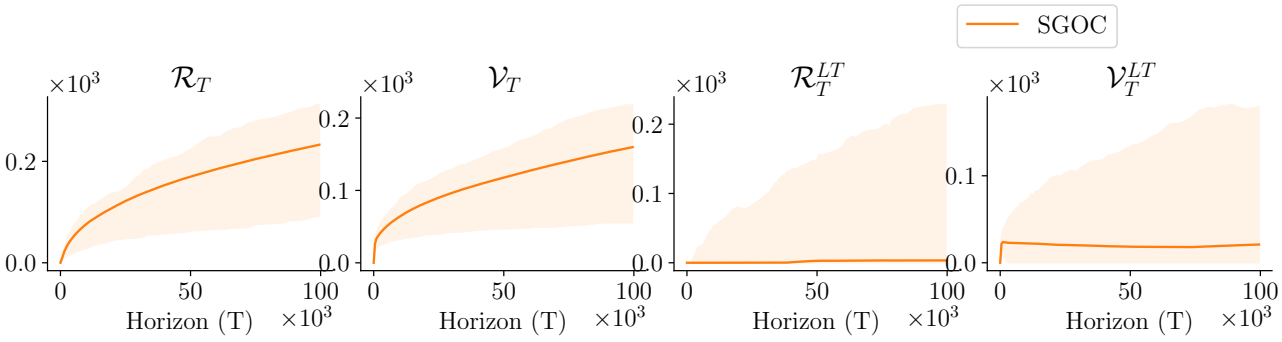

Figure 3: Reproducing the simulation setup from Sinha (2023) focusing only on SGOC for better resolution

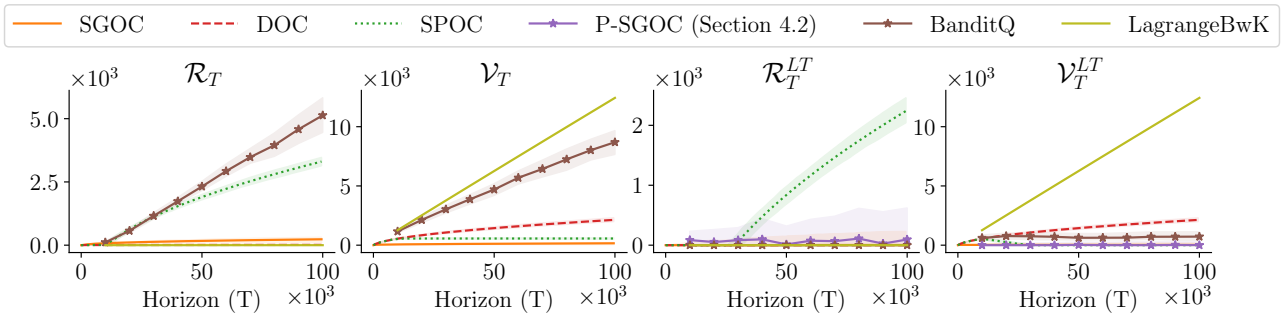

Figure 4: Performance of LagrangeBwK (Slivkins et al., 2022) on the simulation setup from Sinha (2023)