Bandits with Knapsacks beyond the Worst Case

Karthik Abinav Sankararaman
Facebook AI, Menlo Park
karthikabinavs@gmail.com

Aleksandrs Slivkins
Microsoft Research, New York City
slivkins@microsoft.com

Abstract

Bandits with Knapsacks (BwK) is a general model for multi-armed bandits under supply/budget constraints. While worst-case regret bounds for BwK are well-understood, we present three results that go beyond the worst-case perspective. First, we provide upper and lower bounds which amount to a *full characterization* for logarithmic, instance-dependent regret rates. Second, we consider "simple regret" in BwK, which tracks algorithm's performance in a given round, and prove that it is small in all but a few rounds. Third, we provide a general "reduction" from BwK to bandits which takes advantage of some known helpful structure, and apply this reduction to combinatorial semi-bandits, linear contextual bandits, and multinomial-logit bandits. Our results build on the BwK algorithm from Agrawal and Devanur [3], providing new analyses thereof.

1 Introduction

We study multi-armed bandit problems with supply or budget constraints. Multi-armed bandits is a simple model for *exploration-exploitation tradeoff*, *i.e.,* the tension between acquiring new information and making optimal decisions. It is an active research area, spanning computer science, operations research, and economics. Supply/budget constraints arise in many realistic applications, *e.g.,* a seller who dynamically adjusts the prices or product assortment may have a limited inventory, and an algorithm that optimizes ad placement is constrained by the advertisers' budgets. Other motivating examples concern repeated auctions, crowdsourcing markets, and network routing.

We consider a general model called *Bandits with Knapsacks* (*BwK*), which subsumes the examples mentioned above. There are $d \geq 2$ *resources* that are consumed over time, one of which is time itself. Each resource i starts out with budget B_i. In each round t, the algorithm chooses an action (*arm*) $a = a_t$ from a fixed set of K actions. The outcome is a vector in $[0,1]^{d+1}$: it consists of a reward and consumption of each resource. This vector is drawn independently from some distribution over $[0,1]^{d+1}$, which depends on the chosen arm but not on the round, and is not known to the algorithm. The algorithm observes *bandit feedback*, *i.e.,* only the outcome of the chosen arm. The algorithm stops at a known time horizon T, or when the total consumption of some resource exceeds its budget. The goal is to maximize the total reward, denoted REW.

The presence of supply/budget constraints makes the problem much more challenging. First, algorithm's choices constrain what it can do in the future. Second, the algorithm is no longer looking for arms with maximal expected per-round reward (because such arms may consume too much resources). Third, the best fixed distribution over arms can be much better than the best fixed arm. Accordingly, we compete with the *best fixed distribution* benchmark: the total expected reward of the best distribution, denoted $\texttt{OPT}_{\texttt{FD}}$. All this complexity is already present even when $d = 2$, *i.e.,* when there is only one resource other than time, and the minimal budget is $B = \min_i B_i = \Omega(T)$.

BwK were introduced in [14, 16] and extensively studied since then. The optimal worst-case regret rate is well-understood. In particular, it is $\tilde{\mathcal{O}}(\sqrt{KT})$ when $B = \Omega(T)$.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

We present several results that go beyond the worst-case perspective:

1. We provide a full characterization for instance-dependent regret rates. In stochastic bandits, one obtains regret $\mathcal{O}\left(\frac{K}{\Delta}\log T\right)$, where Δ is the the *reward-gap*: the gap in expected reward between the best and the second-best arm. We work out whether, when and how such results extend to BwK.

2. We show that *simple regret*, which tracks algorithm's performance in a given round, can be small in all but a few rounds. Like in stochastic bandits, simple regret can be at least ϵ in at most $\tilde{\mathcal{O}}(K/\epsilon^2)$ rounds, and this is achieved for all $\epsilon > 0$ simultaneously.

3. We improve all results mentioned above for a large number of arms, assuming some helpful structure. In fact, we provide a general "reduction" from BwK to stochastic bandits, and apply this reduction to three well-studied scenarios from stochastic bandits.

Our algorithmic results focus on UcbBwK, a BwK algorithm from [3] which implements the "optimism under uncertainty" paradigm and attains the optimal worst-case regret bound. We provide new analyses of this algorithm along the above-mentioned themes.

Related work. Background on multi-armed bandits can be found in books [23, 54, 42]. *Stochastic bandits* (*i.e.,* BwK without resources) is a basic, well-understood version. The dependence on Δ and ϵ are optimal as stated above [41, 10, 11], and is achieved simultaneously with the optimal worst-case regret $\widetilde{O}(KT)$, *e.g.,* in [10]. Various refinements are known for $O(\log T)$ regret [10, 8, 34, 32, 45]. Most relevant to this paper is $\mathcal{O}\left(\sum_a \log(T)/\Delta(a)\right)$ regret, where $\Delta(a)$ is the gap in expected reward between arm a and the best arm [10]. Improving regret for large / infinite number of arms via a helpful structure is a unifying theme for several prominent lines of work, *e.g.,* linear bandits, convex bandits, Lipschitz bandits, and combinatorial (semi-)bandits.

Bandits with Knapsacks were introduced in [14, 16], and optimally solved in the worst case. Subsequent work extended BwK to a more general notion of rewards/consumptions [3], combinatorial semi-bandits [49], and contextual bandits [15, 6, 4]. Several special cases with budget/supply constraints were studied separately (and inspired a generalization to BwK): dynamic pricing [19, 12, 20, 59], dynamic procurement [13, 52], and dynamic ad allocation [53, 28]. The adversarial version of BwK was studied by [35, 36]. All this work considers worst-case regret bounds.

Several papers achieve $O(\log T)$ regret in BwK, but with substantial caveats that we avoid. [61] assume deterministic consumption, whereas all motivating examples of BwK require stochastic consumption correlated with rewards (*e.g.,* dynamic pricing consumes supply only if a sale happens). They posit $d = 2$ and no other assumptions, whereas we show that "best-arm optimality" is necessary with stochastic consumption. [31] assume "best-arm-optimality" as we do (it is implicit in their version of reward-gap). However, their algorithm inputs an instance-dependent parameter which is "hidden" in BwK. Moreover, their $O(\log T)$ regret bound scales with $c_{\min}$, minimal expected consumption among arms (as $c_{\min}^{-4}$). Their worst-case regret bound is suboptimal, since it also scales with $c_{\min}$ (as $c_{\min}^{-2}$), and only applies for $d = 2$. [58] study a contextual version of BwK with two arms, one of which does nothing; this is meaningless when specialized to BwK. [44], subsequent to our initial draft on arxiv.org, use extra parameters (other than a version of reward-gap), which yield $\geq \sqrt{T}$ regret whenever our lower bounds apply;[1] it is unclear when all their parameters are small. No worst-case regret bounds are provided; their algorithm does not appear to achieve even $o(T)$ regret in the worst case. Finally, [33, 56, 57, 30, 47] posit one constrained resource and $T = \infty$. This is an easier problem, *e.g.,* the best arm is the best distribution over arms.

2 Preliminaries: the problem, linear relaxation and UcbBwK algorithm

The bandits with knapsacks (BwK) problem is as follows. There are K arms, d resources, and T rounds. Initially, each resource $j \in [d]$ is endowed with budget B_j. In each round $t = 1, \ldots, T$, an algorithm chooses an arm a_t, and observes an outcome vector $\boldsymbol{o}_t = (r_t;\ c_{1,t}, \ldots, c_{d,t}) \in [0,1]^{d+1}$, where r_t is the reward, and $c_{j,t}$ is the consumption of each resource j. The algorithm stops when the consumption of some resource j exceeds its budget B_j, or after T rounds, whichever is sooner. We maximize the total reward, $\texttt{REW} = \sum_{t=1}^{\tau} r_t$, where τ is the stopping time. We focus on the stochastic version: for each arm a, there is a distribution $\mathcal{D}_a$ over $[0,1]^{d+1}$ such that each outcome vector $\boldsymbol{o}_t$ is

[1] Conceptually, our assumption of "best-arm-optimality" is replaced with another assumption: a lower bound on the positive entries of the optimal distribution x^* (parameter χ in Section 3.3 of [44]).

an independent draw from distribution $\mathcal{D}_{a_t}$ (which depends only on the chosen arm a_t). A problem instance consists of parameters $(K, d, T;\ B_1, \ldots, B_d)$ and distributions $(\mathcal{D}_a : \text{arms } a)$.

Given a problem instance, the *best dynamic policy* benchmark $\texttt{OPT}_{\texttt{DP}}$ maximizes the total expected reward over all algorithms; it is used in all worst-case regret bounds. The *best fixed distribution* benchmark $\texttt{OPT}_{\texttt{FD}}$, used in some of our results, maximizes the total expected reward over all algorithms that always sample an arm from the same distribution. The worst-case optimal regret rate is [16]:

$$\texttt{OPT}_{\texttt{DP}} - \mathbb{E}[\texttt{REW}] = \tilde{\mathcal{O}}\left(\sqrt{K\ \texttt{OPT}_{\texttt{DP}}} + \texttt{OPT}_{\texttt{DP}} \sqrt{K/B} \right), \quad B = \min_{j\in[d]} B_j. \tag{2.1}$$

Simplifications and notation. Following prior work, we make three assumptions without losing generality. First, all budgets are the same: $B_1 = \ldots = B_d = B$. This is w.l.o.g. because one can divide the consumption of each resource j by $B_j / \min_i B_i$; dependence on the budgets is driven by the smallest B_j. Second, resource d corresponds to time: each arm deterministically consumes B/T units of this resource in each round. It is called the *time resource* and denoted `time`. Third, there is a *null arm*, denoted `null`, whose reward and consumption of all resources except `time` is always 0.[2]

Like most prior work on `BwK`, we use $\mathcal{O}(\cdot)$ notation rather than track explicit constants in regret bounds. This improves clarity and emphasizes the more essential aspects of analyses and results.

For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$ and $\Delta_n = \{\text{all distributions on } [n]\}$. Let $[K]$ and $[d]$ be, resp., the set of all arms and the set of all resources. For each arm a, let $r(a)$ and $c_j(a)$ be, resp., the mean reward and mean resource-j consumption, *i.e.,* $(r(a); c_1(a), \ldots, c_d(a)) := \mathbb{E}_{\boldsymbol{o}\sim\mathcal{D}_a}[\boldsymbol{o}]$. We sometimes write $\boldsymbol{r} = (r(a) : a \in [K])$ and $\boldsymbol{c}_j = (c_j(a) : a \in [K])$ as vectors over arms. Given a function $f : [K] \to \mathbb{R}$, we extend it to distributions $\boldsymbol{X}$ over arms as $f(\boldsymbol{X}) := \mathbb{E}_{a\sim\boldsymbol{X}}[f(a)]$.

Linear Relaxation. Following prior work, we consider a linear relaxation:

$$\begin{array}{lll} \text{maximize} & \boldsymbol{X}\cdot\boldsymbol{r} & \text{such that} \\ & \boldsymbol{X} \in [0,1]^K,\ \boldsymbol{X}\cdot\boldsymbol{1} = 1 & \\ \forall j \in [d] & \boldsymbol{X}\cdot\boldsymbol{c}_j \le B/T. & \end{array} \tag{2.2}$$

Here $\boldsymbol{X}$ is a distributions over arms, the algorithm does not run out of resources in expectation, and the objective is the expected per-round reward. Let $\texttt{OPT}_{\texttt{LP}}$ be the value of this linear program. Then $\texttt{OPT}_{\texttt{LP}} \ge \texttt{OPT}_{\texttt{DP}}/T \ge \texttt{OPT}_{\texttt{FD}}/T$ [16]. The Lagrange function $\mathcal{L} : \Delta_K \times \mathbb{R}^d_+ \to \mathbb{R}$ defined as follows:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\lambda}) := r(\boldsymbol{X}) + \textstyle\sum_{j\in[d]} \lambda_j [\ 1 - {}^{T}\!/_{B}\ c_j(\boldsymbol{X}),\ ]. \tag{2.3}$$

where $\boldsymbol{\lambda}$ corresponds to the dual variables. Then (*e.g.,* by Theorem D.2.2 in [17]):

$$\min_{\boldsymbol{\lambda}\ge 0} \max_{\boldsymbol{X}\in\Delta_K} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\lambda}) = \max_{\boldsymbol{X}\in\Delta_K} \min_{\boldsymbol{\lambda}\ge 0} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\lambda}) = \texttt{OPT}_{\texttt{LP}}. \tag{2.4}$$

The min and max in (2.4) are attained, so that $(\boldsymbol{X}^*, \boldsymbol{\lambda}^*)$ is maximin pair if and only if it is minimax pair; such pair is called a *saddle point*. We'll use $\mathcal{L}(\cdot, \boldsymbol{\lambda}^*)$ to generalize reward-gap to `BwK`.

Algorithm `UcbBwK`. We analyze an algorithm from [3], defined as follows. In the LP (2.2), rescale the last constraint, for each resource $j \neq \texttt{time}$, as $({}^{B}\!/_{T})(1 - \eta_{\text{LP}})$, where

$$\eta_{\text{LP}} := 3 \cdot \left(\sqrt{{}^{K}\!/_{B}\ \log(KdT)} + {}^{K}\!/_{B}\ (\log(KdT))^2 \right). \tag{2.5}$$

We call it the *rescaled LP* (see (C.1)). Its value is $(1 - \eta_{\text{LP}})\ \texttt{OPT}_{\texttt{LP}}$. At each round t, the algorithm forms an "optimistic" version of this LP, upper-bounding rewards and lower-bounding consumption:

$$\begin{array}{lll} \text{maximize} & \sum_{a\in[K]} X(a)\ r_t^+(a) & \text{such that} \\ & \boldsymbol{X} \in [0,1]^K,\quad \sum_{a\in[K]} X(a) = 1 & \\ \forall j \in [d] & \sum_{a\in[K]} X(a)\ c_{j,t}^-(a) \le B(1-\eta_{\text{LP}})/T. & \end{array} \tag{2.6}$$

`UcbBwK` solves (2.6), obtains distribution $\boldsymbol{X}_t$, and samples an arm a_t independently from $\boldsymbol{X}_t$. The algorithm achieves the worst-case optimal regret bound in (2.1). The upper/lower confidence bounds $r_t^+(a),\ c_{j,t}^-(a) \in [0,1]$ are computed in a particular way specified in Appendix B. What matters to this paper is that they satisfy a high-probability event

$$0 \le r_t^+(a) - r(a) \le \text{Rad}_t(a) \text{ and } 0 \le c_j(a) - c_{j,t}^-(a) \le \text{Rad}_t(a), \tag{2.7}$$

[2] Choosing the null arm is equivalent to skipping a round. One can take an algorithm `ALG` that uses `null`, and turn it into an algorithm that doesn't: when `ALG` chooses `null`, just call it again until it doesn't.

for some *confidence radius* $\text{Rad}_t(a)$ specified below. This event holds, simultaneously for all arms a, resources j and rounds t, with probability (say) at least $1 - \frac{\log(KdT)}{T^4}$. For $a \neq \texttt{null}$, we can take

$$\text{Rad}_t(a) = \min(\ 1,\ \sqrt{C_{\texttt{rad}}/N_t(a)} + C_{\texttt{rad}}/N_t(a)\ ), \tag{2.8}$$

where $C_{\texttt{rad}} = 3 \cdot \log(KdT)$ and $N_t(a)$ is the number of rounds before t in which arm a has been chosen. There is no uncertainty on the time resource and the null arm, so we define $c^-_{\texttt{time},\,t}(\cdot) = B/T$ and $\text{Rad}_t(\texttt{null}) = r^+_t(\texttt{null}) = c^-_{j,t}(\texttt{null}) = 0$ for all resources $j \neq \texttt{time}$.

3 Logarithmic instance-dependent regret bounds

We provide upper and lower bounds which amount to *full characterization* of logarithmic, instance-dependent regret rates in BwK. We achieve $O(\log T)$ regret under two assumptions: there is only one resource other than time (*i.e.,* $d = 2$), and the best distribution over arms reduces to the best fixed arm (*best-arm-optimality*). We prove that both assumptions are essentially necessary for any algorithm, deriving complementary $\Omega(\sqrt{T})$ lower bounds if either assumption fails. Both lower bounds hold in a wide range of problem instances; arguably, they represent typical scenarios rather than exceptions.

We achieve $O(\log T)$ regret with UcbBwK algorithm [3], which implies two very desirable properties: the algorithm does not know in advance whether best-arm-optimality holds, and attains the optimal worst-case regret bound for all instances, best-arm-optimal or not. The positive result would have been weaker without either property, although still non-trivial.

