Online Budget Allocation with Censored Semi-Bandit Feedback

François Bachoc
Laboratoire Paul Painlevé
Université de Lille
Lille, France
Institut universitaire de France (IUF)
francois.bachoc@univ-lille.fr

Nicolò Cesa-Bianchi
Department of Computer Science
Università degli Studi di Milano
Milano, Italy
nicolo.cesa-bianchi@unimi.it

Tommaso Cesari
EECS
University of Ottawa
Ottawa, Canada
tcesari@uottawa.ca

Roberto Colomboni
School of Mathematics
University of Bristol
Bristol, United Kingdom
roberto.colomboni@bristol.ac.uk

Abstract

We study a stochastic budget-allocation problem over K tasks. At each round t, the learner chooses an allocation $X_t \in \Delta_K$. Task k succeeds with probability $F_k(X_{t,k})$, where $F_1, \dots, F_K$ are nondecreasing *budget-to-success* curves, and upon success yields a random reward with unknown mean μ_k. The learner observes which tasks succeed, and observes a task's reward *only* upon success (*censored* semi-bandit feedback). This model captures, for instance, splitting payments across crowdsourcing workers or distributing bids across simultaneous auctions, and subsumes stochastic multi-armed bandits and semi-bandits. We design an optimism-based algorithm that operates under censored semi-bandit feedback and assumes access to an exact maximization oracle. Our main result is an instance-dependent bound showing that in *diminishing-returns* regimes, the regret of this algorithm scales polylogarithmically with the horizon T without any *ad hoc* tuning. For general nondecreasing curves, we prove that the same algorithm (with the same tuning) achieves a worst-case regret upper bound of $\widetilde{O}(K\sqrt{T})$. Finally, we establish a matching worst-case regret lower bound of $\Omega(K\sqrt{T})$ that holds even for *full-feedback* algorithms, highlighting the intrinsic hardness of our problem outside diminishing returns.

1 Introduction

The multi-armed bandit, a mathematical model for sequential decision-making with partial feedback, has been applied to a wide range of domains, including clinical trials, product recommendation, online advertising, digital markets, and process optimization. A multi-armed bandit consists of an action space and an unknown reward process associated with each action. At every decision round, a bandit algorithm selects an action and observes the reward produced at that round by the process associated with the chosen action [Lattimore and Szepesvári, 2020]. In this work we focus on a setting where rewards are stochastic and actions have a linear structure. The expected reward is $\langle \mu, F(x) \rangle$, where x is the selected action, $\mu \in [0,1]^K$ are unknown coefficients, and $F\colon \Delta_K \to [0,1]^K$ is some (possibly nonlinear) function defined on the probability simplex Δ_K, with $F(x)$ to be interpreted as the means of K Bernoulli random variables. The realizations of these random variables, which

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the algorithm observes after each round, reveal which of the K components of the chosen action actually contributed to the total reward; for each such contributing component, the algorithm observes the corresponding reward. The total reward is then the sum of the rewards associated with the completed tasks. Hence, compared to the standard bandit feedback in linear bandits [Dani et al., 2008], the feedback we receive—what we call *censored semi-bandit*—is more informative. This natural censored-feedback model goes beyond the settings treated in prior analyses and leads to stronger learning guarantees. To illustrate the scope of the model, we highlight two representative scenarios that fit our framework (with action $x \in \Delta_K$, completion probabilities $F(x)$, and expected reward $\langle \mu, F(x) \rangle$):

- In crowdsourcing, a platform allocates in each round a fixed budget across K workers. Paying worker k a share x_k induces task completion with probability $F_k(x_k)$, naturally increasing with x_k. Thus $F(x) = \big(F_1(x_1), \dots, F_K(x_K)\big)$ models how payments translate into completion. The platform's expected reward in the round is $\langle \mu, F(x) \rangle$, where $\mu_k \in [0,1]$ captures the (unknown) expected value of the work produced upon completion by the k-th worker.
- An autobidding agent spends in each round a fixed budget to bid in K simultaneous auctions. Auctions are won with probabilities $F(x) = \big(F_1(x_1), \dots, F_K(x_K)\big)$ depending on the agent's bids $x = (x_1, \dots, x_K)$ and the competing bids in each auction. In each round, the agent's expected reward $\langle \mu, F(x) \rangle$ depends on the unknown value of each auctioned item, which is modeled via K random variables with means $\mu \in [0,1]^K$.

Other applications include online model selection, where a computing time budget is split among several instances of a learning system (e.g., a LLM trained with different hyperparameter values) and the expected reward is computed based on the quality of the output produced by those instances that terminated training (or inference) within the allocated time—see, e.g., Xia et al. [2024] for work along these lines. Note that our theoretical model is representative of simplified versions of these applicative examples. In crowdsourcing, our model means that payment mainly affects acceptance or completion probability while the output quality is invariant [Tran-Thanh et al., 2014, Kang and Tay, 2020]. In autobidding, our model does not capture global budget constraints affecting multiple time steps [Aggarwal et al., 2025, Avadhanula et al., 2021, Deng et al., 2023]. Similarly, for compute-allocation, our model does not capture multi-round jobs, queuing, and delayed observations [Xia et al., 2024, Li, 2025, Nguyen et al., 2024].

1.1 Setting

For any $n \in \mathbb{N}$, we let $[n]$ be the set $\{1, \dots, n\}$. The action space is the $(K-1)$-dimensional simplex $\Delta_K \coloneqq \{x \in [0,1]^K : \sum_{k \in [K]} x_k = 1\}$, for some natural number $K \geq 2$, which we fix for the remainder of this work. Let $G_{t,k}$ be the stochastic reward associated with the completion of task $k \in [K]$ at time t. For each $k \in [K]$ and $t \in \mathbb{N}$, we model $G_{t,k}$ as an unknown $[0,1]$-valued random variable with unknown expectation μ_k, not depending on t. We further assume the vectors $(G_{t,1}, \dots, G_{t,K})_{t \in \mathbb{N}}$ to be independent. For each task $k \in [K]$, the distribution of $G_{t,k}$ does not depend on t. If we allocate a budget $x_k \in [0,1]$ on a task $k \in [K]$, then the probability of completing the task k is $F_k(x_k)$, where $F_1, \dots, F_K \colon [0,1] \to [0,1]$ are known[1] non-decreasing functions, that we call *budget-to-success* curves. Also, we let $B_{t,k}(x_k)$ be the Bernoulli random variable that indicates the completion of the task k at time t, so that $\mathbb{E}\big[B_{t,k}(x_k)\big] = F_k(x_k)$. We model the family of random functions $(B_{t,k})_{t \in \mathbb{N}, k \in [K]}$ as an independent family, independent of $(G_{t,k})_{t \in \mathbb{N}, k \in [K]}$. [2] For notational convenience, we define: $\mu \coloneqq (\mu_1, \dots, \mu_K)$, $F \colon \Delta_K \to [0,1]^K, (x_1, \dots, x_K) \mapsto \big(F_1(x_1), \dots, F_K(x_K)\big)$, and $G_t \coloneqq (G_{t,1}, \dots, G_{t,K})$. Moreover, for each $t \in \mathbb{N}$, we let B_t be the function that maps each budget allocation $x \coloneqq (x_1, \dots, x_K) \in \Delta_K$ to the vector of random variables

[1] In Appendix A, we investigate the case where the budget-to-success curves are unknown. In this case, we obtain a *strong* impossibility result: even for $K = 2$, when the budget-to-success curves F_1 and F_2 are indicator functions, and the random variables G_t are deterministically equal to $(1,1)$, any algorithm must incur worst-case linear regret. See also Remark 4.

[2] Technically, we need the function $(x, \omega) \mapsto B_{t,k}(x)(\omega)$—as a joint function of the index x and the realization ω—to be jointly measurable as well. This is necessary to avoid measure theoretic pathologies. On the other hand, most common constructions naturally satisfy this assumption. For example, $B_{t,k}(x) \coloneqq \mathbb{I}\{V_{t,k} \leq F_k(x)\}$, where $(V_{t,k})_{t \in \mathbb{N}, k \in [K]}$ is an i.i.d. family of uniform random variables on $[0,1]$. For the sake of simplicity, we shall not discuss measurability issues further in the rest of the paper.

Learning protocol 1: Learning protocol

for *time step* $t = 1, 2, \ldots$ **do**
The learner chooses $X_t \in \Delta_K$;
The learner gains $\langle G_t,\, B_t(X_t)\rangle$;
The learner observes some feedback $\mathscr{F}_t$;

$B_t(x) \coloneqq \big(B_{t,1}(x_1), \ldots, B_{t,K}(x_K)\big)$. Finally, if $a, b \in \mathbb{R}^K$, we denote by $\langle a, b\rangle$ their scalar product $\sum_{k\in[K]} a_k b_k$.

Given these definitions, the online learning interaction is summarized in Algorithm 1.

We focus on the following two feedback models.[3]

- **Censored semi-bandit:** $\mathscr{F}_t$ is $B_t(X_t)$ and $\big(B_{t,1}(X_{t,1})\, G_{t,1},\, \ldots,\, B_{t,K}(X_{t,K})\, G_{t,K}\big)$. In words, the learner observes which tasks have been completed and the rewards of *only* these completed tasks.
- **Full feedback:** $\mathscr{F}_t$ is $B_t(X_t)$ and G_t. In this case, the learner observes the rewards of *all* tasks.

The performance of the learner at any time horizon $T \in \mathbb{N}$ is evaluated with the *regret*, defined as

$$R_T \coloneqq T\left(\sup_{x\in\Delta_K} \big\langle \mu, F(x)\big\rangle\right) - \sum_{t=1}^{T} \mathbb{E}\Big[\big\langle \mu, F(X_t)\big\rangle\Big] \tag{1}$$

i.e., the cumulative difference between the expected reward obtained by the best allocation of the budgets and the expected performance of the learning algorithm generating the sequence $X_1, X_2, \ldots$. If $\sup_{x\in\Delta_K} \langle \mu, F(x)\rangle$ is attained, we will denote one of its maximizers by $x^\star(\mu)$ to stress the dependence on μ, or simply by $x^\star$ if this dependence is clear from context. To give guarantees that are uniform over all possible problem instances, we will also consider the worst-case regret $R_T^\star \coloneqq \sup R_T$, where the supremum is taken over all choices of $F_1, \ldots, F_K$, and joint distributions of $(G_{t,1}, \ldots, G_{t,K})$.

We conclude this section by noting that our model subsumes the stochastic combinatorial semi-bandit model [Chen et al., 2013] with M-sets and $M \in [K]$ (and, in particular, K-armed bandits when $M = 1$). Indeed, let $\mathbb{I}_A$ denote the indicator of a set A, and let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ be the standard basis of $\mathbb{R}^K$. If $F_1 \coloneqq \cdots \coloneqq F_K \coloneqq \mathbb{I}_{[1/M,1]}$, then task k is completed if and only if it receives at least a $1/M$ budget share. In this case, the Pareto-optimal allocations are exactly those of the form $x = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e}_{i_m}$ with $i_1, \ldots, i_M$ distinct in $[K]$, so in each round the learner selects M arms and earns and observes their individual rewards, recovering the standard semi-bandit feedback model with M plays per round.

1.2 Our contributions

We introduce a flexible online learning setting that generalizes the classic semi-bandit framework and models, in particular, crowdsourcing, multi-platform autobidding, and the allocation of compute resources for training large language models (LLMs). We make the following contributions.

- We design an intuitive optimism-based algorithm (Algorithm 2) that operates using only censored semi-bandit feedback observations and assumes access to an exact maximization oracle, as discussed in Remark 3.
- For a practically relevant and technically challenging *diminishing-returns* regime, we prove an instance-dependent bound showing that Algorithm 2 achieves a fast polylogarithmic regret rate *without any ad hoc tuning* (Theorem 1).
- We prove that this fast rate *cannot* be obtained from *bandit* feedback alone: any (linear) bandit algorithm incurs $\Omega(\sqrt{T})$ regret even in the diminishing-returns regime (Theorem 2).

[3] Except in Theorem 2, where we introduce the third model of *bandit* feedback to highlight the gap with the standard linear-bandit literature. Outside of Theorem 2, we establish all *upper bounds* under the realistic and challenging censored semi-bandit model, and we prove the corresponding *lower bounds* under the simpler full-feedback model. In particular, our minimax lower bounds under full feedback apply *a fortiori* to censored feedback, showing that our worst-case guarantees under censored semi-bandit feedback match what is achievable even by powerful algorithms with full-feedback access.

• For general nondecreasing curves, we prove that Algorithm 2 achieves a worst-case regret of $\widetilde{\mathcal{O}}\big(K\sqrt{T}\big)$ (again, without any *ad hoc* tuning; Theorem 3).

• For general nondecreasing curves, we prove that no algorithm can guarantee worst-case regret better than $\Omega\big(K\sqrt{T}\big)$, not even with full-feedback access (Theorem 4).

These results not only establish the worst-case optimality of Algorithm 2, but also highlight its ability to adaptively exploit structure, illustrating its broad applicability in real-world domains.

1.3 Techniques and challenges

Why off-the-shelf bandit techniques fail. Under censored semi-bandit feedback, the learner can compute the realized *total* reward in each round, and therefore has access to standard bandit feedback. Thus, in principle, one could attempt to apply bandit algorithms directly over the (infinite) action set Δ_K.

However, classical bandit guarantees scale with the number of arms H (typically as $\sqrt{HT}$), and in our problem H is not finite. A standard workaround in $\mathcal{X}$-armed bandits is to discretize the action space, provided the expected reward function is sufficiently regular (e.g., Lipschitz). In our setting, this regularity is *not* guaranteed: the budget-to-success curves F_k need not be Lipschitz, and consequently the induced expected reward may exhibit arbitrarily steep variations.

Moreover, even under an L-Lipschitz assumption, discretization suffers from the curse of dimensionality. Since the $(1/n)$-covering number of the unit simplex Δ_K scales as $\Theta(n^{K-1})$, an $(1/n)$-net requires $\Theta(n^{K-1})$ grid points to ensure that every point (in particular, any optimum) lies within distance $1/n$ of the grid. This yields regret guarantees of the form $O\Big(\frac{L}{n}T + \sqrt{n^{K-1}T}\Big)$, which, after optimizing over n, give the nonparametric rate $O\big(L^{(K-1)/(K+1)}\,T^{K/(K+1)}\big)$.

One might hope to do better by exploiting bandit feedback together with the *linear* structure of the realized reward, which can be written as $\langle G_t, B_t(X_t)\rangle$. Yet, as discussed in Appendix G and further shown in Theorem 2, even linear-bandit techniques are suboptimal in our setting.

Diminishing-returns regime. Proving (poly-)logarithmic regret when the budget-to-success curves exhibit strong diminishing returns raises three main challenges.

(i) From estimation error to second-order regret. We first develop a self-bounding argument showing that, when the optimal allocation lies in the interior of the simplex, optimizing the objective computed from estimated task means yields a second-order *instantaneous* regret that scales quadratically with the estimation error in the mean-reward vector.

(ii) Interior optimality without assuming it. A subtler issue is that interior optimality cannot be taken for granted. We show that it is in fact an implication of our regime assumptions: when every task has strictly positive mean reward, the optimal allocation is forced away from the boundary, and no component is starved.

(iii) Quantifying the feedback rate under optimism. Finally, we must control how quickly the mean estimates concentrate under the optimistic allocation rule. Using concentration inequalities, we prove that once the cumulative mass allocated to a task is large enough (logarithmic in T), the number of observed completions is of the same order with high probability. Consequently, each mean can be estimated at the $1/\sqrt{t}$ rate (the same as in full-feedback regimes, despite only having access to censored semi-bandit feedback), which combined with the second-order regret property yields the desired (poly-)logarithmic guarantee.

(iv) Not another standard bandit $\log$ *bound.* Our fast-rate result should not be viewed as a routine analog of classical logarithmic bandit bounds. In finite-armed bandits, logarithmic regret is tied to a positive gap between the best and second-best arm. Here, by contrast, the action space is continuous, so the usual gap-based intuition breaks down: the standard bandit gap can be zero, and yet we still obtain a polylogarithmic rate. The mechanism is fundamentally different, and the fast rate does not come from a classical gap condition, but from a much subtler interplay between curvature, interior optimality induced by $\mu_k > 0$, and the richer observability of censored semi-bandit feedback. This point is further developed in Theorem 2: bandit feedback alone does *not* yield such a fast rate, even in very benign instances where curvature and interior optimality are both present. So the result is

not simply "logarithmic regret with a complicated constant," but rather a qualitatively new fast-rate phenomenon that is specific to this feedback/geometry combination.

Worst-case upper bound. We obtain optimal worst-case regret by exploiting two key features of our model. First, the per-round reward decomposes as a sum of contributions from K tasks. Second, allocating more budget to a task increases its chance of completion and hence the probability of observing its reward; in expectation, the rate at which we collect observations from a task is coupled with the reward we can obtain from it. Together, these properties naturally suggest an optimistic strategy: as soon as a component looks promising, we funnel more budget into it. This increases the flow of feedback from that component, which either confirms the choice or quickly reveals it to be suboptimal, allowing the learner to reallocate before incurring large regret.

We formalize this intuition via an argument reminiscent of the elliptical potential lemma [Lattimore and Szepesvári, 2020, Lemma 19.4]. In particular, under our optimistic allocation rule, the regret attributable to each task can be controlled by the square root of the number of times its reward is observed, which yields an overall worst-case regret of order $K\sqrt{T}$ (Theorem 3).

Worst-case lower bound. To show that $K\sqrt{T}$ is the best possible rate in general—*even with full-feedback access*—we construct a hard instance with $2K$ tasks, arranged into K disjoint pairs. Within each pair, one task is slightly better than the other. The budget-to-success curves $F_1, \dots, F_{2K}$ are designed so that, in each round, any allocation effectively selects a subset of tasks that are *active* (success probability 1) while the remaining tasks are *inactive* (success probability 0). This structure allows us to embed K independent two-armed bandit problems, one per pair.

A key technical difficulty is making this embedding rigorous for *arbitrary* algorithms over the simplex. Indeed, the learner is not restricted to activating exactly one task per pair: it may allocate budget so as to activate none in some pairs and both in others. We prove that, nevertheless, any strategy can be converted into (or compared against) a strategy that induces K independent two-armed decisions without increasing regret. We are not aware of an analogous reduction in the existing literature.

Once this reduction is established, the remaining argument is somewhat closer to the existing literature than for the reduction mentioned above, and follows a classical bandit lower-bound blueprint: since each of the K embedded two-armed problems incurs $\Omega(\sqrt{T})$ regret in the worst case, the overall regret is at least $\Omega(K\sqrt{T})$. However, making this step precise requires a careful information-theoretic analysis tailored to our feedback model, relying on tools from statistical hypothesis testing and information theory.

We conclude this section by remarking that our lower bound is consistent with a standard observation from combinatorial semi-bandits: in the M-set semi-bandit, when $M = \lfloor K/2 \rfloor$ (so that half of the coordinates is observed each round), the minimax worst-case rate is of the same order under semi-bandit and full-feedback. However, this connection is not entirely plug-and-play in our model: semi-bandits arise as a special case only after restricting attention to Pareto-optimal allocations (which, under the threshold curves, correspond exactly to uniform $1/M$ allocations over an M-subset), and our proof makes this reduction explicit.

1.4 Related work

For a more detailed discussion of related work, see Appendix G. The closest related work is the concurrent and independent work of Montanari et al. [2026]. They study a resource-allocation problem with censored feedback and Bernoulli rewards, and analyze learning in a specific smooth parametric threshold model with unknown parameters, which is outside the scope of our impossibility result in Appendix A. On the shared subclass, our $\sqrt{T}$ guarantee is uniform over instances satisfying our oracle assumption (Remark 3), whereas theirs is instance-dependent; Appendix G details the dependence on K. For fast rates, we derive interiority from the behavior of the curves at the origin, whereas they assume it explicitly; the two sets of assumptions are not directly comparable.

Our setting is also related to several other resource-allocation and bandit models. Resource-allocation bandits in [Lattimore et al., 2014, 2015] learn unknown capped-linear completion maps, whereas here the budget-to-success curves are known and the unknown reward means are learned from censored observations. Combinatorial semi-bandits with probabilistically triggered arms [Chen et al., 2016] provide a finite-action analogue, while our action space is continuous and triggering probabilities may become arbitrarily small near the boundary (for small budget). The censoring mechanism in

Algorithm 2:

Input: Time horizon $T \in \mathbb{N}$, confidence parameter $\delta \in (0,1)$

Initialization: Let $\mathrm{UCB}_0^\delta \coloneqq (\mathrm{UCB}_{0,1}^\delta, \dots, \mathrm{UCB}_{0,K}^\delta)$ where, for all $k \in [K]$, $\mathrm{UCB}_{0,k}^\delta \coloneqq \sqrt{\log \frac{2}{\delta}}$

for *time step* $t = 1, \dots, T$ **do**

Play $X_t \in \Delta_K$ such that $X_t \in \operatorname{argmax}_{x \in \Delta_K} \left\langle \mathrm{UCB}_{t-1}^\delta, F(x) \right\rangle$

Update $\mathrm{UCB}_t^\delta \coloneqq (\mathrm{UCB}_{t,1}^\delta, \dots, \mathrm{UCB}_{t,K}^\delta)$, where, for all $k \in [K]$,

$$\mathrm{UCB}_{t,k}^\delta \coloneqq \frac{\sum_{s=1}^t B_{s,k}(X_{s,k}) G_{s,k}}{\sum_{s=1}^t B_{s,k}(X_{s,k})} + \sqrt{\frac{\log \frac{2}{\delta}}{1 + \sum_{s=1}^t B_{s,k}(X_{s,k})}}$$

with the understanding that $\mathrm{UCB}_{t,k}^\delta \coloneqq \mathrm{UCB}_{0,k}^\delta$ if $B_{1,k}(X_{1,k}) = \cdots = B_{t,k}(X_{t,k}) = 0$

Verma et al. [2019] operates in the opposite direction to ours, and Fontaine et al. [2020] consider concave resource allocation with noisy first-order rather than censored reward feedback. Finally, there is related work in online learning with auctions [Hummel and McAfee, 2016, Iyer et al., 2014] but with different objectives, constraints, and feedback models.

2 Algorithm and polylogarithmic regret rate under diminishing returns

First, we introduce Algorithm 2, a UCB-style algorithm designed to operate under censored semi-bandit feedback. At each round, the learner selects an allocation that optimistically maximizes estimated rewards based on past feedback. Since rewards are observed only for completed tasks, the algorithm maintains task-wise upper confidence bounds (UCBs). The confidence width evolves over time as a function of the cumulative completion count and required confidence level.

To lighten the notation, from Algorithm 2 onward, we adopt the convention that $0/0 \coloneqq 0$.

We now investigate the realistic setting where the functions $F_1, \dots, F_K$ exhibit diminishing returns, i.e., are concave. We show that concavity, coupled with the assumption that the components of μ are non-zero, results in instance-dependent polylogarithmic rates. We flesh out the general case in Appendix C, but for the sake of readability, in this section, we will restrict our attention to the special case of *concave power functions*, i.e., of functions $f\colon [0,1] \to [0,1]$ for which there exists $a \in (0,1)$ such that, for all $x \in [0,1]$, $f(x) \coloneqq x^a$. This case suffices to present some of the main ideas of the general analysis in Appendix C while keeping the presentation relatively light. Furthermore, this is a special case of particular interest as it models many important behaviors like *scaling laws* of LLMs.

