Approximately Stationary Bandits with Knapsacks

Giannis Fikioris   GFIKIORIS@CS.CORNELL.EDU
Éva Tardos   EVA.TARDOS@CORNELL.EDU
Cornell University

Abstract

Bandits with Knapsacks (BwK), the generalization of the *Multi-Armed Bandits* problem under global budget constraints, has received a lot of attention in recent years. It has numerous applications, including dynamic pricing, repeated auctions, ad allocation, network scheduling, etc. Previous work has focused on one of the two extremes: *Stochastic* BwK where the rewards and consumptions of the resources of each round are sampled from an i.i.d. distribution, and *Adversarial* BwK where these parameters are picked by an adversary. Achievable guarantees in the two cases exhibit a massive gap: No-regret learning is achievable in the stochastic case, but in the adversarial case only competitive ratio style guarantees are achievable, where the competitive ratio depends either on the budget or on both the time and the number of resources. What makes this gap so vast is that in Adversarial BwK the guarantees get worse in the typical case when the budget is more binding. While "best-of-both-worlds" type algorithms are known (single algorithms that provide the best achievable guarantee in each extreme case), their bounds degrade to the adversarial case as soon as the environment is not fully stochastic.

Our work aims to bridge this gap, offering guarantees for a workload that is not exactly stochastic but is also not worst-case. We define a condition, *Approximately Stationary* BwK, that parameterizes how close to stochastic or adversarial an instance is. Based on these parameters, we explore what is the best competitive ratio attainable in BwK. We explore two algorithms that are oblivious to the values of the parameters but guarantee competitive ratios that smoothly transition between the best possible guarantees in the two extreme cases, depending on the values of the parameters. Our guarantees offer great improvement over the adversarial guarantee, especially when the available budget is small. We also prove bounds on the achievable guarantee, showing that our results are approximately tight when the budget is small.

Keywords: Bandits with Knapsacks, Best of both worlds, Adversarial, Approximately Stationary

1. Introduction

Bandits with Knapsacks (BwK) was first introduced in Badanidiyuru et al. (2013) and models a natural extension of the *Multi-Armed Bandit* (MAB) problem. In MAB a player repeatedly chooses one of many actions, each providing an unknown reward. To maximize her total reward, a player needs to balance exploration and exploitation while picking her actions. In the budgeted version of the problem (BwK), the player has the same objective but also needs to be mindful of different resources: each action consumes some amount of each resource; if any resource is depleted the player cannot get any more rewards. BwK was initially formulated inspired by numerous practical problems where a player wants to maximize her reward with constraints: participating in repeated auctions, dynamic pricing, ad allocation, network routing/scheduling, etc.

Previous work on BwK has focused on two extreme cases of the problem. First, in *Stochastic* BwK the environment (rewards and consumptions) in each round is sampled from a distribution

identical and independent of other rounds. Second, in *Adversarial* `BwK` the environment is picked each round by an adversary. Unlike `MAB`, in `BwK` there is a clear dichotomy in guarantees between the two cases. No-regret learning, i.e., additive error sublinear in the total number of rounds, is achievable in the stochastic case, but not in the adversarial case. Instead, work in Adversarial `BwK` focuses on bounding the achievable *competitive ratio*, i.e., the multiplicative error on the achievable reward. A line of work that tries to connect the adversarial and stochastic cases is "best-of-both-worlds" type results, where an algorithm achieves guarantees in both settings, without knowing if the environment is adversarial or stochastic, e.g., see Castiglioni et al. (2022). However, the guarantee offered degrades to the adversarial guarantee as soon as the setting is not fully stochastic.

In this work, we aim to bridge this vast gap between Stochastic and Adversarial `BwK` and offer performance guarantees that smoothly degrade depending on the deviation from stochasticity, extending the "best-of-both-words" style guarantees to cases between the two extremes. We call our framework *Approximately Stationary* `BwK`, which offers a natural interpolation between Stochastic and Adversarial `BwK`. Adversarial `BwK` is much harder because of the potential for *huge heterogeneity of environments between rounds*. In the Approximately Stationary `BwK` problem, we limit this heterogeneity by assuming that the change in *expected* rewards and consumptions of any arm is limited. We assume that the player is not aware of the parameters limiting the change while running the algorithm. A natural constraint is that if x_t is the expected reward of some action in round t, it must hold that $\min x_t \geq \sigma \max x_t$, with σ limiting the variability of the expectation. In Stochastic `BwK` it must hold that $\sigma = 1$ (actually, distributions are identical across rounds, not just expectations), and in Adversarial `BwK` it can be that $\sigma = 0$.

There are multiple settings where the value of σ is neither of the two extremes. Consider repeated auctions where every round a budget-limited player bids to win a certain item. If the player's value for the item and its price are independently and identically distributed across rounds, then the setting is completely stochastic, and $\sigma = 1$. However, this is rarely the case in practice. The distribution of values might change across rounds (e.g., seasonal differences) or the price might be controlled by other players' bids who change their strategy or by a central entity that lowers or raises the price. This means that $\sigma < 1$, but the values and prices are not adversarial or arbitrary, i.e., the variability of the price is limited. Our goal is to have an algorithm that will have the best guarantees given the value of σ, without knowing its actual value. In this paper, we show that it is possible to achieve performance that degrades gradually depending on the value σ. For the most interesting range of the player's budget, we obtain close to optimal performance for all values of σ without assuming that the player is aware of its value.

Overview of our results. We introduce our model, *Approximately Stochastic* `BwK`, in Section 3. Our model interpolates between Stochastic and Adversarial `BwK` by having two parameters that limit how much the expected value of the rewards and consumptions of any action can change across rounds. More specifically, the parameter $\sigma_{\mathtt{r}} \in [0, 1]$ limits how much the reward of any action can fluctuate across rounds: if $r_t(a)$ is the *expected reward* of action a in round t, we require that $\min_t r_t(a) \geq \sigma_{\mathtt{r}} \max_t r_t(a)$ for all actions a. Note that we apply this definition to the expected reward of an arm since even in Stochastic `BwK` the observed rewards of a single arm can range from 0 to 1. Similarly, the parameter $\sigma_{\mathtt{c}} \in [0, 1]$ limits the consumptions of any action across rounds: if $c_{t,i}(a)$ is the *expected consumption* of resource i by action a in round t, we require that $\min_t c_{t,i}(a) \geq \sigma_{\mathtt{c}} \max_t c_{t,i}(a)$ for all actions a and resources i. A sequence of rewards and consumptions that satisfy the above constraints is called $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-*stationary*. Aside from this restriction,

we make no assumptions, so we think that the adversary can be adaptive, i.e., pick the rewards and consumptions of a round based on the past rounds. We note that in Stochastic BwK the adversary is $(1, 1)$-stationary and in Adversarial BwK the adversary is $(0, 0)$-stationary. Our framework naturally generalizes "best-of-both-worlds" approaches, offering guarantees not only in fully stochastic and adversarial environments but also environments between the two extremes. We note that using two parameters instead of one that constrains both rewards and consumptions is valuable: a small $\sigma_{\mathbf{r}}$ has a much different effect on the achievable guarantees than a small $\sigma_{\mathbf{c}}$.