We identify a suitable instance-dependent parameter, defined via Lagrangians from Eq. (2.3):

$$G_{\texttt{LAG}}(a) := \texttt{OPT}_{\texttt{LP}} - \mathcal{L}(a, \boldsymbol{\lambda}^*) \qquad \textit{(Lagrangian gap of arm } a\textit{)}, \tag{3.1}$$

where $\boldsymbol{\lambda}^*$ is a minimizer in Eq. (2.4). It is a non-obvious generalization of the *reward-gap* from multi-armed bandits, $\Delta(a) = \max_{a'} r(a') - r(a)$. The Lagrangian gap of a problem instance is

$$G_{\texttt{LAG}} := \min_{a \notin \{a^*, \texttt{null}\}} G_{\texttt{LAG}}(a). \tag{3.2}$$

Our regret bound scales as $\mathcal{O}(K G_{\texttt{LAG}}^{-1} \log T)$, which is optimal in $G_{\texttt{LAG}}$, under a mild additional assumption, and as $\mathcal{O}(K G_{\texttt{LAG}}^{-2} \log T)$ otherwise.

We use the best fixed distribution benchmark ($\texttt{OPT}_{\texttt{FD}}$) for our upper and lower bounds. The best dynamic policy benchmark ($\texttt{OPT}_{\texttt{DP}}$) is *too strong* for logarithmic regret. Indeed, a lower bound from [31, Lemma 3] shows that $\sqrt{T}$ regret is broadly unavoidable against $\texttt{OPT}_{\texttt{DP}}$, as long as resource consumption is stochastic. Interestingly, the distinction between $\texttt{OPT}_{\texttt{FD}}$ and $\texttt{OPT}_{\texttt{DP}}$ is unimportant to the worst-case regret analyses, as $\texttt{OPT}_{\texttt{DP}} - \texttt{OPT}_{\texttt{FD}} \leq \tilde{O}(\sqrt{KT})$.

3.1 $O(\log T)$ regret analysis for UcbBwK

We analyze a version of UcbBwK which "prunes out" the null arm, call it PrunedUcbBwK. (This modification can only improve regret, so it retains the worst-case regret (2.1) of UcbBwK.) We provide a new analysis of this algorithm for $d = 2$ and best-arm-optimality. We analyze the sensitivity of the "optimistic" linear relaxation to small perturbations in the coefficients, and prove that the best arm is chosen in all but a few rounds. The key is to connect each arm's confidence term with its Lagrangian gap. This gives us $\mathcal{O}(K G_{\texttt{LAG}}^{-2} \log T)$ regret rate. To improve it to $\mathcal{O}(K G_{\texttt{LAG}}^{-1} \log T)$, we use a careful counting argument which accounts for rewards and consumption of non-optimal arms.

Algorithm PrunedUcbBwK is formally defined as follows: in each round t, call UcbBwK as an oracle, repeat until it chooses a non-null arm a, and set $a_t = a$. (In one "oracle call", UcbBwK outputs an arm and inputs an outcome vector for this arm.) The total number of oracle calls is capped at $N_{\texttt{max}} = \alpha_0 \cdot T^2 \log T$, with a sufficiently large absolute constant α_0 which we specify later in Claim 3.6. Formally, after this many oracle calls the algorithm can only choose the null arm.

Definition 3.1. *An instance of BwK is called* best-arm-optimal *with best arm $a^* \in [K]$ if the following conditions hold: (i) $\texttt{OPT}_{\texttt{LP}} = \frac{B}{T} \cdot r(a^*) / \max_{j \in [d]} c_j(a^*)$, (ii) the linear program (2.2) has a unique optimal solution $\boldsymbol{X}^*$ supported on $\{a^*, \texttt{null}\}$, and (iii) $X^*(a^*) > \frac{3\sqrt{B}\log(KdT)}{T}$.*

Part (ii) here is essentially w.l.o.g.;[3] part (iii) states that the optimal value should not be tiny.

We assume $d = 2$ and best-arm-optimality throughout this section without further mention. In particular, the linear program (2.2) has a unique optimal solution $\boldsymbol{X}^*$, and its support has only one arm $a^* \neq \texttt{null}$. We use $c(a)$ to denote the mean consumption of the non-time resource on arm a. We distinguish two cases, depending on whether $c(a^*)$ is very close to B/T.

Theorem 3.2. *Fix a best-arm optimal problem instance with only one resource other than time* (i.e., $d = 2$)*. Consider Algorithm* PrunedUcbBwK *with parameter $\eta_{\text{LP}} \leq \frac{1}{2}$ in (2.5). Then*

(i) $\texttt{OPT}_{\texttt{FD}} - \mathbb{E}[\texttt{REW}] \leq \mathcal{O}\left(\frac{\texttt{OPT}_{\texttt{FD}}}{B} \cdot \Psi\right)$, *where* $\Psi := \sum_{a \notin \{a^*, \texttt{null}\}} G^{-2}_{\texttt{LAG}}(a) \cdot \log(KdT)$.

(ii) Moreover, if $|c(a^*) - B/T| > \Omega(\Psi/T)$, *then*

$$\texttt{OPT}_{\texttt{FD}} - \mathbb{E}[\texttt{REW}] \leq \mathcal{O}\left(\sum_{a \notin \{a^*, \texttt{null}\}} G^{-1}_{\texttt{LAG}}(a)\ \log(KdT)\right). \tag{3.3}$$

Eq. (3.3) optimally depends on $G_{\texttt{LAG}}(\cdot)$: indeed, it does in the unconstrained case when Lagrangian gap specializes to the reward gap, as per the lower bound in [41]. In particular, Eq. (3.3) holds if $G_{\texttt{LAG}} > T^{-1/4}$ and $|c(a^*) - B/T| > \mathcal{O}(T^{-1/2})$. The constant in $\mathcal{O}(\cdot)$ is 48 in both parts of the theorem; the analysis only suppresses constants from concentration bounds and from Lemma 3.3.

3.1.1 Basic analysis: proof of Theorem 3.2(i)

We analyze UcbBwK in a relaxed version of BwK, where an algorithm runs for exactly $N_{\max}$ rounds, regardless of the time horizon and the resource consumption; call it *Relaxed BwK*. The algorithms are still parameterized by the original B, T, and observe the resource consumption.

We sometimes condition on the high-probability event that (2.7) holds for all rounds $t \in [N_{\max}]$, call it the "clean event". Recall that its probability is at least $1 - \frac{\mathcal{O}(\log(KdT))}{T^2}$.

We prove that the best arm a^* chosen in all but a few rounds. The crux is an argument about sensitivity of linear programs to perturbations. More specifically, we argue about sensitivity of the support of the optimal solution for the linear relaxation (2.2).

Lemma 3.3 (LP-sensitivity)**.** *Consider an execution of* UcbBwK *in Relaxed BwK. Under the "clean event",* $\text{Rad}_t(a) \geq \frac{1}{4} G_{\texttt{LAG}}(a)$ *for each round t and each arm $a \in \text{supp}(\boldsymbol{X}_t) \setminus \{a^*, \texttt{null}\}$.*

Proof Sketch We use a standard result about LP-sensitivity, the details are spelled out in Appendix C. We apply this result via the following considerations. We treat the optimistic LP (2.6) a perturbation of (the rescaled version of) the original LP (2.2). We rely on perturbations being "optimistic" (*i.e.,* upper-bounding rewards and lower-bounding resource consumption). We use the clean event to upper-bound the perturbation size by the confidence radius. Finally, we prove that

$$G_{\texttt{LAG}}(a) = \frac{T}{B} \sum_{j \in [d]}\ \lambda^*_j c_j(a) - r(a), \tag{3.4}$$

and use this characterization to connect Lagrangian gap to the allowed perturbation size. ■

We rely on the following fact which easily follows from the definition of the confidence radius:

Claim 3.4. *Consider an execution of some algorithm in Relaxed BwK. Fix a threshold $\theta > 0$. Then each arm $a \neq \texttt{null}$ can only be chosen in at most $\mathcal{O}\left(\theta^{-2} \log(KdT)\right)$ rounds t with $\text{Rad}_t(a) \geq \theta$.*

Corollary 3.5. *Consider an execution of* UcbBwK *in Relaxed BwK. Under the clean event, each arm $a \notin \{a^*, \texttt{null}\}$ is chosen in at most $N_0(a) := \mathcal{O}\left(G^{-2}_{\texttt{LAG}}(a)\ \log(KdT)\right)$ rounds.*

This follows from Lemma 3.3 and Claim 3.4. Next, the null arm is not chosen too often:

Claim 3.6. *Consider an execution of* UcbBwK *in Relaxed BwK. With probability at least $1 - \mathcal{O}(T^{-3})$, the following happens: the null arm cannot be chosen in any $\alpha_0\, T\, \log(T)$ consecutive rounds, for a large enough absolute constant α_0. Consequently, a non-null arm is chosen in at least T rounds.*

[3] Part (ii) holds almost surely given part (i) if one adds a tiny noise, *e.g.,* ϵ-variance, mean-0 Gaussian for any $\epsilon > 0$, independently to each coefficient in the LP (2.2), as per Prop. 3.1 in [46]. To implement this, an algorithm can precompute the noise terms and add them consistently to observed rewards and consumptions.

Proof Sketch Fix round t, and suppose UcbBwK chooses the null arm in N consecutive rounds, starting from t. No new data is added, so the optimistic LP stays the same throughout. Consequently, the solution $\boldsymbol{X}_t$ stays the same, too. Thus, we have N consecutive independent draws from $\boldsymbol{X}_t$ that return null. It follows that $r(\boldsymbol{X}_t) < 1/T$ with high probability, *e.g.,* by (B.2). On the other hand, assume the clean event. Then $r(\boldsymbol{X}_t) \geq (1-\eta_{\texttt{LP}})\,\texttt{OPT}_{\texttt{LP}}$ by definition of the optimistic LP, and consequently $r(\boldsymbol{X}_t) \geq (1-\eta_{\texttt{LP}})\,\texttt{OPT}_{\texttt{DP}}/T$. We obtain a contradiction. ■

Corollary 3.5 and Claim 3.6 imply a strong statement about the pruned algorithm.

Claim 3.7. *Consider an execution of PrunedUcbBwK in the (original) BwK problem. With probability at least $1-\mathcal{O}(T^{-2})$, each arm $a \notin \{a^*, \texttt{null}\}$ is chosen in at most $N_0(a)$ rounds, and arm a^* is chosen in $T - N_0$ remaining rounds, $N_0 := \sum_{a \notin \{a^*,\texttt{null}\}} N_0(a)$.*

We take a very pessimistic approach to obtain Theorem 3.2(i): we only rely on rewards collected by arm a^*, and we treat suboptimal arms as if they bring no reward and consume the maximal possible amount of resource. We formalize this idea as follows (see Appendix D for details).

For a given arm a, let $\texttt{REW}(a)$ be the total reward collected by arm a in PrunedUcbBwK. Let $\texttt{REW}(a \mid B_0, T_0)$ be the total reward of an algorithm that always plays arm a if the budget and the time horizon are changed to $B_0 \leq B$ and $T_0 \leq T$, respectively. Note that

$$\texttt{LP}(a \mid B_0, T_0) := \mathbb{E}[\texttt{REW}(a \mid B_0, T_0)] = r(a) \cdot \min(\, T_0, \tfrac{B_0}{c(a)}\,). \tag{3.5}$$

is the value of always playing arm a in a linear relaxation with the same constraints. By best-arm-optimality, we have $\mathbb{E}[\texttt{REW}(a^* \mid B, T)] = \texttt{OPT}_{\texttt{FD}}$. We observe that

$$\mathbb{E}[\texttt{REW}(a^* \mid B_0, T_0)] \geq \tfrac{\min\{T_0, B_0\}}{B} \cdot \texttt{OPT}_{\texttt{FD}}. \tag{3.6}$$

By Claim 3.7 there are at least $B_0 = B - N_0$ units of budget and at least $T_0 = T - N_0$ rounds left for arm a^* with high probability. Consequently,

$$\mathbb{E}[\texttt{REW}] \geq \mathbb{E}[\texttt{REW}(a^*)] \geq \mathbb{E}[\texttt{REW}(a^* \mid B_0, T_0)] - \tilde{\mathcal{O}}(1/T). \tag{3.7}$$

We obtain Theorem 3.2(i) by plugging these B_0, T_0 into Eq. (3.6), and then using (3.7).

3.1.2 Tighter computation: proof of Theorem 3.2(ii)

We re-use the basic analysis via Claim 3.7, but perform the final computation more carefully so as to account for the rewards and resource consumption of the suboptimal arms.

Let's do some prep-work. First, we characterize $\texttt{REW}(a^*)$ in a more efficient way compared to Eq. (3.7). Let $B(a), T(a)$ denote, resp., the budget and time consumed by PrunedUcbBwK when playing a given arm a. We use expectations of $B(a)$ and $T(a)$, rather than lower bounds:

$$\begin{aligned}\mathbb{E}[\texttt{REW}(a)] &= r(a)\ \mathbb{E}[T(a)] = r(a)\ \tfrac{\mathbb{E}[B(a)]}{c(a)} \\ &= \texttt{LP}\left(\, a \mid \mathbb{E}[B(a)], \mathbb{E}[T(a)]\,\right) \qquad \text{for each arm } a.\end{aligned} \tag{3.8}$$

We prove Eq. (3.8) via martingale techniques, see Appendix D.5.

Second, we use a tighter version of Eq. (3.6) (see Appendix D.3): for any $B_0 \leq B$, $T_0 \leq T$

$$\texttt{LP}(a^* \mid B_0, T_0)] \geq \texttt{OPT}_{\texttt{FD}} \cdot \tfrac{B_0}{B} \;/\; \left(\max\left\{\, \tfrac{B}{T}, c(a^*)\,\right\} \cdot \max\left\{\, \tfrac{B_0}{T_0}, c(a^*)\,\right\}\right). \tag{3.9}$$

Third, we lower-bound $G_{\texttt{LAG}}(a)$ in a way that removes Lagrange multipliers λ^*:

$$G_{\texttt{LAG}}(a) \geq \begin{cases} \texttt{OPT}_{\texttt{FD}}/T - r(a) & \text{if} \quad c(a^*) < B/T, \\ \texttt{OPT}_{\texttt{FD}} \cdot c(a)/B - r(a) & \text{if} \quad c(a^*) > B/T. \end{cases} \tag{3.10}$$

We derive this from Eq. (3.4) and complementary slackness, see Appendix D.4.

Fourth, let $B_0 = \mathbb{E}[B(a^*)]$ and $T_0 = \mathbb{E}[T(a^*)]$ denote, resp., the expected budget and time consumed by arm a^*. Let $N(a) = \mathbb{E}[T(a)]$ be the expected number of pulls for each arm $a \notin \{a^*, \texttt{null}\}$. In this notation, Eq. (3.8) implies that

$$\mathbb{E}[\texttt{REW}] = \textstyle\sum_{a \notin \{a^*,\texttt{null}\}} N(a)\, r(a) + \texttt{LP}(a^* \mid B_0, T_0). \tag{3.11}$$

Now we are ready for the main computation . We consider four cases, depending on how $c(a^*)$ compares with B/T and B_0/T_0. We prove the desired regret bound when $c(a^*)$ is either larger than both or smaller than both, and we prove that it cannot lie in between. The "in-between" cases is the only place in the analysis where we use the assumption that $c(a^*)$ is close to B/T.

Case 1: $c(a^*) < \min(B/T, B_0/T_0)$. Plugging in Eq. (3.9) into Eq. (3.11) and simplifying,

$$\mathbb{E}[\texttt{REW}] \geq \sum_{a \notin \{a^*, \texttt{null}\}} N(a)\, r(a) + \texttt{OPT}_{\texttt{FD}} \cdot T_0/T. \tag{3.12}$$

Re-arranging, plugging in $T_0 = T - \sum_{a \neq a^*} N(a)$ and simplifying, we obtain

$$\begin{aligned} \texttt{OPT}_{\texttt{FD}} - \mathbb{E}[\texttt{REW}] &\leq \sum_{a \notin \{a^*, \texttt{null}\}} N(a) \left(\tfrac{\texttt{OPT}_{\texttt{FD}}}{T} - r(a) \right) \\ &\leq \sum_{a \notin \{a^*, \texttt{null}\}} N(a)\, G_{\texttt{LAG}}(a) && \textit{(by Eq. (3.10))} \\ &\leq \mathcal{O}\big(\sum_{a \notin \{a^*, \texttt{null}\}} G^{-1}_{\texttt{LAG}}(a)\ \log(KdT) \big) && \textit{(by Claim 3.7).} \end{aligned} \tag{3.13}$$

Case 2: $c(a^*) > \max(B/T, B_0/T_0)$. Plugging in Eq. (3.9) into Eq. (3.11) and simplifying,

$$\mathbb{E}[\texttt{REW}] \geq \sum_{a \notin \{a^*, \texttt{null}\}} N(a)\, r(a) + \texttt{OPT}_{\texttt{FD}} \cdot B_0/B. \tag{3.14}$$

Re-arranging, plugging in $B_0 = B - \sum_{a \neq a^*} N(a)\, c(a)$, and simplifying, we obtain

$$\begin{aligned} \texttt{OPT}_{\texttt{FD}} - \mathbb{E}[\texttt{REW}] &\leq \sum_{a \notin \{a^*, \texttt{null}\}} N(a) \left(\tfrac{\texttt{OPT}_{\texttt{FD}}}{B} \cdot c(a) - r(a) \right) \\ &\leq \sum_{a \notin \{a^*, \texttt{null}\}} N(a)\, G_{\texttt{LAG}}(a) && \textit{(by Eq. (3.10))}, \end{aligned}$$

and we are done by Claim 3.7, just like in Case 1.