Assumption 1. *For any $k \in [K]$, μ_k is non-zero and F_k is a concave power function $x \mapsto x^{a_k}$ for $a_k \in (0,1)$.*

Theorem 1. *Under Assumption 1, for any $T \geq 2$, if we run Algorithm 2 with parameters T and $\delta \coloneqq \frac{1}{(KT)^2}$, letting $\mu_{\inf} \coloneqq \min_{k \in [K]} \mu_k$, $a_{\min} \coloneqq \min_{k \in [K]} a_k$, and $a_{\max} \coloneqq \max_{k \in [K]} a_k$, we have*[4]

$$R_T \leq 800 \left(\frac{K}{a_{\min} \mu_{\inf} (1 - a_{\max})} \right)^{\frac{9}{1 - a_{\max}}} (\log T)^2.$$

Proof sketch Fix any $T \geq 2$ and $a_1, \dots, a_K \in (0,1)$ such that, for all $k \in [K]$ and $x_k \in (0,1)$, $F_k(x_k) = x_k^{a_k}$, and recall $a_{\max} \coloneqq \max_{k \in [K]} a_k$. We prove this result in three steps.

Step 1: (2). The first key observation is that, with high probability, at any time step, the number of times any given task has been completed is at least proportional to a deterministic function of the budget invested in the task so far, assuming that a sufficient amount of budget has been invested in it. More precisely, there exist two numerical constants $c_1, c_2 > 0$ (independent of T and K) such that, with probability at least $1 - K/T$, for all $t \in [T]$ and $k \in [K]$, the following event $\mathcal{G}_T$ holds:

$$\sum_{s=1}^t X_{s,k}^{a_k} \geq c_1 (\log T)^2 \implies \sum_{s=1}^t B_{s,k}(X_{s,k}) \geq c_2 \sum_{s=1}^t X_{s,k}^{a_k}. \tag{2}$$

[4] We remark that we did not optimize the dependence on $K, \mu_{\inf}, a_{\min}$, and $a_{\max}$. The dependence displayed above is deliberately loose in order to obtain a simpler, more readable upper bound.

We prove (2) in Lemma 2 in Appendix B, for the general diminishing-return setting, leveraging the fact that, for all $k \in [K]$, $\left(X_{t,k}^{a_k} - B_{t,k}(X_{t,k})\right)_{t\in\mathbb{N}}$ is a martingale difference sequence, together with some specific properties of our setting and a concentration argument (a version of Bernstein's inequality for martingales).

To be able to exploit the consequent in (2), we must show that its antecedent is satisfied. To this end, we prove in Corollary 3 in Appendix B that one can define a "good event" $\mathcal{E}_T^\delta$, as the event that all the $\mathrm{UCB}_{t,k}^\delta$ estimators of all the μ_k maintained by Algorithm 2 are sufficiently accurate approximations of μ_k (see also (5) in Section 3 for the precise expression of $\mathcal{E}_T^\delta$), such that $\mathcal{E}_T^\delta$ occurs with probability at least $1 - KT\delta$. Consequently, we prove that we can find a $T_0 = \Theta\big((\log T)^2\big)$ such that, under $\mathcal{E}_T^\delta$ and $\mathcal{G}_T$, it holds that $\sum_{s=1}^t X_{s,k}^{a_k} \geq c_1(\log T)^2$ for all $t \geq T_0$. Fix such a T_0 and $\mathcal{E}_T^\delta$.

Step 2: (3). For any $m \coloneqq (m_1, \ldots, m_K) \in [0,\infty)^K$, define the function $\Phi_m \colon \Delta_K \to [0,\infty)$, for all $x \coloneqq (x_1, \ldots, x_K) \in \Delta_K$, by $\Phi_m(x) \coloneqq \sum_{k\in[K]} m_k x_k^{a_k}$. In words, Φ_m is the expected reward function when the underlying mean reward of each task $k \in [K]$ is m_k. Hence, denoting, for any $m \in [0,\infty)^K$, a maximizer (chosen arbitrarily in case of ties) of Φ_m by $x^\star(m)$, we have that $x^\star(\mu)$ is the optimal budget allocation and $x^\star(\mathrm{UCB}_{t-1}^\delta)$ is the action played by Algorithm 2 at time t.

The next claim is the keystone to obtaining a polylogarithmic regret rate. It states that the instantaneous regret due to maximizing a slightly misspecified reward function $\Phi_{\widetilde{m}}$ when the correct one is Φ_m is only quadratic in the distance between m and $\widetilde{m}$, as long as the components of m and $\widetilde{m}$ are bounded away from zero. More precisely, for any $m_{\inf} \in (0,1)$, there exists $c_3 > 0$ (independent of T) such that, for any two vectors $m, \widetilde{m} \in [m_{\inf}, 2]^K$, it holds that

$$\Phi_m\big(x^\star(m)\big) - \Phi_m\big(x^\star(\widetilde{m})\big) \leq c_3 \left\|m - \widetilde{m}\right\|_2^2 . \tag{3}$$

We prove (3) in Lemma 5 in Appendix C for the general diminishing-return setting, leveraging the strong concavity of $F_1, \ldots, F_K$, as well as their smoothness on $[x_{\inf}, 1]$, for any $x_{\inf} > 0$.

Now, to be able to exploit (3), we must show that all components of $x^\star(\mu)$ and $x^\star(\mathrm{UCB}_t^\delta)$, for t large enough, are uniformly bounded below by some $c_4 > 0$ (independent of T, and playing the role of $x_{\inf}$ above). We prove this in Lemma 4 in Appendix C for the general diminishing-return setting.

Step 3: Bounding the regret. Recalling that $\delta \coloneqq \frac{1}{(KT)^2}$, $T_0 = \Theta\big((\log T)^2\big)$ and exploiting (2) and (3), we have, under $\mathcal{E}_T^\delta$ and $\mathcal{G}_T$ (thus, with probability at least $1 - {}^K\!/_T - KT\delta$), that, with the realized regret $\mathcal{R}_T$ in (12) satisfying $\mathbb{E}(\mathcal{R}_T) = R_T$,

$$\begin{aligned}
\mathcal{R}_T \;&\leq\; KT_0 + \sum_{t=T_0}^{T} \Big(\Phi_\mu\big(x^\star(\mu)\big) - \Phi_\mu\big(x^\star(\mathrm{UCB}_{t-1}^\delta)\big)\Big) \overset{(3)}{\leq} KT_0 + \sum_{t=T_0}^{T} c_3 \left\|\mu - \mathrm{UCB}_{t-1}^\delta\right\|_2^2 \\
&\overset{(\circ)}{\leq} KT_0 + 4c_3 \sum_{t=T_0}^{T} \sum_{k=1}^{K} \frac{\log(\frac{2}{\delta})}{1 + \sum_{s=1}^{t-1} B_{s,k}(X_{s,k})} \overset{(2)}{\leq} KT_0 + \frac{4c_3}{c_2} \log\left(\frac{2}{\delta}\right) \sum_{t=T_0}^{T} \sum_{k=1}^{K} \frac{1}{\sum_{s=1}^{t-1} X_{s,k}^{a_k}} \\
&\overset{(*)}{\leq} K(T_0+2) + \frac{4c_3}{c_2 c_4^{a_{\max}}} K \log\left(\frac{2}{\delta}\right) \sum_{t=T_0+2}^{T} \frac{1}{t - T_0} \leq K(T_0+2) + \frac{4c_3}{c_2 c_4^{a_{\max}}} K \log\left(\frac{2}{\delta}\right) \log(T),
\end{aligned}$$

where $(\circ)$ follows from being under the good event $\mathcal{E}_T^\delta$, and $(*)$ from $X_{s,k}^{a_k} \geq c_4^{a_k} \geq c_4^{a_{\max}}$. Finally, Proposition 1 in Appendix C converts this bound into the one stated in Theorem 1. ∎

Remark 1 (Interpretation of the assumptions for the fast-rate results)**.** *The assumptions used here and for the general fast-rate result in Appendix C provide three main ingredients for the proofs: eventual interiority (the selected X_t is bounded away from zero eventually, see Lemma 4 for details), local quadratic stability of the UCB optimizer when the mean vector is changed (see Lemma 5 for details), and a bounded away from zero observation rate (see Lemma 2 for details). Note that eventual interiority is obtained thanks to the infinite derivative of F_k at zero.*

Remark 2 (Open problem: optimality of the $(\log T)^2$ rate)**.** *We do not know if the $(\log T)^2$ dependence in Theorem 1 is exactly optimal. Note that it is possible to provide a lower bound of order $\log T$ in some cases, still leaving a gap of order $\log T$. We do not provide the details for the sake of concision.*

Remark 3 (Approximate optimization of the UCB function)**.** *Note that in Algorithm 2 we assume that the set of exact maximizers of the UCB function is non-empty and that the algorithm selects such*

a maximizer measurably from the available history and its independent internal randomness, in line with many other references [Srinivas et al., 2012, Bect et al., 2019, Iwazaki, 2026]. Hence our results can be interpreted as being information-theoretic. For this algorithm, worst-case regret is taken over instances satisfying this oracle assumption. Nevertheless, our results could be extended to the case where the chosen X_t in Algorithm 2 maximizes the UCB function up to a gap ε_t. If the sequence $(\varepsilon_t)_t$ decays fast enough, one could extend Theorem 1, and also Theorem 3 below, preserving the same rates. Obtaining maximizers up to a gap ε_t can be achieved in a computationally efficient way under suitable assumptions on the F_k's, such as concavity as in Theorem 1.

Remark 4 (Approximately known budget-to-probability functions F_ks)**.** *If the functions F_ks are only known up to a gap ε'_t in the supremum norm, then optimizing the UCB function defined from approximated F_ks boils down to optimizing the real UCB function (defined from the true F_ks) up to an optimization error ε_t. Hence the previous remark applies and theoretical extensions of our results are possible (in settings that exclude Appendix A, of course).*

Comparison with linear bandits The reader may wonder whether the fast rates achieved by our algorithm can also be obtained by existing methods from the bandit literature (including standard linear bandit algorithms), such as those reviewed in Appendix G. We now show that this is *not* the case: even in extremely benign instances among those from diminishing-return regimes, *any* method that relies solely on bandit feedback must incur regret of order $\Omega(\sqrt{T})$. Specifically, the following holds.

Theorem 2. *Suppose $K = 2$ and for every $b \in [0,1]$, the budget-to-success curves are defined by $F_1(b) := \sqrt{b} =: F_2(b)$. Consider any algorithm running on bandit feedback, that is, in Algorithm 1, the feedback $\mathscr{F}_t$ is the gain $\big\langle G_t,\, B_t(X_t)\big\rangle$. Then, for any time horizon $T \geq 4$, there exists a valid instance of our problem with expected reward vector μ such that $\min(\mu_1, \mu_2) \geq \frac{3}{8}$ and $R_T \geq \frac{1}{140}\cdot\sqrt{T}$ over that instance.*

We present here a proof sketch and defer the full proof of Theorem 2 to Appendix F.

Proof sketch Under $F_1(b) = \sqrt{b} = F_2(b)$ and $K = 2$, any budget allocation $x \in \Delta_2$ corresponds to the unit vector $(\sqrt{x_1}, \sqrt{x_2})$ in the first quadrant. Hence, we can parameterize actions by angles $\vartheta \in [0, \pi/2]$ via the relation $(\sqrt{x_1}, \sqrt{x_2}) = (\cos\vartheta, \sin\vartheta)$. Thus, the expected reward is $\langle \mu, (\cos\vartheta, \sin\vartheta)\rangle$, and the optimal action corresponds to the angle whose direction is aligned with the (unknown) mean vector μ.

We construct two hard instances by perturbing $(1/2, 1/2)$:

$$\mu^+ := \left(\tfrac{1}{2} + \varepsilon, \tfrac{1}{2} - \varepsilon\right), \qquad \mu^- := \left(\tfrac{1}{2} - \varepsilon, \tfrac{1}{2} + \varepsilon\right).$$

Their optimal directions lie on opposite sides of $\pi/4$ (indeed, they are aligned with μ^+ and μ^-, respectively). If the learner plays near $\vartheta = \pi/4$ (the direction associated to $(1/2, 1/2)$), then bandit feedback is essentially uninformative: the one-step information to distinguish μ^+ from μ^- is $O\big(\varepsilon^2\,|\vartheta - \pi/4|^2\big)$. Consequently, up to constants, the total KL is controlled by $\varepsilon^2 \sum_{t=1}^{T} \mathbb{E}\Big[|\vartheta_t - \pi/4|^2\Big]$.

Therefore, to gather enough information to distinguish the two instances, the learner must play at angles that are non-negligibly away from $\pi/4$. However, such angles are necessarily far from the optimal direction in at least one of the two instances, and the incurred regret is lower bounded by a quantity proportional to the squared angular distance from that instance's optimum. This creates an "apple-tasting"-type exploration-information tradeoff, with the additional difficulty of a continuum of actions. Balancing the two effects with $\varepsilon = \Theta(T^{-1/4})$ yields a $\Omega(\sqrt{T})$ lower bound. ∎

We stress that, since the instance in Theorem 2 belongs to the diminishing-return class analyzed in Theorem 1, then, in contrast to bandit algorithms, our algorithm achieves polylogarithmic regret guarantee in this setting: when $K = 2$ and F_1, F_2 are as in Theorem 2, there exist universal constants $c, C > 0$ such that for any valid instance with $\min(\mu_1, \mu_2) \geq \frac{3}{8}$, the regret of our algorithm is at most $c + C \cdot (\log T)^2$. For a precise formulation of this claim, see Lemma 6 in Appendix C.

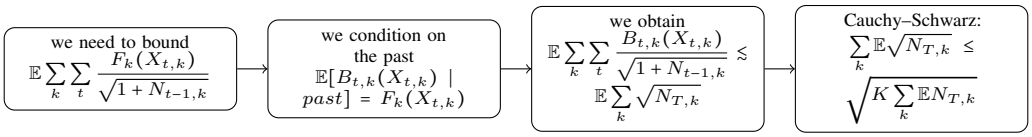

Figure 1: Schematic of the main proof steps for Theorem 3.

3 Worst-case analysis of Algorithm 2 and matching lower bound

We now analyze the worst-case performance of Algorithm 2 for any family of budget-to-success curves and reward distributions. We show that Algorithm 2 is worst-case optimal, both in K and T, up to a logarithmic factor. Due to space constraints, we present here a proof sketch of the next result while deferring the full proof to Appendix E. We also provide a schematic of the main proof steps in Figure 1.

Theorem 3. *For any time horizon T, if we run Algorithm 2 with parameters T, $\delta \coloneqq \frac{1}{(KT)^2}$, its regret satisfies*

$$R_T = \widetilde{\mathcal{O}}\left(1 + \sqrt{K}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\sum_{k=1}^{K} F_k(X_{t,k})\right]}\right).$$

Upper bounding $F_k \le 1$, the worst-case regret of Algorithm 2 satisfies $R_T^\star = \widetilde{\mathcal{O}}\big(K\sqrt{T}\big)$.

Proof sketch Fix any time horizon $T \ge 2$, and let $\delta \coloneqq \frac{1}{(KT)^2}$. For conciseness, we introduce, for each $t \in \{0,\dots,T\}$ and $k \in [K]$, $N_{t,k} \coloneqq \sum_{s=1}^{t} B_{s,k}(X_{s,k})$, and, for all $\tau \in [N_{t,k}]$, we define $\widetilde{G}_{\tau,k} \coloneqq G_{S_\tau,k}$, where S_τ is the random time step $s \in \mathbb{N}$ such that $B_{s,k} = 1$ for the τ-th time, i.e., such that $\sum_{s<S_\tau} B_{s,k} = \tau - 1$ and $\sum_{s\le S_\tau} B_{s,k} = \tau$. Then, note that we have for each $t \in [T]$ and $k \in [K]$,

$$\mathrm{UCB}_{t,k}^{\delta} = \frac{\sum_{\tau=1}^{N_{t,k}} \widetilde{G}_{\tau,k}}{N_{t,k}} + \sqrt{\frac{\log\frac{2}{\delta}}{1+N_{t,k}}}\,. \tag{4}$$

Define the "good event" as

$$\mathcal{E}_T^{\delta} \coloneqq \bigcap_{k\in[K]}\bigcap_{t\in[N_{T,k}]}\left\{\left|\frac{1}{t}\sum_{\tau=1}^{t}\widetilde{G}_{\tau,k} - \mu_k\right| \le \sqrt{\frac{\log\frac{2}{\delta}}{2t}}\right\}. \tag{5}$$

As we prove in Corollary 3 in Appendix B (note that a simple Hoeffding's inequality does not suffice because the random variables $\widetilde{G}_{\tau,k}$ depend on random time steps chosen as a function of the learner's actions), we have $\mathbb{P}\big[\mathcal{E}_T^{\delta}\big] \ge 1 - KT\delta$. Note that under the good event, for all $t \in [T]$, for all $k \in [K]$,

$$\left|\frac{\sum_{\tau=1}^{N_{t,k}}\widetilde{G}_{\tau,k}}{N_{t,k}} - \mu_k\right| \le \sqrt{\frac{\log\frac{2}{\delta}}{1+N_{t,k}}}. \tag{6}$$

When $N_{t,k} = 0$, under the convention $0/0 = 0$, this is indeed equivalent to $|\mu_k| \le \sqrt{\log\frac{2}{\delta}}$ which is true because $\log\frac{2}{\delta} \ge \log 4 > 1$. When $N_{t,k} \ge 1$, the above display holds because $\sqrt{\frac{\log\frac{2}{\delta}}{1+N_{t,k}}} \ge \sqrt{\frac{\log\frac{2}{\delta}}{2N_{t,k}}}$ and because the good event holds. Now, fix an arbitrary $x^\star \in \Delta_K$. We have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\Big[\big\langle\mu, F(x^\star)\big\rangle - \big\langle\mu, F(X_t)\big\rangle\Big] &\le KT\underbrace{\big(1-\mathbb{P}[\mathcal{E}_T^{\delta}]\big)}_{\le KT\delta} + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}}\sum_{t=1}^{T}\big\langle\mu, F(x^\star) - F(X_t)\big\rangle\right] \\
&\le \underbrace{K^2T^2\delta}_{\le 1} + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}}\sum_{t=1}^{T}\underbrace{\big\langle\mathrm{UCB}_{t-1}^{\delta}, F(x^\star) - F(X_t)\big\rangle}_{\le 0 \text{ by definition of } X_t}\right] + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}}\sum_{t=1}^{T}\underbrace{\big\langle\mathrm{UCB}_{t-1}^{\delta} - \mu}_{\ge 0 \text{ in } \mathcal{E}_T^{\delta}}, F(X_t) - F(x^\star)\big\rangle\right] \\
&\le 1 + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}}\sum_{t=1}^{T}\Big\langle\mathrm{UCB}_{t-1}^{\delta} - \mu, \big(F(X_t) - F(x^\star)\big)_+\Big\rangle\right] = 1 + (\star)\,.
\end{aligned}$$

Hence, to complete the proof, it is sufficient to control $(\star)$ as follows:

$$
\begin{aligned}
(\star) &\le \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^\delta} \sum_{t=1}^{T}\sum_{k=1}^{K} 2\sqrt{\frac{\log\frac{2}{\delta}}{1+N_{t-1,k}}}\big(F_k(X_{t,k}) - F_k(x_k^\star)\big)_+\right] \\
&\le 2\sqrt{\log\frac{2}{\delta}}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{F_k(X_{t,k})}{\sqrt{1+N_{t-1,k}}}\right] = 2\sqrt{\log\frac{2}{\delta}}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{B_{t,k}(X_{t,k})}{\sqrt{1+N_{t-1,k}}}\right] \\
&= 2\sqrt{\log\frac{2}{\delta}}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{\tau=1}^{N_{T,k}}\frac{1}{\sqrt{\tau}}\right] \le 4\sqrt{\log\frac{2}{\delta}}\,\mathbb{E}\left[\sum_{k=1}^{K}\sqrt{N_{T,k}}\right] \le 4\sqrt{K\log\frac{2}{\delta}}\,\mathbb{E}\left[\sqrt{\sum_{k=1}^{K}N_{T,k}}\right] \\
&\le 4\sqrt{K\log\frac{2}{\delta}}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[N_{T,k}\right]} = 4\sqrt{K\log\frac{2}{\delta}}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\sum_{k=1}^{K}F_k(X_{t,k})\right]}\,.
\end{aligned}
$$

Here, the first inequality holds because we are on the good event $\mathcal{E}_T^\delta$ and by (6). The first equality follows from the tower rule after conditioning on the history up to the beginning of round t. The second equality rewrites the sum over successful observations: on the τ-th success of task k, the denominator is $\sqrt{\tau}$. The next inequality uses $\sum_{\tau=1}^{n}\tau^{-1/2} \le 2\sqrt{n}$ for all $n \ge 0$. The following inequality is Cauchy's inequality, and the last inequality is Jensen's inequality. ■

As an immediate corollary of Theorem 3, we recover the best known worst-case rate $\widetilde{\mathcal{O}}\big(\sqrt{KMT}\big)$ for combinatorial semi-bandits, see Chen et al. [2013] and Lattimore and Szepesvári [2020, Theorem 30.2], as discussed in Section 1.1.

Corollary 1. *In Theorem 3, if there is $M \in [K]$ such that $F_k = \mathbb{I}_{[1/M,1]}$ for $k \in [K]$, then we have $R_T^\star = \widetilde{\mathcal{O}}\big(\sqrt{KMT}\big)$.*

We now state a worst-case lower bound on the regret, not only for algorithms operating under censored semi-bandit feedback like Algorithm 2 but also for those operating under the richer full feedback. This lower bound is tight up to a logarithmic factor.

Theorem 4. *The worst-case regret of any full-feedback algorithm satisfies $R_T^\star = \Omega\big(K\sqrt{T}\big)$.*

Due to space constraints, we defer the proof of this result to Appendix D.

4 Conclusions, limitations and future work

We introduce a new stochastic bandit framework, which captures applications ranging from crowdsourcing markets to multi-platform autobidding, and to compute allocations in large-scale model training. We design an algorithm and establish its minimax optimality with a regret guarantee of $\widetilde{\Theta}\big(K\sqrt{T}\big)$. The same algorithm with no *ad-hoc* tuning enjoys an instance-dependent polylogarithmic regret bound, when the budget-to-success curves satisfy a natural diminishing-returns property. To the best of our knowledge, these are the first tight guarantees for sequential budget allocation with censored semi-bandit feedback.

Beyond matching the leading-order rates, our analysis yields three structural insights. First, the censored semi-bandit feedback lets the learner decouple exploration across coordinates from reward estimation, giving a clean $\sqrt{T}$ frontier in the worst case. Second, strong concavity acts as an implicit regularizer that yields logarithmic learning rates. Third, even though our logarithmic rate is reminiscent of those in bandits with strongly convex losses, our result is proved through novel, non-trivial techniques. Indeed, we provide a counterexample showing that any bandit algorithm (including linear bandit algorithms) has to suffer $\Omega(\sqrt{T})$ regret even in benign instances of the diminishing-return setting, illustrating that standard bandit techniques are not sufficient in this setting.

A limitation of our theory is that it does not capture other interesting variants: when the per-round budget varies with time (stochastically or adversarially), or when tasks (with deterministic or random completion times) can run for multiple time steps, and the feedback is only observed with some delay, at the completion of the tasks. Lastly, one could consider a contextual version of this problem where some context related to the tasks is revealed to the learner before they are asked to allocate budgets at each time step. We leave these extensions open for future work.

Acknowledgments and Disclosure of Funding

We are grateful to the anonymous reviewers for their constructive comments that led to an improvement of the paper. FB has benefitted from the AI Interdisciplinary Institute ANITI, funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002. NCB acknowledges the financial support from the EU Horizon RIA under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability). TC gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through grant RGPIN-2023-03688 (Discovery Grants Program).

References

A. Agarwal, R. Ghuge, and V. Nagarajan. Semi-bandit learning for monotone stochastic optimization. In *65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1260–1274. IEEE, 2024.

G. Aggarwal, A. Gupta, X. Tan, and M. Zhao. Multi-platform autobidding with and without predictions. In *Proceedings of the ACM on Web Conference 2025*, pages 2850–2859, 2025.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

V. Avadhanula, R. Colini Baldeschi, S. Leonardi, K. A. Sankararaman, and O. Schrijvers. Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference 2021*, pages 2805–2817, 2021.

P. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 335–342, 2008.

J. Bect, F. Bachoc, and D. Ginsbourger. A supermartingale approach to gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1. JMLR Workshop and Conference Proceedings, 2012.

S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127. PMLR, 2018.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.

W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

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*, pages 7617–7644. PMLR, 2023.

X. Fontaine, S. Mannor, and V. Perchet. An adaptive stochastic optimization algorithm for resource allocation. In *Algorithmic Learning Theory*, pages 319–363. PMLR, 2020.

W. T. Huh and P. Rusmevichientong. A non-parametric approach to stochastic inventory planning with lost-sales and censored demand. In *Multi-Echelon Conference, Atlanta, GA*, 2006.

P. Hummel and R. P. McAfee. Machine learning in an auction environment. *Journal of Machine Learning Research*, 17(197):1–37, 2016.

S. Iwazaki. Improved regret bounds for Gaussian process upper confidence bound in Bayesian optimization. *Advances in Neural Information Processing Systems*, 38:96922–96964, 2026.

K. Iyer, R. Johari, and M. Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

Q. Kang and W. P. Tay. Task recommendation in crowdsourcing based on learning preferences and reliabilities. *IEEE Transactions on Services Computing*, 15(4):1785–1798, 2020.

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

T. Lattimore, K. Crammer, and C. Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 477–486, 2014.

T. Lattimore, K. Crammer, and C. Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, 2015.

Y. Li. LLM bandit: Cost-efficient LLM generation via preference-conditioned dynamic routing. *arXiv preprint arXiv:2502.02743*, 2025.

G. Montanari, C. Fiegel, C. Pla, A. Saha, and V. Perchet. Learning to allocate resources with censored feedback. *arXiv preprint arXiv:2602.06565v2*, 2026.

Q. H. Nguyen, D. C. Hoang, J. Decugis, S. Manchanda, N. V. Chawla, and K. D. Doan. MetaLLM: A high-performant and cost-efficient dynamic framework for wrapping LLMs. *arXiv preprint arXiv:2407.10834*, 2024.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE transactions on information theory*, 58 (5):3250–3265, 2012.

B. Thananjeyan, K. Kandasamy, I. Stoica, M. Jordan, K. Goldberg, and J. Gonzalez. Resource allocation in multi-armed bandit exploration: Overcoming sublinear scaling with adaptive parallelism. In *International Conference on Machine Learning*, pages 10236–10246. PMLR, 2021.

L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111, 2014.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

U. ul Hassan and E. Curry. Efficient task assignment for spatial crowdsourcing: A combinatorial fractional optimization approach with semi-bandit learning. *Expert Systems with Applications*, 58: 36–56, 2016.

D. Van Der Hoeven, C. Pike-Burke, H. Qiu, and N. Cesa-Bianchi. Trading-off payments and accuracy in online classification with paid stochastic experts. In *International Conference on Machine Learning*, pages 34809–34830. PMLR, 2023.

G. Van Ryzin and J. McGill. Revenue management without forecasting or optimization: An adaptive algorithm for determining airline seat protection levels. *Management Science*, 46(6):760–775, 2000.

A. Verma, M. Hanawal, A. Rajkumar, and R. Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. Xia, F. Kong, T. Yu, L. Guo, R. A. Rossi, S. Kim, and S. Li. Which LLM to play? Convergence-aware online model selection with time-increasing bandits. In *Proceedings of the ACM Web Conference 2024*, pages 4059–4070, 2024.

H. Zhang, Y. Ma, and M. Sugiyama. Bandit-based task assignment for heterogeneous crowdsourcing. *Neural computation*, 27(11):2447–2475, 2015.

A What if the budget-to-success curves are unknown?

In this section we investigate the setting in which the learner does not know the budget-to-success curves in advance. We prove a particularly strong impossibility result: even when $K = 2$, the learner knows that F_1 and F_2 are indicator functions, and the rewards are deterministic with $G_t \equiv (1,1)$, the worst-case regret is linear.

Theorem 5 (Unknown budget-to-success curves imply linear minimax regret)**.** *Let $K \coloneqq 2$ and deterministic rewards $G_{t,1} = G_{t,2} \equiv 1$ for all t. Let $\Delta_2 = \{x \in [0,1]^2 : x_1 + x_2 = 1\}$. For each $b \in [0,1]$, define the budget-to-success curves*

$$F_1^{(b)} \coloneqq \mathbb{I}_{[b,1]}\,, \qquad F_2^{(b)} \coloneqq \mathbb{I}_{[1-b,1]}\,.$$

Suppose that the learner knows that (F_1, F_2) belongs to the family $\left((F_1^b, F_2^b)\right)_{b\in[0,1]}$ and that $G_{t,1} = G_{t,2} \equiv 1$ for all t. Then any (randomized) learning algorithm incurs linear worst-case regret:

$$\sup_{b\in[0,1]} R_T(F_1^b, F_2^b) = T, \qquad \text{for all } T \in \mathbb{N}.$$

Proof. Fix a (randomized) learning algorithm α. We assume that α has access sequentially to i.i.d. $[0,1]$-valued uniform random variables $U_1, \dots, U_T$ as random seeds.

Fix $b \in [0,1]$ and consider the instance with deterministic rewards $G_t \equiv (1,1)$ and budget-to-success curves $F_1^{(b)} = \mathbb{I}_{[b,1]}$ and $F_2^{(b)} = \mathbb{I}_{[1-b,1]}$. If at time t the learner plays an allocation $x \in \Delta_2$, then the completion feedback is

$$B_t(x) = (B_{t,1}(x_1), B_{t,2}(x_2))\,,$$

and the reward feedback is

$$(G_{t,1}B_{t,1}(x_1), G_{t,2}B_{t,2}(x_2))\,.$$

Since $G_t \equiv (1,1)$, these two vectors coincide, so the reward part carries no additional information beyond the completion vector. Moreover, since $F_1^{(b)}$ and $F_2^{(b)}$ are indicator functions, the corresponding Bernoulli variables are degenerate, and we have

$$B_{t,1}(x_1) = \mathbb{I}\{b \le x_1\}\,, \qquad B_{t,2}(x_2) = \mathbb{I}\{1 - b \le x_2\} = \mathbb{I}\{x_1 \le b\}\,. \tag{7}$$

In particular,

$$B_t(x) = (B_{t,1}(x_1), B_{t,2}(x_2)) \in \{(0,1),(1,0),(1,1)\}\,.$$

Hence, without loss of generality, we can assume that α is represented by a sequence of functions $\alpha_1, \dots, \alpha_T$ such that

$$\alpha_t : ([0,1] \times \{(0,1),(1,0),(1,1)\})^{t-1} \times [0,1] \to \Delta_2\,,$$

and that for each $b \in [0,1]$ the generated allocations $(X_t^{(b)})_{t\in[T]}$ satisfy $X_1^{(b)} \coloneqq \alpha_1(U_1)$ and, for $t \ge 2$,

$$X_t^{(b)} \coloneqq \alpha_t\left(U_1, B_1(X_1^{(b)}), \dots, U_{t-1}, B_{t-1}(X_{t-1}^{(b)}), U_t\right)\,.$$

From (7) it follows that for the instance indexed by b the unique allocation for which both tasks are completed is

$$\bar{x}(b) \coloneqq (b, 1-b)\,.$$

In particular, the optimal (expected) reward per round equals 2, because

$$\sup_{x\in\Delta_2} \left\langle \mu, F^{(b)}(x)\right\rangle = \max_{x\in\Delta_2} \left\langle \mu, F^{(b)}(x)\right\rangle = \left\langle (1,1),(1,1)\right\rangle = 2\,.$$

For a generic allocation $x \in \Delta_2$, (7) implies that exactly one task is completed unless $x_1 = b$, in which case both are completed. Therefore the reward at time t as a function of the allocated budget $x \in \Delta_2$ is

$$r_t^{(b)}(x) \coloneqq G_{t,1}B_{t,1}(x_1) + G_{t,2}B_{t,2}(x_2) = 1 + \mathbb{I}\{x_1 = b\}\,,$$

and for the sequence $(X_t^{(b)})_{t\in[T]}$ we obtain that the regret, when the underlying instance has parameter $b \in [0,1]$, is

$$R_T^b \coloneqq 2T - \sum_{t=1}^{T} \mathbb{E}\left[r_t^{(b)}\left(X_t^{(b)}\right)\right] = T - \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{I}\left\{X_{t,1}^{(b)} = b\right\}\right]\,. \tag{8}$$

Consequently, using Fubini/Tonelli's theorem, we have

$$\sup_{b\in[0,1]} R_T^b \geq \int_0^1 R_T^b \,\mathrm{d}b = T - \sum_{t=1}^{T} \mathbb{E}\left[\int_0^1 \mathbb{I}\left\{X_{t,1}^{(b)} = b\right\} \,\mathrm{d}b\right] .$$

Thus it suffices to prove that for every fixed $t \in [T]$ ($\mathbb{P}$-almost surely),

$$\int_0^1 \mathbb{I}\left\{X_{t,1}^{(b)} = b\right\} \,\mathrm{d}b = 0 . \tag{9}$$

Fix $t \in [T]$ and fix arbitrary seeds $u_1, \dots, u_T \in [0,1]$. With these seeds fixed, the entire trajectory is deterministic for each parameter b. We denote by $x_t^{(b)}(u_1, \dots, u_t) \in \Delta_2$ the allocation at time t when the instance parameter is b. Define the set of fixed points

$$A_t(u_1, \dots, u_t) \coloneqq \left\{b \in [0,1] : x_{t,1}^{(b)}(u_1, \dots, u_t) = b\right\} .$$

We claim that $A_t(u_1, \dots, u_t)$ is finite. This would imply

$$\int_0^1 \mathbb{I}\left\{x_{t,1}^{(b)}(u_1, \dots, u_t) = b\right\} \,\mathrm{d}b = 0 ,$$

and since the seeds $(u_1, \dots, u_T)$ were arbitrary, (9) would follow from

$$\int_0^1 \mathbb{I}\left\{X_{t,1}^{(b)} = b\right\} \,\mathrm{d}b = \int_0^1 \mathbb{I}\left\{x_{t,1}^{(b)}(U_1, \dots, U_t) = b\right\} \,\mathrm{d}b = 0 .$$

It remains to prove that $A_t(u_1, \dots, u_t)$ is finite. We do so by induction on t. For $t = 1$, $x_{1,1}^{(b)}(u_1) = (\alpha_1(u_1))_1$ does not depend on b, hence $A_1(u_1)$ is a singleton.

Assume now that the claim holds up to time $t-1$ and fix $(u_1, \dots, u_t)$. Define

$$S_{t-1}(u_1, \dots, u_{t-1}) \coloneqq \bigcup_{s=1}^{t-1} A_s(u_1, \dots, u_s) .$$

Since each $A_s(u_1, \dots, u_s)$ is finite by the induction hypothesis, the set $S_{t-1}(u_1, \dots, u_{t-1})$ is finite as well.

Consider any interval I contained in $[0,1] \setminus S_{t-1}(u_1, \dots, u_{t-1})$. For every $b \in I$ and every $s \leq t-1$ we have $b \neq x_{s,1}^{(b)}(u_1, \dots, u_s)$, hence by (7) the completion vector

$$B_s\left(x_s^{(b)}(u_1, \dots, u_s)\right) = \left(\mathbb{I}\left\{b \leq x_{s,1}^{(b)}(u_1, \dots, u_s)\right\}, \mathbb{I}\left\{x_{s,1}^{(b)}(u_1, \dots, u_s) \leq b\right\}\right)$$

takes values in $\{(1,0),(0,1)\}$ and is constant over I (as a function of b). Therefore the entire history

$$\left(u_1, B_1\left(x_1^{(b)}(u_1)\right), \dots, u_{t-1}, B_{t-1}\left(x_{t-1}^{(b)}(u_1, \dots, u_{t-1})\right)\right)$$

is constant over $b \in I$, and since α_t is deterministic given this history and u_t, it follows that $x_{t,1}^{(b)}(u_1, \dots, u_t)$ is constant over $b \in I$.

We have shown that the map $b \mapsto x_{t,1}^{(b)}(u_1, \dots, u_t)$ is piecewise constant on $[0,1]$, with possible breakpoints contained in the finite set $S_{t-1}(u_1, \dots, u_{t-1})$. On any piece where $b \mapsto x_{t,1}^{(b)}(u_1, \dots, u_t)$ is constant, the equation $x_{t,1}^{(b)}(u_1, \dots, u_t) = b$ can hold for at most one value of b. Since there are finitely many pieces, this shows that $A_t(u_1, \dots, u_t)$ is finite, completing the induction.

Combining the above with (9), we obtain $\int_0^1 R_T^b \,\mathrm{d}b = T$, and therefore $\sup_{b\in[0,1]} R_T^b \geq T$. To prove that this is in fact an equality, we notice that (8) immediately implies that for every $b \in [0,1]$

$$R_T^b \leq T ,$$

and hence $\sup_{b\in[0,1]} R_T^b \leq T$, from which it follows

$$\sup_{b\in[0,1]} R_T^b = T .$$

□

B Preliminary results

Let U encode the learner's internal randomization, independently of the environment, and write $D_t = (B_{t,1}(X_{t,1}), \dots, B_{t,K}(X_{t,K}))$. For the analysis, define

$$\mathcal{F}_0 \coloneqq \sigma(U), \qquad \mathcal{F}_t \coloneqq \sigma\big(U, (D_s, G_s)_{s=1}^t\big), \quad t \in \mathbb{N}. \tag{10}$$

This filtration includes past unobserved rewards only for the analysis and does not change the learner's feedback. Since the learner chooses X_t from its internal randomization and past feedback, X_t is $\mathcal{F}_{t-1}$-measurable.

The next lemma and its two corollaries show that successive rewards observed for a given task can be controlled as if they were i.i.d. variables, despite their observation times depending on past choices of the algorithm.

Lemma 1. *Fix $k \in [K]$. Let $N \in [T]$ be deterministic and such that*

$$\mathbb{P}\left[\sum_{t=1}^{\infty} B_{t,k}(X_{t,k}) \geq N\right] = 1.$$

Let $(Y_{1,k}, \dots, Y_{N,k}) \coloneqq (G_{t_1,k}, \dots, G_{t_N,k})$ where $t_1 < \cdots < t_N$ are the (random) N first indices $t \in \mathbb{N}$ such that $B_{t,k}(X_{t,k}) = 1$. Then $Y_{1,k}, \dots, Y_{N,k}$ are i.i.d. and distributed as $G_{1,k}$.

Proof. Let ψ be a bounded continuous function on $[0,1]$. Let $A_1 = \big\{B_{1,k}(X_{1,k}) = 1\big\} = \{t_1 = 1\}$ and for $t \geq 2$, $A_t = \big\{B_{1,k}(X_{1,k}) = 0, \dots, B_{t-1,k}(X_{t-1,k}) = 0, B_{t,k}(X_{t,k}) = 1\big\} = \{t_1 = t\}$. We have, using the law of total expectation,

$$\mathbb{E}\big[\psi(Y_{1,k})\big] = \sum_{t=1}^{\infty} \mathbb{E}\big[\psi(Y_{1,k}) \mid A_t\big]\mathbb{P}[A_t] = \sum_{t=1}^{\infty} \mathbb{E}\big[\psi(G_{t,k}) \mid A_t\big]\mathbb{P}[A_t]. \tag{11}$$

Then, using $\mathcal{F}_t$ from (10), noting that A_t is a measurable event of $B_{t,k}(X_{t,k})$ and the variables in $\mathcal{F}_{t-1}$, and using the law of total expectation,

$$\mathbb{E}\big[\psi(G_{t,k}) \mid A_t\big] = \mathbb{E}\Big[\mathbb{E}\big[\psi(G_{t,k}) \mid \mathcal{F}_{t-1}, B_{t,k}(X_{t,k})\big] \mid A_t\Big].$$

Since $G_{t,k}$ is independent of $B_{t,k}(X_{t,k})$ and $\mathcal{F}_{t-1}$, and has the same distribution as $G_{1,k}$, we have

$$\mathbb{E}\big[\psi(G_{t,k}) \mid \mathcal{F}_{t-1}, B_{t,k}(X_{t,k})\big] = \mathbb{E}\big[\psi(G_{1,k})\big].$$

Hence

$$\mathbb{E}\big[\psi(G_{t,k}) \mid A_t\big] = \mathbb{E}\big[\psi(G_{1,k})\big].$$

It follows from (11) that

$$\mathbb{E}\big[\psi(Y_{1,k})\big] = \sum_{t=1}^{\infty} \mathbb{E}\big[\psi(G_{1,k})\big]\mathbb{P}[A_t] = \mathbb{E}\big[\psi(G_{1,k})\big].$$

Hence $Y_{1,k}$ is distributed as $G_{1,k}$.

Next, for $\ell \in [N]$, let

$$\mathcal{S}_\ell = \big\{(s_1, \dots, s_\ell) \in \mathbb{N}^\ell : s_1 < \cdots < s_\ell\big\}.$$

Note that $\mathcal{S}_{n+1}$ is the set of values for $(t_1, \dots, t_{n+1})$ that can be reached. Consider any two bounded measurable functions $\phi\colon [0,1]^n \to \mathbb{R}$ and $\psi\colon [0,1] \to \mathbb{R}$. Define

$$T(s_1, \dots, s_{n+1}) \coloneqq \{(t_1, \dots, t_{n+1}) = (s_1, \dots, s_{n+1})\}.$$

Then we have, using the law of total expectation,

$$\begin{aligned}
&\mathbb{E}\big[\phi(Y_{1,k}, \dots, Y_{n,k})\psi(Y_{n+1,k})\big] \\
&= \sum_{(s_1,\dots,s_{n+1}) \in \mathcal{S}_{n+1}} \mathbb{E}\big[\phi(Y_{1,k}, \dots, Y_{n,k})\psi(Y_{n+1,k}) \mid T(s_1, \dots, s_{n+1})\big]\mathbb{P}\big[T(s_1, \dots, s_{n+1})\big] \\
&= \sum_{(s_1,\dots,s_{n+1}) \in \mathcal{S}_{n+1}} \mathbb{E}\big[\phi(G_{s_1,k}, \dots, G_{s_n,k})\psi(G_{s_{n+1},k}) \mid T(s_1, \dots, s_{n+1})\big]\mathbb{P}\big[T(s_1, \dots, s_{n+1})\big] \\
&= \sum_{(s_1,\dots,s_{n+1}) \in \mathcal{S}_{n+1}} \mathbb{E}\Big[\mathbb{E}\big[\phi(G_{s_1,k}, \dots, G_{s_n,k})\psi(G_{s_{n+1},k}) \mid \mathcal{F}_{s_{n+1}-1}, B_{s_{n+1},k}(X_{s_{n+1},k})\big] \\
&\qquad\qquad \mid T(s_1, \dots, s_{n+1})\Big] \times \mathbb{P}\big[T(s_1, \dots, s_{n+1})\big],
\end{aligned}$$

where the last equality holds by the law of total expectation since $T(s_1, \dots, s_{n+1})$ is a measurable event of $B_{s_{n+1},k}(X_{s_{n+1},k})$ and the variables in $\mathcal{F}_{s_{n+1}-1}$.

Then, conditional on $\mathcal{F}_{s_{n+1}-1}$ and $B_{s_{n+1},k}(X_{s_{n+1},k})$, only $G_{s_{n+1},k}$ remains random (and keeps its unconditional distribution). Hence, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\phi(Y_{1,k}, \dots, Y_{n,k})\psi(Y_{n+1,k})\big] \\
&= \sum_{(s_1,\dots,s_{n+1})\in\mathcal{S}_{n+1}} \mathbb{E}\Big[\phi(G_{s_1,k}, \dots, G_{s_n,k})\mathbb{E}\big[\psi(G_{s_{n+1},k}) \mid \mathcal{F}_{s_{n+1}-1}, B_{s_{n+1},k}(X_{s_{n+1},k})\big] \\
&\qquad\qquad \mid T(s_1, \dots, s_{n+1})\Big] \times \mathbb{P}\big[T(s_1, \dots, s_{n+1})\big] \\
&= \mathbb{E}\big[\psi(G_{1,k})\big] \sum_{(s_1,\dots,s_{n+1})\in\mathcal{S}_{n+1}} \mathbb{E}\big[\phi(G_{s_1,k}, \dots, G_{s_n,k}) \mid T(s_1, \dots, s_{n+1})\big]\mathbb{P}\big[T(s_1, \dots, s_{n+1})\big] \\
&= \mathbb{E}\big[\psi(G_{1,k})\big] \sum_{(s_1,\dots,s_{n+1})\in\mathcal{S}_{n+1}} \mathbb{E}\big[\phi(Y_{1,k}, \dots, Y_{n,k}) \mid T(s_1, \dots, s_{n+1})\big]\mathbb{P}\big[T(s_1, \dots, s_{n+1})\big] \\
&= \mathbb{E}\big[\psi(G_{1,k})\big]\mathbb{E}\big[\phi(Y_{1,k}, \dots, Y_{n,k})\big],
\end{aligned}$$

where the last equality follows from the law of total expectation.

Hence, by induction it follows that for any bounded measurable functions $\psi_1, \dots, \psi_N$ from $[0, 1]$ to $\mathbb{R}$,

$$\mathbb{E}\left[\prod_{i=1}^{N} \psi_i(Y_{i,k})\right] = \prod_{i=1}^{N} \mathbb{E}\left[\psi_i(G_{1,k})\right].$$

This concludes the proof. □

Corollary 2. *Fix $N, T \in \mathbb{N}$ with $N \le T$. Fix $k \in [K]$. Let $E_{N,T}$ be the event*

$$E_{N,T} \coloneqq \left\{\sum_{t=1}^{T} B_{t,k}(X_{t,k}) \ge N\right\}.$$

Under the event $E_{N,T}$, define $t_1, \dots, t_N$ as the sequence of the N first random times $t \in [T]$ where $B_{t,k}(X_{t,k}) = 1$, with $1 \le t_1 < \cdots < t_N \le T$. Then, for any Borel subset $A \subset [0, 1]^N$,

$$\mathbb{P}\big[E_{N,T} \cap \{(G_{t_1,k}, \dots, G_{t_N,k}) \in A\}\big] \le \mathbb{P}\big[(G_{1,k}, \dots, G_{N,k}) \in A\big].$$

Proof. Consider the budget-to-success curves $F_1, \dots, F_K$. Consider new (time-dependent) functions with, for $t \in \mathbb{N}$, $k \in [K]$, and $x \in [0, 1]$,

$$\widetilde{F}_{t,k}(x) = \begin{cases} F_k(x) & \text{if } t \le T \\ 1 \text{ (constant function)} & \text{if } t \ge T + 1. \end{cases}$$

These new $\widetilde{F}_{t,1}, \dots \widetilde{F}_{t,K}$ define an auxiliary sampling process, for which we denote probabilities with $\widetilde{\mathbb{P}}$. We keep the reward process unchanged and extend the learner's allocation rule arbitrarily after round T. The proof of Lemma 1 also applies to this construction, since it does not require the completion probabilities to be constant over time. The behavior of the gains and algorithm choices is unchanged from times $t = 1$ to $t = T$ between the two settings. Hence

$$\mathbb{P}\big[E_{N,T} \cap \{(G_{t_1,k}, \dots, G_{t_N,k}) \in A\}\big] = \widetilde{\mathbb{P}}\big[E_{N,T} \cap \{(G_{t_1,k}, \dots, G_{t_N,k}) \in A\}\big].$$

Consider $N \in [T]$. Under the new setting, each arm is infinitely explored over times $t \in \mathbb{N}$ and so we can apply Lemma 1 (with $t_1, \dots, t_N$ defined as in this lemma with probability one) yielding

$$\begin{aligned}
\widetilde{\mathbb{P}}\big[E_{N,T} \cap \{(G_{t_1,k}, \dots, G_{t_N,k}) \in A\}\big] &\le \widetilde{\mathbb{P}}\big[(G_{t_1,k}, \dots, G_{t_N,k}) \in A\big] \\
&= \widetilde{\mathbb{P}}\big[(G_{1,k}, \dots, G_{N,k}) \in A\big] \\
&= \mathbb{P}\big[(G_{1,k}, \dots, G_{N,k}) \in A\big].
\end{aligned}$$

This concludes the proof. □

Corollary 3. *With the notation of Corollary 2, for any $\delta \in (0,1)$, we have, for $N, T \in \mathbb{N}$, $N \leq T$,*

$$\mathbb{P}\left[E_{N,T} \cap \left\{\left|\frac{1}{N}\sum_{i=1}^{N} G_{t_i,k} - \mu_k\right| > \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}\right\}\right] \leq \delta.$$

Proof. We apply Corollary 2 to the set

$$A = \left\{(g_1, \ldots, g_N) \in [0,1]^N : \left|\frac{1}{N}\sum_{i=1}^{N} g_i - \mu_k\right| > \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}\right\}$$

and then we apply Hoeffding's inequality. □

The next crucial lemma shows that the number of times a task has been completed is not much smaller than the total amount of probability invested in the task (once at least a polylogarithmic amount of probability has been invested in the task).