As is standard we assume without loss of generality that the rewards and consumptions are non-negative and bounded by 1 and every resource has budget B. In adversarial BwK the best competitive ratio depends on the player's average (or per round) budget. We will use $\rho = B/T$ to denote the player's per-round budget. Given that the consumptions are bounded by 1 each round, $\rho = 1$ means that the player is not budget limited; in more typical cases players have budgets only for a small fraction of the items available. Our goal is to design algorithms that guarantee a fraction of the optimal solution when the player's per-round budget is ρ and the adversary is $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-stationary. We denote this fraction with $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$. There are algorithms that have $\alpha_\rho(1, 1) = 1$ and $\alpha_\rho(0, 0) = \rho$, and these are best possible when ρ is a constant independent of the time horizon, but nothing is known for intermediate values. This effectively means that previous work can only guarantee $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}}) = \rho$ when $\sigma_{\mathbf{r}} < 1$ or $\sigma_{\mathbf{c}} < 1$. This is an enormous and unnatural gap, especially in the typical case when ρ is small. We study algorithms that are oblivious to the values of $\sigma_{\mathbf{r}}$ and $\sigma_{\mathbf{c}}$ and achieve a fraction $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$ of the optimal solution that is continuous and increasing in both arguments and satisfies $\alpha_\rho(1, 1) = 1$ and $\alpha_\rho(0, 0) = \rho$.

All of our guarantees are against an *adaptive adversary*, an adversary that is restricted to be $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-stationary, but beyond this restriction, is allowed to pick the distributions of rewards and consumptions of each round based on outcomes in previous rounds. This is in contrast to an *oblivious adversary*, who picks all the rewards and consumptions upfront. Allowing an adaptive adversary is important: it makes our guarantees apply when the algorithm is used in a multi-agent game setting, e.g. in repeated auctions, where prices depend on other agents' bids, who are all adaptive to the history of play. All the guarantees we present in this paper are against an adaptive adversary, which, to the best of our knowledge, are the first such guarantees for Adversarial BwK.

In Section 5 we present our first guarantee. We show that we can achieve a guarantee of $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}}) \geq \rho + \sigma_{\mathbf{r}}(\sigma_{\mathbf{c}} - \rho)^+$ (Theorem 4). The most interesting range of parameters is when ρ is small (and therefore the gap between Stochastic and Adversarial BwK is biggest) and $\sigma_{\mathbf{r}}\sigma_{\mathbf{c}}$ is much bigger than ρ. In this case, our guarantee becomes approximately $\sigma_{\mathbf{r}}\sigma_{\mathbf{c}}$. This is in stark contrast to the guarantee that previous work would suggest, ρ. An alternate, naive approach would be to use an algorithm assuming a fully stochastic setting, which would yield a guarantee of $\approx \sigma_{\mathbf{r}}^2\sigma_{\mathbf{c}}^2$ (see Section 5), two orders of magnitude smaller. Our results show that even if the player has a small budget, as long as the environment is not completely adversarial, good guarantees are achievable. We note that a small budget is indeed the most common case: Typical budgets are far from sufficient for all items, so ρ is small, and the environment is often less variable and independent of the player's budget. For example, in repeated auctions, we expect that expected item prices fluctuate across rounds but this fluctuation to not depend on the player's small budget.

In Section 6 we provide a bound on the achievable guarantees any algorithm can get. The upper bound of Theorem 6 on $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$ is using only one resource and an oblivious adversary. When ρ is much smaller than $\sigma_{\mathbf{r}}\sigma_{\mathbf{c}}^2$, Theorem 6 shows that $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}}) \lesssim \sigma_{\mathbf{r}}\sigma_{\mathbf{c}}$, making the result of Theorem 4 almost tight in that case. Theorem 6 also shows that when $\sigma_{\mathbf{r}} = O(\rho)$ then $\alpha_\rho(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}}) \leq O(\rho)$,

making Theorem 4 (and any algorithm with the guarantee $\alpha_\rho(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}}) \geq \rho$) tight up to constant factors. Another interesting corollary of our theorem is that $\alpha_\rho(0,0) \leq \rho$, i.e., that a ρ fraction of the optimal reward is the best possible in Adversarial BwK[1]. The bound of Theorem 6 relies on that any algorithm needs to conserve its budget because of the uncertainty of the future. Without knowing the values of $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$ it is not possible to utilize the reward of the optimal action in the initial rounds, in fear of the adversary being $(0,0)$-stationary.

Our first guarantee in Theorem 4 is close to optimal when $\sigma_{\mathtt{c}}$ is much bigger than ρ. When $\sigma_{\mathtt{c}}$ is similar to ρ the guarantee is very small: if $\sigma_{\mathtt{c}} \leq \rho$ the theorem only provides the adversarial guarantee of $\alpha_\rho(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}}) \geq \rho$. In Section 7 we offer better guarantees for the case when $\sigma_{\mathtt{c}}$ and ρ are small, but $\sigma_{\mathtt{r}}$ is larger. Our improved guarantee requires an additional condition to approximate stationarity: the total per-round change of the consumption of any action is sublinear in the total number of rounds. Under this assumption and $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationarity, we provide improved bounds in Theorem 8. The improvement is most impressive when there is only one resource (e.g., money), $\sigma_{\mathtt{c}}$ is close to ρ, and $\sigma_{\mathtt{r}}$ is much larger, in which case the guarantee is close to the best possible (see Theorem 6).