Case 3: $B_0/T_0 \leq c(a^*) \leq B/T$**.** Let us write out B_0 and T_0:

$$\begin{aligned} c(a^*) \geq \frac{B_0}{T_0} &= \frac{B - \sum_{a \notin \{a^*, \texttt{null}\}} N(a)\, c(a)}{T - \sum_{a \notin \{a^*, \texttt{null}\}} N(a)} \geq \frac{B}{T} \left(1 - \frac{1}{B} \cdot \sum_{a \notin \{a^*, \texttt{null}\}} N(a) \right) \\ &\geq B/T - O(\Psi/T), \text{ where } \Psi \text{ is as in Theorem 3.2} && \textit{(by Claim 3.7).} \end{aligned}$$

Since $c(a^*) \leq B/T$, we have $0 \leq B/T - c(a^*) \leq O(\Psi/T)$ which contradicts the premise.

Case 4: $B/T \leq c(a^*) \leq B_0/T_0$**.** The argument is similar to Case 3. Writing out B_0, T_0, we have

$$c(a^*) \leq \frac{B_0}{T_0} = \frac{B - \sum_{a \notin \{a^*, \texttt{null}\}} N(a) c(a)}{T - \sum_{a \notin \{a^*, \texttt{null}\}} N(a)} \leq \frac{B}{T(1 - \frac{1}{T} \cdot \sum_{a \notin \{a^*, \texttt{null}\}} N(a))}.$$

By Claim 3.7, $c(a^*) \leq B/T\ (1 + O(\Psi/T))$. Therefore, $0 \leq c(a^*) - B/T \leq O(\Psi/T)$, contradiction.

3.2 Lower Bounds (for arbitrary algorithms)

We provide two lower bounds to complement Theorem 3.2: we argue that regret $\Omega(\sqrt{T})$ is essentially inevitable if a problem instance is far from best-arm-optimal or if there are $d > 2$ resources.

We consider problem instances with three arms $\{A_1, A_2, \texttt{null}\}$, Bernoulli rewards, and $d \geq 2$ resources, one of which is time; call them $3 \times d$ *instances*. Each lower bound constructs two similar problem instances $\mathcal{I}, \mathcal{I}'$ such that any algorithm incurs high regret on at least one of them.[4] The two instances have the same parameters T, K, d, B, and the mean reward and the mean consumption for each arm and each resource differ by at most ϵ; we call them ϵ*-perturbation* of each other.

We start with an "original" problem instance $\mathcal{I}_0$ and construct problem instances $\mathcal{I}, \mathcal{I}'$ that are small perturbations of $\mathcal{I}_0$. This is a fairly general result: unlike many bandit lower bounds that focus on a specific pair $\mathcal{I}, \mathcal{I}'$, we allow a wide range for $\mathcal{I}_0$, as per the assumption below.

Assumption 3.8. *There exists an absolute constant* $c_{\texttt{LB}} \in (0, 1/3)$ *such that:*

[4] A standard approach for lower-bounding regret in multi-armed bandits is to construct multiple problem instances. A notable exception is the celebrated $\Omega(\log T)$ lower bound in Lai and Robbins [41], which considers one (arbitrary) problem instance, but makes additional assumptions on the algorithm.

1. $r(A_i),\ c_j(A_i) \in [c_{\texttt{LB}},\ 1 - c_{\texttt{LB}}]$ *for each arm* $i \in \{1, 2\}$ *and each resource* j.
2. $r(A_2) - r(A_1) \geq c_{\texttt{LB}}$ *and* $c_j(A_2) - c_j(A_1) \geq c_{\texttt{LB}} + G_{\texttt{LAG}}$ *for every resource* $j \in [d]$.
3. $B \leq c_{\texttt{LB}} \cdot T \leq \texttt{OPT}_{\texttt{FD}}$.
4. *Lagrangian gap is not extremely small:* $G_{\texttt{LAG}} \geq c_{\texttt{LB}}/\sqrt{T}$.

For a concrete example, let us construct a family of $3 \times d$ problem instances that satisfy these assumptions. Fix some absolute constants $\epsilon, c_{\texttt{LB}} \in (0, 1/3)$ and time horizon T. The problem instance is defined as follows: budget $B = c_{\texttt{LB}}\, T$, mean rewards $r(A_1) = \frac{1-c_{\texttt{LB}}}{2}$ and $r(A_2) = 1 - c_{\texttt{LB}} - \epsilon$, mean consumptions $c(A_1) = c_{\texttt{LB}} - \epsilon$ and $c(A_2) = 2c_{\texttt{LB}}$. Parts (1-4) of Assumption 3.8 hold trivially. One can work out that $G_{\texttt{LAG}} = \epsilon$, so part (4) holds as long as $\epsilon \geq c_{\texttt{LB}}/\sqrt{T}$.

Theorem 3.9. *Posit an arbitrary time horizon T, budget B, and d resources (including time). Fix any $3 \times d$ problem instance $\mathcal{I}_0$ which satisfies Assumption 3.8. In part (a), assume that $d = 2$ and $\mathcal{I}_0$ is far from being best-arm-optimal, in the sense that*

$$\text{There exists an optimal solution } \boldsymbol{X}^* \text{ such that } X(A_1) > 2c_{\texttt{LB}}^4/\sqrt{T} \text{ and } X(A_2) \geq c_{\texttt{LB}}. \tag{3.15}$$

In part (b), assume that $d > 2$. For both parts, there exist problem instances $\mathcal{I}, \mathcal{I}'$, which are $\mathcal{O}\left(1/\sqrt{T}\right)$-perturbations of $\mathcal{I}_0$, such that

$$\text{Any algorithm incurs regret } \texttt{OPT}_{\texttt{FD}} - \mathbb{E}[\texttt{REW}] \geq \Omega\left(\, c_{\texttt{LB}}^4\, \sqrt{T}\, \right) \text{ on } \mathcal{I} \text{ or } \mathcal{I}' \tag{3.16}$$

For part (a), instance $\mathcal{I}$ has the same expected outcomes as $\mathcal{I}_0$ (but possibly different outcome distributions); we call such problem instances *mean-twins*. For part (b), one can take $\mathcal{I}_0$ to be best-arm-optimal. For both parts, the problem instances $\mathcal{I}, \mathcal{I}'$ require randomized resource consumption.

Both parts follow from a more generic lower bound which focuses on linear independence of per-resource consumption vectors $\boldsymbol{c}_j := (\, c_j(A_1),\, c_j(A_2),\, c_j(\texttt{null})\,) \in [0, 1]^3$, resources $j \in [d]$.

Theorem 3.10. *Posit an arbitrary time horizon T, budget B, and $d \geq 2$ resources (including time). Fix any $3 \times d$ problem instance $\mathcal{I}_0$ that satisfies Assumption 3.8 and Eq. (3.15). Assume that the consumption vectors $\boldsymbol{c}_j$, $j \in [d]$ are linearly independent. Then there are instances $\mathcal{I}, \mathcal{I}'$ which are ϵ-perturbations of $\mathcal{I}_0$, with $\epsilon = 2\, c_{\texttt{LB}}^2/\sqrt{T}$, which satisfy (3.16). In fact, $\mathcal{I}$ is a mean-twin of $\mathcal{I}_0$.*

Proof Sketch (see Appendix E for full proof). Let $r(a)$ and $\boldsymbol{c}(a) \in [0, 1]^d$ be, resp., the mean reward and the mean resource consumption vector for each arm a for instance $\mathcal{I}_0$. Let $\epsilon = c_{\texttt{LB}}/\sqrt{T}$.

Problem instances $\mathcal{I}, \mathcal{I}'$ are constructed as follows. For both instances, the rewards of each non-null arm $a \in \{A_1, A_2\}$ are deterministic and equal to $r(a)$. Resource consumption vector for arm A_1 is deterministic and equals $\boldsymbol{c}(A_1)$. Resource consumption vector of arm A_2 in each round t, denoted $\boldsymbol{c}_{(t)}(A_2)$, is a carefully constructed random vector whose expectation is $c(A_2)$ for instance $\mathcal{I}$, and slightly less for instance $\mathcal{I}'$. Specifically, $\boldsymbol{c}_{(t)}(A_2) = \boldsymbol{c}(A_2) \cdot W_t/(1 - c_{\texttt{LB}})$, where W_t is an independent Bernoulli random variable which correlates the consumption of all resources. We posit $\mathbb{E}[W_t] = 1 - c_{\texttt{LB}}$ for instance $\mathcal{I}$, and $\mathbb{E}[W_t] = 1 - c_{\texttt{LB}} - \epsilon$ for instance $\mathcal{I}'$.

Because of the small differences between $\mathcal{I}, \mathcal{I}'$, any algorithm will choose a sufficiently "wrong" distribution over arms sufficiently often. The assumption in Eq. (3.15) and the linear independence condition are needed to ensure that "wrong" algorithm's choices result in large regret. ■

The corollaries are obtained as follows. For Theorem 3.9(a), problem instance $\mathcal{I}_0$ trivially satisfies all preconditions in Theorem 3.10. Indeed, letting time be resource 1, the per-resource vectors are $\boldsymbol{c}_1 = (0, 0, 1)$ and $\boldsymbol{c}_2 = (\,\cdot\,,\,\cdot\,,\,0)$, hence they are linearly independent. For Theorem 3.9(b), we use some tricks from the literature to transform the original problem instance $\mathcal{I}_0$ to another instance $\widetilde{\mathcal{I}}_0$ which satisfies Eq. (3.15) and the linear independence condition. The full proof is in Section F.

4 Simple regret of UcbBwK algorithm

We define *simple regret* in a given round t as $\texttt{OPT}_{\texttt{DP}}/T - r(\boldsymbol{X}_t)$, where $\boldsymbol{X}_t$ is the distribution over arms chosen by the algorithm. The benchmark $\texttt{OPT}_{\texttt{DP}}/T$ generalizes the best-arm benchmark from stochastic bandits. If each round corresponds to a user and the reward is this user's utility, then $\texttt{OPT}_{\texttt{DP}}/T$ is the "fair share" of the total reward. We prove that with UcbBwK, all but a few users receive close to their fair share. This holds if $B > \Omega(T) \gg K$, without any other assumptions.

Theorem 4.1. *Consider* UcbBwK*. Assume* $B \geq \Omega(T)$ *and* $\eta_{\texttt{LP}} \leq \frac{1}{2}$*. With probability* $\geq 1 - O(T^{-3})$*, for each* $\epsilon > 0$*, there are at most* $N_\epsilon = \mathcal{O}\left(\frac{K}{\epsilon^2} \log KTd\right)$ *rounds* t *such that* $\texttt{OPT}_{\texttt{DP}}/T - r(\boldsymbol{X}_t) \geq \epsilon$*.*

To prove Theorem 4.1, we consider another generalization of the "reward-gap", which measures the difference in LP-value compared to $\texttt{OPT}_{\texttt{LP}}$. For distribution $\boldsymbol{X}$ over arms, the *LP-gap* of $\boldsymbol{X}$ is

$$G_{\texttt{LP}}(\boldsymbol{X}) := \texttt{OPT}_{\texttt{LP}} - V(\boldsymbol{X}), \text{ where } V(\boldsymbol{X}) := (B/T) \cdot r(\boldsymbol{X}) / \left(\max_{j\in[d]} c_j(\boldsymbol{X})\right). \tag{4.1}$$

Here, $V(\boldsymbol{X})$ is the value of $\boldsymbol{X}$ in the LP (2.2) after rescaling, so that $\texttt{OPT}_{\texttt{LP}} = \sup_{\boldsymbol{X}} V(\boldsymbol{X})$. Note that $\boldsymbol{X}$ does not need to be feasible for (2.2). It suffices to study the LP-gap because $r(\boldsymbol{X}_t) \geq V(\boldsymbol{X}_t)(1 - \eta_{\texttt{LP}})$ for each round t with high probability. This holds under the "clean event" in (2.7), because $\boldsymbol{X}_t$ being the solution to the optimistic LP implies $\max_j c_j(\boldsymbol{X}_t) \geq B/T\ (1 - \eta_{\texttt{LP}})$.

Thus, we upper-bound the number of rounds t in which $G_{\texttt{LP}}(\boldsymbol{X}_t)$ is large. We do this in two steps, focusing on the confidence radius $\text{Rad}_t(\boldsymbol{X}_t)$ as defined in (2.8). First, we upper-bound the number of rounds t with large $\text{Rad}_t(\boldsymbol{X}_t)$. A crucial argument concerns *confidence sums*:

$$\sum_{t\in S} \text{Rad}_t(a_t) \quad \text{and} \quad \sum_{t\in S} \text{Rad}_t(\boldsymbol{X}_t), \tag{4.2}$$

the sums of confidence radii over a given subset of rounds $S \subset [T]$, for, resp., actions a_t and distributions $\boldsymbol{X}_t$ chosen by the algorithm. Second, we upper-bound $G_{\texttt{LP}}(\boldsymbol{X}_t)$ in terms of $\text{Rad}_t(\boldsymbol{X}_t)$. The details are spelled out in Appendix G.

5 Reduction from BwK to stochastic bandits

We improve all regret bounds for UcbBwK algorithm, from worst-case regret to logarithmic regret to simple regret, when the problem instance has some helpful structure. In fact, we provide a general *reduction* which translates insights from stochastic bandits into results on BwK. This reduction works as follows: if prior work on a particular scenario in stochastic bandits provides an improved upper bound on the confidence sums (4.2), this improvement propagates throughout the analyses of UcbBwK. Specifically, suppose $\sum_{t\in S} \text{Rad}_t(a_t) \leq \sqrt{\beta\,|S|}$ for all algorithms, all subsets of rounds $S \subset [T]$, and some instance-dependent parameter $\beta \ll K$, then UcbBwK satisfies

(i) worst-case regret $\texttt{OPT}_{\texttt{DP}} - \mathbb{E}[\texttt{REW}] \leq O(\sqrt{\beta T})(1 + \texttt{OPT}_{\texttt{DP}}/B)$.
(ii) Theorem 3.2 holds with $\Psi = \beta\, G_{\texttt{LAG}}^{-2}$ and regret $\mathcal{O}\left(\beta\, G_{\texttt{LAG}}^{-1}\right)$ in part (ii).
(iii) Theorem 4.1 holds with $N_\epsilon = \mathcal{O}\left(\beta\, \epsilon^{-2}\right)$.

Conceptually, this works because confidence sum arguments depend only on the confidence radii, rather than the algorithm that chooses arms, and are about stochastic bandits rather than BwK. The analyses of UcbBwK in [3] and the previous sections use $\beta = K$, the number of arms. The confidence sum bound with $\beta = K$ and results (i, ii, iii) for stochastic bandits follow from the analysis in [10].

We apply this reduction to three well-studied scenarios in stochastic bandits: combinatorial semi-bandits [*e.g.,* 25, 40, 39], linear contextual bandits [*e.g.,* 9, 29, 43, 27, 2], and multinomial-logit (MNL) bandits [*e.g.,* 7, 48, 51, 24]. The confidence-sum bounds are implicit in prior work on stochastic bandits, and we immediately obtain the corresponding extensions for BwK. To put this in perspective, each scenario has lead to a separate paper on BwK [resp., 49, 5, 26], for the worst-case regret bounds alone. We essentially match the worst-case regret bounds from prior work, and obtain new bounds on logarithmic regret and simple regret.[5] The details are spelled out in Appendix H.

Another reduction from BwK to bandits, found in [35], is very different from ours. It requires a much stronger premise (a regret bound against an adaptive adversary), and only yields worst-case regret bounds. Moreover, it reuses a bandit algorithm as a subroutine, whereas ours reuses a lemma.

6 Discussion: significance and novelty

Characterizing (poly-)logarithmic regret rates is a very natural question, and we give a complete answer. The answer consists of positive and negative parts: the positive part requires substantial assumptions, and these assumptions are necessary. The positive result comes "for free" despite the assumptions: it is achieved via UcbBwK and without sacrificing the worst-case performance.

[5] However, we do not provide a generic computationally efficient implementation.

The $O(\log T)$ regret result is well-motivated on its own, even though it requires $d = 2$ and best-arm-optimality and a reasonably small K = #arms. Indeed, problems with $d = 2$ and small K arise in many motivating applications of BwK (see Appendix A), and capture the three challenges of BwK discussed in the Introduction. Moreover, best-arm-optimality is a typical, non-degenerate case. [6]

For lower bounds in terms of Lagrangian gap $G_{\texttt{LAG}}$, we rely on the $\Omega(1/G \cdot \log T)$ regret bound for bandits [41], where G is the reward-gap (since $G_{\texttt{LAG}}$ generalizes reward-gap). In particular, $1/G_{\texttt{LAG}}$ scaling is optimal. No other instance-dependent lower bounds are known for BwK. However, Theorem 3.9 implies $\Omega(\sqrt{T})$ regret for some "proper" instances of BwK (*i.e.,* ones with resource consumption) that have small $G_{\texttt{LAG}}$.