Lemma 2. *For any $T \geq 2$, with probability at least $1 - \frac{K}{T}$, the following event holds: For $t \in [T]$ and $k \in [K]$, if*

$$\sum_{s=1}^{t} F_k(X_{s,k}) \geq 100 \log(T)^2$$

then

$$\sum_{s=1}^{t} B_{s,k}(X_{s,k}) \geq \frac{\sum_{s=1}^{t} F_k(X_{s,k})}{2}.$$

Proof. For $k \in [K]$, the sequence

$$\big(F_k(X_{t,k}) - B_{t,k}(X_{t,k})\big)_{t \in \mathbb{N}}$$

is a martingale difference sequence, together with the filtration $(\mathcal{F}_t)_{t \geq 0}$ defined in (10). Indeed, for all $t \in \mathbb{N}$, $F_k(X_{t,k}) - B_{t,k}(X_{t,k})$ is $\mathcal{F}_t$-measurable and

$$\mathbb{E}\big[F_k(X_{t,k}) - B_{t,k}(X_{t,k}) \mid \mathcal{F}_{t-1}\big] = 0$$

since conditional on $\mathcal{F}_{t-1}$, $F_k(X_{t,k})$ is fixed and $B_{t,k}(X_{t,k})$ follows a Bernoulli distribution with parameter $F_k(X_{t,k})$.

For the same reason,

$$\mathrm{Var}\big[F_k(X_{t,k}) - B_{t,k}(X_{t,k}) \mid \mathcal{F}_{t-1}\big] = F_k(X_{t,k})\big(1 - F_k(X_{t,k})\big).$$

Let us define

$$v_t = \sum_{s=1}^{t} \mathrm{Var}\big[F_k(X_{s,k}) - B_{s,k}(X_{s,k}) \mid \mathcal{F}_{s-1}\big] \leq \sum_{s=1}^{t} F_k(X_{s,k}).$$

Hence from Corollary 16 in Cesa-Bianchi et al. [2005] (a version of Bernstein's inequality for martingales), we have, for $\delta \in (0,1)$ and $t \in [T]$,

$$\mathbb{P}\left[\sum_{s=1}^{t} F_k(X_{s,k}) - \sum_{s=1}^{t} B_{s,k}(X_{s,k}) > \sqrt{2(v_t+1)\log(t/\delta)} + \frac{\sqrt{2}}{3}\log(t/\delta)\right] \leq \delta.$$

Let us take $\delta = \frac{1}{T^2}$ and replace v_t by its upper bound yielding

$$\mathbb{P}\left[\sum_{s=1}^{t} F_k(X_{s,k}) - \sum_{s=1}^{t} B_{s,k}(X_{s,k}) > \sqrt{2\left(\sum_{s=1}^{t} F_k(X_{s,k}) + 1\right)3\log(T)} + \frac{3\sqrt{2}}{3}\log(T)\right] \leq \frac{1}{T^2}.$$

If $\sum_{s=1}^{t} F_k(X_{s,k}) \geq 100\log(T)^2$, we have

$$\begin{aligned}\sqrt{2\left(\sum_{s=1}^{t} F_k(X_{s,k}) + 1\right)3\log(T)} + \frac{3\sqrt{2}}{3}\log(T) &\leq \left(3 + \frac{3\sqrt{2}}{3}\right)\sqrt{\sum_{s=1}^{t} F_k(X_{s,k})}\log(T) \\ &\leq 5\log(T)\sqrt{\sum_{s=1}^{t} F_k(X_{s,k})}.\end{aligned}$$

Hence,

$$\mathbb{P}\left[\begin{array}{l}\sum_{s=1}^{t} F_k(X_{s,k}) \geq 100\log(T)^2 \\ \text{and } \sum_{s=1}^{t} F_k(X_{s,k}) - \sum_{s=1}^{t} B_{s,k}(X_{s,k}) > 5\log(T)\sqrt{\sum_{s=1}^{t} F_k(X_{s,k})}\end{array}\right] \leq \frac{1}{T^2}.$$

Then, if

$$\sum_{s=1}^{t} F_k(X_{s,k}) \geq 100\log(T)^2,$$

we have

$$\frac{\sum_{s=1}^{t} F_k(X_{s,k})}{2} \geq 5\log(T)\sqrt{\sum_{s=1}^{t} F_k(X_{s,k})}\,.$$

Hence we have

$$\mathbb{P}\left[\sum_{s=1}^{t} F_k(X_{s,k}) \geq 100\log(T)^2 \text{ and } \sum_{s=1}^{t} B_{s,k}(X_{s,k}) < \frac{\sum_{s=1}^{t} F_k(X_{s,k})}{2}\right] \leq \frac{1}{T^2}.$$

We conclude the proof with a union bound. □

C Polylogarithmic regret with diminishing returns: missing details

We now introduce two key assumptions that greatly generalize Assumption 1 (in Section 2), but will still allow us to prove a polylogarithmic regret in the diminishing-returns setting. The first one is a mild regularity assumption on $F_1, \ldots, F_K$. The second one is a concavity assumption weaker than the classic strong concavity assumption.

Assumption 2.

1. *For each $k \in [K]$, F_k is continuously differentiable on $(0,1]$ with $\lim_{u\to 0^+} F_k'(u) = +\infty$.*
2. *There is a function $\alpha : (0,1) \to (0,\infty)$ such that for each $k \in [K]$ and for each $\epsilon \in (0,1)$, F_k is $\alpha(\epsilon)$-strongly concave on $[\epsilon, 1]$.*

Note that Assumption 2 is strictly weaker than Assumption 1, since it is implied by but it does not imply Assumption 1. Note also that, under Assumption 2:

• There exists a function $M\colon (0,1) \to (0,\infty)$ such that for each $k \in [K]$ and for each $\epsilon \in (0,1)$, F_k is $M(\epsilon)$-Lipschitz on $[\epsilon, 1]$.

• There exists a function $D_{\min}\colon (0,1] \to [0,\infty)$ defined, for all $\epsilon \in (0,1]$, by $D_{\min}(\epsilon) \coloneqq \min_{k\in[K]} \inf_{u\in(0,\epsilon]} F_k'(u)$.

• There exists a function $D_{\max}\colon (0,1] \to [0,\infty)$ defined, for all $\epsilon \in (0,1]$, by $D_{\max}(\epsilon) \coloneqq \max_{k\in[K]} \sup_{u\in[\epsilon,1]} F_k'(u)$.

• There exists a function $D_{\min}^{-1}\colon \big[D_{\min}(1), \infty\big) \to (0,1]$, defined, for all $y \in \big[D_{\min}(1), \infty\big)$, by $D_{\min}^{-1}(y) \coloneqq \inf\big\{\epsilon \in (0,1] : D_{\min}(\epsilon) \leq y\big\}$ —indeed, the set in the inf before is non-empty (it contains 1) and $D_{\min}^{-1}(y) \in (0,1]$ because $\lim_{\epsilon\to 0^+} D_{\min}(\epsilon) = \infty$.

For the remainder of this section, we assume fixed such functions M, $D_{\min}$, $D_{\max}$, and $D_{\min}^{-1}$.

We now introduce some notation that will be used to state and prove the next theorem, that is our main result of this section. As in Section 2, for each $m \in [0,\infty)^K$, define $\Phi_m\colon [0,1]^K \to [0,\infty)$, $x \mapsto \sum_{k\in[K]} m_k F_k\big(x_k\big)$. Also as in Section 2, we use the notation $x^\star(m)$ to denote a maximizer (chosen arbitrarily if there are more than one of them) of Φ_m over Δ_K and we remark that, for any $t \geq 1$, Algorithm 2 selects allocations

$$(X_{t,1}, \ldots, X_{t,K}) = x^\star(\mathrm{UCB}_{t-1}^{\delta}),$$

where $\mathrm{UCB}_{t-1}^{\delta}$ is defined in the algorithm. We can now state our generalized version of the polylogarithmic rate of Algorithm 2 in the general diminishing-returns setting.

Theorem 6. *Under Assumption 2, for any time horizon $T \geq 3$ and mean vector $\mu = (\mu_1, \ldots, \mu_K) \in (0,1]^K$, define $\mu_{\inf} \coloneqq \min_{k\in[K]} \mu_k$,*

$$u_{\inf} \coloneqq D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)\right)$$

and the realized regret

$$\mathcal{R}_T \coloneqq \sum_{t=1}^{T}\sum_{k=1}^{K}\left(\mu_k F_k\left(\left(x^\star(\mu)\right)_k\right) - \mu_k F_k(X_{t,k})\right). \tag{12}$$

If we run Algorithm 2 with parameters T and $\delta \in \left[2e^{-50(\log T)^2}, \frac{1}{2}\right)$, then its realized regret satisfies, with probability at least $1 - \frac{K}{T} - KT\delta$,

$$\mathcal{R}_T \leq \frac{64M(u_{\inf})^2 \log\left(\frac{2}{\delta}\right) K}{\mu_{\inf}\alpha(u_{\inf}) \min_{k\in[K]} F_k(u_{\inf})} \log T + 3K^2 + \frac{4K^2\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)}. \tag{13}$$

Consequently, if we run Algorithm 2 with parameters T and $\delta = (KT)^{-2}$, then the expected regret satisfies

$$R_T \leq \frac{64M(u_{\inf})^2 \log\left(2K^2T^2\right) K}{\mu_{\inf}\alpha(u_{\inf}) \min_{k\in[K]} F_k(u_{\inf})} \log T + 5K^2 + \frac{4K^2\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)}. \tag{14}$$

Note that in (13), taking $\delta = T^{-2}$ gives $\log(2/\delta) = \log\left(2T^2\right) = O(\log T)$ and the good event has probability at least

$$1 - \frac{K}{T} - KT\delta = 1 - \frac{2K}{T}.$$

Thus, for fixed problem parameters, (13) yields a $\mathrm{polylog}(T)$ realized-regret bound with probability at least $1 - \frac{2K}{T}$.

To prove Theorem 6, we will need three technical lemmas, which we state and prove first. The first lemma is a first step toward obtaining quantitative lower bounds on the optimal budget allocation as well as the budget allocations selected by Algorithm 2.

Lemma 3. *Let Assumption 2 hold. For any $m \in (0,+\infty)^K$, for any $k, \ell \in [K]$, $k \neq \ell$, we have*

$$\left(x^\star(m)\right)_\ell \geq D_{\min}^{-1}\left(\frac{m_k}{m_\ell} D_{\max}\Big(\left(x^\star(m)\right)_k\Big)\right).$$

Proof. Since $\lim_{u\to 0^+} F_k'(u) = +\infty$ and the components of m are non-zero, we can see that all the components of $x^\star(m)$ are non-zero. Hence decreasing $\left(x^\star(m)\right)_k$ by a quantity $\gamma > 0$, increasing $\left(x^\star(m)\right)_\ell$ by the same $\gamma > 0$, and taking the limit for $\gamma \to 0^+$, leads to a modification of $\sum_{r=1}^K m_r F_r\Big(\left(x^\star(m)\right)_r\Big)$ satisfying

$$0 \geq F_\ell'\Big(\left(x^\star(m)\right)_\ell\Big) m_\ell - F_k'\Big(\left(x^\star(m)\right)_k\Big) m_k. \tag{15}$$

Hence

$$F_\ell'\Big(\left(x^\star(m)\right)_\ell\Big) \leq \frac{m_k}{m_\ell} F_k'\Big(\left(x^\star(m)\right)_k\Big),$$

and thus

$$D_{\min}\Big(\left(x^\star(m)\right)_\ell\Big) \leq \frac{m_k}{m_\ell} D_{\max}\Big(\left(x^\star(m)\right)_k\Big).$$

This yields the lemma. □

Let us use the same notation $N_{T,k}$ and $\widetilde{G}_{\tau,k}$ as in (4) in Section 3. For $\delta \in (0, 1/2]$, consider a first "good event", as the event in (5) in Section 3,

$$\mathcal{E}_T^\delta \coloneqq \bigcap_{k\in[K]} \bigcap_{t\in[N_{T,k}]} \left\{\left|\frac{1}{t}\sum_{s=1}^{t} \widetilde{G}_{s,k} - \mu_k\right| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2t}}\right\}. \tag{16}$$

Then, define a second good event $\mathcal{G}_T$ as the event of Lemma 2 (that holds with probability at least $1 - \frac{K}{T}$). The next lemma shows that each arm is explored at least a linear function of t, when t is larger than some logarithmic function of T, under these good events.

Lemma 4. *Let Assumption 2 hold. Assume $\mu_{\inf} \coloneqq \min_{k\in[K]} \mu_k > 0$. Let $T \geq 3$. Consider $t \in [T]$ with*

$$t \geq \frac{400K(\log T)^2}{\min_{k\in[K]} F_k\left(\frac{1}{K^2}\right)}. \tag{17}$$

Assume that the two good events $\mathcal{E}_T^\delta$ in (16) *and $\mathcal{G}_T$ hold. For $\mathcal{E}_T^\delta$ assume that $\delta \geq 2e^{-50(\log T)^2}$. Then for each arm $\ell \in [K]$ it holds*

$$\sum_{s=1}^{t} F_\ell(X_{s,\ell}) \geq \left(\frac{t}{4K} - \frac{1}{2}\right) F_\ell\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right). \tag{18}$$

In addition, if we also have

$$t \geq 2K + \frac{4K\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)} \tag{19}$$

it also holds that, for each $\ell \in [K]$,

$$X_{t,\ell} \geq D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)\right). \tag{20}$$

Proof. For the time t under consideration, we consider k such that arm k received the maximum budget $\sum_{s=1}^{t} X_{s,k}$. We also consider any other arm ℓ than k. We let

$$S_t = \#\left\{s \in [t] : X_{s,k} \geq \frac{1}{K^2}\right\}.$$

Then $S_t + \frac{t}{K^2} \geq \sum_{s=1}^{t} X_{s,k} \geq \frac{t}{K}$ leads to $S_t \geq \frac{t}{K} - \frac{t}{K^2}$ and thus

$$S_t \geq \frac{t}{2K}.$$

In the list $s_1, \ldots, s_{S_t}$ of times $s \in [t]$ for which $X_{s,k} \geq \frac{1}{K^2}$, take an element s in the second half of it (composed of r elements if $S_t = 2r + 1$ is odd).

Using that $\mathcal{E}_T^\delta$ in (16) holds we obtain that $\mathrm{UCB}_{s-1,\ell}^\delta \geq \mu_\ell$. Also,

$$\sum_{s'=1}^{s-1} F_k(X_{s',k}) \geq \frac{S_t}{2} F_k\left(\frac{1}{K^2}\right) \geq \frac{t}{4K} F_k\left(\frac{1}{K^2}\right). \tag{21}$$

From (17), we thus have

$$\sum_{s'=1}^{s-1} F_k(X_{s',k}) \geq 100(\log T)^2. \tag{22}$$

Hence using that the event of Lemma 2 holds,

$$\sum_{s'=1}^{s-1} B_{s',k}(X_{s',k}) \geq \frac{\sum_{s'=1}^{s-1} F_k(X_{s',k})}{2} \geq 50(\log T)^2.$$

Let us recall the notation $N_{t,k} = \sum_{s=1}^{t} B_{s,k}(X_{s,k})$ and $\widetilde{G}_{\tau,k}$ as in (4). This yields, with the same understanding that $0/0 \coloneqq 0$,

$$\mathrm{UCB}_{s-1,k}^\delta = \frac{\sum_{s'=1}^{N_{s-1,k}} \widetilde{G}_{s',k}}{N_{s-1,k}} + \sqrt{\frac{\log(\frac{2}{\delta})}{1 + N_{s-1,k}}} \leq 1 + \sqrt{\frac{\log(\frac{2}{\delta})}{50(\log T)^2}}.$$

From $\delta \geq 2e^{-50(\log T)^2}$, we obtain,

$$\mathrm{UCB}_{s-1,k}^\delta \leq 2.$$

Hence applying Lemma 3 to $m = \mathrm{UCB}^{\delta}_{s-1}$, with $\mathrm{UCB}^{\delta}_{s-1,\ell} \geq \mu_\ell$ and $\mathrm{UCB}^{\delta}_{s-1,k} \leq 2$, we get

$$X_{s,\ell} \geq D^{-1}_{\min}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right). \tag{23}$$

Since (23) is true for all s in the second half of the list described above, we obtain

$$\begin{aligned}\sum_{s=1}^{t} F_\ell(X_{s,\ell}) &\geq \frac{S_t - 1}{2} F_\ell\left(D^{-1}_{\min}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right) \\ &\geq \left(\frac{t}{4K} - \frac{1}{2}\right) F_\ell\left(D^{-1}_{\min}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right).\end{aligned}$$

This yields (18) for all $\ell \neq k$. From (21), one can see that (18) also holds for $\ell = k$.

Let us now show (20), using Lemma 3. Consider any $\ell \in [K]$ and t such that (19) holds. There is $\ell' \in [K]$ such that $X_{t,\ell'} \geq \frac{1}{K}$. We have

$$\mathrm{UCB}^{\delta}_{t-1,\ell'} = \frac{\sum_{s=1}^{N_{t-1,\ell'}} \widetilde{G}_{s,\ell'}}{N_{t-1,\ell'}} + \sqrt{\frac{\log(\frac{2}{\delta})}{1 + N_{t-1,\ell'}}} \leq 1 + \sqrt{\frac{\log(\frac{2}{\delta})}{1 + N_{t-1,\ell'}}}.$$

From (18) and then (19), we have

$$\sum_{s=1}^{t-1} F_{\ell'}(X_{s,\ell'}) \geq \left(\frac{t}{4K} - \frac{1}{2}\right) F_{\ell'}\left(D^{-1}_{\min}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right) - 1 \geq 100(\log T)^2. \tag{24}$$

Hence, using that the event of Lemma 2 holds,

$$\mathrm{UCB}^{\delta}_{t-1,\ell'} \leq 1 + \sqrt{\frac{2\log(\frac{2}{\delta})}{\sum_{s=1}^{t-1} F_{\ell'}(X_{s,\ell'})}} \leq 2,$$

using $\delta \geq 2e^{-50(\log T)^2}$. Finally, under the good event $\mathcal{E}^{\delta}_T$ in (16), we have $\mathrm{UCB}^{\delta}_{t-1,\ell} \geq \mu_\ell$. Hence from Lemma 3, (20) holds. □

The last crucial lemma establishes the stability of the average reward under reward mean m, when the budget allocation goes from the optimal $x^\star(m)$ to the suboptimal $x^\star(\widetilde{m})$, computed from a presumed $\widetilde{m} \neq m$.

Lemma 5. *Let Assumption 2 hold. Let $u_{\inf} \in (0, \frac{1}{K}]$ and $0 < m_{\inf} < m_{\sup} < \infty$. Consider mean vectors $m, \widetilde{m} \in [m_{\inf}, m_{\sup}]^K$. Assume that $x^\star(m) \in [u_{\inf}, 1]^K$ and $x^\star(\widetilde{m}) \in [u_{\inf}, 1]^K$. Then*

$$\Phi_m\big(x^\star(m)\big) - \Phi_m\big(x^\star(\widetilde{m})\big) \leq \frac{2M(u_{\inf})^2}{m_{\inf}\alpha(u_{\inf})} \left\|m - \widetilde{m}\right\|_2^2.$$

Proof. For conciseness, write $\alpha = \alpha(u_{\inf})$ and $M = M(u_{\inf})$. Since $m \in [m_{\inf}, m_{\sup}]^K$, the function Φ_m is $m_{\inf}\alpha$-strongly concave on $[u_{\inf}, 1]^K$ with respect to the Euclidean norm. Indeed, for each $u, v \in [u_{\inf}, 1]^K$,

$$\begin{aligned}\Phi_m(v) &= \sum_{k\in[K]} m_k F_k\big(v_k\big) \\ &\leq \sum_{k\in[K]} m_k\Big(F_k\big(u_k\big) + F_k'\big(u_k\big)\big(v_k - u_k\big) - \frac{\alpha}{2}\big(v_k - u_k\big)^2\Big) \\ &\leq \Phi_m(u) + \sum_{k\in[K]} m_k F_k'\big(u_k\big)\big(v_k - u_k\big) - \frac{m_{\inf}\alpha}{2}\left\|v - u\right\|_2^2.\end{aligned}$$

Similarly, $\Phi_{\widetilde{m}}$ is $m_{\inf}\alpha$-strongly concave on $[u_{\inf}, 1]^K$ with respect to the Euclidean norm.

For conciseness, write $\Phi^\star = \Phi_m$ and $\widetilde{\Phi} = \Phi_{\widetilde{m}}$. Write also $x^\star = x^\star(m)$ and $\widetilde{u} = x^\star(\widetilde{m})$. For $v \in [0, \frac{1}{K}]$, write

$$\Delta_K(v) = \Delta_K \cap [v, 1]^K.$$

Since $x^\star(m) \in [u_{\inf}, 1]^K$ and $x^\star(\widetilde{m}) \in [u_{\inf}, 1]^K$, we have

$$x^\star \in \operatorname*{argmax}_{u \in \Delta_K(u_{\inf})} \Phi^\star(u) \qquad \text{and} \qquad \widetilde{u} \in \operatorname*{argmax}_{u \in \Delta_K(u_{\inf})} \widetilde{\Phi}(u).$$

Since $u_{\inf} > 0$, the point $x^\star$ is in the relative interior of Δ_K. Moreover, $x^\star$ maximizes $\Phi^\star$ on Δ_K. Therefore, for each $u \in \Delta_K$,

$$\left\langle \nabla \Phi^\star(x^\star), u - x^\star \right\rangle = 0. \tag{25}$$

Leveraging first the $m_{\inf}\alpha$-strong concavity of $\Phi^\star$, and then the M-Lipschitzness of $F_1, \dots, F_K$ on $[u_{\inf}, 1]$, we have

$$\begin{aligned}
&\frac{m_{\inf}\alpha}{2} \left\| \widetilde{u} - x^\star \right\|_2^2 && (26)\\
&\quad \le \Phi^\star(x^\star) + \underbrace{\left\langle \nabla \Phi^\star(x^\star), \widetilde{u} - x^\star \right\rangle}_{=0 \text{ by } (25)} - \Phi^\star(\widetilde{u}) \\
&\quad = \Phi^\star(x^\star) - \Phi^\star(\widetilde{u}) && (27)\\
&\quad = \Phi^\star(x^\star) - \widetilde{\Phi}(x^\star) + \underbrace{\widetilde{\Phi}(x^\star) - \widetilde{\Phi}(\widetilde{u})}_{\le 0} + \widetilde{\Phi}(\widetilde{u}) - \Phi^\star(\widetilde{u}) \\
&\quad \le \Phi^\star(x^\star) - \widetilde{\Phi}(x^\star) + \widetilde{\Phi}(\widetilde{u}) - \Phi^\star(\widetilde{u}) \\
&\quad = \sum_{k \in [K]} \big(m_k - \widetilde{m}_k\big)\big(F_k(x_k^\star) - F_k(\widetilde{u}_k)\big) \\
&\quad \le M \left\| m - \widetilde{m} \right\|_2 \left\| x^\star - \widetilde{u} \right\|_2. && (28)
\end{aligned}$$

From (26)–(28), we get

$$\left\| x^\star - \widetilde{u} \right\|_2 \le \frac{2M}{m_{\inf}\alpha} \left\| m - \widetilde{m} \right\|_2. \tag{29}$$

Putting (27)–(28) together with (29), we obtain

$$\Phi^\star(x^\star) - \Phi^\star(\widetilde{u}) \le M \left\| m - \widetilde{m} \right\|_2 \left\| x^\star - \widetilde{u} \right\|_2 \le \frac{2M^2}{m_{\inf}\alpha} \left\| m - \widetilde{m} \right\|_2^2.$$

This concludes the proof. □

We now have all the tools to prove our polylogarithmic regret bound under concavity, i.e., to complete the proof of Theorem 6.