Theorem 4 is based on a simplified version of the algorithm in Castiglioni et al. (2022) (see Algorithm 1 in Section 4), while to achieve the guarantee of Theorem 8 we need to break the time horizon into smaller batches and restart the algorithm periodically (see Algorithm 2 in Section 7). The main technical lemmas leading to the guarantees in both cases follow an interesting and novel approach. They prove that the rewards of our algorithms have no-regret against the total reward of any action *across all rounds* but scaled down to observe the average consumption bound in each iteration. This property is not useful in Stochastic BwK, where one knows that the optimal arm has low consumption, or in Adversarial BwK, where the rewards of the optimal arm can become 0 after a certain round. In Approximately Stochastic BwK, where the reward of the optimal arm cannot become 0 (unless $\sigma_{\mathtt{r}} = 0$) but its consumption can increase, guaranteeing a fraction of the reward from the entire period is a very useful property. The auxiliary lemma we use to prove the improved guarantee of Theorem 8 scales down the reward of each round based on the consumption of the optimal arm *in that round*. This makes its guarantee stronger when $\sigma_{\mathtt{c}}$ is small compared to the lemma used for Theorem 4 where the reward needs to be scaled down using the maximum consumption across rounds.

Related work. There is a vast amount of literature on online learning and regret minimization; we refer the reader to textbooks like Hazan (2016) and Slivkins (2020) for background. The most commonly used algorithms for MAB are Hedge for full information feedback (Freund and Schapire (1997)) and EXP3 for bandit feedback (Auer et al. (2002)). A "best-of-both-worlds" type result providing guarantees for both stochastic and adversarial MAB was first proven in Bubeck and Slivkins (2012).

The BwK framework was first introduced in Badanidiyuru et al. (2013) in the stochastic setting. Following this work, various extensions have been studied including concave rewards and convex consumptions (Agrawal and Devanur (2014)), combinatorial semi-bandits with knapsacks (Sankararaman and Slivkins (2018)), contextual bandits (Badanidiyuru et al. (2014)), and negative consumptions that replenish the player's budget (Kumar and Kleinberg (2022)).

---

1. This bound was previously proven by Balseiro and Gur (2017) when the benchmark used for the optimal solution is the best sequence of actions. Our bound uses the best distribution of actions which is the standard benchmark in BwK since the previous one is attainable only in special cases.

The first guarantees for Adversarial BwK were by Immorlica et al. (2019) and then by Kesselheim and Singla (2020). The first work guarantees a $1/O(d \log T)$ fraction of the optimal solution and the second a $1/O(\log d \log T)$ fraction, where d is the number of resources. The second guarantee is tight when the average budget is $\rho = O(T^{-\varepsilon})$ for any constant $\varepsilon > 0$. Balseiro and Gur (2017) study repeated second-price auctions with budgets, a special case of BwK; they assume $\rho = \Omega(1)$, and get a "best-of-both-worlds" guarantee: a constant ρ fraction of the optimal solution in adversarial settings and no-regret in stochastic ones. They also prove that even in this restricted second-price setting, their adversarial result is tight when ρ is a constant and the comparator is the best sequence of actions, an extremely strong benchmark that an algorithm can be competitive against only in special cases of BwK, like second-price auctions. Castiglioni et al. (2022) generalize the previous "best-of-both-worlds" guarantee for general BwK. Rangi et al. (2019) study a variant of BwK where there is no time horizon and only one resource whose consumption is strictly positive; the game stops when that resource is depleted. Their variant is much easier than general BwK, evident by their guarantees of $\operatorname{poly} \log T$ and $O(\sqrt{T})$ regret bounds for the stochastic and adversarial cases, respectively; in general BwK these types of guarantees are not achievable.

There have been other models that interpolate between Stochastic and Adversarial MAB. Lykouris et al. (2018) study Stochastic MAB with Corruptions where the environment is stationary except for C rounds that are corrupted by an adversary; their no-regret guarantee interpolates between the stochastic and adversarial guarantees depending on how large C is. In Restless Bandits, e.g., studied by Tekin and Liu (2012) and Wang et al. (2020), the environment each round is generated by a Markov Chain whose state changes each round. Algorithms have no-regret when the Markov chain has size much smaller than T, making the environments of two rounds approximately independent if they are far enough apart. Both of these models interpolate between the adversarial and stochastic case but in a fundamentally different way than ours: Before round 1 both have an "expected environment" (the uncorrupted one in Stochastic MAB with Corruptions and the one generated by the stationary distribution of the Markov chain in Restless bandits) which the environments of most rounds are "close" to in expectation. Instead, our $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-adversary does not have to conform to this restriction and can vary environments significantly at each time step.

Previous work has considered MAB with close to stationarity constraints similar to ours. Gur et al. (2014) impose a constraint on the rewards of the actions similar to the one we consider in Section 7: they bound the total per-round change of the reward of any action by a parameter V. They prove regret bounds of order $\tilde{\Theta}(V^{1/3} T^{2/3})$ against an oblivious adversary. Besbes et al. (2015) generalize the previous results for general action spaces and convex reward functions. In contrast, having only this constraint for both the rewards and consumptions of the actions in BwK, even with $V = 1$, does not improve over Adversarial BwK: Theorem 6 satisfies the above constraint with $\sigma_{\mathbf{r}} = \sigma_{\mathbf{c}} = 0$ which shows that only a ρ fraction of the optimal solution is achievable.

No-regret guarantees have been studied for BwK under stationarity conditions similar the one above. Liu et al. (2022) design no-regret algorithms when the reward and consumptions of every action change slowly over time and remain close to their average value (with known bounds for these constraints). Concurrently and independently with our work, Slivkins et al. (2023) study stochastic *contextual bandits with linear constraints* (a generalization of contextual BwK) and design no-regret algorithms for this setting. They also consider environments that are mostly stationary with a small number of changes in the distribution of rewards and consumptions. In this environment, if the number of changes is not too big, they achieve no-regret against a benchmark similar to the one we use in Lemma 7 but make no guarantees against the standard benchmark we use in our theorems.

The results of both Liu et al. (2022) and Slivkins et al. (2023) are incomparable to the ones we make for Approximately Stationary BwK: even if $\sigma_{\mathtt{r}} = \sigma_{\mathtt{c}} = 1 - \varepsilon$ for some small constant ε, the variability in the environment can be too high to yield meaningful regret guarantees.