Simple regret is a standard performance measure in stochastic bandits, previously not studied for BwK. While our result requires $B > \Omega(T) \gg K$, this is the main "parameter regime" of interest in most/all prior work on BwK, and a necessity in an important subset of this work [19, 20, 59, 35]. In contrast with stochastic bandits, Theorem 4.1 does not imply logarithmic regret, as per our lower bounds.

The "reduction" result is conceptual rather than technical. We make the point that regret bounds for many extensions of BwK can be derived seamlessly, and identify a mathematical structure which drives these extensions (namely, a bound on confidence sums). In a way, we formalize the intuition that analyses of "optimism under uncertainty" are likely to carry over from stochastic bandits to BwK.

We introduce several new concepts and techniques: *Lagrangian gap* (3.1) for logarithmic regret, *LP-gap* (E.2) for analyzing simple regret, and the abstraction of *confidence sums* (4.2). Also, LP-sensitivity arguments appear new in bandit analyses. Both new notions of "gap" satisfy the natural desiderata: they generalize reward-gap, do not depend on T (fixing the B/T ratio), and are "productive", leading to improved results. However, neither notion captures *all* BwK instances with low regret.[7]

[6] To make this point formal, we focus on $d = 2$ and observe that best-arm-optimality arises with probability at least p, for some absolute constant $p > 0$, if expected rewards and expected resource consumptions are drawn independently and uniformly at random. This is a generic fact about LPs, which follows, *e.g.,* from the definition of primal degeneracy in Section 2 of [46], combined with Proposition 2.7.2 in [55].

[7] This should not be surprising per se, as reward-gap does not capture all "nice" bandit instances either. *E.g.,* , problem instances with small reward-gap admit $O(\log T)$ regret if they have a likewise small best reward.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NIPS)*, 24:2312–2320, 2011.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *25th Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.

[3] Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *15th ACM Conf. on Economics and Computation (ACM-EC)*, 2014.

[4] Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *29th Advances in Neural Information Processing Systems (NIPS)*, 2016.

[5] Shipra Agrawal and Nikhil R Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 2016.

[6] Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *29th Conf. on Learning Theory (COLT)*, 2016.

[7] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019. Prelinimary version in *ACM EC 2016*.

[8] J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

[9] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. of Machine Learning Research (JMLR)*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.

[10] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[11] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. Preliminary version in *36th IEEE FOCS*, 1995.

[12] Moshe Babaioff, Shaddin Dughmi, Robert D. Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Trans. on Economics and Computation*, 3(1):4, 2015. Special issue for *13th ACM EC*, 2012.

[13] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Yaron Singer. Learning on a budget: posted price mechanisms for online procurement. In *13th ACM Conf. on Electronic Commerce (ACM-EC)*, pages 128–145, 2012.

[14] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *54th IEEE Symp. on Foundations of Computer Science (FOCS)*, 2013.

[15] Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *27th Conf. on Learning Theory (COLT)*, 2014.

[16] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. of the ACM*, 65(3):13:1–13:55, 2018. Preliminary version in *FOCS 2013*.

[17] Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[18] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. 1997.

[19] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

[20] Omar Besbes and Assaf J. Zeevi. Blind network revenue management. *Operations Research*, 60(6):1537–1550, 2012.

[21] Jean Bourgain, Van H Vu, and Philip Matchett Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[23] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122, 2012. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1204.5721`.

[24] Felipe Caro and Jérémie Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

[25] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *20th Intl. Conf. on Machine Learning (ICML)*, pages 151–159, 2013.

[26] Wang Chi Cheung and David Simchi-Levi. Assortment optimization under unknown multinomial logit choice models, 2017. Technical report, available at `http://arxiv.org/abs/1704.00108`.

[27] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual Bandits with Linear Payoff Functions. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[28] Richard Combes, Chong Jiang, and Rayadurgam Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1): 245–257, 2015.

[29] Varsha Dani, Thomas P. Hayes, and Sham Kakade. Stochastic Linear Optimization under Bandit Feedback. In *21th Conf. on Learning Theory (COLT)*, pages 355–366, 2008.

[30] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[31] Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

[32] Aurélien Garivier and Olivier Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *24th Conf. on Learning Theory (COLT)*, 2011.

[33] András György, Levente Kocsis, Ivett Szabó, and Csaba Szepesvári. Continuous time associative bandit problems. In *20th Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 830–835, 2007.

[34] Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *23rd Conf. on Learning Theory (COLT)*, 2010.

[35] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. of the ACM*, 2021. To appear. Preliminary version in *60th IEEE FOCS*, 2019.

[36] Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *33rd Conf. on Learning Theory (COLT)*, pages 2286–2305, 2020.

[37] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pages 681–690, 2008.

[38] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *J. of the ACM*, 66(4):30:1–30:77, May 2019. Merged and revised version of conference papers in *ACM STOC 2008* and *ACM-SIAM SODA 2010*. Also available at `http://arxiv.org/abs/1312.1277`.

[39] Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *13th Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 420–429, 2014.

[40] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *18th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[41] Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[42] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020. Versions available at `https://banditalgs.com/` since 2018.

[43] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th Intl. World Wide Web Conf. (WWW)*, 2010.

[44] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *38th Intl. Conf. on Machine Learning (ICML)*, 2021.

[45] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *24th Conf. on Learning Theory (COLT)*, 2011.

[46] Nimrod Megiddo and R Chandrasekaran. On the-perturbation method for avoiding degeneracy. In *Tech. Report, IBM Almaden Research Center, San Jose, CA*, 1988.

[47] Anshuka Rangi, Massimo Franceschetti, and Long Tran-Thanh. Unifying the stochastic and the adversarial bandits with knapsack. In *28th Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 3311–3317, 2019.

[48] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

[49] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 1760–1770, 2018.

[50] Denis Sauré and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[51] Denis Sauré and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[52] Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd Intl. World Wide Web Conf. (WWW)*, pages 1167–1178, 2013.

[53] Aleksandrs Slivkins. Dynamic ad allocation: Bandits with budgets. A technical report on `arxiv.org/abs/1306.0155`, June 2013.

[54] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`. Latest online revision: June 2021.

[55] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[56] Long Tran-Thanh, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. ϵ-first policies for budget-limited multi-armed bandits. In *24th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1211–1216, 2010.

[57] Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *26th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1134–1140, 2012.

[58] Alberto Vera, Siddhartha Banerjee, and Itai Gurvich. Online allocation and pricing: Constant regret via bellman inequalities. *Operations Research*, 2020.

[59] Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

[60] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122, 2015.

[61] Huasen Wu, R. Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *28th Advances in Neural Information Processing Systems (NIPS)*, 2015.

Bandits with Knapsacks beyond the Worst Case (Supplementary Materials)

Contents

A Motivating examples with $d = 2$ and small number of arms

We provide direct motivation for Theorem 3.2, our positive result for $O(\log T)$ regret. Recall that Theorem 3.2 only holds with $d = 2$ resources, and is only meaningful with a reasonably small number of arms K (because the regret bounds are linear in K). Such problems arise in many motivating applications of BwK, *e.g.,* as listed in [14, 16]. Below we spell out several stylized examples.

In *dynamic assortment* [50, 7, 26], an algorithm is a seller which chooses among possible assortments of products. In each round, a customer arrives, the algorithm chooses an assortment, and offers this assortment for sale at an exogenously fixed price. If a sale happens, the algorithm receives revenue and consumes some amount of inventory. The following version features $d = 2$ and non-huge K: there are K possible offerings for sale, and a limited amount of "raw material" used to manufacture them. Each offering, if sold, consumes some pre-fixed amount of this raw material.[8]

The "inverted" dynamic assortment problem takes the procurement perspective. An algorithm is a budget-limited contractor which chooses among K possible types of offers, *e.g.,* different items to procure from vendors, or different tasks to complete in an online labor market. In each round, a new agent arrives, the algorithm chooses an offer and presents it to the customer at an exogenously fixed price. If the offer is accepted, the contractor receives some utility (*i.e.,* reward) and spends the corresponding amount of money.

In *dynamic pricing* [19, 12, 20, 59] an algorithm is a seller with limited supply of some product, and chooses a price in each round. If this price is accepted, a sale happens, and algorithm receives revenue and spends inventory. Of our interest is the case when the set of possible prices is small and exogenously fixed, *e.g.,* there are a few possible discount levels. Likewise, in *dynamic procurement* [13, 52, 14], an algorithm is a budget-limited contractor who continuously procures some product or service. The algorithm chooses a price in each round. If this price is accepted, a transaction happens, so that the algorithm receives an "item" (*i.e.,* reward of 1) and spends the corresponding amount of money. We focus on the case when there are only a few possible prices, *e.g.,* exogenously fixed levels of premium or surcharge.

Our last example concerns fault-tolerance in systems. Consider a system, either physical or computational, which experiments with different possible policies to process incoming requests. In each time step, it chooses one of the possible policies, and observes the outcome (and there are no lingering effects, *e.g.,* no persistent "system state" that changes over time). The outcome consists of utility for performance-as-usual (*i.e.,* reward), and penalty for various mistakes or faults. Fault-tolerance requirement is expressed as a a "budget" on the total penalty accrued by the algorithm.

B Confidence bounds in UcbBwK

Let us fill in the exact specification of the confidence bounds in the UcbBwK algorithm. (This is for the sake of completeness only; as pointed out in Preliminaries, these details do not affect our analysis.)

Confidence radius. Given an unknown quantity μ and its estimator $\widehat{\mu}$, a *confidence radius* is an observable high-confidence upper bound on $|\mu - \widehat{\mu}|$. More formally, it is some quantity $\text{Rad} \in \mathbb{R}_{\geq 0}$ such that it is computable from the algorithm's observations, and $|\mu - \widehat{\mu}| \leq \text{Rad}$ with probability (say) at least $1 - 1/T^3$. Throughout, the estimator $\widehat{\mu}$ is a sample average over all available observations pertaining to μ, unless specified otherwise.

Following the prior work on BwK [12, 16, 3], we use the confidence radius from [38]:

$$f_{\texttt{rad}}(\widehat{\mu}, N) := \min\left(1,\ \sqrt{\tfrac{C_{\texttt{rad}}\,\widehat{\mu}}{\max(1,N)}} + \tfrac{C_{\texttt{rad}}}{\max(1,N)}\right),\ \text{where } C_{\texttt{rad}} = 3 \cdot \log(KdT), \tag{B.1}$$

and N is the number of samples. If $\widehat{\mu}$ is a sample average of N independent random variables with support in $[0, 1]$, and $\mu = \mathbb{E}[\mu]$, then with probability at least $1 - (Kdt)^{-2}$ we have

$$|\widehat{\mu} - \mu| \leq f_{\texttt{rad}}(\widehat{\mu}, N) \leq 3\ f_{\texttt{rad}}(\mu, N). \tag{B.2}$$

For each arm, we use this confidence radius separately for expected reward of this arm, and expected consumption of each resource.x

[8] This framing with raw material(s) — BwK formulations of revenue management problems in which products being sold are separate from raw material(s) being consumed — traces back to Besbes and Zeevi [20].

Confidence bounds. Fix arm $a \neq \texttt{null}$, round t, and resource $j \neq \texttt{time}$.

Let $S_t(a) = \{s < t : a_s = a\}$ be the set of all previous rounds in which this arm has been chosen, and let $N_t(a) = |S_t(a)|$. Let

$$\hat{r}_t(a) := \tfrac{1}{t} \sum_{s \in S_t(a)} r_s(a) \quad \text{and} \quad \hat{c}_{j,t}(a) := \tfrac{1}{t} \sum_{s \in S_t(a)} c_{j,s}(a) \tag{B.3}$$

denote, resp., the sample average of reward and resource-j consumption of this arm so far.

Define the confidence radii $\mathrm{Rad}_{0,t}(a)$ and $\mathrm{Rad}_{j,t}(a)$ for, resp., expected reward $r(a)$ and resource consumption $c_j(a)$, and the associated upper/lower confidence bounds:

$$\begin{aligned} r^{\pm}_t(a) &= \mathrm{proj}\left(\hat{r}_t(a) \pm \mathrm{Rad}_{0,t}(a) \right), & \mathrm{Rad}_{0,t}(a) &:= f_{\texttt{rad}}(\hat{r}_t(a), N_t(a)), \\ c^{\pm}_{j,t}(a) &= \mathrm{proj}\left(\hat{c}_{j,t}(a) \pm \mathrm{Rad}_{j,t}(a) \right), & \mathrm{Rad}_{j,t}(a) &:= f_{\texttt{rad}}(\hat{c}_{j,t}(a), N_t(a)), \end{aligned} \tag{B.4}$$

where $\mathrm{proj}(x) := \arg\min_{y \in [0,1]} |y - x|$ denotes the projection into $[0,1]$. Then, the event

$$r(a) \in [r^-_t(a),\ r^+_t(a)] \text{ and } c_j(a) \in [c^-_{j,t}(a),\ c^+_{j,t}(a)], \quad \forall a \in [K], j \in [d-1]. \tag{B.5}$$

holds for each round t with probability (say) at least $1 - \frac{\log(KdT)}{T^4}$ [12].

Note that all confidence radii in (B.4) are upper-bounded by

$$\mathrm{Rad}_t(a) := f_{\texttt{rad}}(1, N_t(a)), \tag{B.6}$$

which is a version of a more standard confidence radius $\widetilde{O}(1/\sqrt{N_t(a)})$.

There is no uncertainty on the time resource and the null arm. So, we set $\mathrm{Rad}_{\texttt{time},t}(\cdot) = 0$ and $c^{\pm}_{\texttt{time},t}(\cdot) = B/T$, and $\mathrm{Rad}_{0,t}(\texttt{null}) = \mathrm{Rad}_{j,t}(\texttt{null}) = r^{\pm}(\texttt{null}) = c^{\pm}_{j,t}(\texttt{null}) = 0$.

C LP Sensitivity: proof of Lemma 3.3

We focus on the sensitivity *of the support of the optimal solution*. We build on some well-known results, which we state below in a convenient form (and provide a proof for completeness). We use the textbook material from Bertsimas and Tsitsiklis [18].

Throughout this appendix, we consider a best-arm-optimal problem instance with best arm a^*. Let $\boldsymbol{X}^*$ denote the optimal solution for the linear program (2.2). Recall that the support of X^* is either $\{a^*\}$ or $\{a^*, \texttt{null}\}$. We consider perturbations in the *rescaled LP*:

$$\begin{array}{lll} \text{maximize} & \boldsymbol{X} \cdot \boldsymbol{r} & \text{such that} \\ & \boldsymbol{X} \in [0,1]^K & \\ & \boldsymbol{X} \cdot \boldsymbol{1} = 1 & \\ \forall j \in [d-1] & \boldsymbol{X} \cdot \boldsymbol{c}_j \leq (B/T)(1 - \eta_{\texttt{LP}}) & \\ & \boldsymbol{X} \cdot \boldsymbol{c}_d \leq B/T. & \end{array} \tag{C.1}$$

Recall that $\boldsymbol{r}, \boldsymbol{c}_j \in [0,1]^K$ are vectors of expected rewards and expected consumption of resource j. The d-th resource is time. The rescaling parameter $\eta_{\texttt{LP}}$ is given in Eq. (2.5).

Let $\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}}$ denote the value of this LP; it is easy to see that $\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}} = (1 - \eta_{\texttt{LP}})\ \texttt{OPT}_{\texttt{LP}}$.

We observe that a^* is the best arm for the rescaled LP, too, because $G_{\texttt{LAG}}$ is large enough. Call a distribution over arms *null-degenerate* if its support includes exactly one non-null arm.

Claim C.1. *The rescaled LP (C.1) has a null-degenerate optimal solution with non-null arm a^*.*

Proof. From the theory in [18, Ch.5], if the optimal basis to LP (2.2) remains *feasible* to the rescaled LP (C.1) then the basis is also optimal to this LP. This is because LP (C.1) is obtained by a small perturbation to the right-hand side values in LP (2.2). Let $\boldsymbol{X}^*$ denote the optimal solution to LP (2.2). From assumption this is a null-degenerate optimal solution. Using the same analysis in [18, Ch. 4.4] we only have to show that the perturbation is smaller than $X^*(a^*)$. Since the perturbation is $\frac{B\eta_{\texttt{LP}}}{T} \leq \frac{3\sqrt{B}\log(KTd)}{T}$ while $X^*(a^*) > \frac{3\sqrt{B}\log(KTd)}{T}$, this perturbation does not change the basis. Thus, the rescaled LP has a null-degenerate optimal solution. □

Claim C.2. *Let $\boldsymbol{\lambda}^*$ denote the vector of the optimal dual solution to the* LP (2.2)*. Then*

$$G_{\texttt{LAG}}(a) = \frac{T}{B}\sum_{j\in[d]} \lambda_j^* c_j(a) - r(a). \tag{C.2}$$

Proof. From Eq. (3.1) we have the following.

$$\begin{aligned} G_{\texttt{LAG}}(a) &:= \mathcal{L}(\boldsymbol{X}^*, \boldsymbol{\lambda}^*) - \mathcal{L}(\boldsymbol{X}_a, \boldsymbol{\lambda}^*) \\ &= \boldsymbol{r}(\boldsymbol{X}^*) - \frac{T}{B}\sum_{j\in[d]} \lambda_j^*\, \boldsymbol{c}_j(\boldsymbol{X}^*) + \frac{T}{B}\sum_{j\in[d]} \lambda_j^* c_j(a) - r(a). \end{aligned}$$

Consider the dual of the LP (2.2). It can be seen that the objective of this dual is $\sum_{j\in[d]} \lambda_j$. It follows that $\texttt{OPT}_{\texttt{LP}} = \sum_{j\in[d]} \lambda_j^*$ by strong duality [22, Section 5.2.3]. As proved in [35], $\mathcal{L}(\boldsymbol{X}^*, \boldsymbol{\lambda}^*) = \texttt{OPT}_{\texttt{LP}}$. Thus,

$$\sum_{j\in[d]} \lambda_j^* = \texttt{OPT}_{\texttt{LP}} = \mathcal{L}(\boldsymbol{X}^*, \boldsymbol{\lambda}^*) = \boldsymbol{r}(\boldsymbol{X}^*) - \frac{T}{B}\sum_{j\in[d]} \lambda_j^*\, \boldsymbol{c}_j(\boldsymbol{X}^*) + \sum_{j\in[d]} \lambda_j^*.$$

Therefore, $\boldsymbol{r}(\boldsymbol{X}^*) = \frac{T}{B}\sum_{j\in[d]} \lambda_j^*\, \boldsymbol{c}_j(\boldsymbol{X}^*)$, which implies (C.2). □

Claim 3.3 easily follows from the following standard result by letting $\delta(a) = \mathrm{Rad}_t(a)$.