Proof of Theorem 6. Let $\delta \in \left[2e^{-50(\log T)^2}, \frac{1}{2}\right)$ be the confidence parameter used by Algorithm 2. Thus the algorithm selects $X_t = x^\star\left(\mathrm{UCB}_{t-1}^\delta\right)$ at every time t. The aim is to use Lemma 5 with

$$m_{\inf} = \mu_{\inf}, \qquad m_{\sup} = 2, \qquad u_{\inf} = D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)\right).$$

For conciseness, write $\alpha = \alpha(u_{\inf})$ and $M = M(u_{\inf})$.

By Corollary 3 and the union bound over $k \in [K]$ and over at most T possible positive values of the completion counter,

$$\mathbb{P}\left[\mathcal{E}_T^\delta\right] \ge 1 - KT\delta.$$

Moreover, by Lemma 2, $\mathbb{P}\left[\mathcal{G}_T\right] \ge 1 - \frac{K}{T}$. Hence

$$\mathbb{P}\left[\mathcal{E}_T^\delta \cap \mathcal{G}_T\right] \ge 1 - \frac{K}{T} - KT\delta.$$

We work on the event $\mathcal{E}_T^\delta \cap \mathcal{G}_T$ for the rest of the proof.

Define

$$v_0 \coloneqq D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)$$

and

$$S(T) \coloneqq 2K + \left\lceil \frac{4K\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(v_0\right)} \right\rceil.$$

This is the quantity appearing in (19). We first note that the definition of $S(T)$ implies the conditions needed to apply Lemma 4. Indeed, by the monotonicity properties of $D_{\min}^{-1}$ and since $D_{\min}\left(\frac{1}{K^2}\right) \le D_{\max}\left(\frac{1}{K^2}\right)$, we have

$$v_0 \le \frac{1}{K^2}.$$

Since the F_k are non-decreasing, this gives

$$\min_{k\in[K]} F_k(v_0) \le \min_{k\in[K]} F_k\left(\frac{1}{K^2}\right).$$

Therefore, $t \ge S(T)$ implies both (17) and (19). By (20) in Lemma 4, for every $t \ge S(T)$ and every $k \in [K]$,

$$X_{t,k} = \left(x^\star\left(\mathrm{UCB}_{t-1}^{\delta}\right)\right)_k \ge u_{\inf}.$$

We next check the remaining hypotheses of Lemma 5. First, $\mu \in [\mu_{\inf}, 2]^K$. Second, on $\mathcal{E}_T^{\delta}$, for every $t \in [T]$ and every $k \in [K]$,

$$\mathrm{UCB}_{t,k}^{\delta} \ge \mu_k \ge \mu_{\inf}.$$

Moreover, for every $t \ge S(T)$ and every $k \in [K]$, the argument leading to (24) in the proof of Lemma 4 gives

$$\sum_{s=1}^{t-1} F_k(X_{s,k}) \ge 100(\log T)^2.$$

Since $\mathcal{G}_T$ holds, this implies

$$N_{t-1,k} = \sum_{s=1}^{t-1} B_{s,k}(X_{s,k}) \ge \frac{1}{2}\sum_{s=1}^{t-1} F_k(X_{s,k}) \ge 50(\log T)^2.$$

Therefore, using also $\delta \ge 2e^{-50(\log T)^2}$,

$$\mathrm{UCB}_{t-1,k}^{\delta} \le 1 + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + N_{t-1,k}}} \le 1 + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{50(\log T)^2}} \le 2.$$

Thus, for every $t \ge S(T)$,

$$\mathrm{UCB}_{t-1}^{\delta} \in [\mu_{\inf}, 2]^K.$$

It remains to lower bound the optimal allocation $x^\star(\mu)$. Choose $j \in [K]$ such that $\left(x^\star(\mu)\right)_j \ge \frac{1}{K}$. For any $\ell \in [K]$, Lemma 3, applied with $m = \mu$, gives

$$\left(x^\star(\mu)\right)_\ell \ge D_{\min}^{-1}\left(\frac{\mu_j}{\mu_\ell} D_{\max}\left(\left(x^\star(\mu)\right)_j\right)\right).$$

Since $\mu_j/\mu_\ell \le 1/\mu_{\inf} \le 2/\mu_{\inf}$ and $\left(x^\star(\mu)\right)_j \ge 1/K$, the monotonicity properties of $D_{\max}$ and $D_{\min}^{-1}$ yield

$$\left(x^\star(\mu)\right)_\ell \ge D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)\right) = u_{\inf}.$$

Hence both $x^\star(\mu)$ and $x^\star\left(\mathrm{UCB}_{t-1}^{\delta}\right)$ have all coordinates at least $u_{\inf}$ for every $t \ge S(T)$, and the hypotheses of Lemma 5 are satisfied.

If $S(T) > T$, then the deterministic bound $\mathcal{R}_T \le KT \le KS(T)$ already gives the desired high-probability bound. We may therefore assume $S(T) \le T$. The regret incurred before time $S(T)$ is at most $KS(T)$. Therefore,

$$\mathcal{R}_T \le KS(T) + \sum_{t=S(T)}^{T} \left(\Phi_\mu\left(x^\star(\mu)\right) - \Phi_\mu\left(x^\star\left(\mathrm{UCB}_{t-1}^{\delta}\right)\right)\right).$$

Applying Lemma 5 with $m = \mu$ and $\widetilde{m} = \mathrm{UCB}_{t-1}^{\delta}$ gives

$$\begin{aligned}\mathcal{R}_T &\le KS(T) + \sum_{t=S(T)}^{T} \frac{2M^2}{\mu_{\inf}\alpha} \left\|\mu - \mathrm{UCB}_{t-1}^{\delta}\right\|_2^2 \\ &\le KS(T) + \sum_{t=S(T)}^{T} \frac{4M^2}{\mu_{\inf}\alpha} \sum_{k=1}^{K} \frac{4\log\left(\frac{2}{\delta}\right)}{1 + \sum_{s=1}^{t-1} B_{s,k}(X_{s,k})}.\end{aligned}$$

The last inequality uses $\mathcal{E}_T^{\delta}$, which implies

$$0 \le \mathrm{UCB}_{t-1,k}^{\delta} - \mu_k \le 2\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + N_{t-1,k}}}$$

for every t and k.

Using $\mathcal{G}_T$ and the lower bound $\sum_{s=1}^{t-1} F_k(X_{s,k}) \ge 100(\log T)^2$ from above, we get

$$\sum_{s=1}^{t-1} B_{s,k}(X_{s,k}) \ge \frac{1}{2} \sum_{s=1}^{t-1} F_k(X_{s,k}).$$

Moreover, for every $t \ge S(T)$ and every $k \in [K]$,

$$\sum_{s=1}^{t-1} F_k(X_{s,k}) \ge \sum_{s=1}^{S(T)-1} F_k(X_{s,k}) + \sum_{s=S(T)}^{t-1} F_k(X_{s,k}) \ge 100(\log T)^2 + \big(t - S(T)\big) F_k(u_{\inf}).$$

Since $F_k(u_{\inf}) \le 1$ and $T \ge 3$, this implies

$$\sum_{s=1}^{t-1} F_k(X_{s,k}) \ge \big(t - S(T) + 1\big) F_k(u_{\inf}).$$

Hence

$$\begin{aligned}\mathcal{R}_T &\le KS(T) + \frac{16M^2 \log\left(\frac{2}{\delta}\right)}{\mu_{\inf}\alpha} \sum_{t=S(T)}^{T} \sum_{k=1}^{K} \frac{2}{\sum_{s=1}^{t-1} F_k(X_{s,k})} \\ &\le KS(T) + \frac{32M^2 \log\left(\frac{2}{\delta}\right) K}{\mu_{\inf}\alpha \min_{k\in[K]} F_k(u_{\inf})} \sum_{r=1}^{T-S(T)+1} \frac{1}{r} \\ &\le KS(T) + \frac{64M^2 \log\left(\frac{2}{\delta}\right) K}{\mu_{\inf}\alpha \min_{k\in[K]} F_k(u_{\inf})} \log T.\end{aligned}$$

Substituting the definition of $S(T)$ yields

$$\begin{aligned}\mathcal{R}_T &\le \frac{64M(u_{\inf})^2 \log\left(\frac{2}{\delta}\right) K}{\mu_{\inf}\alpha(u_{\inf}) \min_{k\in[K]} F_k(u_{\inf})} \log T + 2K^2 + K\left\lceil \frac{4K\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)} \right\rceil \\ &\le \frac{64M(u_{\inf})^2 \log\left(\frac{2}{\delta}\right) K}{\mu_{\inf}\alpha(u_{\inf}) \min_{k\in[K]} F_k(u_{\inf})} \log T + 3K^2 + \frac{4K^2\left(100(\log T)^2 + 1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)}.\end{aligned}$$

This proves the high-probability bound on an event of probability at least $1 - \frac{K}{T} - KT\delta$ as stated in (13).

We now prove (14).

Set $\delta = (KT)^{-2}$. If $\delta \ge 2e^{-50(\log T)^2}$, the previous bound applies. Since the reward function is $[0, K]$-valued, $\mathcal{R}_T \le KT$ always, and therefore

$$R_T = \mathbb{E}\big[\mathcal{R}_T\big] \le \mathrm{RHS}(\delta) + KT\left(\frac{K}{T} + KT\delta\right) \le \mathrm{RHS}(\delta) + 2K^2,$$

where $\mathrm{RHS}(\delta)$ denotes the right-hand side of (13). This gives (14).

If $\delta < 2e^{-50(\log T)^2}$, then

$$K > \frac{e^{25(\log T)^2}}{\sqrt{2}T}.$$

For $T \geq 3$, the right-hand side is at least $T/5$, and hence $KT \leq 5K^2$. Since always $R_T \leq KT$, the bound (14) follows trivially in this case. Thus (14) holds in all cases. □

In the next proposition, we make explicit the upper bound on R_T of Theorem 6, for the special case of concave power functions $F_1, \dots, F_K$, as in Theorem 1.

Proposition 1. *Consider the case where there are $a_1, \dots, a_K \in (0,1)$ such that, for all $k \in [K]$ and $x \in [0,1]$, $F_k(x) = x^{a_k}$. Define $a_{\min} \coloneqq \min_{k\in[K]} a_k$ and $a_{\max} \coloneqq \max_{k\in[K]} a_k$. Define $\mu_{\inf} \coloneqq \min_{k\in[K]} \mu_k$ and assume $\mu_{\inf} > 0$. If we run Algorithm 2 with parameters $T \geq 3$ and confidence parameter $\delta = \frac{1}{(KT)^2}$, and if*

$$\frac{1}{(KT)^2} \geq 2e^{-50(\log T)^2},$$

then

$$R_T \leq \frac{64K^{1+\frac{(2(1-a_{\min})+a_{\max})(1-a_{\min})}{1-a_{\max}}}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{2(1-a_{\min})+a_{\max}}{1-a_{\max}}} \mu_{\inf} \min_{k\in[K]} a_k(1-a_k)} \log(2K^2T^2)\log(T) + 5K^2$$
$$+ \frac{4\left(100(\log T)^2+1\right)K^{2+\frac{2a_{\max}(1-a_{\min})}{1-a_{\max}}}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{a_{\max}}{1-a_{\max}}}}. \tag{30}$$

In particular, for any $T \geq 2$, if we run Algorithm 2 with parameters T and confidence parameter $\delta = \frac{1}{(KT)^2}$, then

$$R_T \leq 800\left(\frac{K}{a_{\min}\mu_{\inf}(1-a_{\max})}\right)^{\frac{9}{1-a_{\max}}} (\log T)^2. \tag{31}$$

Proof. First, assume that $T \geq 3$ and

$$\frac{1}{(KT)^2} \geq 2e^{-50(\log T)^2}.$$

As already discussed, Assumption 2 holds, and we can apply Theorem 6 with $\delta = (KT)^{-2}$. Let us first bound the individual quantities in (14).

Since $F_k'(x) = a_k x^{a_k-1}$, the function F_k is $a_k\varepsilon^{a_k-1}$-Lipschitz on $[\varepsilon,1]$ and thus ε^{a_k-1}-Lipschitz on $[\varepsilon,1]$. Hence, M can be chosen such that, for $\varepsilon \in (0,1]$,

$$M(\varepsilon) \leq \varepsilon^{a_{\min}-1}.$$

Also, for $x \in (0,\varepsilon]$, $F_k'(x) \geq a_k\varepsilon^{a_k-1} \geq a_{\min}\varepsilon^{a_{\max}-1}$. Hence, for $\varepsilon \in (0,1]$,

$$D_{\min}(\varepsilon) \geq a_{\min}\varepsilon^{a_{\max}-1}.$$

Then, for $x \in [\varepsilon,1]$, $F_k'(x) \leq a_k\varepsilon^{a_k-1} \leq a_{\max}\varepsilon^{a_{\min}-1}$. Hence, for $\varepsilon \in (0,1]$,

$$D_{\max}(\varepsilon) \leq a_{\max}\varepsilon^{a_{\min}-1}.$$

Next, for $\varepsilon, y > 0$ such that

$$\varepsilon < \left(\frac{a_{\min}}{y}\right)^{\frac{1}{1-a_{\max}}},$$

we have $D_{\min}(\varepsilon) \geq a_{\min}\varepsilon^{a_{\max}-1} > y$. Hence, for $y \in [D_{\min}(1), \infty)$,

$$D_{\min}^{-1}(y) \geq \left(\frac{a_{\min}}{y}\right)^{\frac{1}{1-a_{\max}}}.$$

Then,

$$u_{\inf} = D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)\right) \geq \left(\frac{a_{\min}}{\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K}\right)}\right)^{\frac{1}{1-a_{\max}}}$$
$$\geq \left(\frac{a_{\min}}{\frac{2}{\mu_{\inf}} a_{\max}\left(\frac{1}{K}\right)^{a_{\min}-1}}\right)^{\frac{1}{1-a_{\max}}} = \left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{1}{1-a_{\max}}} K^{\frac{a_{\min}-1}{1-a_{\max}}}.$$

Similarly,

$$D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right) \geq \left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{1}{1-a_{\max}}} K^{\frac{2(a_{\min}-1)}{1-a_{\max}}}.$$

Also, for $\varepsilon \in (0,1]$,

$$\min_{k\in[K]} F_k(\varepsilon) \geq \varepsilon^{a_{\max}}.$$

Finally, for $k \in [K]$, for $\varepsilon \in (0,1]$ and $x \in (\varepsilon, 1]$,

$$\frac{\partial^2}{\partial x^2} x^{a_k} = a_k(a_k-1)x^{a_k-2} \leq a_k(a_k-1).$$

Thus, F_k is $a_k(1-a_k)$-strongly concave on $(\varepsilon,1]$. Hence, we can choose α such that, for $\varepsilon \in (0,1]$,

$$\alpha(\varepsilon) \geq \min_{k\in[K]} a_k(1-a_k).$$

Now, let us consider the regret upper bound in (14) and use the individual bounds above:

$$\begin{aligned}
R_T \leq & \frac{64M(u_{\inf})^2 \log(2K^2T^2)K}{\mu_{\inf}\alpha(u_{\inf})\left(\min_{k\in[K]} F_k(u_{\inf})\right)}\log(T) + 5K^2 + \frac{4K^2\left(100(\log T)^2+1\right)}{\min_{k\in[K]} F_k\left(D_{\min}^{-1}\left(\frac{2}{\mu_{\inf}} D_{\max}\left(\frac{1}{K^2}\right)\right)\right)} \\
\leq & \frac{64u_{\inf}^{2(a_{\min}-1)-a_{\max}}K}{\mu_{\inf}\min_{k\in[K]} a_k(1-a_k)}\log(2K^2T^2)\log(T) + 5K^2 + \frac{4K^2\left(100(\log T)^2+1\right)}{\left(\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{1}{1-a_{\max}}} K^{\frac{2(a_{\min}-1)}{1-a_{\max}}}\right)^{a_{\max}}} \\
\leq & \frac{64\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{2(a_{\min}-1)-a_{\max}}{1-a_{\max}}} K^{1+\frac{(2(a_{\min}-1)-a_{\max})(a_{\min}-1)}{1-a_{\max}}}}{\mu_{\inf}\min_{k\in[K]} a_k(1-a_k)}\log(2K^2T^2)\log(T) + 5K^2 \\
& + 4\left(100(\log T)^2+1\right)\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{a_{\max}}{a_{\max}-1}} K^{2+\frac{2a_{\max}(1-a_{\min})}{1-a_{\max}}} \\
= & \frac{64K^{1+\frac{(2(1-a_{\min})+a_{\max})(1-a_{\min})}{1-a_{\max}}}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{2(1-a_{\min})+a_{\max}}{1-a_{\max}}}\mu_{\inf}\min_{k\in[K]} a_k(1-a_k)}\log(2K^2T^2)\log(T) + 5K^2 \\
& + \frac{4\left(100(\log T)^2+1\right)K^{2+\frac{2a_{\max}(1-a_{\min})}{1-a_{\max}}}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^{\frac{a_{\max}}{1-a_{\max}}}}.
\end{aligned}$$

This proves (30).

We now derive the simpler estimate (31) in the present regime, i.e., when $T \geq 3$ and

$$\frac{1}{(KT)^2} \geq 2e^{-50(\log T)^2}.$$

Let

$$b \coloneqq 1 - a_{\max} \qquad \text{and} \qquad Q \coloneqq \frac{K}{a_{\min}\mu_{\inf} b}.$$

We use this notation only for the rest of the derivation. Since $K \geq 2$, $a_{\min} \leq a_{\max} < 1$, and $\mu_{\inf} \leq 1$, we have $Q \geq K \geq 2$.

First, for $T \geq 3$ and $K \geq 2$,

$$\log(2K^2T^2) = \log 2 + 2\log K + 2\log T \leq 6\log K \log T.$$

Hence,

$$\log(2K^2T^2)\log T \leq 6\log K(\log T)^2.$$

Moreover,

$$100(\log T)^2 + 1 \leq 101(\log T)^2.$$

We now bound the first term in (30). Set

$$p \coloneqq \frac{2(1-a_{\min}) + a_{\max}}{1-a_{\max}}.$$

Since $1 - a_{\min} \leq 1$ and $a_{\max} \leq 1$, we have

$$p \leq \frac{3}{b}.$$

Also,

$$p(1-a_{\min}) \leq p \leq \frac{3}{b}.$$

Furthermore,

$$\min_{k\in[K]} a_k(1-a_k) \geq a_{\min}(1-a_{\max}) = a_{\min}b,$$

and

$$\frac{2a_{\max}}{a_{\min}\mu_{\inf}} \leq \frac{K}{a_{\min}\mu_{\inf}b} = Q.$$

Therefore, using also $\log K \leq K \leq Q$ and

$$\frac{1}{a_{\min}\mu_{\inf}b} = \frac{Q}{K} \leq Q,$$

we obtain

$$\begin{aligned}\frac{K^{1+p(1-a_{\min})}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^p \mu_{\inf}\min_{k\in[K]} a_k(1-a_k)}\log K &\leq K^{1+p(1-a_{\min})}\left(\frac{2a_{\max}}{a_{\min}\mu_{\inf}}\right)^p \frac{\log K}{\mu_{\inf}a_{\min}b} \\ &\leq Q^{1+p(1-a_{\min})}Q^pQQ = Q^{3+p+p(1-a_{\min})} \\ &\leq Q^{3+\frac{6}{b}} \leq Q^{\frac{9}{b}},\end{aligned}$$

where the last inequality uses $b \leq 1$. Thus, the first term in (30) is at most

$$64 \cdot 6\, Q^{\frac{9}{b}}(\log T)^2 = 384\, Q^{\frac{9}{b}}(\log T)^2.$$

Next, consider the third term in (30). Set

$$r \coloneqq \frac{2a_{\max}(1-a_{\min})}{1-a_{\max}} \qquad \text{and} \qquad s \coloneqq \frac{a_{\max}}{1-a_{\max}}.$$

Then

$$r \leq \frac{2}{b}, \qquad s \leq \frac{1}{b}.$$

Hence,

$$\frac{K^{2+r}}{\left(\frac{a_{\min}\mu_{\inf}}{2a_{\max}}\right)^s} \leq K^{2+r}\left(\frac{2a_{\max}}{a_{\min}\mu_{\inf}}\right)^s \leq Q^{2+r+s} \leq Q^{2+\frac{3}{b}} \leq Q^{\frac{5}{b}} \leq Q^{\frac{9}{b}}.$$

Therefore, the third term in (30) is at most

$$4 \cdot 101\, Q^{\frac{9}{b}}(\log T)^2 = 404\, Q^{\frac{9}{b}}(\log T)^2.$$

Finally, since $Q \geq K \geq 2$, $b \leq 1$, and $T \geq 3$,

$$5K^2 \leq 5Q^2 \leq 5Q^{\frac{9}{b}}(\log T)^2.$$

Combining the three estimates gives

$$R_T \le (384 + 404 + 5) Q^{\frac{9}{b}} (\log T)^2 \le 800 Q^{\frac{9}{b}} (\log T)^2.$$

Substituting back

$$Q = \frac{K}{a_{\min}\mu_{\inf}(1 - a_{\max})} \qquad \text{and} \qquad b = 1 - a_{\max}$$

yields

$$R_T \le 800 \left(\frac{K}{a_{\min}\mu_{\inf}(1 - a_{\max})} \right)^{\frac{9}{1-a_{\max}}} (\log T)^2.$$

It remains only to prove (31) in the cases not covered above. Since the per-round reward lies in $[0, K]$, we always have

$$R_T \le KT.$$

Moreover, since the rewards are $[0, 1]$-valued, $\mu_{\inf} \le 1$. Therefore,

$$\frac{K}{a_{\min}\mu_{\inf}(1 - a_{\max})} \ge K, \qquad \frac{9}{1 - a_{\max}} \ge 9.$$

If $T = 2$, then, since $K \ge 2$,

$$R_T \le KT = 2K \le K^2 \le 800 \left(\frac{K}{a_{\min}\mu_{\inf}(1 - a_{\max})} \right)^{\frac{9}{1-a_{\max}}} (\log T)^2.$$

Finally, suppose that $T \ge 3$ and

$$\frac{1}{(KT)^2} < 2e^{-50(\log T)^2}.$$

Taking logarithms gives

$$-2\log K - 2\log T < \log 2 - 50(\log T)^2,$$

and therefore

$$\log K > 25(\log T)^2 - \log T - \frac{\log 2}{2}.$$

For $T \ge 3$, the right-hand side is larger than $\log T$. Hence $K > T$, and thus

$$R_T \le KT \le K^2 \le 800 \left(\frac{K}{a_{\min}\mu_{\inf}(1 - a_{\max})} \right)^{\frac{9}{1-a_{\max}}} (\log T)^2.$$

This concludes the proof. □

The next lemma makes explicit the sharper upper bound of Proposition 1 for the example of Theorem 2.