Online allocation problems are similar to BwK except that the player gets to observe the rewards and consumptions of a round *before* she picks an action for that round and so the only uncertainty comes from future rewards and consumptions. Devanur et al. (2019) prove no-regret bounds for this problem in the stochastic setting. They also study the adversarial case where their algorithm guarantees no-regret against a specific benchmark as long as their algorithm knows the value of this benchmark. More specifically, they compare against $\min_t \mathtt{OPT}_t$, where $\mathtt{OPT}_t$ is the maximum achievable reward if the rewards and consumptions of all rounds were distributed similarly to those of round t (which can be adversarially picked). Balseiro et al. (2020) study the same problem and provide guarantees similar to Castiglioni et al. (2022) in the stochastic and adversarial cases. Additionally, they study models that interpolate the two extremes, similar to Stochastic MAB with Corruptions and Restless Bandits. Their results, similar to Devanur et al. (2019), focus on getting no-regret guarantees in those models.

2. Bandits with Knapsacks

In this section, we formally define the Bandits with Knapsack (BwK) framework. Our notation and definitions are similar to (Slivkins, 2020, Chapter 10). We introduce some additional notation to help distinguish between the realized reward/consumptions of an action and their expectation.

There are T rounds, d *resources* (denoted with $[d]$), and a *budget* per resource which w.l.o.g. we assume is the same for every resource $i \in [d]$, B. We denote $\rho = B/T$. The player has a set of K *actions*, $[K]$. In every round $t \in [T]$ the adversary chooses $(d + 1)K$ random (and potentially dependent on previous rounds and of each other) variables: $R_t, C_{t,1}, \ldots, C_{t,d} : [K] \to [0, 1]$. For an action $a \in [K]$, $R_t(a)$ is the *reward* the player receives on round t if they play action a, and for a resource $i \in [d]$, $C_{t,i}(a)$ is the *consumption* of that resource in round t by a. Because consumptions are upper bounded by 1 we can assume that $\rho \leq 1$. As is standard, we also assume that there is an action, called the *null action*, with 0 reward and 0 consumption of every resource.

Every round t the player chooses a (potentially randomized) action $A_t \in [K]$ without any knowledge of the reward or consumptions of that round or their distribution. The game ends either after round T or in the round that any resource is depleted. We define $T_{\mathtt{A}}$ to be the last round the player receives a reward: $T_{\mathtt{A}} = \max \left\{ t \in [T] : \forall i \in [d], \sum_{\tau=1}^{t} C_{\tau,i}(A_\tau) \leq B \right\}$.

We denote with $\mathcal{H}_t$ the *history* up to round t. This includes the realization of the actions of the player up to round t, $A_1, \ldots, A_t$, as well as the realization of the rewards and consumptions up to round t, $R_1, \ldots, R_t$ and $\{C_{1,i}\}_i, \ldots, \{C_{t,i}\}_i$. We generally assume that the player has bandit knowledge, i.e., only knows the realization of rewards and consumptions of the actions she took in previous rounds. As we note later, our results can be improved if the player has full information knowledge.

If for every round t the distributions of the functions $R_t, C_{t,1}, \ldots, C_{t,d}$ are independent of $\mathcal{H}_{t-1}$ (which allows dependence on the algorithm used by the player but not the realization of its actions) then the adversary is called *oblivious*; otherwise, the adversary is called *adaptive*. Additionally, if the aforementioned functions are not time dependant, the adversary is called *stochastic*.

For any action $a \in [K]$, we denote with $r_t(a)$ the expected reward of action a in round t *conditioned on the history of the previous rounds*: $r_t(a) = \mathbb{E}\left[R_t(a)|\mathcal{H}_{t-1}\right]$. Similarly, we define the conditional expected cost of action a and resource i: $c_{t,i}(a) = \mathbb{E}\left[C_{t,i}(a)|\mathcal{H}_{t-1}\right]$.

3. Approximately Stationary Bandits with Knapsacks

In this section, we present our model, *Approximately Stationary* BwK. Our model interpolates between Stationary and Adversarial BwK, providing guarantees that smoothly improve as expectations change less across time or equivalently as the setting is less adversarial. This generalizes "best-of-both-worlds" results by providing guarantees for the whole spectrum, not just the extremes.

As we mentioned in the introduction, Adversarial BwK is hard because the rewards and consumptions of an arm can oscillate between extreme values. The issue with Adversarial BwK becomes apparent by looking at the impossibility results of Balseiro and Gur (2017) and Immorlica et al. (2019). Both have similar structure: there is only one action with positive reward every round. However, even if the player knows which action that is, she cannot fully utilize it, in case some other action has a much larger reward in a later round. Our model limits these extreme-case examples by constraining the expectation of the rewards and consumptions the adversary can pick. We focus on expected rewards and consumptions since even in Stochastic BwK the range of the realized rewards can be the entire interval $[0, 1]$. Our definition uses two parameters. The first, $\sigma_{\mathbf{r}}$, bounds the multiplicative difference in the maximum and minimum expected reward of any action. The second, $\sigma_{\mathbf{c}}$, similarly constrains the expected consumption of any resource by any action.

Definition 1 *An adversary in* BwK *is called* $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-*stationary if for any action* a, *resource* i, *and history* $\mathcal{H}_T$ *it holds that* $\min_t r_t(a) \geq \sigma_{\mathbf{r}} \max_t r_t(a)$ *and* $\min_t c_{t,i}(a) \geq \sigma_{\mathbf{c}} \max_t c_{t,i}(a)$.

Remark 2 *First, note that our definition bounds the relative variation of a sequence and not the absolute one, e.g.,* $\max_t r_t(a) - \min_t r_t(a) \leq \varepsilon$. *Bounding the relative size rather than the absolute difference makes sense, making the result invariant on the scale of the rewards. Second, we notice that if* $\sigma_{\mathbf{r}} = \sigma_{\mathbf{c}} = 0$ *then the setting is completely adversarial. If the setting is Stochastic* BwK, *we get* $\sigma_{\mathbf{r}} = \sigma_{\mathbf{c}} = 1$. *Third, since the reward constraint is applied to the expected rewards of an action given the history of the previous rounds, our adversary can be adaptive.*

4. Benchmarks and Algorithm

In this section, we present the benchmark that we use to compare the quality of our algorithm, as well as the algorithm that we use that provides our first guarantee.