Theorem C.3 (perturbation)**.** *Posit only one resource other than time* (i.e., $d = 2$)*. Consider a perturbation of the rescaled LP* (C.1)*, where the reward vector $\boldsymbol{r}$ is replaced with $\tilde{\boldsymbol{r}}$, and the consumption vector $\boldsymbol{c}_1$ for the non-time resource is replaced with $\tilde{\boldsymbol{c}}_1$. Let $\tilde{\boldsymbol{X}}^*$ be its optimal solution. Assume $0 \le \tilde{\boldsymbol{r}} - \boldsymbol{r} \le \boldsymbol{\delta}$ and $0 \le \boldsymbol{c}_1 - \tilde{\boldsymbol{c}}_1 \le \boldsymbol{\delta}$, for some vector $\boldsymbol{\delta} \in [0,1]^K$. Then for each arm $a \neq a^*$,*

$$\delta(a) > G_{\texttt{LAG}}(a) \quad \textit{if} \quad a \in \mathrm{supp}(\tilde{\boldsymbol{X}}^*).$$

Proof. Let $\lambda_1^* \ge 0$ denote the dual variable corresponding to the single resource. Note that since $\texttt{OPT}_{\texttt{LP}} \le 1$ and the dual vector $\boldsymbol{\lambda}^* \ge \boldsymbol{0}$ coordinate wise, we have $\lambda_1^* \le 1$. From [18, Ch. 5.1] on local sensitivity when non-basic column of A is changed, we have that the maximum allowable change to any single column $\delta(a) \le \frac{\tilde{c}(a)}{\lambda_1^*}$ where $\tilde{c}(a)$ is the reduced-cost for the simplex algorithm, as defined in [18]. We will show that $\tilde{c}(a) = G_{\texttt{LAG}}(a)$. Thus, if $\delta(a) \le \frac{\tilde{c}(a)}{\lambda_1^*} = \frac{G_{\texttt{LAG}}(a)}{\lambda_1^*}$ we have that the basis remains unchanged. Likewise from Bertsimas and Tsitsiklis [18, Ch. 5], the maximum allowed perturbation $\delta(a)$ on the reward $r(a)$ for the basis to remain unchanged is $\delta(a) \le \tilde{c}(a)$. Combining these two we get the "*if*" part of the theorem.

It remains to prove that the reduced cost $\tilde{c}(a) = G_{\texttt{LAG}}(a)$. After converting the linear program to the standard form as required in [18], the reduced-cost $\tilde{c}(a)$ is given by the expression $\frac{T}{B(1-\eta_{\texttt{LP}})}\sum_{j\in[d]} c_j(a)\tilde{\lambda}_j^* - r(a)$ where $\tilde{\boldsymbol{\lambda}}^*$ is the optimal dual solution to LP (C.1). Note that $\boldsymbol{\lambda}^* := \left(\frac{1}{1-\eta_{\texttt{LP}}}\right)\tilde{\boldsymbol{\lambda}}^*$ is an optimal solution to the dual of the LP (2.2). Thus, plugging it into the definition of reduced cost and combining it with Claim C.2 we have that

$$\tilde{c}(a) = \frac{T}{B}\sum_{j\in[d]} \lambda_j^* c_j(a) - r(a) = G_{\texttt{LAG}}(a).$$ □

D Various technicalities from Sections 3

D.1 Standard tools

We rely on some standard tools, which we state below for the sake of convenience.

Theorem D.1 (Wald's identity). *Let $X_i : i \in \mathbb{N}$ be i.i.d. real-valued random variables, adapted to filtration $\mathcal{F}_i : i \in \mathbb{N}$. Let N be a stopping time relative to the same filtration. Then*

$$\mathbb{E}[X_1 + X_2 + \ldots + X_N] = \mathbb{E}[X_i] \cdot \mathbb{E}[N].$$

Theorem D.2 (Optimal Stopping Theorem). *Let $X_i : i \in \mathbb{N}$ be a martingale sequence with $\mathbb{E}[X_0] = 0$ adapted to filtration $\mathcal{F}_i : i \in \mathbb{N}$. Let N be a stopping time relative to the same filtration. Then we have that $\mathbb{E}[X_N] = 0$.*

Theorem D.3 ([37, 12]). *Let $Z_1, Z_2, , \ldots, Z_T$ be a martingale w.r.t. filtration $(\mathcal{F}_t)_{t\in[T]}$, such that $|Z_t| \leq c$ for all $t \in [T]$. Let $\mu := \frac{1}{T}\sum_{t\in[T]} \mathbb{E}[Z_t \mid \mathcal{F}_{t-1}]$. Then,*

$$\Pr\left[\left|\textstyle\sum_{t\in[T]} Z_t - \mu T\right| > \sqrt{2\mu T c^2 \ln\frac{T}{\delta}}\right] \leq \delta.$$

D.2 Proof of Eq. (3.6)

Let τ denote the stopping time of the algorithm that chooses arm a^* in every time-step, given that the total budget is B_0, T_0 on the two resources. From definition we have $\texttt{REW}(a^* \mid B_0, T_0) = \sum_{t\in[\tau]} r_t(a^*)$. Using Wald's identity (Theorem D.1), we have that $\mathbb{E}[\texttt{REW}(a^* \mid B_0, T_0)] = \mathbb{E}[\tau]\, r(a^*)$.

Let B_0, T_0 denote the budget remaining for the two resources. By definition, we have that $\tau \geq T_0$ and $\sum_{t\in[\tau]} c_t(a^*) \geq B_0$. Using the Wald's identity (Theorem D.1) we have that $\mathbb{E}[\sum_{t\in[\tau]} c_t(a^*)] = \mathbb{E}[\tau]c(a^*)$. Thus, we have $\mathbb{E}[\tau] \geq \min\left\{T_0, \frac{B_0}{c(a^*)}\right\} \geq \min\{T_0, B_0\}$. Therefore, we obtain the following.

$$\mathbb{E}[\texttt{REW}(a^* \mid B_0, T_0)] = \mathbb{E}[\tau] r(a^*) > \left(\frac{\min\{T_0, B_0\}}{\max\{\frac{B}{T}, c(a^*)\}}\right) r(a^*), \quad \text{and} \tag{D.1}$$

$$\mathbb{E}[\texttt{REW}(a^* \mid B)] = \mathbb{E}[\tau_B] r(a^*) \leq \left(\frac{B}{\max\{\frac{B}{T}, c(a^*)\}}\right) r(a^*). \tag{D.2}$$

Combining Equations (D.1) and (D.2), we get Eq. (3.6).

D.3 Proof of Eq. (3.9)

We now modify the above proof to get the tighter lower-bound in Eq. (3.9). Let T_0, B_0 denote the expected remaining time and budget (respectively) and let τ denote the (random) stopping time of the algorithm that chooses arm a^* in every time-step given T_0 time-steps and B_0 budget. This implies that we have, $\mathbb{E}[\sum_{t\in[\tau]} c_t(a^*)] \geq B_0$ and $\mathbb{E}[\tau] \geq T_0$. From Theorem D.1, this implies that we have $\mathbb{E}[\tau]c(a^*) \geq B_0$ and $\mathbb{E}[\tau] \geq T_0$. This implies that $\mathbb{E}[\tau] \geq \min\{T_0, \frac{B_0}{c(a^*)}\}$.

Similar to Eq. (D.1) and Eq. (D.2) we obtain the following.

$$\mathbb{E}[\texttt{REW}(a^* \mid B_0, T_0)] = \mathbb{E}[\tau] r(a^*) > \min\{T_0, \tfrac{B_0}{c(a^*)}\} r(a^*), \quad \text{and} \tag{D.3}$$

$$\mathbb{E}[\texttt{REW}(a^* \mid B_0 = B, T_0 = T)] = \texttt{OPT}_{\texttt{FD}} \leq \left(\frac{B}{\max\{\frac{B}{T}, c(a^*)\}}\right) r(a^*). \tag{D.4}$$

Combining Equations (D.3) and (D.4), we get Eq. (3.9).

D.4 Lower bound on Lagrange gap: Proof of Eq. (3.10)

We will use Eq. (3.4) and some standard properties of linear programming.

Assume $c(a^*) < \frac{B}{T}$. Using complementary slackness theorem on LP (2.2), this implies that $\lambda_1^* = 0$. Moreover, note that the objective in the dual of LP (2.2) is $\lambda_0^* + \lambda_1^* = \lambda_0^*$. The optimal value of the primal LP (2.2) is $r(a^*)$ since, $X(a^*) = 1$ is the optimal solution to the LP. This implies that $\lambda_0^* = r(a^*) \geq \frac{\texttt{OPT}_{\texttt{FD}}}{T}$. Substituting this into Eq. (3.4) gives the first inequality in Eq. (3.10).

Now assume $c(a^*) > \frac{B}{T}$. Again, as above complementary slackness theorem on LP (2.2), this implies that $\lambda_0^* = 0$. Thus, $G_{\texttt{LAG}}(a) = \frac{T}{B} \cdot \lambda_1^* \cdot c(a) - r(a)$. Using the dual objective function $\lambda_0^* + \lambda_1^* = \lambda_1^*$ combined with strong duality, this implies that $\lambda_1^* = \frac{\texttt{OPT}_{\texttt{LP}}}{T} \geq \frac{\texttt{OPT}_{\texttt{FD}}}{T}$. Plugging this back into Eq. (3.4) gives the second inequality in Eq. (3.10).

D.5 Martingale arguments: Proof of Eq. (3.8)

For the proof of Eq. (3.8), we use the well-known theorem on optimal stopping time of martingales (Theorem D.2). Fix an arm $a \in [K]$. For any subset $S \subseteq [T]$ of rounds let $N_S(a)$, $r_S(a)$ and $c_S(a)$ denote the number of times arm a is chosen, the total realized rewards for arm a and the total realized consumption of arm a, respectively. Let τ denote the (random) stopping time of a BwK algorithm with (random) budget B and time T. Then we have the following claim.

Claim D.4. *For a random stopping time τ, for every arm $a \in [K]$ we have the following.*

$$\mathbb{E}\left[r_{[\tau]}(a)\right] = r(a) \cdot \mathbb{E}[N_{[\tau]}(a)]. \tag{D.5}$$

$$\mathbb{E}\left[c_{[\tau]}(a)\right] = c(a) \cdot \mathbb{E}[N_{[\tau]}(a)]. \tag{D.6}$$

Proof. We will prove the equality in Eq. (D.5); the one in Eq. (D.6) follows. Consider $r_{[\tau]}(a)$. By definition this is equal to $\sum_{t \in [\tau]} r_t(a) \cdot \mathbb{I}[a_t = a]$. Let $A_t := \mathbb{I}[a_t = a]$ denote the random variable corresponding to the event that arm a is chosen at time t. Define the random variable

$$Y_t := \sum_{t' \leq t} A_{t'} r_{t'}(a) - \mathop{\mathbb{E}}_{t'}\left[A_{t'} r_{t'}(a)\right],$$

where $\mathbb{E}_t[.]$ denotes the conditional expectation given the random variables $A_1, A_2, \ldots, A_{t-1}$. It is easy to see that the sequences $\{X_t\}_{t \in [\tau]}$, $\{Y_t\}_{t \in [\tau]}$ and $\{Z_t\}_{t \in [\tau]}$ forms a martingale sequence. Thus, we will apply the optimal stopping theorem (Theorem D.2) at time τ, we have the following.

$$\mathbb{E}\left[Y_\tau\right] = \mathbb{E}\left[\sum_{t' \leq \tau} A_{t'} r_{t'}(a)\right] - \mathbb{E}\left[\sum_{t' \leq \tau} \mathop{\mathbb{E}}_{t'}\left[A_{t'} r_{t'}(a)\right]\right] = 0. \tag{D.7}$$

Consider the term $\mathbb{E}\left[\sum_{t' \leq \tau} \mathbb{E}_{t'}\left[A_{t'} r_{t'}(a)\right]\right]$ in Eq. (D.7). This can be simplified to $\mathbb{E}\left[\sum_{t' \leq \tau} r(a) \cdot \Pr[a_{t'} = a]\right]$. Consider the following random variable

$$Z_t := \sum_{t' \leq t} \Pr[a_{t'} = a] - \mathop{\mathbb{E}}_{t'}[\Pr[a_{t'} = a]].$$

Note that $\sum_{t' \leq t} \mathbb{E}_{t'}[\Pr[a_{t'} = a]] = N_{[t]}(a)$. Thus, using Theorem D.2 on the sequence Z_t at the stopping time τ, we obtain $\mathbb{E}\left[\sum_{t' \leq \tau} \Pr[a_{t'} = a]\right] = \mathbb{E}[N_{[\tau]}(a)]$.

Thus, the term $\mathbb{E}\left[\sum_{t' \leq \tau} \mathbb{E}_{t'}\left[A_{t'} r_{t'}(a)\right]\right]$ in Eq. (D.7) simplifies to $r(a) \cdot N_{[\tau]}(a)$ which gives the required equality in Eq. (D.5). □

We will now use Claim D.4 to prove Eq. (3.8). Recall that $\texttt{REW}(a \mid B(a), T(a))$ denotes the total contribution to the reward by the BwK algorithm by playing arm a with a (random) resource consumption of $B(a)$ and time steps of $T(a)$. Let τ be the (random) stopping time of this algorithm.

By definition we have that $N_{[\tau]}(a) = T(a)$. Thus, $\mathbb{E}[N_{[\tau]}(a)] = \mathbb{E}[T(a)$. From Eq. (D.6), we also have that $\mathbb{E}[N_{[\tau]}(a)] = \frac{\mathbb{E}[c_{[\tau]}(a)]}{c(a)}$. From the definition of $B(a)$ we have, $B(a) = c_{[\tau]}(a)$ and thus, $\mathbb{E}[B(a)] = \mathbb{E}[c_{[\tau]}(a)]$. Thus, this implies that $\mathbb{E}[N_{[\tau]}(a)] = \min\{T(a), \frac{\mathbb{E}[B(a)]}{c(a)}\}$.

Consider $\mathbb{E}[\texttt{REW}(a)] = \mathbb{E}[\texttt{REW}(a \mid B(a), T(a))]$.

$$\begin{aligned}\mathbb{E}[\texttt{REW}(a \mid B(a), T(a))] &= \mathbb{E}\left[r_{[\tau]}(a)\right] \\ &= r(a) \cdot \mathbb{E}[N_{[\tau]}(a)] && \textit{(From Eq. (D.5))} \\ &= r(a) \cdot \min\{T(a), \tfrac{\mathbb{E}[B(a)]}{c(a)}\} && \text{(D.8)}\end{aligned}$$

Now, consider $\texttt{LP}(a \mid \mathbb{E}[B(a)], \mathbb{E}[T(a)])$. This value is equal to,

$$\begin{aligned}\mathbb{E}[\texttt{REW}(a \mid \mathbb{E}[B(a)], \mathbb{E}[T(a)])] &= \frac{r(a)}{\max\{\mathbb{E}[B(a)]/\,\mathbb{E}[T(a)], c(a)\}} \cdot \tfrac{\mathbb{E}[B(a)]}{\mathbb{E}[T(a)]} \\ &= r(a) \cdot \min\left\{\mathbb{E}[T(a)], \tfrac{\mathbb{E}[B(a)]}{c(a)}\right\}.\end{aligned}$$

Note that the last equality is same as the RHS in Eq. (D.8).

E Proof of Theorem 3.10: generic $\sqrt{T}$ lower bound

Preliminaries. We rely on a well-known information-theoretic result for multi-armed bandits: essentially, no algorithm can reliably tell apart two bandit instances at time T if they differ by at most $O(1/\sqrt{T})$.[9] We formulate this result in a way that is most convenient for our applications.