Lemma 6. *Consider the setting of Proposition 1 and assume further that $K = 2$, $a_1 = a_2 = \frac{1}{2}$, $\mu_1 \ge \frac{1}{4}$, and $\mu_2 \ge \frac{1}{4}$. Then, for any $T \ge 3$,*

$$R_T \le 2097152\sqrt{2}\log(8T^2)\log(T) + 20 + 256\left(100(\log T)^2 + 1\right).$$

Proof. For $K = 2$, the technical condition in Proposition 1 reads

$$\frac{1}{4T^2} \ge 2e^{-50(\log T)^2}.$$

Equivalently,

$$50(\log T)^2 \ge 2\log T + \log 8.$$

This holds for all $T \ge 3$. Therefore, we may apply (30) in Proposition 1.

We have $a_{\min} = a_{\max} = \frac{1}{2}$ and $\mu_{\inf} \geq \frac{1}{4}$. Thus, from (30),

$$\begin{aligned}
R_T &\leq \frac{64 \cdot 2^{1+\frac{(2(1-\frac{1}{2})+\frac{1}{2})(1-\frac{1}{2})}{1-\frac{1}{2}}}}{\left(\frac{\frac{1}{2}\frac{1}{4}}{2\frac{1}{2}}\right)^{\frac{2(1-\frac{1}{2})+\frac{1}{2}}{1-\frac{1}{2}}} \frac{1}{4}\frac{1}{4}} \log(8T^2)\log(T) + 20 + \frac{4\left(100(\log T)^2+1\right)2^{2+\frac{2\frac{1}{2}(1-\frac{1}{2})}{1-\frac{1}{2}}}}{\left(\frac{\frac{1}{2}\frac{1}{4}}{2\frac{1}{2}}\right)^{\frac{\frac{1}{2}}{1-\frac{1}{2}}}} \\
&= \frac{64 \cdot 2^{1+\frac{3}{2}}}{\left(\frac{1}{8}\right)^3 \frac{1}{16}} \log(8T^2)\log(T) + 20 + \frac{4\left(100(\log T)^2+1\right)2^3}{\frac{1}{8}} \\
&= 2 \cdot 16 \cdot 512 \cdot 32 \cdot 4\sqrt{2}\log(8T^2)\log(T) + 20 + 256\left(100(\log T)^2+1\right) \\
&= 2097152\sqrt{2}\log(8T^2)\log(T) + 20 + 256\left(100(\log T)^2+1\right).
\end{aligned}$$

This concludes the proof. □

D Worst-case lower bound: proof of Theorem 4

Proof of Theorem 4. Recall the notation in Section 1.1. Assume $K \geq 2$ and $T \geq 16$. Let

$$m \coloneqq \lfloor K/2 \rfloor.$$

We construct a hard instance using the first $2m$ tasks, grouped into m pairs. If K is odd, the remaining task is a dummy task. For each $k \in [2m]$, define

$$F_k \coloneqq \mathbb{I}_{[1/m,1]}.$$

If K is odd, define $F_K \equiv 0$ for the dummy task.

Let $\varepsilon \coloneqq T^{-1/2}$. Let $\mathcal{B}(q)$ be the Bernoulli distribution with parameter q for $q \in [0,1]$. For any $j_1, \ldots, j_m \in [2]$, let $\mathbb{P}_{j_1,\ldots,j_m}$ be a probability measure under which the rewards are independent over time and tasks, and such that, for each $t \in [T]$ and each $i \in [m]$, if $j_i = 1$, then $G_{t,2i-1}$ has distribution $\mathcal{B}(\frac{1}{2}+\varepsilon)$ and $G_{t,2i}$ has distribution $\mathcal{B}(\frac{1}{2})$, while, if $j_i = 2$, then $G_{t,2i-1}$ has distribution $\mathcal{B}(\frac{1}{2})$ and $G_{t,2i}$ has distribution $\mathcal{B}(\frac{1}{2}+\varepsilon)$. If K is odd, the dummy reward is deterministically zero. Throughout the proof, for any $j_1, \ldots, j_m$ under consideration, we write $\mathbb{P} = \mathbb{P}_{j_1,\ldots,j_m}$ and denote the corresponding expectation by $\mathbb{E}$. We also write $R_{j_1,\ldots,j_m,T}$ for the regret up to horizon T under $\mathbb{P}_{j_1,\ldots,j_m}$.

To prove the theorem, it is enough to show that there exist $j_1, \ldots, j_m \in [2]$ such that

$$R_{j_1,\ldots,j_m,T} = \Omega\left(m\sqrt{T}\right). \tag{32}$$

Indeed, since $m = \lfloor K/2 \rfloor \geq K/3$ for $K \geq 2$, this implies

$$R_T^\star = \Omega\left(K\sqrt{T}\right).$$

Write

$$I_{t,k} \coloneqq F_k(X_{t,k})$$

for $t \in [T]$ and $k \in [2m]$. We first show that, without loss of generality, we may restrict attention to algorithms satisfying

$$\sum_{k=1}^{2m} I_{t,k} = m \qquad \text{for every } t \in [T].$$

Indeed, consider any full-feedback algorithm. At any time t, replace every active coordinate $X_{t,k} \geq 1/m$, $k \in [2m]$, by $1/m$, replace every inactive coordinate and the dummy coordinate, if present, by 0, and then choose $m - \sum_{k=1}^{2m} I_{t,k}$ inactive indices among $[2m]$ and set their allocations to $1/m$. This produces a feasible action and can only increase the realized reward, because it preserves all previously active non-dummy tasks and activates additional non-dummy tasks with nonnegative rewards. Moreover, since the completion indicators are deterministic functions of the played allocation and the algorithm has full-feedback access to the entire reward vector G_t, the transformed algorithm can internally simulate the history that the original algorithm would have observed. Thus the transformation defines a valid online algorithm and does not reduce the information available to the

simulation. It is therefore sufficient to prove the lower bound for algorithms satisfying $\sum_{k=1}^{2m} I_{t,k} = m$ almost surely.

We next transform such an algorithm into one satisfying

$$I_{t,2i-1} + I_{t,2i} = 1 \qquad \text{for every } t \in [T] \text{ and } i \in [m].$$

Fix a time t. If there are pairs with both tasks active, let $\{u_1 < \cdots < u_N\}$ be the set of indices $i \in [m]$ for which $I_{t,2i-1} = I_{t,2i} = 1$. Let $\{v_1 < \cdots < v_N\}$ be the set of indices $i \in [m]$ for which $I_{t,2i-1} = I_{t,2i} = 0$. These two sets have the same cardinality because $\sum_{k=1}^{2m} I_{t,k} = m$. For each $\ell \in [N]$, draw U_ℓ uniformly from $\{2u_\ell - 1, 2u_\ell\}$ and V_ℓ uniformly from $\{2v_\ell - 1, 2v_\ell\}$, independently of everything else conditionally on the past. Then set $X_{t,U_\ell} = 0$ and $X_{t,V_\ell} = 1/m$.

We claim that this transformation preserves the expected reward under every instance $\mathbb{P}_{j_1,\dots,j_m}$. Fix $j_1, \dots, j_m \in [2]$. Let $\mathcal{G}_{t-1}$ denote the history up to the moment before the auxiliary random variables $U_1, V_1, \dots, U_N, V_N$ are sampled. For each $\ell \in [N]$, let $(a_\ell, b_\ell) = (2u_\ell - 1, 2u_\ell)$ if $j_{u_\ell} = 1$ and $(a_\ell, b_\ell) = (2u_\ell, 2u_\ell - 1)$ if $j_{u_\ell} = 2$. Similarly, let $(a'_\ell, b'_\ell) = (2v_\ell - 1, 2v_\ell)$ if $j_{v_\ell} = 1$ and $(a'_\ell, b'_\ell) = (2v_\ell, 2v_\ell - 1)$ if $j_{v_\ell} = 2$. Thus a_ℓ and a'_ℓ are the better tasks in their respective pairs. Conditionally on $\mathcal{G}_{t-1}$,

$$\mathbb{E}\big[\mu_{V_\ell} - \mu_{U_\ell} \mid \mathcal{G}_{t-1}\big] = \frac{1}{2}\left(\frac{1}{2} + \varepsilon + \frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2} + \varepsilon + \frac{1}{2}\right) = 0.$$

Therefore the expected reward is unchanged by the pair-balancing transformation. As above, the transformed algorithm can internally simulate the original one because full feedback reveals the entire reward vector and the completion indicators are deterministic functions of the played allocation. Hence it is sufficient to prove (32) for full-feedback algorithms satisfying

$$I_{t,2i-1} + I_{t,2i} = 1 \qquad \text{for every } t \in [T] \text{ and } i \in [m].$$

We now prove that, for every $j_1, \dots, j_m \in [2]$, the regret of any such full-feedback algorithm satisfies

$$R_{j_1,\dots,j_m,T} \geq \frac{\sqrt{T}}{2} \sum_{i=1}^{m} p_{j_1,\dots,j_m}(i), \tag{33}$$

where

$$p_{j_1,\dots,j_m}(i) \coloneqq \begin{cases} \mathbb{P}_{j_1,\dots,j_m}[E_i^c] & \text{if } j_i = 1, \\ \mathbb{P}_{j_1,\dots,j_m}[E_i] & \text{if } j_i = 2, \end{cases}$$

and

$$E_i \coloneqq \Big\{\sum_{t=1}^{T} I_{t,2i-1} \geq \frac{T}{2}\Big\}.$$

Fix $j_1, \dots, j_m \in [2]$. Since the optimal allocation activates the better task in every pair, we have

$$\max_{x \in \Delta_K} \big\langle F(x), \mu \big\rangle = m\left(\frac{1}{2} + \varepsilon\right).$$

Therefore,

$$\begin{aligned} R_{j_1,\dots,j_m,T} &= Tm\left(\frac{1}{2} + \varepsilon\right) - \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i=1}^{m}\big(I_{t,2i-1}\mu_{2i-1} + I_{t,2i}\mu_{2i}\big)\Big] \\ &= \varepsilon \sum_{i=1}^{m} \mathbb{E}\Big[\sum_{t=1}^{T}\big(I_{t,2i-1}\mathbb{I}\{j_i = 2\} + I_{t,2i}\mathbb{I}\{j_i = 1\}\big)\Big]. \end{aligned}$$

If $j_i = 2$, then the inferior task in pair i is $2i - 1$, and hence

$$\mathbb{E}\Big[\sum_{t=1}^{T} I_{t,2i-1}\Big] \geq \frac{T}{2}\mathbb{P}\big[E_i\big].$$

If $j_i = 1$, then the inferior task is $2i$. Moreover, since $I_{t,2i-1} + I_{t,2i} = 1$ for every t, the event E_i^c implies

$$\sum_{t=1}^{T} I_{t,2i} \geq \frac{T}{2}.$$

Thus

$$\mathbb{E}\Big[\sum_{t=1}^{T} I_{t,2i}\Big] \ge \frac{T}{2}\mathbb{P}\big[E_i^c\big].$$

Combining these inequalities and using $\varepsilon = T^{-1/2}$ gives (33).

We next recall the elementary bound that, for $p, q \in [1/4, 3/4]$,

$$\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) \le 8(p-q)^2. \tag{34}$$

Indeed, for fixed p, the function

$$q \mapsto \mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$$

is convex, is minimized at $q = p$, and has second derivative

$$\frac{p}{q^2} + \frac{1-p}{(1-q)^2} \le 16$$

on $[1/4, 3/4]$. A Taylor expansion gives (34).

Consider two vectors $j_1, \dots, j_m, j_1', \dots, j_m' \in [2]$ that differ only in coordinate $k \in [m]$. Since $T \ge 16$, both $\frac{1}{2}$ and $\frac{1}{2} + \varepsilon$ belong to $[1/4, 3/4]$. By the chain rule for KL, and because full feedback reveals both rewards in the changed pair at every round,

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{j_1,\dots,j_m}, \mathbb{P}_{j_1',\dots,j_m'}\right) \\
&\quad \le T\left[\mathrm{KL}\left(\mathcal{B}\left(\frac{1}{2}+\varepsilon\right), \mathcal{B}\left(\frac{1}{2}\right)\right) + \mathrm{KL}\left(\mathcal{B}\left(\frac{1}{2}\right), \mathcal{B}\left(\frac{1}{2}+\varepsilon\right)\right)\right] \\
&\quad \le T\left(8\varepsilon^2 + 8\varepsilon^2\right) = 16.
\end{aligned}$$

Hence, by the Bretagnolle–Huber inequality [Lattimore and Szepesvári, 2020, Theorem 14.2], there is a universal constant $c_0 > 0$ such that

$$\begin{aligned}
&p_{j_1,\dots,j_m}(k) + p_{j_1',\dots,j_m'}(k) \\
&\quad = \mathbb{I}\{j_k = 1\}\left(\mathbb{P}_{j_1,\dots,j_m}\big[E_k^c\big] + \mathbb{P}_{j_1',\dots,j_m'}\big[E_k\big]\right) \\
&\qquad + \mathbb{I}\{j_k = 2\}\left(\mathbb{P}_{j_1,\dots,j_m}\big[E_k\big] + \mathbb{P}_{j_1',\dots,j_m'}\big[E_k^c\big]\right) \ge c_0.
\end{aligned} \tag{35}$$

Using (33) and averaging over all instances,

$$\begin{aligned}
&\frac{1}{2^m}\sum_{j_1=1}^{2}\cdots\sum_{j_m=1}^{2} R_{j_1,\dots,j_m,T} \\
&\quad \ge \frac{\sqrt{T}}{2}\frac{1}{2^m}\sum_{j_1=1}^{2}\cdots\sum_{j_m=1}^{2}\sum_{i=1}^{m} p_{j_1,\dots,j_m}(i) \\
&\quad = \frac{\sqrt{T}}{2}\frac{1}{2^m}\sum_{i=1}^{m}\sum_{j_1=1}^{2}\cdots\sum_{j_{i-1}=1}^{2}\sum_{j_{i+1}=1}^{2}\cdots\sum_{j_m=1}^{2}\Big(p_{j_1,\dots,j_{i-1},1,j_{i+1},\dots,j_m}(i) \\
&\qquad\qquad + p_{j_1,\dots,j_{i-1},2,j_{i+1},\dots,j_m}(i)\Big) \\
&\quad \ge \frac{\sqrt{T}}{2}\frac{1}{2^m}\sum_{i=1}^{m}\sum_{j_1=1}^{2}\cdots\sum_{j_{i-1}=1}^{2}\sum_{j_{i+1}=1}^{2}\cdots\sum_{j_m=1}^{2} c_0 \\
&\quad = \frac{c_0}{4} m\sqrt{T}.
\end{aligned}$$

Therefore, there exists $j_1, \dots, j_m \in [2]$ such that

$$R_{j_1,\dots,j_m,T} \ge \frac{c_0}{4} m\sqrt{T}.$$

Since $m = \lfloor K/2 \rfloor \ge K/3$ for $K \ge 2$, this proves

$$R_T^\star = \Omega\big(K\sqrt{T}\big).$$

□

E Worst-case upper bound: proof of Theorem 3

Recall that, if the learner plays $X_1, X_2, \cdots \in \Delta_K$, then, for any $t \in \mathbb{N}$, $k \in [K]$, and $\delta \in (0,1)$, we defined

$$\mathrm{UCB}_{t,k}^{\delta} \coloneqq \frac{\sum_{s=1}^{t} B_{s,k}(X_{s,k}) G_{s,k}}{\sum_{s=1}^{t} B_{s,k}(X_{s,k})} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + \sum_{s=1}^{t} B_{s,k}(X_{s,k})}} \tag{36}$$

with the understanding that $0/0 \coloneqq 0$. Also, for conciseness, we use the same notation $N_{t,k} = \sum_{s=1}^{t} B_{s,k}(X_{s,k})$ and $\widetilde{G}_{\tau,k}$ as in (4) in Section 3. We then have, for each $t \in \mathbb{N}$, $k \in [K]$, and $\delta \in (0,1)$,

$$\mathrm{UCB}_{t,k}^{\delta} = \frac{\sum_{\tau=1}^{N_{t,k}} \widetilde{G}_{\tau,k}}{N_{t,k}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + N_{t,k}}}. \tag{37}$$

For $\delta \in (0, 1/2]$, recall the definition of the "good event" $\mathcal{E}_T^{\delta}$ in (5) in Section 3. From Corollary 3 and a union bound, we have

$$\mathbb{P}\left[\mathcal{E}_T^{\delta}\right] \geq 1 - KT\delta.$$

With the convention that $\sum_{t=1}^{0} = 0$, and recalling that $\mathrm{UCB}_{0,k}^{\delta} = \sqrt{\log \frac{2}{\delta}}$, note that under the good event, for all $t \in \{0, \ldots, T\}$, for all $k \in [K]$,

$$\left| \frac{\sum_{\tau=1}^{N_{t,k}} \widetilde{G}_{\tau,k}}{N_{t,k}} - \mu_k \right| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + N_{t,k}}}. \tag{38}$$

When $N_{t,k} = 0$, this is indeed equivalent to $|\mu_k| \leq \sqrt{\log\left(\frac{2}{\delta}\right)}$, which is true because $\log\left(\frac{2}{\delta}\right) \geq \log(4) > 1$. When $N_{t,k} \geq 1$, (38) holds because

$$\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1 + N_{t,k}}} \geq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N_{t,k}}}$$

and because the good event holds.

We then recall that, for $t \geq 1$, the allocation $X_t \in \Delta_K$ played by the algorithm at time t is such that

$$\sup_{x \in \Delta_K} \left\langle \mathrm{UCB}_{t-1}^{\delta}, F(x) \right\rangle = \left\langle \mathrm{UCB}_{t-1}^{\delta}, F(X_t) \right\rangle. \tag{39}$$

Let $x^\star \in \Delta_K$. We have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\Big[\big\langle \mu, F(x^\star)\big\rangle - \big\langle \mu, F(X_t)\big\rangle\Big] &\leq KT\left(1 - \mathbb{P}\left[\mathcal{E}_T^{\delta}\right]\right) + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \Big(\big\langle \mu, F(x^\star)\big\rangle - \big\langle \mu, F(X_t)\big\rangle\Big)\right] \\
&\leq K^2T^2\delta + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \Big(\big\langle \mu, F(x^\star)\big\rangle - \big\langle \mu, F(X_t)\big\rangle\Big)\right].
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \Big(\big\langle \mu, F(x^\star)\big\rangle - \langle \mu, F(X_t)\rangle\Big)\right] \\
&\quad = \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \big\langle \mathrm{UCB}_{t-1}^{\delta}, F(x^\star) - F(X_t)\big\rangle\right] + \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \big\langle \mu - \mathrm{UCB}_{t-1}^{\delta}, F(x^\star) - F(X_t)\big\rangle\right].
\end{aligned}$$

Above, the first expectation is upper bounded by 0 by definition of the algorithm. Furthermore, if for any $y \in \mathbb{R}^K$ we define $y_+ \coloneqq \big((y_1)_+, \ldots, (y_K)_+\big)$ and, for all $z \in \mathbb{R}$, $z_+ \coloneqq \max\{z, 0\}$, recalling that for each $k \in [K]$ and $t \in [T]$, when $\mathcal{E}_T^{\delta}$ holds, we have $\mathrm{UCB}_{t-1,k}^{\delta} \geq \mu_k$, we get

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \Big(\big\langle \mu, F(x^\star)\big\rangle - \big\langle \mu, F(X_t)\big\rangle\Big)\right] &\leq \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \big\langle \mu - \mathrm{UCB}_{t-1}^{\delta}, F(x^\star) - F(X_t)\big\rangle\right] \\
&= \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \big\langle \mathrm{UCB}_{t-1}^{\delta} - \mu, F(X_t) - F(x^\star)\big\rangle\right] \\
&\leq \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^{\delta}} \sum_{t=1}^{T} \Big\langle \mathrm{UCB}_{t-1}^{\delta} - \mu, \big(F(X_t) - F(x^\star)\big)_+\Big\rangle\right].
\end{aligned}$$

Using then (38), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^\delta}\sum_{t=1}^{T}\left\langle \mathrm{UCB}_{t-1}^{\delta}-\mu,\left(F(X_t)-F(x^\star)\right)_+\right\rangle\right]\\
&\quad\le \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^\delta}\sum_{t=1}^{T}\sum_{k=1}^{K}2\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\left(F_k(X_{t,k})-F_k(x_k^\star)\right)_+\right]\\
&\quad\le 2\sqrt{\log\left(\frac{2}{\delta}\right)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\left(F_k(X_{t,k})-F_k(x_k^\star)\right)_+}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\right]\\
&\quad\le 2\sqrt{\log\left(\frac{2}{\delta}\right)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{F_k(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\right].
\end{aligned}$$

We have, recalling the filtration $\mathcal{F}_t$ from (10), and recalling that given $\mathcal{F}_{t-1}$, $B_{t,k}(X_{t,k})$ has expectation $F_k(X_{t,k})$,

$$\begin{aligned}
\mathbb{E}\left[\frac{F_k(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\right] &= \mathbb{E}\left[\frac{1}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\mathbb{E}\left[B_{t,k}(X_{t,k})\mid\mathcal{F}_{t-1}\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{B_{t,k}(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\middle|\mathcal{F}_{t-1}\right]\right]\\
&= \mathbb{E}\left[\frac{B_{t,k}(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\right].
\end{aligned}$$

Hence, we get

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^\delta}\sum_{t=1}^{T}\left\langle \mathrm{UCB}_{t-1}^{\delta}-\mu,\left(F(X_t)-F(x^\star)\right)_+\right\rangle\right]\le 2\sqrt{\log\left(\frac{2}{\delta}\right)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{B_{t,k}(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}\right].$$

Recall the definition of $N_{T,k}$ above, and note that since $B_{s,k}(X_{s,k})\in\{0,1\}$ for $k\in[K]$ and $s\in[T]$, we have

$$\sum_{t=1}^{T}\frac{B_{t,k}(X_{t,k})}{\sqrt{1+\sum_{s=1}^{t-1}B_{s,k}(X_{s,k})}}=\sum_{\tau=1}^{N_{T,k}}\frac{1}{\sqrt{\tau}}\le 2\sqrt{N_{T,k}}=2\sqrt{\sum_{t=1}^{T}B_{t,k}(X_{t,k})}.$$

The inequality above is also valid when $N_{T,k}=0$, since both sides are equal to zero.

Hence we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_T^\delta}\sum_{t=1}^{T}\left(\left\langle\mu,F(x^\star)\right\rangle-\left\langle\mu,F(X_t)\right\rangle\right)\right]\\
&\le 4\sqrt{\log\left(\frac{2}{\delta}\right)}\mathbb{E}\left[\sum_{k=1}^{K}\sqrt{\sum_{t=1}^{T}B_{t,k}(X_{t,k})}\right]\\
&\le 4\sqrt{K\log\left(\frac{2}{\delta}\right)}\mathbb{E}\left[\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{T}B_{t,k}(X_{t,k})}\right]\\
&\le 4\sqrt{K\log\left(\frac{2}{\delta}\right)}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[B_{t,k}(X_{t,k})\right]}\\
&= 4\sqrt{K\log\left(\frac{2}{\delta}\right)}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[F_k(X_{t,k})\right]}.
\end{aligned}$$

The second inequality follows from Cauchy's inequality. The third inequality follows from Jensen's inequality.

Hence

$$T\big\langle \mu, F(x^\star)\big\rangle - \sum_{t=1}^{T} \mathbb{E}\Big[\big\langle \mu, F(X_t)\big\rangle\Big] = \sum_{t=1}^{T} \mathbb{E}\Big[\big\langle \mu, F(x^\star)\big\rangle - \big\langle \mu, F(X_t)\big\rangle\Big]$$
$$\leq K^2T^2\delta + 4\sqrt{K\log\left(\frac{2}{\delta}\right)}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\big[F_k(X_{t,k})\big]}\,.$$

Since $x^\star$ was chosen arbitrarily, we can take the supremum over $x^\star \in \Delta_K$. Substituting $\delta \coloneqq 1/(KT)^2$ gives $K^2T^2\delta = 1$, and therefore

$$R_T \leq 1 + 4\sqrt{K\log\left(\frac{2}{\delta}\right)}\sqrt{\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\big[F_k(X_{t,k})\big]},$$

which is the claimed bound, up to logarithmic factors.