Benchmark. The benchmark we use is the standard *best-fixed distribution of actions in hindsight*. Its reward $\mathrm{OPT}_{\mathrm{FD}}$ is equal to the reward of the best distribution of actions $A^* \in \Delta([K])$, up to the round when it runs out of budget. For simplicity of presentation, we define $\mathrm{OPT}_{\mathrm{FD}}$ using the expected rewards and consumptions r, c:

$$
\mathrm{OPT}_{\mathrm{FD}} = \max_{\substack{T^* \in [T] \\ A^* \in \Delta([K])}} \sum_{t=1}^{T^*} \mathbb{E}_{a \sim A^*}\left[r_t(a)\right]
$$

$$
\text{such that} \quad \sum_{t=1}^{T^*} \mathbb{E}_{a \sim A^*}\left[c_{t,i}(a)\right] \leq B, \quad \forall i \in [d]
\tag{1}
$$

We note that in $\mathbb{E}_{a \sim A^*}[\cdot]$ the expectation is taken only over the action $a \sim A^*$ and not any choices the player or adversary make, i.e., $\mathbb{E}_{a \sim A^*}[r_t(a)] = \sum_a \mathbb{P}[A^* = a] r_t(a)$, where $r_t(a)$ is the expected reward of action a for the actual history of the play, and not for the history of playing action distribution A^* each round. This means that $\mathtt{OPT_{FD}}$ depends on the realization of the random choices of the game, i.e., the player's and adversary's actions. This is similar to benchmarks in $\mathtt{MAB}$ with an adaptive adversary, where the optimal reward used as a comparator for no-regret depends on the actions the player takes. We denote with (T^*, A^*) the solution to the above optimization problem.

Optimization problem (1) is simplified when the expectations of the rewards and consumptions are the same every round $r_t(\cdot) = r(\cdot)$ and $c_{t,i}(\cdot) = c_i(\cdot)$ for all t, i. In this case, (1) becomes

$$\max_{A^* \in \Delta([K])} \quad T \mathop{\mathbb{E}}_{a \sim A^*}[r(a)] \qquad \text{such that} \quad T \mathop{\mathbb{E}}_{a \sim A^*}[c_i(a)] \leq B, \quad \forall i \in [d] \qquad (2)$$

where we can drop the dependence on T^* because of the null action: For every feasible solution $(\hat{T}, \hat{A})$ there is a feasible solution $(T, \hat{A}')$ with the same reward. More specifically, $\hat{A}'$ is the same distribution as $\hat{A}$ with probability $\hat{T}/T$ and the null action otherwise.

We use $\mathtt{REW}$ to denote the total reward of the player: $\mathtt{REW} = \sum_{t=1}^{T_A} R_t(A_t)$. We focus on high-probability bounds. The player has competitive ratio $\gamma \geq 1$ and regret $\mathtt{Reg}$ against $\mathtt{OPT_{FD}}$ with probability $1 - \delta$ if $\mathbb{P}\left[\mathtt{REW} \geq \frac{\mathtt{OPT_{FD}} - \mathtt{Reg}}{\gamma}\right] \geq 1 - \delta$.

As mentioned in the related work, in Adversarial $\mathtt{BwK}$ the player can guarantee competitive ratio $\gamma = \min\{1/\rho, O(\log T)\}$ and sublinear regret. Without any additional assumptions, this result is tight. By assuming that the adversary is $(\sigma_\mathbf{r}, \sigma_\mathbf{c})$-stationary, we prove greatly improved guarantees for the competitive ratio γ. Our guarantees provide a smooth interpolation between Adversarial and Stochastic $\mathtt{BwK}$.

Algorithm based on Lagrangian maximization/minimization. Next, we present a simplified version of the algorithm of Castiglioni et al. (2022) that achieves a "best-of-both-worlds" guarantee: competitive ratio of 1 in stochastic environments and $1/\rho$ in adversarial ones. Our first guarantee against a $(\sigma_\mathbf{r}, \sigma_\mathbf{c})$-adversary in Theorem 4 is based on this algorithm and provides a competitive ratio that ranges between 1 and $1/\rho$ depending on the values of $\sigma_\mathbf{r}$ and $\sigma_\mathbf{c}$.

The algorithm is inspired by the Lagrangian of (2), $\mathcal{L}(a, \boldsymbol{\lambda}) = r(a) + \sum_{i \in [d]} \lambda_i(\rho - c_i(a))$ where $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^d$. The importance of this function can be seen by the fact that in the stochastic case

$$\frac{\mathtt{OPT_{FD}}}{T} = \max_{A \in \Delta([K])} \min_{\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^d} \mathop{\mathbb{E}}_{a \sim A}[\mathcal{L}(a, \boldsymbol{\lambda})] = \min_{\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^d} \max_{A \in \Delta([K])} \mathop{\mathbb{E}}_{a \sim A}[\mathcal{L}(a, \boldsymbol{\lambda})] \qquad (3)$$

as shown by by Immorlica et al. (2019). Castiglioni et al. (2022) improve (3) by restricting the domain of $\boldsymbol{\lambda}$: (3) also holds if the minimum is over $\boldsymbol{\lambda} \in \mathcal{D}$ where $\mathcal{D} = \{\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^d : \sum_i \lambda_i \leq 1/\rho\}$.

Even though the original inspiration comes from Stochastic $\mathtt{BwK}$, previous work designed Adversarial and Stochastic $\mathtt{BwK}$ algorithms that aim to find a saddle point of a Lagrangian. However, the player does not know the expected values of the rewards and consumptions, so their realized values are used instead. Additionally, since the Lagrangian is linear in $\boldsymbol{\lambda}$, similarly to Immorlica et al. (2019), we replace the domain $\mathcal{D}$ with its extreme points: instead of $\boldsymbol{\lambda}$ the second argument becomes $i \in [d] \cup \{0\}$ where $i = 0$ corresponds to the zero vector and $i > 0$ corresponds to the all-zero vector with $1/\rho$ in its i-th position. Handling this function is easier since it is defined over

---

Algorithm 1: BwK Algorithm based on Lagrangian maximization/minimization.

Input: Maximization algorithm $\mathtt{Alg_{max}}$ and minimization algorithm $\mathtt{Alg_{min}}$

1 for *rounds $t = 1, 2, \ldots$ and while budget lasts* **do**

2   Receive A_t from $\mathtt{Alg_{max}}$ and I_t from $\mathtt{Alg_{min}}$.

3   Play action A_t.

4   Receive reward $R_t(A_t)$ and consumptions $C_{t,1}(A_t), \ldots, C_{t,d}(A_t)$.

5   Update $\mathtt{Alg_{max}}$ with reward $\mathcal{L}_t(A_t, I_t) = R_t(A_t) + \frac{1}{\rho}\mathbb{1}\left[I_t \neq 0\right](\rho - C_{t,I_t}(A_t))$.