Lemma E.1. *Consider multi-armed bandits with Bernoulli rewards. Fix $\epsilon > 0$ and two problem instances $\mathcal{I}, \mathcal{I}'$ such that the mean reward of each arm differs by at most ϵ between $\mathcal{I}$ and $\mathcal{I}'$. Suppose some bandit algorithm outputs distribution $\boldsymbol{Y}_t$ over arms at time $t \leq c/\epsilon^2$, for a sufficiently small absolute constant c. Let H be an arbitrary Lebesgue-measurable set of distributions over arms. Then either $\Pr[\boldsymbol{Y}_t \in H \mid \mathcal{J}_t = \mathcal{I}] > 1/4$ or $\Pr[\boldsymbol{Y}_t \notin H \mid \mathcal{J}_t = \mathcal{I}'] > 1/4$ holds.*

Applying Lemma E.1 to bandits with knapsacks necessitates some subtlety. First, the rewards in the lemma will henceforth be called *quasi-rewards*, as they may actually correspond to consumption of a particular resource. Second, while a BwK algorithm receives multi-dimensional feedback in each round, the feedback other than the quasi-rewards will be the same (in distribution) for both problem instances, and hence can be considered a part of the algorithm. Third, distribution $\boldsymbol{Y}_t$ will be the conditional distribution over arms chosen by the BwK algorithm in round t given the algorithm's observations so far; we will assume this without further mention. Fourth, we will need to specify the set H of distributions (which will depend on a particular application).

Consider the rescaled LP (C.1) with $\eta_{\texttt{LP}} := 6 * \texttt{OPT}_{\texttt{LP}} \sqrt{\frac{\log dT}{B}}$; we use this $\eta_{\texttt{LP}}$ throughout this proof. Let $\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}}$ be the value of this LP. We prove the lower bound using $\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}}$ as a benchmark. This suffices by the following claim from prior work: [10]

Claim E.2 (Immorlica et al. [35]). *$\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}} \leq \texttt{OPT}_{\texttt{FD}}$ for $\eta_{\texttt{LP}} := 6 \cdot \texttt{OPT}_{\texttt{LP}} \sqrt{\frac{\log dT}{B}}$.*

Problem instances. Let $\boldsymbol{r}(a)$ and $\boldsymbol{c}(a) \in [0,1]^d$ be, resp., the mean reward and the mean resource consumption vector for each arm a for instance $\mathcal{I}_0$. Let $\epsilon = c_{\texttt{LB}}/\sqrt{T}$.

Problem instances $\mathcal{I}, \mathcal{I}'$ are constructed as specified in the proof sketch; we repeat it here for the sake of convenience. For both instances, the rewards of each non-null arm $a \in \{A_1, A_2\}$ are deterministic and equal to $r(a)$. Resource consumption vector for arm A_1 is deterministic and equals $\boldsymbol{c}(A_1)$. Resource consumption vector of arm A_2 in each round t, denoted $\boldsymbol{c}_{(t)}(A_2)$, is a carefully

[9] This strategy for proving lower bounds in multi-armed bandits goes back to Auer et al. [11]. Lemma E.1 is implicit in Auer et al. [11], see Slivkins [54, Lemma 2.9] for exposition.

[10] Claim E.2 is a special case of Lemma 8.6 in Immorlica et al. [35] for $\tau^* = T$ and the reward/consumption for each arm, each resource and each time-step replaced with the mean reward/consumption.

constructed random vector whose expectation is $c(A_2)$ for instance $\mathcal{I}$, and slightly less for instance $\mathcal{I}'$. Specifically, $\boldsymbol{c}_{(t)}(A_2) = \boldsymbol{c}(A_2) \cdot W_t/(1 - c_{\texttt{LB}})$, where W_t is an independent Bernoulli random variable which correlates the consumption of all resources. We posit $\mathbb{E}[W_t] = 1 - c_{\texttt{LB}}$ for instance $\mathcal{I}$, and $\mathbb{E}[W_t] = 1 - c_{\texttt{LB}} - \epsilon$ for instance $\mathcal{I}'$.

Main derivation. From the premise of the theorem (Eq. (3.15)), problem instance $\mathcal{I}$ admits an optimal solution $\boldsymbol{X}^*$ that is substantially supported on both non-null arms. Let $\boldsymbol{X}^*_{\mathcal{I}}$, $\boldsymbol{X}^*_{\mathcal{I}'}$ denote the optimal solutions to the scaled LP, instantiated for instances $\mathcal{I}, \mathcal{I}'$ respectively.

The proof proceeds as follows. We first prove certain properties of distributions $\boldsymbol{X}^*_{\mathcal{I}}$ and $\boldsymbol{X}^*_{\mathcal{I}'}$. We then use these properties and apply Lemma E.1 with suitable quasi-rewards to complete the proof of the lower-bounds.

Since we modify the mean consumption of all resources for one arm in $\mathcal{I}'$ this implies that $\boldsymbol{X}^*_{\mathcal{I}} \neq \boldsymbol{X}^*_{\mathcal{I}'}$. From assumption 3.8-(3.8) we have that $G_{\texttt{LAG}} \geq c_{\texttt{LB}}/\sqrt{T}$. From the premise of the theorem, we have that the mean vector of consumptions for the resources $j \in [d]$ are all linearly independent. Thus, we can apply sensitivity theorem C.3 to conclude that the support of the solution $\boldsymbol{X}^*_{\mathcal{I}'}$ is same as $\boldsymbol{X}^*_{\mathcal{I}}$.

Moreover, from the linear independence of the consumption vectors and Eq. (3.15). combined with standard LP theory (see chapter 4 on duality in [18]) we have that there exists a resource $j^* \in [d]$ such that the optimal solution $\boldsymbol{X}^*_{\mathcal{I}}$ satisfies the resource constraint with equality.

In what follows, we denote the vector $\boldsymbol{c}$ as a shorthand for $\boldsymbol{c}_{j^*}$ (*i.e.,* we drop the index j^*). Note that from the perturbation we have that $c(A_1) < c(A_2)$. Thus, for some $\delta > 0$ we have $X^*_{\mathcal{I}'}(A_1) = X^*_{\mathcal{I}}(A_1) - \delta$ and $X^*_{\mathcal{I}'}(A_2) = X^*_{\mathcal{I}}(A_2) + \delta$. Let $\|\boldsymbol{X}\|$ denote the ℓ_1-norm of a given distribution $\boldsymbol{X}$. Thus, we have

$$\|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{X}^*_{\mathcal{I}'}\| = 2\delta. \tag{E.1}$$

Given any distribution $\boldsymbol{Y}$ over the arms, let

$$V_{\texttt{sc}}(\boldsymbol{Y}) := (1 - \eta_{\texttt{LP}}) \cdot {}^{B}\!/_{T} \cdot r(\boldsymbol{Y}) / \left(\max_{j \in [d]} c_j(\boldsymbol{Y}) \right). \tag{E.2}$$

This is the value of $\boldsymbol{Y}$ in the rescaled LP (C.1), where $\boldsymbol{Y}$ itself is rescaled to make it LP-feasible (and as large as possible). Note that $V_{\texttt{sc}}(\boldsymbol{Y}) = (1 - \eta_{\texttt{LP}})\, V(\boldsymbol{Y})$, where $V(\boldsymbol{Y})$ is the value of the original LP, as defined in (E.2). Also, $\texttt{OPT}^{\texttt{sc}}_{\texttt{LP}} = \sup_{\boldsymbol{Y}} V_{\texttt{sc}}(\boldsymbol{Y})$.

By a slight abuse of notation, let $V_{\texttt{sc}}(\boldsymbol{Y}), V'_{\texttt{sc}}(\boldsymbol{Y})$ be the value of $V_{\texttt{sc}}(\boldsymbol{Y})$ corresponding to instances $\mathcal{I}$ and $\mathcal{I}'$ respectively.

We use the following two claims in the proof of our lower-bound. Claim E.3 states that if a distribution is close to the optimal distribution for instance $\mathcal{I}$ then it is also far from the optimal distribution for $\mathcal{I}'$. Claim E.4 states that if a distribution is far from the optimal distribution, then playing from that distribution also incurs large instantaneous regret. Both claims have nothing to do with particular algorithms.

Claim E.3. *Fix distribution $\boldsymbol{Y} \in \Delta^3$ and $\epsilon < 1$. If $\|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\| < \epsilon \cdot c^2_{\texttt{LB}}$ then $\|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\texttt{LB}}$.*

Claim E.4. *Fix distribution $\boldsymbol{Y} \in \Delta^3$ and $\epsilon < 1$. If $\|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\texttt{LB}}$ then $V_{sc}(\boldsymbol{X}^*_{\mathcal{I}}) - V_{sc}(\boldsymbol{Y}) \geq \epsilon \cdot \frac{c^3_{\texttt{LB}}}{2}$. Likewise, if $\|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\texttt{LB}}$ then $V'_{sc}(\boldsymbol{X}^*_{\mathcal{I}'}) - V'_{sc}(\boldsymbol{Y}) \geq \epsilon \cdot \frac{c^3_{\texttt{LB}}}{2}$.*

We now invoke Lemma E.1 with the quasi-rewards at each time-step determined by the consumption of the resource j^*.

Define the set,

$$\mathcal{H} := \left\{ \boldsymbol{Y} : \|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\texttt{LB}} \right\}, \tag{E.3}$$

to complete the proof Theorem 3.10. Consider an arbitrary algorithm ALG. We consider two cases: $\mathcal{J} = \mathcal{I}$ and $\mathcal{J} = \mathcal{I}'$, which denote the instance that satisfies the conclusion of this lemma for at least $\frac{T}{2}$ rounds for $T := \frac{c_{\texttt{LB}}}{\epsilon^2}$.

Let $\mathcal{J} = \mathcal{I}$. Let $\mathcal{T}$ denote the set of time-steps $t \in [T]$ such that $\mathcal{J}_t = \mathcal{I}$ and $\boldsymbol{Y}_t \in \mathcal{H}$. Then, the expected regret of ALG can be lower-bounded by,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{T}} V_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}}) - V_{\mathtt{sc}}(\boldsymbol{Y}_t)\right] &= \mathbb{E}\left[\sum_{t\in\mathcal{T}:\ \|\boldsymbol{X}^*_{\mathcal{I}}-\boldsymbol{Y}_t\|\geq\epsilon\cdot c^2_{\mathtt{LB}}} V_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}}) - V_{\mathtt{sc}}(\boldsymbol{Y}_t)\right] && \textit{(by Eq. (E.3))}\\
&\geq \mathbb{E}\left[\textstyle\sum_{t\in\mathcal{T}} \epsilon\cdot\frac{c^3_{\mathtt{LB}}}{2}\right] && \textit{(by Eq. (E.4))}\\
&\geq T/4\cdot\epsilon\cdot\frac{c^3_{\mathtt{LB}}}{2} && \textit{(by Lemma E.1)}\\
&\geq O\left(c^4_{\mathtt{LB}}\cdot\sqrt{T}\right). && \textit{(Since } \epsilon = \tfrac{c_{\mathtt{LB}}}{\sqrt{T}}\textit{)}
\end{aligned}$$

We use a similar argument when $\mathcal{J} = \mathcal{I}'$. Let $\mathcal{T}'$ denote the set of time-steps $t \in [T]$ such that $\mathcal{J}_t = \mathcal{I}'$ and $\|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{Y}_t\| \geq \epsilon \cdot c^2_{\mathtt{LB}}$. The expected regret of ALG can be lower-bounded by,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{T}'} V'_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}'}) - V'_{\mathtt{sc}}(\boldsymbol{Y}_t)\right] &= \mathbb{E}\left[\sum_{t\in\mathcal{T}':\ \|\boldsymbol{X}^*_{\mathcal{I}'}-\boldsymbol{Y}_t\|\geq\epsilon\cdot c^2_{\mathtt{LB}}} V'_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}'}) - V'_{\mathtt{sc}}(\boldsymbol{Y}_t)\right]\\
&\geq \mathbb{E}\left[\sum_{t\in\mathcal{T}':\ \|\boldsymbol{X}^*_{\mathcal{I}}-\boldsymbol{Y}_t\|<\epsilon\cdot c^2_{\mathtt{LB}}} V'_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}'}) - V'_{\mathtt{sc}}(\boldsymbol{Y}_t)\right] && \textit{(by Claim E.3)}\\
&= \mathbb{E}\left[\sum_{t\in\mathcal{T}':\ \boldsymbol{Y}_t\notin\mathcal{H}} V'_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}'}) - V'_{\mathtt{sc}}(\boldsymbol{Y}_t)\right] && \textit{(by Eq. (E.3))}\\
&\geq \mathbb{E}\left[\sum_{t\in[T]:\ \boldsymbol{Y}_t\notin\mathcal{H}} \epsilon\cdot\frac{c^3_{\mathtt{LB}}}{2}\right] && \textit{(by Eq. (E.4))}\\
&\geq T/4\cdot\epsilon\cdot\frac{c^3_{\mathtt{LB}}}{2} && \textit{(by Lemma E.1)}\\
&\geq O\left(c^4_{\mathtt{LB}}\cdot\sqrt{T}\right). && \textit{(Since } \epsilon = \tfrac{c_{\mathtt{LB}}}{\sqrt{T}}\textit{).}
\end{aligned}$$

Proof of Claim E.3. Let $c(A_1), c(A_2)$ denote the expected consumption of arms A_1 and A_2 respectively in instance $\mathcal{I}$. Define $\zeta := \frac{\epsilon c(A_1)}{1-c_{\mathtt{LB}}}$. By definition, this implies that the expected consumption of arm A_2 in instance $\mathcal{I}'$ is $c(A_2) - \zeta$. Additionally, since the support contains two arms, we have that the following holds: $c(A_1)X^*_{\mathcal{I}}(A_1) + c(A_2)X^*_{\mathcal{I}}(A_2) = B/T * (1 - \eta_{\mathtt{LP}})$ and $c(A_1)X^*_{\mathcal{I}'}(A_1) + c(A_2)X^*_{\mathcal{I}'}(A_2) - \zeta X^*_{\mathcal{I}'}(A_2) = B/T * (1 - \eta_{\mathtt{LP}})$. Thus, we have

$$c(A_1)X^*_{\mathcal{I}}(A_1)+c(A_2)X^*_{\mathcal{I}}(A_2) = c(A_1)X^*_{\mathcal{I}}(A_1)+c(A_2)X^*_{\mathcal{I}}(A_2)+\delta(C(A_2)-c(A_1)-\zeta)-\zeta X^*_{\mathcal{I}}(A_2).$$

Rearranging and using the assumptions in 3.8, we get that

$$\delta = \frac{\zeta X^*_{\mathcal{I}}(A_2)}{c(A_2)-c(A_1)-\zeta} \geq \frac{\epsilon c_{\mathtt{LB}}}{1-c_{\mathtt{LB}}}\cdot\frac{c_{\mathtt{LB}}}{1-2c_{\mathtt{LB}}-\frac{\epsilon\cdot c_{\mathtt{LB}}}{1-c_{\mathtt{LB}}}} \geq \epsilon\cdot c^2_{\mathtt{LB}}. \tag{E.4}$$

Consider $\|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{Y}\|$. This can be rewritten as

$$\begin{aligned}
&= \|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{Y} - \boldsymbol{X}^*_{\mathcal{I}} + \boldsymbol{X}^*_{\mathcal{I}}\|\\
&\geq \big|\|\boldsymbol{X}^*_{\mathcal{I}'} - \boldsymbol{X}^*_{\mathcal{I}}\| - \|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\|\big| && \textit{(Triangle inequality)}\\
&\geq 2\delta - \epsilon\cdot c^2_{\mathtt{LB}} && \textit{(Premise of the claim and Eq. (E.1))}\\
&\geq \epsilon\cdot c^2_{\mathtt{LB}}. && \textit{(From Eq. (E.4))}
\end{aligned}$$

Proof of Claim E.4. We will prove the statement $\|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\mathtt{LB}} \implies V_{\mathtt{sc}}(\boldsymbol{X}^*_{\mathcal{I}}) - V_{\mathtt{sc}}(\boldsymbol{Y}) \geq \epsilon \cdot \frac{c^3_{\mathtt{LB}}}{2}$. The exact same argument holds by replacing $\boldsymbol{X}^*_{\mathcal{I}}$ with $\boldsymbol{X}^*_{\mathcal{I}'}$ and $V_{\mathtt{sc}}(.)$ with $V'_{\mathtt{sc}}(.)$.

Consider $V_{\texttt{sc}}(\boldsymbol{X}^*_{\mathcal{I}}) - V_{\texttt{sc}}(\boldsymbol{Y})$. By definition, this equals,

$$r(\boldsymbol{X}^*_{\mathcal{I}}) - \frac{r(\boldsymbol{Y})}{\max\{\frac{B'}{T}, c(\boldsymbol{Y})\}} \cdot \frac{B'}{T}, \tag{E.5}$$

where B' is the scaled budget.