F $\sqrt{T}$ bandit feedback counterexample

This section is devoted to providing the full setting of Theorem 2 and proving it.

Setting. Set $\gamma \coloneqq \frac{1}{6}$ and fix a time horizon $T \geq 4$. Set $\varepsilon \coloneqq \gamma \cdot T^{-1/4}$. Notice that with our choice of γ and T, we have that $\varepsilon \leq \frac{1}{8}$, a fact that we will use several times in what follows. Define the two instances

$$\mu^+ \coloneqq \left(\tfrac{1}{2}+\varepsilon,\ \tfrac{1}{2}-\varepsilon\right), \qquad \mu^- \coloneqq \left(\tfrac{1}{2}-\varepsilon,\ \tfrac{1}{2}+\varepsilon\right).$$

For each $\sigma \in \{+,-\}$, consider two i.i.d. sequences $\big(Y_t^\sigma(1)\big)_{t\in\mathbb{N}}$ and $\big(Y_t^\sigma(2)\big)_{t\in\mathbb{N}}$, with

$$Y_t^\sigma(1) \sim \text{Bernoulli}\left(\tfrac{1}{2}+\sigma\varepsilon\right), \qquad Y_t^\sigma(2) \sim \text{Bernoulli}\left(\tfrac{1}{2}-\sigma\varepsilon\right),$$

which we further build to be independent of each other.

Independently of $\big(Y_t^\sigma(1), Y_t^\sigma(2)\big)_{t\in\mathbb{N}}$, let $(V_{t,k})_{t\in\mathbb{N},k\in\{1,2\}}$ be an i.i.d. family of uniform random variables on $[0,1]$. For all $t \in \mathbb{N}$, $k \in \{1,2\}$, and $b \in [0,1]$, define

$$B_{t,k}(b) \coloneqq \mathbb{I}\{V_{t,k} \leq \sqrt{b}\}.$$

Then $B_{t,k}(b) \sim \text{Bernoulli}(\sqrt{b})$, and the random functions $B_{t,k}$ are independent across (t,k) and jointly measurable in (b,ω). Since we want to allow the learner to use randomized algorithms, we also assume that they can have sequential access to an i.i.d. sequence of $[0,1]$-uniforms $(U_t)_{t\in\mathbb{N}}$ used as random seeds, generated independently of the previous families of random variables. Recall that our budget-to-success curves are defined by $F_1(b) \coloneqq \sqrt{b} =: F_2(b)$, for each $b \in [0,1]$, and hence, each budget allocation $x \in \Delta_2$ satisfies

$$\Big(F_1\big(x_1\big)\Big)^2 + \Big(F_2\big(x_2\big)\Big)^2 = x_1 + x_2 = 1\,,$$

so there exists $\vartheta \in [0,\pi/2]$ such that $F_1\big(x_1\big) = \cos\vartheta$ and $F_2\big(x_2\big) = \sin\vartheta$. Vice versa, for each $\vartheta \in [0,\pi/2]$, if we set $x_1 = (\cos\vartheta)^2$ and $x_2 = (\sin\vartheta)^2$ we have that $x \in \Delta_2$ and $F_1\big(x_1\big) = \cos\vartheta$ and $F_2\big(x_2\big) = \sin\vartheta$. Hence, we see that there is a 1-to-1 correspondence between choices of $\vartheta \in [0,\pi/2]$ and $x \in \Delta_2$ with a correspondence given by $\vartheta \mapsto \big((\cos\vartheta)^2, (\sin\vartheta)^2\big)$.

It follows that, when the learner interacts with the environment generated by $\sigma \in \{+,-\}$, the learner choosing budget allocations is equivalent to choosing angles $\vartheta_t^\sigma \in \left[0, \frac{\pi}{2}\right]$ and allocating

$$X_t^\sigma \coloneqq \big((\cos\vartheta_t^\sigma)^2, (\sin\vartheta_t^\sigma)^2\big)\,.$$

We assume that the learner has access only to bandit feedback, i.e., at the end of each time t, they observe only the reward associated to the action they played:

$$O_t^\sigma \coloneqq Y_t^\sigma(1)\cdot B_{t,1}\big(\cos^2(\vartheta_t^\sigma)\big) + Y_t^\sigma(2)\cdot B_{t,2}\big(\sin^2(\vartheta_t^\sigma)\big) \in \{0,1,2\}.$$

For each natural number $t \geq 2$, define the history space $\mathcal{H}_t \coloneqq ([0,1] \times \{0,1,2\})^{t-1} \times [0,1]$ and denote by $H_t^\sigma \coloneqq (U_1, O_1^\sigma, \ldots, U_{t-1}, O_{t-1}^\sigma, U_t)$ the history observed by the learner just before receiving the feedback at time t when the underlying scenario is $\sigma \in \{+,-\}$. Define $\mathcal{H}_1 \coloneqq [0,1]$ and $H_1^\sigma \coloneqq U_1$.

Hence, the angles at time t are generated via the relation $\vartheta_t^\sigma \coloneqq \alpha_t(H_t^\sigma)$ for some deterministic function $\alpha_t \colon \mathcal{H}_t \to \left[0, \frac{\pi}{2}\right]$ (the algorithm).

Regret. For $\sigma \in \{+,-\}$, the regret of the algorithm up to time T when the underlying instance is determined by σ is

$$R_T^\sigma \coloneqq T \cdot \max_{\vartheta \in [0,\pi/2]} \langle \mu^\sigma, (\cos\vartheta, \sin\vartheta)\rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle \mu^\sigma, (\cos\vartheta_t^\sigma, \sin\vartheta_t^\sigma)\rangle\right].$$

Per-round law. Now, under instance $\sigma \in \{+,-\}$ and for each $\vartheta \in \left[0, \frac{\pi}{2}\right]$, define

$$p_\sigma(\vartheta) \coloneqq \left(\tfrac{1}{2} + \sigma\varepsilon\right) \cdot \cos\vartheta, \qquad q_\sigma(\vartheta) \coloneqq \left(\tfrac{1}{2} - \sigma\varepsilon\right) \cdot \sin\vartheta.$$

Then, conditional on ϑ_t^σ, $Y_t^\sigma(1) B_{t,1}(\cos^2\vartheta_t^\sigma) \sim \mathrm{Bernoulli}(p_\sigma(\vartheta_t^\sigma))$ and $Y_t^\sigma(2) B_{t,2}(\sin^2\vartheta_t^\sigma) \sim \mathrm{Bernoulli}(q_\sigma(\vartheta_t^\sigma))$, independently, hence O_t^σ is a sum of two independent Bernoulli random variables of parameters $p_\sigma(\vartheta_t^\sigma)$ and $q_\sigma(\vartheta_t^\sigma)$. Therefore

$$\begin{aligned}
\mathbb{P}[O_t^\sigma = 2 \mid \vartheta_t^\sigma = \vartheta] &= p_\sigma(\vartheta) q_\sigma(\vartheta) = \left(\tfrac{1}{4} - \varepsilon^2\right) \cos\vartheta \sin\vartheta \eqqcolon \eta(\vartheta), \\
\mathbb{P}[O_t^\sigma = 1 \mid \vartheta_t^\sigma = \vartheta] &= p_\sigma(\vartheta) + q_\sigma(\vartheta) - 2p_\sigma(\vartheta) q_\sigma(\vartheta), \\
\mathbb{P}[O_t^\sigma = 0 \mid \vartheta_t^\sigma = \vartheta] &= 1 - p_\sigma(\vartheta) - q_\sigma(\vartheta) + p_\sigma(\vartheta) q_\sigma(\vartheta).
\end{aligned}$$

One-step KL expansion. For $\vartheta \in [0, \frac{\pi}{2}]$, define $P_{t,\vartheta}^\sigma \coloneqq \mathbb{P}_{O_t^\sigma \mid \vartheta_t^\sigma = \vartheta}$ as the law of O_t^σ conditioned on $\vartheta_t^\sigma = \vartheta$. Then

$$\mathrm{KL}\left(P_{t,\vartheta}^+, P_{t,\vartheta}^-\right) = \sum_{k \in \{0,1,2\}} \mathbb{P}[O_t^+ = k \mid \vartheta_t^+ = \vartheta] \log \frac{\mathbb{P}[O_t^+ = k \mid \vartheta_t^+ = \vartheta]}{\mathbb{P}[O_t^- = k \mid \vartheta_t^- = \vartheta]}.$$

Lemma 7 (Per-round KL upper bound.)**.** *We have that*

$$\mathrm{KL}\left(P_{t,\vartheta}^+, P_{t,\vartheta}^-\right) \leq 84 \cdot \varepsilon^2 \cdot \left|\vartheta - \tfrac{\pi}{4}\right|^2.$$

Proof. Fix $\vartheta \in \left[0, \frac{\pi}{2}\right]$. For brevity, write

$$a \coloneqq \cos\vartheta, \qquad b \coloneqq \sin\vartheta, \qquad p_\sigma \coloneqq p_\sigma(\vartheta), \qquad q_\sigma \coloneqq q_\sigma(\vartheta), \qquad \eta \coloneqq \eta(\vartheta).$$

Set, for $k \in \{0,1,2\}$,

$$r_k \coloneqq P_{t,\vartheta}^+[\{k\}], \qquad s_k \coloneqq P_{t,\vartheta}^-[\{k\}].$$

Then $r_2 = s_2 = \eta$ and $r_0 + r_1 = s_0 + s_1 = 1 - \eta$. Define:

$$\alpha \coloneqq \frac{r_1}{1-\eta}, \qquad \beta \coloneqq \frac{s_1}{1-\eta}.$$

Step 1: Reducing to Bernoulli KL. By direct expansion,

$$\begin{aligned}
\mathrm{KL}\left(P_{t,\vartheta}^+, P_{t,\vartheta}^-\right) &= \sum_{k=0}^{2} r_k \log \frac{r_k}{s_k} \\
&= r_2 \log \frac{r_2}{s_2} + r_1 \log \frac{r_1}{s_1} + r_0 \log \frac{r_0}{s_0} \\
&= r_1 \log \frac{r_1}{s_1} + r_0 \log \frac{r_0}{s_0} \qquad (\text{since } r_2 = s_2) \\
&= (1-\eta)\left(\alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}\right) \\
&= (1-\eta)\,\mathrm{kl}(\alpha \| \beta),
\end{aligned}$$

where kl denotes the Bernoulli KL divergence.

Step 2: bound $\mathrm{kl}(\alpha\|\beta)$ by a χ^2-type inequality. Recall that, for Bernoulli laws, for all $\alpha, \beta \in (0,1)$, from [Tsybakov, 2008, Lemma 2.7],

$$\mathrm{kl}(\alpha\|\beta) \le \chi^2(\alpha\|\beta) = \frac{(\alpha-\beta)^2}{\beta(1-\beta)} \,. \tag{40}$$

Thus,

$$\mathrm{KL}\left(P^+_{t,\vartheta}, P^-_{t,\vartheta}\right) \le (1-\eta)\cdot\frac{(\alpha-\beta)^2}{\beta(1-\beta)} \,. \tag{41}$$

Step 3: compute $\alpha-\beta$ explicitly. Since $r_1 = p_+ + q_+ - 2\eta$ and $s_1 = p_- + q_- - 2\eta$,

$$r_1 - s_1 = (p_+ + q_+) - (p_- + q_-).$$

But

$$p_+ - p_- = 2\varepsilon a, \qquad q_+ - q_- = -2\varepsilon b,$$

hence

$$r_1 - s_1 = 2\varepsilon(a-b).$$

Therefore,

$$\alpha - \beta = \frac{r_1 - s_1}{1-\eta} = \frac{2\varepsilon(a-b)}{1-\eta}. \tag{42}$$

Plugging (42) into (41) yields

$$\mathrm{KL}\left(P^+_{t,\vartheta}, P^-_{t,\vartheta}\right) \le \frac{4\varepsilon^2(a-b)^2}{1-\eta}\cdot\frac{1}{\beta(1-\beta)}. \tag{43}$$

Step 4: uniform lower bounds on $1-\eta$ and on $\beta(1-\beta)$.

Lower bound on $1-\eta$. We have $ab \le \frac{1}{2}$ because $a^2+b^2=1$ and $a,b\ge 0$ (maximum at $a=b=1/\sqrt{2}$). Thus, recalling $\varepsilon \le \frac{1}{8}$, we have

$$\eta = \left(\tfrac{1}{4} - \varepsilon^2\right)ab \le \tfrac{1}{4}\cdot\tfrac{1}{2} = \tfrac{1}{8}, \qquad \text{hence} \qquad 1-\eta \ge \tfrac{7}{8} \,.$$

Therefore

$$\frac{1}{1-\eta} \le \frac{8}{7}. \tag{44}$$

Lower bound on $\beta(1-\beta)$. Recall $\beta = s_1/(1-\eta)$ and

$$s_1 = p_- + q_- - 2\eta \,.$$

First, we lower bound s_1. We have

$$p_- + q_- - \left(\tfrac{1}{2} - \varepsilon\right) = \left(\tfrac{1}{2} - \varepsilon\right)(a-1) + \left(\tfrac{1}{2} + \varepsilon\right)b = -\left(\tfrac{1}{2} - \varepsilon\right)(1-a) + \left(\tfrac{1}{2} + \varepsilon\right)b \,.$$

Using $1-a = 1-\cos\vartheta = 2\left(\sin(\vartheta/2)\right)^2$ and $b = \sin\vartheta = 2\sin(\vartheta/2)\cos(\vartheta/2)$, this becomes

$$p_- + q_- - \left(\tfrac{1}{2} - \varepsilon\right) = 2\sin\left(\tfrac{\vartheta}{2}\right)\left(\left(\tfrac{1}{2} + \varepsilon\right)\cos\left(\tfrac{\vartheta}{2}\right) - \left(\tfrac{1}{2} - \varepsilon\right)\sin\left(\tfrac{\vartheta}{2}\right)\right) \,.$$

Since $\vartheta/2 \in [0, \pi/4]$ we have $\cos(\vartheta/2) \ge \sin(\vartheta/2)$, and thus

$$\left(\tfrac{1}{2} + \varepsilon\right)\cos\left(\tfrac{\vartheta}{2}\right) - \left(\tfrac{1}{2} - \varepsilon\right)\sin\left(\tfrac{\vartheta}{2}\right) = \left(\tfrac{1}{2} - \varepsilon\right)\cdot\left(\cos\left(\tfrac{\vartheta}{2}\right) - \sin\left(\tfrac{\vartheta}{2}\right)\right) + 2\varepsilon\cdot\cos\left(\tfrac{\vartheta}{2}\right) \ge 0 \,.$$

Therefore

$$p_- + q_- \ge \tfrac{1}{2} - \varepsilon \,,$$

and hence

$$s_1 = (p_- + q_-) - 2\eta \ge \left(\tfrac{1}{2} - \varepsilon\right) - 2\cdot\tfrac{1}{8} \ge \tfrac{1}{4} - \varepsilon \ge \tfrac{1}{8} \,.$$

Second, we upper bound s_1. Since $s_1 = p_- + q_- - 2\eta$ and $\eta = \left(\frac{1}{4} - \varepsilon^2\right) ab$, we can rewrite

$$s_1 = \left(\tfrac{1}{2} - \varepsilon\right) a + \left(\tfrac{1}{2} + \varepsilon\right) b - \left(\tfrac{1}{2} - 2\varepsilon^2\right) ab = \left(\tfrac{1}{2} + \varepsilon\right) (a + b - ab) + \varepsilon a\big((1 + 2\varepsilon)\, b - 2\big) .$$

First, notice that

$$(1 + 2\varepsilon)\, b - 2 \le (1 + 2\varepsilon) - 2 = -1 + 2\varepsilon \le -\tfrac{3}{4} ,$$

so the correction term $\varepsilon a((1 + 2\varepsilon)\, b - 2)$ is non-positive and, since for $a, b \in [0, 1]$ we have $a + b - ab = 1 - (1 - a)(1 - b) \le 1$, recalling that $\varepsilon \le 1/8$, we get

$$s_1 \le \left(\tfrac{1}{2} + \varepsilon\right) (a + b - ab) \le \tfrac{1}{2} + \varepsilon \le \tfrac{5}{8} .$$

Putting the upper and lower bounds on s_1 together, we get

$$\frac{5}{7} = \frac{5/8}{7/8} \ge \frac{s_1}{7/8} \ge \frac{s_1}{1 - \eta} = \beta = \frac{s_1}{1 - \eta} \ge s_1 \ge \tfrac{1}{8} .$$

Thus $\beta \in [1/8, 5/7]$, and since $\beta(1 - \beta)$ is concave on $[0, 1]$, its minimum on this interval is attained at an endpoint:

$$\beta(1 - \beta) \ge \min\left(\tfrac{1}{8} \cdot \tfrac{7}{8},\ \tfrac{5}{7} \cdot \tfrac{2}{7}\right) = \min\left(\tfrac{7}{64},\ \tfrac{10}{49}\right) = \tfrac{7}{64} .$$

Therefore

$$\frac{1}{\beta(1 - \beta)} \le \frac{64}{7}. \tag{45}$$

Step 5: conclude the explicit constant. Combining (43), (44), and (45) gives

$$\mathrm{KL}\left(P^+_{t,\vartheta}, P^-_{t,\vartheta}\right) \le 4\varepsilon^2 (a - b)^2 \cdot \frac{8}{7} \cdot \frac{64}{7} = \frac{2048}{49} \cdot \varepsilon^2 \cdot (a - b)^2 \le 42 \cdot \varepsilon^2 \cdot (\cos\vartheta - \sin\vartheta)^2 .$$

Step 6: relate $(\cos\vartheta - \sin\vartheta)^2$ to $|\vartheta - \pi/4|^2$. Note that

$$\cos\vartheta - \sin\vartheta = \sqrt{2} \cdot \sin\left(\frac{\pi}{4} - \vartheta\right),$$

and since $|\sin x| \le |x|$, we get

$$(\cos\vartheta - \sin\vartheta)^2 \le 2 \cdot \left|\vartheta - \frac{\pi}{4}\right|^2 .$$

We conclude that

$$\mathrm{KL}\left(P^+_{t,\vartheta}, P^-_{t,\vartheta}\right) \le 42 \cdot \varepsilon^2 \cdot (\cos\vartheta - \sin\vartheta)^2 \le 84 \cdot \varepsilon^2 \cdot \left|\vartheta - \frac{\pi}{4}\right|^2 .$$

This concludes the proof. □

Lemma 8 (KL chain rule)**.** *It holds that*

$$\mathrm{KL}\left(\mathbb{P}_{(H^+_T, O^+_T)}, \mathbb{P}_{(H^-_T, O^-_T)}\right) \le 84 \cdot \varepsilon^2 \cdot \sum_{t=1}^{T} \mathbb{E}\left[\left|\vartheta^+_t - \tfrac{\pi}{4}\right|^2\right]. \tag{46}$$

Proof.

$$
\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{(H_T^+,O_T^+)},\mathbb{P}_{(H_T^-,O_T^-)}\right)\\
&\quad=\mathrm{KL}\left(\mathbb{P}_{H_T^+},\mathbb{P}_{H_T^-}\right)+\int_{\mathcal{H}_T}\mathrm{KL}\left(\mathbb{P}_{O_T^+|H_T^+=h},\mathbb{P}_{O_T^-|H_T^-=h}\right)\mathrm{d}\mathbb{P}_{H_T^+}(h)\\
&\quad=\mathrm{KL}\left(\mathbb{P}_{(H_{T-1}^+,O_{T-1}^+,U_T)},\mathbb{P}_{(H_{T-1}^-,O_{T-1}^-,U_T)}\right)+\mathbb{E}\left[\left[\mathrm{KL}\left(P_{T,\vartheta}^+,P_{T,\vartheta}^-\right)\right]_{\vartheta=\alpha_T(H_T^+)}\right]\\
&\quad=\mathrm{KL}\left(\mathbb{P}_{(H_{T-1}^+,O_{T-1}^+)},\mathbb{P}_{(H_{T-1}^-,O_{T-1}^-)}\right)+\mathbb{E}\left[\left[\mathrm{KL}\left(P_{T,\vartheta}^+,P_{T,\vartheta}^-\right)\right]_{\vartheta=\alpha_T(H_T^+)}\right]\\
&\quad\le\mathrm{KL}\left(\mathbb{P}_{(H_{T-1}^+,O_{T-1}^+)},\mathbb{P}_{(H_{T-1}^-,O_{T-1}^-)}\right)+84\cdot\varepsilon^2\cdot\mathbb{E}\left[\left|\vartheta_T^+-\tfrac{\pi}{4}\right|^2\right]\\
&\quad\le\cdots\le\mathrm{KL}\left(\mathbb{P}_{(H_1^+,O_1^+)},\mathbb{P}_{(H_1^-,O_1^-)}\right)+84\cdot\varepsilon^2\cdot\sum_{t=2}^{T}\mathbb{E}\left[\left|\vartheta_t^+-\tfrac{\pi}{4}\right|^2\right]\\
&\quad=\mathrm{KL}\left(\mathbb{P}_{(U_1,O_1^+)},\mathbb{P}_{(U_1,O_1^-)}\right)+84\cdot\varepsilon^2\cdot\sum_{t=2}^{T}\mathbb{E}\left[\left|\vartheta_t^+-\tfrac{\pi}{4}\right|^2\right]\\
&\quad=\mathrm{KL}\left(\mathbb{P}_{U_1},\mathbb{P}_{U_1}\right)+\int_{[0,1]}\mathrm{KL}\left(\mathbb{P}_{O_1^+|U_1=u},\mathbb{P}_{O_1^-|U_1=u}\right)\mathrm{d}\mathbb{P}_{U_1}(u)+84\cdot\varepsilon^2\cdot\sum_{t=2}^{T}\mathbb{E}\left[\left|\vartheta_t^+-\tfrac{\pi}{4}\right|^2\right]\\
&\quad=0+\mathbb{E}\left[\left[\mathrm{KL}\left(P_{1,\vartheta}^+,P_{1,\vartheta}^-\right)\right]_{\vartheta=\alpha_1(U_1)}\right]+84\cdot\varepsilon^2\cdot\sum_{t=2}^{T}\mathbb{E}\left[\left|\vartheta_t^+-\tfrac{\pi}{4}\right|^2\right]\\
&\quad\le 84\cdot\varepsilon^2\cdot\sum_{t=1}^{T}\mathbb{E}\left[\left|\vartheta_t^+-\tfrac{\pi}{4}\right|^2\right].
\end{aligned}
$$

This concludes the proof. □

We are now ready to prove Theorem 2, leveraging Lemma 8.