6   Update $\mathtt{Alg_{min}}$ with costs $\mathcal{L}_t(A_t, i) = R_t(A_t) + \frac{1}{\rho}\mathbb{1}\left[i \neq 0\right](\rho - C_{t,i}(A_t))$ for $i \in [d] \cup \{0\}$.

7 end

---

a discrete set. Putting all of these together, we define for every $t \in [T]$, $a \in [K]$, and $i \in [d] \cup \{0\}$:

$$\mathcal{L}_t(a, i) = R_t(a) + \frac{1}{\rho}\mathbb{1}\left[i \neq 0\right](\rho - C_{t,i}(a)).$$

The algorithms of Immorlica et al. (2019) and Castiglioni et al. (2022) focus on finding a saddle point of functions similar to the above[2]. More specifically, they use two online algorithms, one that tries to maximize $\mathcal{L}_t(a, i)$ over a and one that tries to minimize it over i. We follow a similar approach and since the domain of both arguments of $\mathcal{L}_t(a, i)$ is discrete, we use two MAB algorithms with no-regret guarantees: In round t, $\mathtt{Alg_{max}}$ chooses an action A_t and $\mathtt{Alg_{min}}$ chooses a resource or the number 0, I_t. Then, $\mathtt{Alg_{max}}$ receives reward $\mathcal{L}_t(A_t, I_t)$ and $\mathtt{Alg_{min}}$ incurs cost $\mathcal{L}_t(A_t, I_t)$. We note that the choices of A_t and I_t are made without knowledge of the rewards and consumptions of round t. We also note that the player can provide $\mathtt{Alg_{max}}$ with bandit information only, i.e., it knows only $\mathcal{L}(A_t, I_t)$. In contrast, the player can give $\mathtt{Alg_{min}}$ full information since it knows the value of $\mathcal{L}(A_t, i)$ for all i. Our full algorithm can be found in Algorithm 1.

We will use algorithms $\mathtt{Alg_{max}}$ and $\mathtt{Alg_{min}}$ that guarantee no-regret with high probability. We use EXP3.P from Auer et al. (2002) as $\mathtt{Alg_{max}}$ which guarantees that for all $\delta > 0$ with probability at least $1 - \delta$ it holds that for all T'

$$\max_{a \in [K]} \sum_{t=1}^{T'} \mathcal{L}_t(a, I_t) - \sum_{t=1}^{T'} \mathcal{L}_t(A_t, I_t) \leq \mathtt{Reg_{max}}(T, \delta) := O\left(\frac{1}{\rho}\sqrt{KT\log(T/\delta)}\right). \qquad (4)$$

Using Hedge from Freund and Schapire (1997) as $\mathtt{Alg_{min}}$ guarantees that for all $\delta > 0$ with probability at least $1 - \delta$ it holds that for all T'

$$\sum_{t=1}^{T'} \mathcal{L}_t(A_t, I_t) - \min_{i \in [d] \cup \{0\}} \sum_{t=1}^{T'} \mathcal{L}_t(A_t, i) \leq \mathtt{Reg_{min}}(T, \delta) := O\left(\frac{1}{\rho}\sqrt{T\log(Td/\delta)}\right). \qquad (5)$$

For BwK with full information, we can use Hedge for $\mathtt{Alg_{max}}$ to get the improved regret guarantee $\mathtt{Reg_{max}}(T, \delta) = O\left(\frac{1}{\rho}\sqrt{T\log(TK/\delta)}\right)$ in (4).

---

2. Castiglioni et al. (2022) use the initial formulation with $\boldsymbol{\lambda} \in \mathcal{D}$. In Immorlica et al. (2019) the second argument of the function considers only the d non-zero extreme points, i.e., the domain of i is $[d]$ instead of $[d] \cup \{0\}$; even if every consumption is less than ρ (in which case the algorithm is not budget constrained) the choice of action a still takes into account the consumptions making potentially sub-optimal choices. They fix this by picking a slightly different function $\mathcal{L}_t$.

5. Guarantees of Algorithm 1 in Approximately Stationary BwK

In this section, we prove our guarantee for Algorithm 1 against a stationary adversary. For any values of $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$ our algorithm achieves the competitive ratio of $1/\rho$ that is guaranteed in Adversarial BwK. As $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$ increase, the competitive ratio of Algorithm 1 smoothly improves and becomes 1 when $\sigma_{\mathtt{r}} = \sigma_{\mathtt{c}} = 1$. In the most interesting range of parameters, when ρ is much smaller than $\sigma_{\mathtt{r}}\sigma_{\mathtt{c}}$ (which also implies that the gap in guarantees between Stochastic and Adversarial BwK is largest), our algorithm achieves a close to tight $\sigma_{\mathtt{r}}\sigma_{\mathtt{c}}$ fraction of the optimal solution. This is a huge improvement over the ρ fraction that is guaranteed in Adversarial BwK.

We start with a lemma comparing the rewards achieved by Algorithm 1 against a distribution of actions whose maximum expected consumption of any resource is at most ρ. This lemma is true for any $\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}}$, and adaptive adversary. It also easily proves the two guarantees of Castiglioni et al. (2022) about Stochastic and Adversarial BwK and extends the second for adaptive adversaries.

Lemma 3 *Let $A \in \Delta([K])$ be a distribution of actions such that $\max_{i,t} \mathbb{E}_{a \sim A}\left[c_{i,t}(a)\right] \leq \rho$. Then for any adversary and $\delta > 0$, with probability at least $1 - \delta$ Algorithm 1 achieves*

$$\mathtt{REW} \geq \sum_{t=1}^{T} \mathbb{E}_{a \sim A}\left[r_t(a)\right] - \mathtt{Reg}_{\max}(T, \delta) - \mathtt{Reg}_{\min}(T, \delta) \tag{6}$$

Using Lemma 3 it is easy to prove the two guarantees of Castiglioni et al. (2022). First, if the adversary is stochastic we can prove a 1 competitive ratio and sublinear regret by noticing that the optimal action A^* in (2) satisfies the conditions of the lemma. Second, against an adaptive adversary, we notice that the action distribution that plays with probability ρ the best-fixed *unbudgeted* action, and the null action otherwise satisfies Lemma 3. This proves a $1/\rho$ competitive ratio against $\max_a \sum_t r_t(a)$ with high probability, which also proves the same guarantee against $\mathtt{OPT}_{\mathtt{FD}}$ as without a budget, the optimal action is fixed.