We have two cases. In case 1, let $\max\{\frac{B'}{T}, c(\boldsymbol{Y})\} = \frac{B'}{T}$. Thus, Eq. (E.5) simplifies to,

$$\begin{aligned}
&= r(\boldsymbol{X}^*_{\mathcal{I}}) - r(\boldsymbol{Y}) \\
&= r(A_1)[X^*_{\mathcal{I}}(A_1) - Y(A_1)] + r(A_2)[X^*_{\mathcal{I}}(A_2) - Y(A_2)]
\end{aligned}$$

Note that since $\max\{\frac{B'}{T}, c(\boldsymbol{Y})\} = \frac{B'}{T}$, this implies that $Y(\texttt{null}) = 0$. Since $\boldsymbol{X}^*_{\mathcal{I}}$ is an optimal solution and $r(A_2) > r(A_1)$, this implies that we have $Y(A_1) = X^*_{\mathcal{I}}(A_1) + \zeta$ and $Y(A_2) = X^*_{\mathcal{I}}(A_2) - \zeta$. Thus, we have,

$$\begin{aligned}
r(A_1)[X^*_{\mathcal{I}}(A_1) - Y(A_1)] + r(A_2)[X^*_{\mathcal{I}}(A_2) - Y(A_2)] &\geq [r(A_2) - r(A_1)]\zeta \\
&\geq c_{\texttt{LB}} \cdot \|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\|/2 \\
&\geq \epsilon \cdot \tfrac{c^3_{\texttt{LB}}}{2}.
\end{aligned}$$

Consider case 2 where $\max\{\frac{B'}{T}, c(\boldsymbol{Y})\} = c(\boldsymbol{Y})$. Then, Eq. (E.5) simplifies to,

$$\begin{aligned}
&= r(\boldsymbol{X}^*_{\mathcal{I}}) - \tfrac{B'}{T} \cdot \tfrac{r(\boldsymbol{Y})}{c(\boldsymbol{Y})} \\
&\geq r(\boldsymbol{X}^*_{\mathcal{I}}) - \max_{\boldsymbol{Y} \in \Delta_3 : \|\boldsymbol{X}^*_{\mathcal{I}} - \boldsymbol{Y}\| \geq \epsilon \cdot c^2_{\texttt{LB}}} \tfrac{B(1-\eta_{\texttt{LP}})}{T} \cdot \tfrac{r(\boldsymbol{Y})}{c(\boldsymbol{Y})}
\end{aligned}$$

The maximization happens when the distribution $\boldsymbol{Y}$ is such that $Y(A_1) = X^*_{\mathcal{I}} - \epsilon \cdot c^2_{\texttt{LB}}/2$ and $Y(A_2) = X^*_{\mathcal{I}} - \epsilon \cdot c^2_{\texttt{LB}}/2$. Plugging this into the expression we get the RHS is at least,

$$\begin{aligned}
&\geq r(\boldsymbol{X}^*_{\mathcal{I}}) - \tfrac{B(1-\eta_{\texttt{LP}})}{T} \cdot \frac{r(\boldsymbol{X}^*_{\mathcal{I}}) + \epsilon \cdot c^2_{\texttt{LB}}/2 \cdot (r(A_2) - r(A_1))}{c(\boldsymbol{X}^*_{\mathcal{I}}) + \epsilon \cdot c^2_{\texttt{LB}}/2 \cdot (c(A_2) - c(A_1))} \\
&\geq r(\boldsymbol{X}^*_{\mathcal{I}}) - c_{\texttt{LB}}(1 - \eta_{\texttt{LP}}) \cdot \frac{r(\boldsymbol{X}^*_{\mathcal{I}}) + \epsilon \cdot c^2_{\texttt{LB}}/2 \cdot (r(A_2) - r(A_1))}{c(\boldsymbol{X}^*_{\mathcal{I}}) + \epsilon \cdot c^2_{\texttt{LB}}/2 \cdot (c(A_2) - c(A_1))} \\
&\geq r(\boldsymbol{X}^*_{\mathcal{I}}) - (1 - \eta_{\texttt{LP}}) \cdot \frac{r(\boldsymbol{X}^*_{\mathcal{I}}) + \epsilon \cdot c^2_{\texttt{LB}}/2 \cdot (r(A_2) - r(A_1))}{1 + \epsilon \cdot c^2_{\texttt{LB}}/2} \\
&\geq \tfrac{\eta_{\texttt{LP}}}{2} \cdot r(\boldsymbol{X}^*_{\mathcal{I}}) \geq \epsilon \cdot \tfrac{c^3_{\texttt{LB}}}{2}.
\end{aligned}$$

The last two inequality follows from Assumption 3.8-(3.8), the value of $\eta_{\texttt{LP}}$ and the fact that $\epsilon = \frac{c_{\texttt{LB}}}{\sqrt{T}}$, respectively. Combining the two cases we get the claim.

F Proof of Theorem 3.9(b): $\sqrt{T}$ lower bound for $d > 2$

We first show that for any given instance $\mathcal{I}_0$, for a given $0 < \delta_1 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ we can obtain a δ_1-perturbation of this instance, denoted by $\mathcal{I}'_0$, that satisfies Eq. (3.15). Given instance $\mathcal{I}_0$ we construct the δ_1-perturbation as follows. We construct instance $\mathcal{I}'_0$ by decreasing the mean consumption on arm A_i and resource j by ζ^j_1. We keep the mean rewards the same. Let $\boldsymbol{X}$ denote the optimal solution to instance $\mathcal{I}$. As a notation we denote the matrix $\boldsymbol{C} \in [0,1]^{d \times 3}$ as the matrix of mean consumption. Let $\boldsymbol{B}$ denote the sub-matrix of $\boldsymbol{C}$ such that, $\boldsymbol{X}$ satisfies the constraints in the scaled LP (C.1) with equality. Thus, we have $\boldsymbol{C} \cdot \boldsymbol{X} = \boldsymbol{b}$, where every co-ordinate of b is $\frac{B(1-\eta_{\texttt{LP}})}{T}$. Thus, the perturbation is equivalent to perturbing the vector $\boldsymbol{b}$, such that the j^{th} entry has an additive perturbation of ζ^j. From Proposition 3.1 in [46], this linear program has a non-degenerate primal optimal solution, in the sense that it satisfies Eq. (3.15).

Next, we show that given an instance $\mathcal{I}'_0$ we can obtain a δ_2 perturbation of $\mathcal{I}'_0$ for a given $0 < \delta_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$, such that the consumption vectors are linearly independent. Define a random matrix

$\boldsymbol{D} \in [-\zeta_2, \zeta_2]^{d\times 3}$ such that every entry in $\boldsymbol{D}$ is generated uniformly at random from the set $[-\zeta_2, \zeta_2]$. We claim that the vectors $\boldsymbol{c}_j - \boldsymbol{d}_j$ are all linearly independent, where $\boldsymbol{d}_j$ is the j^{th} row of $\boldsymbol{D}$ with probability at least 0.6. In other words, decreasing each of the mean consumption by a uniformly random value chosen from the set $[-\zeta_2, \zeta_2]$ implies that there exists a realization of $\boldsymbol{D}$ such that the vectors $\boldsymbol{c}_j - \boldsymbol{d}_j$ are all linearly independent.

The proof of this claim proceeds as follows. As before define $\boldsymbol{C} \in [0,1]^{d\times 3}$ to be the matrix of mean consumption. From definition of linear independence we need to show that the smallest singular value of the matrix $\boldsymbol{C} - \boldsymbol{D}$ is non-zero. Note that every entry in the matrix $\boldsymbol{C} - \boldsymbol{D}$ is chosen independently. Thus, using the bound on the probability of singularity in Theorem 2.2 of [21] we have that the probability that the smallest singular value is 0 is at most $\frac{1}{2\sqrt{2}}$. Thus, with probability at least $1 - \frac{1}{2\sqrt{2}} > 0.6$ we have that the matrix $\boldsymbol{C} - \boldsymbol{D}$ is singular.

Thus, for $\delta := \delta_1 + \delta_2$, we have that there exists a δ-perturbed instance $\widetilde{\mathcal{I}}_0$, that satisfies all the assumptions in 3.8 and linear independence condition required in the premise of Theorem 3.10.

G Simple regret: proof of Theorem 4.1

For convenience, let us restate the theorem:

Theorem. *Assume $B \geq \Omega(T)$ and $\eta_{\text{LP}} \leq \frac{1}{2}$. With probability at least $1 - O(T^{-3})$, for each $\epsilon > 0$, there are at most $N_\epsilon = \mathcal{O}\left(\frac{K}{\epsilon^2}\log KTd\right)$ rounds t such that $\texttt{OPT}_{\texttt{DP}}/T - r(\boldsymbol{X}_t) \geq \epsilon$.*

The proof consists of two major steps: we argue about confidence sums, and we upper-bound simple regret in terms of the confidence radius.

G.1 Confidence sums

The following arguments depend only on the definition of the confidence radius, and work for any algorithm ALG. Suppose in each round t, this algorithm chooses a distribution $\boldsymbol{Y}_t$ over arms and samples arm a_t independently $\boldsymbol{Y}_t$. We upper-bound the number of rounds t with large $\text{Rad}_t(\boldsymbol{Y}_t)$:

Lemma G.1. *Fix the threshold $\theta_0 > 0$, and let S be the set of all rounds $t \in [T]$ such that $\text{Rad}_t(\boldsymbol{Y}_t) \geq \theta_0$. Then $|S| \leq \mathcal{O}\left(\theta_0^{-2} \cdot K\log(KdT)\right)$ with probability at least $1 - O(T^{-3})$.*

To prove the lemma, we study *confidence sums*: for a subset $S \subset [T]$ of rounds, define

$$W_{\texttt{act}}(S) := \textstyle\sum_{t\in S} \text{Rad}_t(a_t) \qquad \textit{(action-confidence sum of ALG)},$$
$$W_{\texttt{dis}}(S) := \textstyle\sum_{t\in S} \text{Rad}_t(\boldsymbol{Y}_t) \qquad \textit{(distribution-confidence sum of ALG)}.$$

First, a standard argument (*e.g.,* implicit in [10], see Section G.4) implies that

$$W_{\texttt{act}}(S) \leq \mathcal{O}\left(\sqrt{K\,|S|\,C_{\texttt{rad}}} + K\cdot\ln|S|\cdot C_{\texttt{rad}}\right) \quad \text{for any fixed subset } S\subset[T]. \tag{G.1}$$

Second, note that $W_{\texttt{dis}}(S)$ is close to $W_{\texttt{act}}(S)$: for any fixed subset $S \subset [T]$,

$$|W_{\texttt{dis}}(S) - W_{\texttt{act}}(S)| \leq \mathcal{O}(\sqrt{|S|\,\log T}) \quad \text{with probability at least } 1 - T^{-3}. \tag{G.2}$$

This is by Azuma-Hoeffding inequality, since $(\,\text{Rad}_t(a_t) - \text{Rad}_t(\boldsymbol{Y}_t) :\ t \in S\,)$ is a martingale difference sequence. We extend this observation to *random* sets S. A random set $S \subset [T]$ is called *time-consistent* if the event $\{t \in S\}$ does not depend on the choice of arm a_t or anything that happens afterwards, for each round t. (But it *can* depend on the choice of distribution $\boldsymbol{Y}_t$.)

Claim G.2. *For any any time-consistent random set $S \subset [T]$,*

$$|W_{\texttt{dis}}(S) - W_{\texttt{act}}(S)| \leq \mathcal{O}\left(\sqrt{|S|\,\log T} + \log T\right) \quad \textit{with probability at least } 1 - T^{-3}. \tag{G.3}$$

Proof. By definition of time-consistent set, for each round t,

$$\mathbb{E}[\mathbf{1}_{\{t\in S\}}\cdot\text{Rad}_t(a_t) \mid (\boldsymbol{Y}_1, a_1)\,,\ \ldots\ ,(\boldsymbol{Y}_{t-1}, a_{t-1}), \boldsymbol{Y}_t] = \mathbf{1}_{\{t\in S\}}\cdot\text{Rad}_t(\boldsymbol{Y}_t).$$

Thus, $\mathbf{1}_{\{t\in S\}}\,\text{Rad}_t(a_t) - \text{Rad}_t(\boldsymbol{Y}_t)$, $t \in [T]$ is martingale difference sequence. Claim G.2 follows from a concentration bound from prior work (Theorem D.3). □

We complete the proof of Lemma G.1 as follows. Fix $\delta > 0$. Since S is a time-consistent random subset of $[T]$, by Eq. (G.1) and Claim G.2, with probability at least $1-\delta$ it holds that

$$\theta_0 \cdot |S| \le W_{\texttt{dis}}(S) \le \mathcal{O}\left(\sqrt{|S|KC_{\texttt{rad}}} + K\,C_{\texttt{rad}} + \sqrt{|S|\,\log T} + \log T\right).$$

We obtain the Lemma by simplifying and solving this inequality for $|S|$.

G.2 Connecting LP-gap and the confidence radius

In what follows, let $B_{\texttt{sc}} = B(1-\eta_{\texttt{LP}})$ be the budget in the rescaled LP.

Lemma G.3. *Fix round $t \in [T]$, and assume the "clean event" in (2.7). Then*

$$G_{LP}(\boldsymbol{X}_t) \le (2 + {}^{T}/_{B_{\texttt{sc}}})\,\mathrm{Rad}_t(\boldsymbol{X}_t).$$

Proof. Let $\alpha := B_{\texttt{sc}}/T$. For any distribution $\boldsymbol{X}$, let

$$V_+(\boldsymbol{X}) := {}^{B_{\texttt{sc}}}/_{T}\ \cdot r(\boldsymbol{X}) / \max_{j\in[d]} c_j^-(\boldsymbol{X}).$$

denote the value of $\boldsymbol{X}$ in the optimistic LP (2.6), after proper rescaling. Let $\boldsymbol{X}^*$ be an optimal solution to the (original) LP (2.2). Then

$$G_{\texttt{LP}}(\boldsymbol{X}_t) = V(\boldsymbol{X}^*) - V(\boldsymbol{X}_t) - V_+(\boldsymbol{X}_t) + V_+(\boldsymbol{X}_t). \tag{G.4}$$

Since $V_+(\boldsymbol{X}_t)$ is the optimal solution to the optimistic LP (2.6),

$$V_+(\boldsymbol{X}_t) \ge V_+(\boldsymbol{X}^*).$$

Moreover, since $\boldsymbol{X}^*$ is feasible to the optimistic LP (2.6) with the scaled budget $B_{\texttt{sc}}$,

$$V_+(\boldsymbol{X}^*) \ge V(\boldsymbol{X}^*).$$

It follows that Eq. (G.4) an be upper-bounded as

$$G_{\texttt{LP}}(\boldsymbol{X}_t) \le V_+(\boldsymbol{X}_t) - V(\boldsymbol{X}_t). \tag{G.5}$$

We will now upper-bound the right-hand side in the above. Denote

$$c_{\max}(\boldsymbol{X}_t) := \max_{j\in[d]} \sum_{a\in[K]} c_{j,t}(a) X_t(a)$$

$$c^-_{\max}(\boldsymbol{X}_t) := \max_{j\in[d]} \sum_{a\in[K]} c^-_{j,t}(a) X_t(a).$$

By definition of the value of a linear program, we can continue Eq. (G.5) as follows:

$$\begin{aligned} G_{\texttt{LP}}(\boldsymbol{X}_t) &\le V_+(\boldsymbol{X}_t) - V(\boldsymbol{X}_t) \\ &\le \alpha \cdot \frac{\hat{r}(\boldsymbol{X}_t) + \mathrm{Rad}_t(\boldsymbol{X}_t)}{c^-_{\max}(\boldsymbol{X}_t)} - \alpha \cdot \frac{r(\boldsymbol{X}_t)}{c_{\max}(\boldsymbol{X}_t)}. \end{aligned} \tag{G.6}$$

Under the clean event in Eq. (2.7), we continue Eq. (G.6) as follows:

$$\le \alpha \left(\frac{2\,\mathrm{Rad}_t(\boldsymbol{X}_t) + r(\boldsymbol{X}_t)}{c^-_{\max}(\boldsymbol{X}_t)} - \frac{r(\boldsymbol{X}_t)}{c_{\max}(\boldsymbol{X}_t)} \right). \tag{G.7}$$

Since time is one of the resources, $c^-_{\max}(\boldsymbol{X}_t) \ge \frac{B_{\texttt{sc}}}{T}$. Thus, we continue Eq. (G.7) as follows:

$$\begin{aligned} &\le 2\,\mathrm{Rad}_t(\boldsymbol{X}_t) + \alpha r(\boldsymbol{X}_t) \left(\frac{1}{c^-_{\max}(\boldsymbol{X}_t)} - \frac{1}{c_{\max}(\boldsymbol{X}_t)} \right) \\ &= 2\,\mathrm{Rad}_t(\boldsymbol{X}_t) + \alpha r(\boldsymbol{X}_t) \left(\frac{\mathrm{Rad}_t(\boldsymbol{X}_t)}{c^-_{\max}(\boldsymbol{X}_t) \cdot c_{\max}(\boldsymbol{X}_t)} \right) \\ &\le 2\,\mathrm{Rad}_t(\boldsymbol{X}_t) + \frac{\mathrm{Rad}_t(\boldsymbol{X}_t)}{c^-_{\max}(\boldsymbol{X}_t)} \qquad \text{(G.8)} \\ &\le \left(2 + \tfrac{T}{B_{\texttt{sc}}}\right) \mathrm{Rad}_t(\boldsymbol{X}_t) \qquad \text{(G.9)} \end{aligned}$$