Proof of Theorem 2. **Step 1: optimal angle and a one-sided gap at $\pi/4$.** For $\sigma\in\{+,-\}$ and $\vartheta\in\left[0,\frac{\pi}{2}\right]$, define

$$f_\sigma(\vartheta)\coloneqq\langle\mu^\sigma,(\cos\vartheta,\sin\vartheta)\rangle\,.$$

The maximizer $\vartheta_\sigma^\star$ is attained at the unit vector pointing in the direction of μ^σ, i.e.,

$$(\cos\vartheta_\sigma^\star,\sin\vartheta_\sigma^\star)=\frac{\mu^\sigma}{\|\mu^\sigma\|_2}.$$

Equivalently, using $\tan\vartheta_\sigma^\star=\mu_2^\sigma/\mu_1^\sigma$, we get

$$\vartheta_\sigma^\star=\arctan\left(\frac{\frac12-\sigma\varepsilon}{\frac12+\sigma\varepsilon}\right)=\frac{\pi}{4}-\sigma\cdot\arctan(2\varepsilon)\,. \tag{47}$$

Moreover,

$$\max_{\vartheta\in[0,\pi/2]}f_\sigma(\vartheta)=\|\mu^\sigma\|_2=\frac{\sqrt2}{2}\cdot\sqrt{1+4\varepsilon^2}\,,$$

while, at $\vartheta=\frac{\pi}{4}$, we have

$$f_\sigma\left(\tfrac{\pi}{4}\right)=\left(\tfrac12+\sigma\varepsilon\right)\cdot\tfrac{\sqrt2}{2}+\left(\tfrac12-\sigma\varepsilon\right)\cdot\tfrac{\sqrt2}{2}=\tfrac{\sqrt2}{2}\,.$$

We define the gap (which is independent of $\sigma\in\{+,-\}$)

$$g(\varepsilon)\coloneqq\|\mu^\sigma\|_2-f_\sigma\left(\tfrac{\pi}{4}\right)=\frac{\sqrt2}{2}\cdot\left(\sqrt{1+4\varepsilon^2}-1\right)=\frac{\sqrt2}{2}\cdot\frac{4\varepsilon^2}{\sqrt{1+4\varepsilon^2}+1}\,,$$

and, since $\sqrt{1+4\varepsilon^2}\le1+2\varepsilon^2\le1+\frac{1}{32}=\frac{33}{32}$ (because $\varepsilon\le\frac18$), we have

$$g(\varepsilon)\ge\frac{\sqrt2}{2}\cdot\frac{4\varepsilon^2}{\frac{33}{32}+1}=\frac{\sqrt2}{2}\cdot\frac{128}{65}\cdot\varepsilon^2=\frac{64\sqrt2}{65}\cdot\varepsilon^2\,.$$

Step 2: playing on the wrong side costs at least $g(\varepsilon)$. For $\vartheta \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right]$,

$$f_+'(\vartheta) = -\left(\tfrac{1}{2}+\varepsilon\right)\sin\vartheta + \left(\tfrac{1}{2}-\varepsilon\right)\cos\vartheta \le -\left(\tfrac{1}{2}+\varepsilon\right)\cos\vartheta + \left(\tfrac{1}{2}-\varepsilon\right)\cos\vartheta = -2\varepsilon\cos\vartheta \le 0\ ,$$

so f_+ is nonincreasing on $\left[\frac{\pi}{4}, \frac{\pi}{2}\right]$, hence

$$\vartheta \ge \frac{\pi}{4} \quad \Longrightarrow \quad \|\mu^+\|_2 - f_+(\vartheta) \ge \|\mu^+\|_2 - f_+\left(\tfrac{\pi}{4}\right) = g(\varepsilon).$$

Similarly, f_- is increasing on $\left[0, \frac{\pi}{4}\right]$, hence

$$\vartheta \le \frac{\pi}{4} \quad \Longrightarrow \quad \|\mu^-\|_2 - f_-(\vartheta) \ge g(\varepsilon)\ .$$

Therefore, if we consider the instantaneous pointwise regret

$$\Delta_t^\sigma \coloneqq \|\mu^\sigma\|_2 - \langle \mu^\sigma, (\cos\vartheta_t^\sigma, \sin\vartheta_t^\sigma)\rangle$$

we have that

$$\Delta_t^+ \ge g(\varepsilon)\cdot \mathbb{I}\left\{\vartheta_t^+ \ge \tfrac{\pi}{4}\right\}\ , \qquad \Delta_t^- \ge g(\varepsilon)\cdot \mathbb{I}\left\{\vartheta_t^- < \tfrac{\pi}{4}\right\}.$$

Step 3: change-of-measure inequality. Consider the events

$$A_t \coloneqq \left\{(u_1,k_1,\dots,u_{T-1},k_{T-1},u_T,k_T) \in \mathcal{H}_T \times \{0,1,2\} \mid \alpha_t(u_1,k_1,\dots,u_{t-1},k_{t-1},u_t) < \frac{\pi}{4}\right\}\ ,$$

for $t = 1, 2, \dots, T$. Recall that for any two probability measures μ and ν over some sigma algebra $\mathcal{G}$ and any event $G \in \mathcal{G}$, we have that

$$|\mu[G] - \nu[G]| \le \|\mu - \nu\|_{\mathrm{TV}}\ ,$$

and hence, if we define the two push-forward probability measures on the sample space $\mathcal{H}_T \times \{0,1,2\}$ as $\mathbb{P}_+ \coloneqq \mathbb{P}_{(H_T^+, O_T^+)}$ and $\mathbb{P}_- \coloneqq \mathbb{P}_{(H_T^-, O_T^-)}$, we have that

$$\mathbb{P}_+[A_t^c] + \mathbb{P}_-[A_t] = 1 - (\mathbb{P}_+[A_t] - \mathbb{P}_-[A_t]) \ge 1 - \left\|\mathbb{P}_{(H_T^+, O_T^+)} - \mathbb{P}_{(H_T^-, O_T^-)}\right\|_{\mathrm{TV}}\ .$$

Combining with Step 2, summing over t, using Pinsker inequality [Tsybakov, 2008], and Lemma 8, we get

$$\begin{aligned}
R_T^+ + R_T^- &= \sum_{t=1}^T \mathbb{E}[\Delta_t^+] + \sum_{t=1}^T \mathbb{E}[\Delta_t^-] \ge g(\varepsilon)\cdot \sum_{t=1}^T \left(\mathbb{P}_+[A_t^c] + \mathbb{P}_-[A_t]\right) \\
&\ge \frac{64\sqrt{2}}{65}\cdot\varepsilon^2\cdot\sum_{t=1}^T \left(\mathbb{P}_+[A_t^c] + \mathbb{P}_-[A_t]\right) \\
&\ge \frac{64\sqrt{2}}{65}\cdot\varepsilon^2\cdot T\cdot\left(1 - \left\|\mathbb{P}_{(H_T^+, O_T^+)} - \mathbb{P}_{(H_T^-, O_T^-)}\right\|_{\mathrm{TV}}\right) \\
&\ge \frac{64\sqrt{2}}{65}\cdot\varepsilon^2\cdot T\cdot\left(1 - \sqrt{\tfrac{1}{2}\cdot \mathrm{KL}\left(\mathbb{P}_{(H_T^+, O_T^+)}, \mathbb{P}_{(H_T^-, O_T^-)}\right)}\right) \\
&\ge \frac{64\sqrt{2}}{65}\cdot\varepsilon^2\cdot T\cdot\left(1 - \sqrt{42\cdot\varepsilon^2\cdot\sum_{t=1}^T \mathbb{E}\left[\left|\vartheta_t^+ - \tfrac{\pi}{4}\right|^2\right]}\right)\ .
\end{aligned} \tag{48}$$

Step 4: angles to regret. Next, let's relate $\sum_{t=1}^T \mathbb{E}\left[|\vartheta_t^+ - \frac{\pi}{4}|^2\right]$ to R_T^+.

For $\vartheta \in \left[0, \frac{\pi}{2}\right]$, let $z(\vartheta) = (\cos\vartheta, \sin\vartheta)$ and $z_+^\star = z(\vartheta_+^\star) = \mu^+ / \|\mu^+\|_2$. Then

$$\Delta_t^+ = \|\mu^+\|_2 \cdot \left(1 - \langle z(\vartheta_t^+), z_+^\star\rangle\right) = \|\mu^+\|_2 \cdot (1 - \cos(\vartheta_t^+ - \vartheta_+^\star)) = 2\cdot\|\mu^+\|_2\cdot\left(\sin\left(\tfrac{\vartheta_t^+ - \vartheta_+^\star}{2}\right)\right)^2.$$

Since $\left|\sin\left(\frac{x}{2}\right)\right| \ge \frac{|x|}{\pi}$ for $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, leveraging that $|\vartheta_t^+ - \vartheta_+^\star| \le \frac{\pi}{2}$, we have

$$\Delta_t^+ \ge 2\cdot\|\mu^+\|_2\cdot\left(\frac{|\vartheta_t^+ - \vartheta_+^\star|}{\pi}\right)^2 \quad \Longrightarrow \quad |\vartheta_t^+ - \vartheta_+^\star|^2 \le \frac{\pi^2}{2\cdot\|\mu^+\|_2}\cdot\Delta_t^+\ .$$

Using $\|\mu^+\|_2 \geq \frac{\sqrt{2}}{2}$, we then get

$$|\vartheta_t^+ - \vartheta_+^\star|^2 \leq \frac{\pi^2\sqrt{2}}{2} \cdot \Delta_t^+ .$$

Also, since $\vartheta_+^\star = \frac{\pi}{4} - \arctan(2\varepsilon)$, from (47), we get

$$\left|\vartheta_+^\star - \tfrac{\pi}{4}\right| = \arctan(2\varepsilon) \leq 2\varepsilon .$$

Therefore,

$$\left|\vartheta_t^+ - \tfrac{\pi}{4}\right|^2 \leq 2\left|\vartheta_t^+ - \vartheta_+^\star\right|^2 + 2\left|\vartheta_+^\star - \tfrac{\pi}{4}\right|^2 \leq \pi^2\sqrt{2} \cdot \Delta_t^+ + 8 \cdot \varepsilon^2.$$

Summing over t and taking expectations yields

$$\sum_{t=1}^{T} \mathbb{E}\left[\left|\vartheta_t^+ - \tfrac{\pi}{4}\right|^2\right] \leq \pi^2\sqrt{2} \cdot R_T^+ + 8 \cdot \varepsilon^2 \cdot T.$$

Plugging back into (48), and recalling that $\varepsilon = \gamma \cdot T^{-1/4}$ and that $\gamma = \frac{1}{6}$, we get

$$\begin{aligned} R_T^+ + R_T^- &\geq \frac{64\sqrt{2}}{65} \cdot \varepsilon^2 \cdot T \cdot \left(1 - \sqrt{42 \cdot \varepsilon^2 \cdot \left(\pi^2\sqrt{2} \cdot R_T^+ + 8 \cdot \varepsilon^2 \cdot T\right)}\right) \\ &= \frac{64\sqrt{2}}{65} \cdot \gamma^2 \cdot \left(1 - \gamma \cdot \sqrt{42 \cdot \left(\pi^2\sqrt{2} \cdot \frac{R_T^+}{\sqrt{T}} + 8 \cdot \gamma^2\right)}\right) \cdot \sqrt{T} \\ &= \frac{16\sqrt{2}}{585} \cdot \left(1 - \frac{1}{6} \cdot \sqrt{42 \cdot \left(\pi^2\sqrt{2} \cdot \frac{R_T^+}{\sqrt{T}} + \frac{2}{9}\right)}\right) \cdot \sqrt{T}. \end{aligned} \tag{49}$$

Step 5: conclusion. Now, if $R_T^+ \geq \frac{1}{140} \cdot \sqrt{T}$ then the conclusion follows. Otherwise, $R_T^+ < \frac{1}{140} \cdot \sqrt{T}$. Then, from (49), we have that

$$R_T^- \geq \left(\frac{16\sqrt{2}}{585} \cdot \left(1 - \frac{1}{6} \cdot \sqrt{42 \cdot \left(\pi^2\sqrt{2} \cdot \frac{1}{140} + \frac{2}{9}\right)}\right) - \frac{1}{140}\right) \cdot \sqrt{T} \geq \frac{1}{140} \cdot \sqrt{T} .$$

This concludes the proof. □

G Related work

Concurrent and independent work by Montanari et al. [2026] studies a related censored resource-allocation problem with Bernoulli rewards. The models overlap when threshold distributions are known and rewards are independent Bernoulli variables. Their feedback additionally reveals the threshold whenever success occurs. In the shared setting with known threshold distributions and independent Bernoulli rewards, this observation can be simulated from our feedback. Indeed, success means that the Bernoulli reward equals one and the threshold does not exceed the allocated budget $X_{t,k}$. By independence, the conditional distribution of the threshold is therefore simply its known distribution truncated to $[0, X_{t,k}]$. Whenever $B_{t,k}(X_{t,k})G_{t,k} = 1$, we can draw a fresh sample from this truncated distribution and supply it to their algorithm together with the success indicator. On failure, no threshold value is required. The resulting feedback has the same conditional distribution as in their model. We now make a comparison between their stated results and ours in the shared setting, considering first the $\sqrt{T}$ rates. While our optimism-based analysis yields *worst-case* $K\sqrt{T}$ guarantees that hold for arbitrary[5] known nondecreasing curves, including discontinuous ones, their stated bound is instance-dependent, with a reported leading term of order $\widetilde{\mathcal{O}}(K^{3/2}\sqrt{T})$ and an additive $\mathcal{O}(K^2 \log T)$ term. Additional dependence remains through instance-specific quantities, so the displayed powers of K do not constitute a uniform comparison of the two bounds. Their Theorem 3.1 also establishes an $\Omega(\sqrt{KT})$ lower bound with known uniform threshold distributions. After normalizing the budget, these correspond to $F_k(x) = x$. On this subclass, $\sum_k F_k(x_k) \leq 1$, so our refined bound in Theorem 3 yields $\widetilde{\mathcal{O}}(\sqrt{KT})$. Our $\Omega(K\sqrt{T})$ lower bound concerns the larger

[5] Up to the existence of the argmax in Algorithm 2, see Remark 3.

class of arbitrary nondecreasing curves. We now make a comparison between our fast-rate regime and theirs. Our fast-rate analysis uses the divergence of the slope of the known budget-to-success curves at the origin to *prove* that optimal allocations are in the interior of the simplex, and then combines this structural property with optimism to obtain sufficiently many censored semi-bandit observations. By contrast, their stated polylogarithmic bound (Theorem 4.6) assumes an interior optimum, activation probabilities bounded away from zero, and a globally smooth, uniformly strongly concave surrogate satisfying parameter regularity conditions (Assumption 3). The theorem also requires the oracle to select a budget-saturating maximizer, and the algorithm uses explicit cyclic boosted exploration. Consequently, their and our fast-rate assumptions are not directly comparable: theirs do not cover singular power-law curves $x \mapsto x^{\alpha_i}$, $\alpha_i \in (0,1)$, which are central examples in our fast-rate regime, while ours do not cover finite-slope exponential CDFs such as $x \mapsto 1 - e^{-\lambda x}$, which are central in their parametric threshold setting.

Our model is also related to stochastic linear bandits; see, e.g., Lattimore and Szepesvári [2020, Chapter 19]. In the linear bandit setting, the learner chooses an action X_t and observes a scalar reward $\langle \mu, X_t \rangle + \eta_t$, where μ is an unknown parameter and η_t is a zero-mean subgaussian noise term, so that $\langle \mu, X_t \rangle$ is the expected reward. In our model, the expected reward has the analogous linear form $\langle \mu, F(X_t) \rangle$, where $X_t \in \Delta_K$ is chosen by the learner and F is known. Equivalently, the learner chooses the feature vector $F(X_t)$, and the dependence on μ is linear as in stochastic linear bandits.

The feedback, however, is fundamentally different. Linear bandits provide exactly one scalar observation per round, which is an unbiased estimate of a linear functional of μ. In contrast, under censored semi-bandit feedback our learner observes a *random number* of scalar rewards per round (one per completed task), and this number can have expectation much larger than one depending on $F_1, \dots, F_K$. Thus the feedback can be substantially more informative than in linear bandits (and the full-feedback model is richer still). This difference underlies some of our main results. In particular, our fast regret bound (Theorem 1) arises in regimes where the feedback is "rich" with high probability; see Lemma 4 in Appendix C. In Theorem 2, we show that an analogous fast rate does *not* occur in the corresponding linear bandit instance.

For our worst-case upper bound (Theorem 3), another subtle, but fundamental distinction from linear bandits arises. In our model the per-round reward lies in $[0, K]$, and we obtain worst-case regret of order $K\sqrt{T}$. In the linear bandit literature (e.g., Bartlett et al. [2008], Dani et al. [2008]), the corresponding worst-case regret scales as $K\sqrt{T}$ under a standardized reward range such as $[-1, 1]$. If one rescales that range to length $\Theta(K)$, then those bounds incur an additional factor K relative to ours. This gap intuitively reflects the different feedback: when the budget-to-success curves $F_1, \dots, F_K$ are small, both the expected reward $\langle \mu, F(X_t) \rangle$ and the effective regret range are small, whereas when they are large the regret range grows but the expected number of observed scalar rewards per round (one per completed task) can exceed one, unlike in linear bandits.

Stochastic linear bandits are traditionally formulated with X and μ lying in Euclidean balls ("(ℓ_2, ℓ_2) geometry"). More generally, regret bounds for arbitrary (ℓ_p, ℓ_q) geometries (with p and q being conjugate exponents) were obtained in Bubeck et al. [2012, 2018]. In the adversarial *bandits with experts* model of Auer et al. [2002], the geometry is (ℓ_1, ℓ_∞) as in our setting. However, the feedback there is a single coordinate θ_{t,I_t}, where θ_t is the adversarial reward vector at time t and $I_t \in [K]$ is drawn according to the chosen distribution $X_t \in \Delta_K$.

Closely related resource-allocation models were studied by Lattimore et al. [2014] and Lattimore et al. [2015]. The former consider capped-linear completion curves $F_k(x_k) = \min\{1, x_k/\nu_k\}$, where the cutoff parameters ν_k are unknown, and all completed jobs have unit value ($\mu_k = 1$). The latter extends this line of work to multiple resource types: task k receives a resource vector $M_{t,k}$ and succeeds with probability $\min\{1, \langle M_{t,k}, \nu_k \rangle\}$, where ν_k is unknown. These settings are complementary to ours. They learn the resource-to-completion map within a clipped-linear parametric family, whereas we assume that the budget-to-success curves F_k are known and learn the reward means μ_k from censored reward observations. The models overlap in special cases; for example, when there is a single resource type, $F_k(x) = x$, and completed rewards are Bernoulli with mean μ_k, the product of completion and reward has mean $x\mu_k$, matching their clipped-linear form after coarsening our feedback. However, our model allows general known nondecreasing curves and observes completions separately from rewards. Accordingly, the mechanisms and techniques are different: their analyses rely on learning clipped-linear completion parameters, using weighted least squares in the multi-

resource case, whereas our fast-rate result comes from curvature, interior optimality, and censored observations of task rewards.

Another related line of work studies online resource allocation with diminishing returns under first-order feedback. Fontaine et al. [2020] consider repeated allocation over K resources with unknown concave, nondecreasing reward functions and noisy gradient observations, obtaining adaptive rates depending on Łojasiewicz-type regularity. Our fast-rate regime is related through the separable concave objective over the simplex, but differs in the information structure: in our model F_k describes a known allocation-dependent completion probability, while the reward associated with a completed task has unknown mean μ_k and is observed only upon completion. Thus, the effective reward function is $\mu_k F_k(x_k)$, but it is learned through censored semi-bandit observations rather than direct noisy gradients. In smooth interior regimes, these observations can be importance-weighted to estimate gradients, e.g., via $\frac{F_k'(x_k)}{F_k(x_k)} B_{t,k}(x_k) G_{t,k}$, but this reduction may have unbounded variance near the boundary, and does not apply to non-smooth or threshold curves. A related setting has been investigated by Thananjeyan et al. [2021] for best arm identification. They assume the player has access to a divisible unit resource which is used to pull the arms in each round. The time taken to execute a single pull using a fraction α of this resource is $\lambda\big(\frac{1}{\alpha}\big)$, where λ is a monotone increasing known function. For example, the player can pull m arms in parallel in time $\lambda(m)$. Their goal is to probabilistically identify the best arm while minimizing the total time. Another related model is combinatorial bandits with probabilistically triggered arms [Chen et al., 2016]. Finite discretizations of our action space Δ_K can be encoded in this framework: each allocation x acts as a super-action that triggers base arm k with probability $F_k(x_k)$ and, upon triggering, reveals a reward sample with mean μ_k. A difference is that in CMAB-T the triggering probabilities may be unknown and are part of the stochastic triggering mechanism, whereas in our setting the curves F_k are known. This makes the optimistic oracle in our problem explicit, but existing CMAB-T regret guarantees still do not directly imply our results. Their distribution-dependent bounds are finite-action and gap-based, while their distribution-independent bounds retain dependence on triggering-probability and smoothness parameters. In our continuous model, the probabilities $F_k(x_k)$ can become arbitrarily small near the boundary, and reducing Δ_K to a finite grid would require additional regularity assumptions on the curves F_k; even under Lipschitzness, such a discretization suffers from the usual curse of dimensionality. Our analysis instead works directly over the continuous simplex and exploits the coupling between allocation, completion probability, and reward observability. Our model is also reminiscent of the binary-classification-with-paid-experts setting studied in Van Der Hoeven et al. [2023] (see also Zhang et al. 2015 for earlier work). That line considers payments $x \in [0,1]^K$ and unknown Lipschitz functions $F_1, \dots, F_K$ mapping them to accuracies $(\varepsilon_1, \dots, \varepsilon_K)$ via $\varepsilon_k = F_k(x_k)$, where ε_k is the probability that expert k correctly predicts the t-th bit of an unknown bit sequence. The resulting objective is not of the form $\langle \mu, F(x) \rangle$, so their guarantees do not directly apply to our problem.

Prior bandit formulations for crowdsourcing [Tran-Thanh et al., 2014], multi-platform autobidding [Aggarwal et al., 2025], and multi-channel advertising [Avadhanula et al., 2021, Deng et al., 2023] typically impose *global* (across-round) budget constraints. By contrast, we consider *per-round* budget constraints; replacing them with knapsack-style constraints changes the nature of the problem and makes it closer to constrained online optimization with bandit feedback. Moreover, per-round constraints better capture some applications. For example, in cost-efficient inference for LLMs [Li, 2025, Nguyen et al., 2024], one selects which model to query for each input prompt by trading off generation cost against output quality on a per-prompt basis. In crowdsourcing, stochastic models of acceptance probability and reliability were investigated by ul Hassan and Curry [2016] without providing performance guarantees. A setting closer to ours is studied in Kang and Tay [2020], where $F_1, \dots, F_K$ are restricted to be linear. In that work, these linear functions are known only for some $k \in [K]$, and the learner cannot observe whether tasks with unknown F_k complete successfully. They prove a $\sqrt{T}$-type regret bound for a notion of reward different from ours.

Regarding censored feedback in the literature, the reference Agarwal et al. [2024] addresses a semi-bandit setting, that can be censored, similarly as in our work. Otherwise, the setting there is different from ours, since it is based on a decision tree, and the feedback depends on which path on the tree is sampled/selected, but not on additional random variables that would play the role of our variables $B_{t,k}(x_k)$. Also Van Ryzin and McGill [2000], Huh and Rusmevichientong [2006] study revenue management with censored demand, again with settings different from ours. Related *threshold-based* resource-allocation models with censored semi-bandit feedback were studied by

Verma et al. [2019], but their censoring mechanism differs from ours: in their model allocating enough resource suppresses loss and censors the observation, whereas in our model allocating more budget increases the probability of task completion and hence reward observation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theorems correspond to all claims (with clear statements and proofs).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a section "Conclusions, limitations and future work", where we discuss variants of our setting that we do not capture.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All claims are clearly stated (including assumptions) and all proofs are present.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: There are no experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include any dataset or code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: There are no experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: There are no experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: There are no experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work is theoretical, and thus not impacted by data-related and human-related concerns. For societal impact, we refer to the next question.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper is theoretical and focuses on regret guarantees for online budget allocation under censored semi-bandit feedback. It motivates the model through applications such as online platforms and crowdsourcing, but it does not include a separate discussion of positive and negative societal impacts. Any societal impact would depend on future deployment choices and application domains, which are outside the scope of the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Being theoretical work, we are not releasing any data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We do not use any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not introduce any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Large language models were used as an auxiliary tool for algebraic manipulation, checking intermediate computations, and drafting or proofreading portions of some theorem proofs. All mathematical statements, proof steps, and final text were reviewed and verified by the authors, who take full responsibility for the content. No LLM-generated text, computation, or citation was accepted without author inspection.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.