Proof sketch The lemma's proof is based on the guarantees of the two algorithms, $\mathtt{Alg}_{\max}$ and $\mathtt{Alg}_{\min}$, found in (4) and (5), respectively, up to the stopping round of the algorithm, $T_{\mathtt{A}}$. On the one hand, we compare $\sum_{t \leq T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t)$ with $\min_i \sum_{t \leq T_{\mathtt{A}}} \mathcal{L}_t(A_t, i)$: this lower bounds the reward of the algorithm, $\mathtt{REW}$, using an additive term that boosts $\mathtt{REW}$ as $T - T_{\mathtt{A}}$ becomes bigger: if the algorithm runs out of budget fast, the $C_{t,i}(A_t)$ terms in $\mathcal{L}(A_t, i)$ become larger making this bound better. On the other hand, we compare $\sum_{t \leq T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t)$ with $\sum_{t \leq T_{\mathtt{A}}} \mathbb{E}_{a \sim A}\left[\mathcal{L}(a, I_t)\right]$ (where A is the distribution defined in Lemma 3): this contains the total reward of distribution A up to round $T_{\mathtt{A}}$ and an additive error that is not too high because on expectation $C_{t,i}(a) = c_{t,i}(a)$ and the second term is low by the properties of A. We defer the detailed proof to Appendix A. ∎

We now move to the main theorem of this section. Our theorem guarantees a fraction of the optimal solution with high probability against an adaptive $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationary adversary.

Theorem 4 *Against an adaptive $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationary adversary, the reward of Algorithm 1 satisfies for any $\delta > 0$ with probability at least $1 - \delta$*

$$\mathtt{REW} \geq \left(\rho + \sigma_{\mathtt{r}}(\sigma_{\mathtt{c}} - \rho)^+\right) \mathtt{OPT}_{\mathtt{FD}} - \mathtt{Reg}_{\max}(T, \delta) - \mathtt{Reg}_{\min}(T, \delta).$$

The proof of the theorem is deferred to Appendix A. The idea is to use Lemma 3. To take advantage of (6) we need to lower bound the reward of the optimal distribution A^* after its stopping

time T^* and upper bound its maximum consumption. These two quantities depend on T^*: as T^* becomes larger, both the reward of A^* after T^* and the maximum consumption become smaller (the first depending on $\sigma_{\mathtt{r}}$ and the second on $\sigma_{\mathtt{c}}$). Carefully examining these effects and choosing the T^* (as a function of $\rho, \sigma_{\mathtt{r}}, \sigma_{\mathtt{c}}$) that yields the worst guarantee for the algorithm gets us the theorem.

Remark 5 *When $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$ are much larger than ρ a naive approach would be to use an algorithm assuming a fully stochastic setting. Such an algorithm would lead to much weaker results. An algorithm that is based on the classic Arm Elimination algorithm would guarantee only a $\sigma_{\mathtt{r}}^2\sigma_{\mathtt{c}}^2$ fraction of $\mathtt{OPT_{FD}}$: it might eliminate an action because it identified it as sub-optimal if it is a bit worse than the one that it has identified as optimal. However, after that round, an $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationary adversary might make the previously optimal action worse by a factor of $\sigma_{\mathtt{r}}\sigma_{\mathtt{c}}$ and the sub-optimal one better by the same factor. This would result in a sub-optimality factor of $\sigma_{\mathtt{r}}^2\sigma_{\mathtt{c}}^2$.*

6. Impossibility results for Approximately Stationary BwK

In this section, we show a bound on the guarantee any algorithm can achieve against an $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationary adversary. The bound applies to algorithms that are oblivious to the values of $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$ and guarantee a fraction of at least ρ against an oblivious adversary. Additionally, we get this bound when there is only one resource. Our theorem has many interesting corollaries. First, it proves that a ρ fraction of $\mathtt{OPT_{FD}}$ is the best possible in fully adversarial cases, improving over the result of Balseiro and Gur (2017) that a ρ fraction of the best sequence of actions is the best possible. Second, when ρ is much smaller than $\sigma_{\mathtt{r}}\sigma_{\mathtt{c}}^2$, it makes the guarantee of Theorem 4 approximately tight (in fact if $\rho \leq \sigma_{\mathtt{r}}\sigma_{\mathtt{c}}^2$ the two bounds are within a factor of 2 of each other). Third, when $\sigma_{\mathtt{r}} = O(\rho)$, it proves that any algorithm can guarantee at most $O(\rho)$ fraction of $\mathtt{OPT_{FD}}$, making Algorithm 1 optimal up to constant factors.

Theorem 6 *Fix any algorithm that achieves sublinear regret and an $\alpha_\rho(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$ fraction of the optimal solution against an oblivious $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-adversary. If the algorithm is oblivious to the values of $\sigma_{\mathtt{r}}$ and $\sigma_{\mathtt{c}}$, $\rho = \Theta(1)$, and $\alpha_\rho(0, 0) \geq \rho$ (i.e., guarantees at least a ρ fraction of the optimal solution), then*

$$\alpha_\rho(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}}) \leq \begin{cases} \sigma_{\mathtt{r}} + \rho(1 - \sigma_{\mathtt{r}}), & \text{if } \sigma_{\mathtt{r}} \leq \rho \\ 2\sqrt{\sigma_{\mathtt{r}}\rho} - \sigma_{\mathtt{r}}\rho, & \text{if } \rho \leq \sigma_{\mathtt{r}} \leq \frac{\rho}{\sigma_{\mathtt{c}}^2} \\ \sigma_{\mathtt{r}}\sigma_{\mathtt{c}} + \rho(1/\sigma_{\mathtt{c}} - \sigma_{\mathtt{r}}), & \text{if } \sigma_{\mathtt{r}} \geq \frac{\rho}{\sigma_{\mathtt{c}}^2} \end{cases}$$

In the example we use to prove the theorem, there are $K + 1$ different outcomes (the oblivious adversary selects which outcome the algorithm faces). Additionally, the T rounds are divided into K batches. In outcome $q \in [K]$, the optimal action has positive reward only in the rounds of the q-th batch. The algorithm cannot distinguish if it is facing the q-th outcome on the $(q-1)$-th batch, so it needs to act conservatively and not spend its resources to guarantee a ρ fraction of the optimal solution. Because of this behavior, any algorithm is guaranteed to miss out on a large part of the optimal solution in the $(K + 1)$-instance, which is the one that is $(\sigma_{\mathtt{r}}, \sigma_{\mathtt{c}})$-stationary. The proof of the theorem is deferred to Appendix B.