Eq. (G.8) uses the fact that $\alpha \frac{r(\boldsymbol{X}_t)}{c_{\max}(\boldsymbol{X}_t)} \le \frac{B}{T} \frac{r(\boldsymbol{X}_t)}{c_{\max}(\boldsymbol{X}_t)} = V(\boldsymbol{X}_t) \le 1$. Eq. (G.9) uses the fact that time is one of the resources and thus, $c^-_{\max}(\boldsymbol{X}_t) \ge \frac{B_{\texttt{sc}}}{T}$. □

G.3 Finishing the proof of Theorem 4.1

Claim G.4. *Fix round t, and assume the "clean event" in (2.7). Then*

$$\texttt{OPT}_{\texttt{DP}}/T - r(\boldsymbol{X}_t) \leq G_{\texttt{LP}}(\boldsymbol{X}_t) + \eta_{\texttt{LP}}.$$

Proof. By (2.7) and because $\boldsymbol{X}_t$ is the solution to the optimistic LP, we have

$$\max_{j\in d} c_j(\boldsymbol{X}_t) \geq \max_{j\in d} c_j^-(\boldsymbol{X}_t) = {}^{B}\!/_{T}\,(1-\eta_{\texttt{LP}}).$$

It follows that $r(\boldsymbol{X}_t) \geq V(\boldsymbol{X}_t)(1-\eta_{\texttt{LP}})$. Finally, we know that $\texttt{OPT}_{\texttt{LP}} \geq \texttt{OPT}_{\texttt{DP}}/T$. □

Condition on (2.7), and the high-probability event in Lemma G.1. (Take the union bound in Lemma G.1 over all thresholds $\theta_0 \geq 1/\sqrt{T}$, *e.g.,* over an exponential scale.) Fix $\epsilon > 0$. By Claim G.4 and Lemma G.3, any round t with simple regret at least ϵ satisfies

$$\epsilon \leq \texttt{OPT}_{\texttt{DP}}/T - r(\boldsymbol{X}_t) \leq \eta_{\texttt{LP}} + (2 + {}^{T}\!/_{B_{\texttt{sc}}})\,\mathrm{Rad}_t(\boldsymbol{X}_t).$$

Therefore, $\mathrm{Rad}_t(\boldsymbol{X}_t) \geq \theta_0$, where $\theta_0 = \frac{\epsilon-\eta_{\texttt{LP}}}{(2+T/B_{\texttt{sc}})} \geq \Theta(\epsilon)$ when $\epsilon \geq 2\eta_{\texttt{LP}}$. Now, the theorem follows from Lemma G.1. Note, when $\epsilon < 2\eta_{\texttt{LP}}$, then the total number of rounds in the theorem is larger than T and hence not meaningful.

G.4 The standard confidence-sum bound: proof of Eq. (G.1)

Let us prove Eq. (G.1) for the sake of completeness. By definition of $\mathrm{Rad}_t(a_t)$ from Eq. (2.8),

$$\mathrm{Rad}_t(a_t) = f(n) := \min\left(1,\ \sqrt{C_{\texttt{rad}}/n} + C_{\texttt{rad}}/n\right),$$

where $N_t(a)$ is the number of times arm a was chosen before round t. Therefore:

$$\begin{aligned}\sum_{t\in S} \mathrm{Rad}_t(a_t) &\leq \sum_{a\in[K]} \sum_{n=1}^{|S|/K} f(n)\\ &\leq \sum_{a\in[K]} \int_{x=1}^{|S|/K} f(x)\,\mathrm{d}x \leq 3\left(\sqrt{K|S|\,C_{\texttt{rad}}} + K\cdot\ln|S|\cdot C_{\texttt{rad}}\right).\end{aligned}$$

H Reduction from BwK to bandits

We extend our results to any problem which can be cast as a special case of BwK and admits an upper bound on action-confidence sums, in the style of (G.1), for a suitably defined confidence radius.

To state the general result, let us define an abstract notion of "confidence radius". For each round t, a *formal confidence radius* is a mapping $\mathrm{Rad}_t(a)$ from algorithm's history and arm a to $[0,1]$ such that with probability at least $1 - O(T^{-4})$ it holds that

$$|r(a) - \hat{r}_t(a)| \leq \mathrm{Rad}_t(a) \quad \text{and} \quad |c_j(a) - \hat{c}_{j,t}(a)| \leq \mathrm{Rad}_t(a)$$

for each resource j, where $\hat{r}_t(a)$ and $\hat{c}_{j,t}(a)$ denote average reward and resource consumption, as defined in Eq. (B.3). Such $\mathrm{Rad}_t(a)$ induces a version of UcbBwK with confidence bounds

$$r_t^+(a) = \min(1, \hat{r}_t(a) + \mathrm{Rad}_t(a)\ ) \quad \text{and} \quad c_{j,t}^-(a) = \max(\ 0, \hat{c}_{j,t}(a) - \mathrm{Rad}_t(a)\ ).$$

We allow the algorithm to observe auxiliary feedback before and/or after each round, depending on a particular problem formulation, and this feedback may be used to compute the confidence radii.

We replace Eq. (G.1) with a generic bound on the action-confidence sum, for some β that can depend on the parameters in the problem instance, but not on S:

$$\textstyle\sum_{t\in S} \mathrm{Rad}_t(a_t) \leq \sqrt{|S|\,\beta}, \quad \text{for any algorithm and any subset } S \subset [T]. \tag{H.1}$$

Theorem H.1. *Consider an instance of BwK with time horizon T. Let $\mathrm{Rad}_t(\cdot)$ be a formal confidence radius which satisfies (H.1) for some β. Consider the induced algorithms UcbBwK and PrunedUcbBwK with rescaling parameter $\eta_{\texttt{LP}} = \frac{2}{B}\sqrt{\beta T}$.*

(i) Both algorithms obtain regret $\texttt{OPT}_{\texttt{DP}} - \mathbb{E}[\texttt{REW}] \leq O(\sqrt{\beta T})(1 + \texttt{OPT}_{\texttt{DP}}/B)$.
(ii) Theorem 3.2 holds with $\Psi = \beta\, G_{\texttt{LAG}}^{-2}$ *and regret* $\mathcal{O}\left(\beta\, G_{\texttt{LAG}}^{-1}\right)$ *in part (ii).*
(iii) Theorem 4.1 holds with $N_\epsilon = \mathcal{O}\left(\beta\, \epsilon^{-2}\right)$.

Proof Sketch For part (i), the analysis in [3] explicitly relies on (G.1). For part (ii), we modify the proof of Theorem 3.2 so as to use (G.1) instead of Claim 3.4. For part (iii), our proof of Theorem 4.1 uses (G.1) explicitly. In all three parts, we replace (G.1) with (H.1), and trace how the latter propagates through the respective proof. ■

We apply this general result to three specific scenarios: linear contextual bandits with knapsacks (LinCBwK) [5], combinatorial semi-bandits with knapsacks (SemiBwK) [49], and multinomial-logit bandits with knapsacks (MnlBwK) [26]. In all three applications, the confidence-sum bound (H.1) is implicit in prior work on the respective problem without resources. The guarantees in part (i) match those in prior work referenced above, up to logarithmic factors, and are optimal when $B = \Omega(T)$; in fact, we obtain an improvement for MnlBwK. Parts (ii) and (iii) – the results for logarithmic regret and simple regret – did not appear in prior work.

H.1 Linear Contextual Bandits with Knapsacks (LinCBwK)

In *Contextual Bandits with Knapsacks* (CBwK), we have K actions, d resources, budget B and time horizon T, like in BwK, and moreover we have a set $\mathcal{X}$ of possible contexts. At each round $t \in [T]$, the algorithm first obtains a context $\boldsymbol{x}_t \in X$. The algorithm then chooses an action $a_t \in [K]$ and obtains an outcome $\boldsymbol{o}_t(a_t) \in [0,1]^{d+1}$ like in BwK. The tuple $(\, \boldsymbol{x}_t; \boldsymbol{o}_t(a) : a \in [K]\,)$ is drawn independently from some fixed but unknown distribution. The algorithm continues until some resource, including time, is exhausted. One compares against a given a set Π of *policies*: mappings from contexts to actions. We can formally interpret CBwK as an instance of BwK in which actions correspond to policies in Π. This interpretation defines the benchmarks $\texttt{OPT}_{\texttt{DP}}$ and $\texttt{OPT}_{\texttt{FD}}$ that we compete with.

LinCBwK is a special case of CBwK in which the context space is $\mathcal{X} = [0,1]^{K\times m}$, for some parameter $m \in \mathbb{N}$, so that each context $\boldsymbol{x}_t$ is in fact a tuple $\boldsymbol{x}_t = (\, \boldsymbol{x}_t(a) \in [0,1]^m : a \in [K]\,)$. We have a linearity assumption: for some unknown matrix $\boldsymbol{W}_* \in [0,1]^{m\times(d+1)}$ and each arm $a \in [K]$,

$$\mathbb{E}\left[\, \boldsymbol{o}_t(a) \mid \boldsymbol{x}_t(a) \,\right] = \boldsymbol{W}_*^{\mathrm{T}} \cdot \boldsymbol{x}_t(a).$$

The policy set Π consists of all possible policies.

Linear contextual bandits, studied in prior work [*e.g.,* 9, 29, 43, 27, 2], is the special case without resources. Much of the complexity of linear contextual bandits (resp., LinCBwK) is captured by the special case of of *linear bandits* (resp., *linear BwK*) where the context is the same in each round.

The general theme in the work on linear bandits (contextual or not) to replace the dependence on the number of arms K in the regret bound with the dependence on the dimension m and, if applicable, avoid the dependence on $|\Pi|$. This is what we accomplish, too.

Corollary H.2. *For LinCBwK, Theorem H.1 holds with* $\beta = \mathcal{O}(m^2 d^2 \log(mTd))$.

Proof. Combining Lemma 13 of [9] and Theorem 2 of [1], it follows that the confidence-sum bound Eq. (H.1) holds with $\beta = \mathcal{O}(m^2 d^2 \log mTd)$. □

H.2 Combinatorial Semi-bandits with Knapsacks (SemiBwK)

SemiBwK is a version of BwK, where actions correspond to subsets of some fixed ground set $[N]$ (whose elements are called *atoms*). There is a fixed family $\mathcal{F} \subset 2^{[N]}$ of feasible actions. In each round t, the algorithm chooses a subset $A_t \in \mathcal{F}$ and observes the outcome $\boldsymbol{o}_t(a) \in [0, 1/n]^d$ for each atom $a \in A_t$, where $n = \max_{A\in\mathcal{F}} |A|$. The outcome for a given subset $A \in \mathcal{F}$ is defined as the sum

$$\boldsymbol{o}_t(A) = \textstyle\sum_{a\in A} \boldsymbol{o}_t(a) \in [0,1]^{d+1}. \tag{H.2}$$

The outcome matrix $(\, \boldsymbol{o}_t(a) : a \in [N]\,)$ is drawn independently from some fixed but unknown distribution. The algorithm continues until some resource, including time, is exhausted.

Combinatorial semi-bandits, the problem studied in prior work [*e.g.,* 25, 40, 39], is the special case without resources. Note that the number of feasible actions can be exponential in N. The general

theme in this line of work is to replace the dependence on $|\mathcal{F}|$ in the regret bound with the dependence on N, or, even better, on n. We extend this to `SemiBwK`.

Corollary H.3. *For `SemiBwK`, Theorem H.1 holds with* $\beta = \mathcal{O}(n \log(NdT))$.

Proof. Using Lemma 4 in [60] we immediately obtain the confidence-sum bound Eq. (H.1) with $\beta = n \log KdT$. □

H.3 Multinomial-logit Bandits with Knapsacks (`MnlBwK`)

In the `MnlBwK` problem, the setup starts like in `SemiBwK`. There is a ground set of N *atoms*, and a fixed family $\mathcal{F} \subset 2^{[N]}$ of feasible actions. In each round, each atom a has an outcome $\boldsymbol{o}_t(a) \in [0,1]^{d+1}$, and the outcome matrix $(\boldsymbol{o}_t(a) : a \in [N])$ is drawn independently from some fixed but unknown distribution. The aggregate outcome is formed in a different way: when a given subset $A_t \in \mathcal{F}$ is chosen by the algorithm in a given round t, at most one atom $a_t \in A_t$ is chosen stochastically by "nature", and the aggregate outcome is then $\boldsymbol{o}_t(A_t) := \boldsymbol{o}_t(a)$; otherwise, the algorithm skips this round. A common interpretation is that the atoms correspond to products, the chosen action $A_t \in \mathcal{F}$ is the bundle of products offered to the customer, and at most one product from this bundle is actually purchased. As usual, the algorithm continues until some resource (incl. time) is exhausted.

The selection probabilities are defined via the multinomial-logit model. For each atom a there is a hidden number $v_a \in [0,1]$, interpreted as the customers' valuation of the respective product, and the

$$\Pr[\text{ atom } a \text{ is chosen } \mid A_t] = \begin{cases} \frac{v_a}{1+\sum_{a' \in A_t} v_{a'}} & \text{if } a \in A_t \\ 0 & \text{otherwise.} \end{cases}$$

The set $\mathcal{F}$ of possible bundles is

$$\mathcal{F} = \{ A \subset [N] : \ \boldsymbol{M} \cdot x(A) \leq \boldsymbol{b} \},$$

for some (known) totally unimodular matrix $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ and a vector $\boldsymbol{b} \in \mathbb{R}^N$, where $x(A) \in \{0,1\}^N$ represents set A as a binary vector over atoms.

Multinomial-logit bandits, the problem studied in prior work [*e.g.,* 7, 48, 51, 24], is the special case without resources. We derive the following corollary from the analysis of MNL-bandits in Agrawal et al. [7], which analyzes the confidence sum for the v_a's.

Corollary H.4. *Consider `MnlBwK` and denote* $V := \sum_{a \in [N]} v_a$. *Theorem H.1 holds with*

$$\beta = \mathcal{O}\left(\left(\frac{\ln T}{\ln(1+1/V)} \right)^2 \left(N\sqrt{\ln(NT)} + \ln(NT) \right) \right) = \widetilde{O}\left(N^3 \right).$$

Proof. The proof is implicit in the analysis in Agrawal et al. [7]. As in their paper, let n_ℓ denote the number of time-steps in phase ℓ. Let $V_\ell = \sum_{a \in S_\ell} v_a$. Recall that n_ℓ is a geometric random variable with mean $\frac{1}{1+V_\ell}$. Using Chernoff-Hoeffding bounds we obtain that with probability at least $1 - \frac{1}{T^2}$, $n_\ell \leq \frac{\ln T}{\ln(1+1/v_\ell)}$.

Consider a random subset S. Summing the LHS and RHS in Lemma 4.3, we get that $\sum_{t \in S} \mathrm{Rad}_t(a_t) \leq \sum_{a \in [N]} \sum_{\ell : t \in \mathcal{T}_a(\ell)} \tilde{R}_a(S_\ell)$. Using Lemma 4.3 in [7] we have, $\sum_{a \in [N]} \sum_{\ell : t \in \mathcal{T}_a(\ell)} \tilde{R}_a(S_\ell) \leq \sum_{a \in [N]} \sum_{\ell : t \in \mathcal{T}_a(\ell)} n_\ell \sqrt{\frac{v_a \ln \sqrt{NT}}{T_a(\ell)}} + \frac{\ln \sqrt{NT}}{T_a(\ell)}$. Note that $v_a \leq 1$. Using the upper bound on n_ℓ derived above combined with the argument used to obtain (A.19) in [7] we get the desired value of β. □

The worst-case regret bound from Corollary H.4 improves over prior work [26]. In particular, consider the worst-case dependence on N, the number of atoms. Our regret bound scales as $N^{3/2}$, whereas the regret bound in [26] scales as $N^{7/2}$ (while both scale as $\sqrt{T}$).

H.4 Computational issues

We do not provide a generic computationally efficient implementation for UcbBwK in our reduction. The algorithm constructs and solves a linear program in each round, with one variable per arm in the reduction. So, even if the regret is fairly small, the number of LP variables may be very large: indeed, it may be exponential in the number of atoms in SemiBwK and MnlBwK, arbitrarily large compared to the other parameters in linear BwK, or even infinite as in LinCBwK. The corresponding LPs have a succinct representation in all these applications, but we do not provide a generic implementation. However, such (or very similar) linear programs may be computationally tractable via application-specific implementations, and indeed this is the case in LinCBwK [5] and SemiBwK [49]. In the prior work on MnlBwK [26], the $\sqrt{T}$-regret algorithm is not computationally efficient, same as ours; there is, however, a computationally efficient algorithm with regret $T^{2/3}$.