7. Improved guarantee for one resource

In this section, we provide an algorithm that improves the guarantee of Algorithm 1 when $\sigma_{\mathtt{c}}$ is small, that is, the variability in expected consumption is high. The improved guarantee requires an additional assumption, that the sum of the differences in the expected consumptions of the actions is sublinear. With this additional guarantee we can ensure sublinear regret. The bound in the previous section shows that the guarantee of Theorem 4 is close to optimal when ρ and $\sigma_{\mathtt{r}}$ are small and $\sigma_{\mathtt{c}}$ is significantly larger. In contrast, when $\sigma_{\mathtt{c}} \leq \rho \leq \sigma_{\mathtt{r}}$ Theorem 4 provides the adversarial guarantee of ρ while Theorem 6 suggests that much better guarantees may be achievable.

Our improved algorithm uses Algorithm 1 as a subroutine and a parameter $T_{\mathtt{res}}$. It runs and restarts Algorithm 1 every $T_{\mathtt{res}}$ rounds. We allocate each run of Algorithm 1 a budget of $\rho T_{\mathtt{res}} - 1$. This way the per-round budget in every run is approximately ρ. It also guarantees that every run of the algorithm uses at most $\rho T_{\mathtt{res}}$ of every resource since Algorithm 1 would have terminated when going above budget and not get the last item; however, when simulating the algorithm and using the actions it suggests, the player has to know to terminate it herself before it uses more than the desired budget which we achieve by allocating the algorithm 1 less budget than that.

Algorithm 2: Restarting BwK Algorithm based on Algorithm 1

Input: Inputs needed for Algorithm 1 and parameter $T_{\mathtt{res}}$

1   Split rounds into $\lceil \frac{T}{T_{\mathtt{res}}} \rceil$ disjoint batches $[T] = \mathcal{T}_1 \cup \ldots \cup \mathcal{T}_{\lceil T/T_{\mathtt{res}} \rceil}$, each batch having $T_{\mathtt{res}}$ rounds (except maybe for the last one).

2   **for** *each batch* $j = 1, \ldots, \lceil \frac{T}{T_{\mathtt{res}}} \rceil$ **do**

3 | Independently of previous rounds, run Algorithm 1 on rounds $\mathcal{T}_j$ with budget $\rho|\mathcal{T}_j| - 1$.

4   **end**

We now show a lemma for Algorithm 2, which will lead to the promised improved guarantee. The main ingredient of our guarantee for Algorithm 1 was Lemma 3 that bounds the reward against a distribution A that satisfies $\max_{i,t} \mathbb{E}_{a \sim A}\left[c_{i,t}(a)\right] \leq \rho$. To ensure this condition for an arbitrary distribution A, we need to scale it down by playing it only with probability ρ divided by the above maximum over the whole time horizon, and playing the null action with the remaining probability. The new lemma is structured similarly to Lemma 3: it shows that the reward of Algorithm 2 is at least the reward of any action distribution across all rounds but scaled down as a function of the consumptions of that distribution. In contrast to Lemma 3 however, the deterioration of the reward is much more fine-grained. Instead of scaling down the whole reward by the maximum consumption, the reward of each round is scaled down by the *consumptions of that round*.

Lemma 7 *For any $A \in \Delta([k])$, Algorithm 2 guarantees reward that for every $\delta > 0$,*

$$\mathbb{P}\left[\texttt{REW} \geq \sum_{t=1}^{T} \mathbb{E}_{a \sim A}\left[r_t(a)\right] \min\left\{1, \frac{\rho}{\max_i \mathbb{E}_{a \sim A}\left[c_{t,i}(a)\right]}\right\} - \texttt{Reg}\right] \geq 1 - \delta$$

where using $\mathcal{E} \geq \sum_{t=1}^{T-1} \max_{i \in [d]} \left| \mathbb{E}_{a \sim A}\left[c_{t,i}(a) - c_{t+1,i}(a)\right] \right|$ we have

$$\texttt{Reg} = \frac{T}{T_{\mathtt{res}}}\left(\texttt{Reg}_{\max}(T_{\mathtt{res}}, \delta T/T_{\mathtt{res}}) + \texttt{Reg}_{\min}(T_{\mathtt{res}}, \delta T/T_{\mathtt{res}})\right) - \frac{T_{\mathtt{res}}}{\rho}\mathcal{E}$$
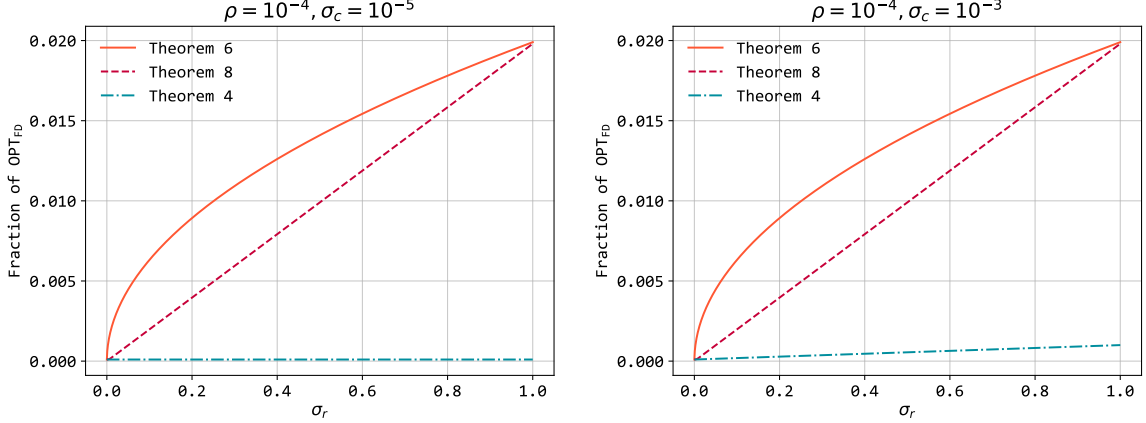
Figure 1: The bounds of Theorems 4, 6, and 8 on the fraction of the optimal solution for $d = 1$ and small ρ and $\sigma_\mathtt{c}$ (mentioned above each graph).

If $T_{\mathtt{res}} = \Theta\left((\rho T/\mathcal{E})^{\frac{2}{3}}\right)$ and ignoring dependence on K, d then $\mathtt{Reg} = O(T^{\frac{2}{3}}\mathcal{E}^{\frac{1}{3}}\log(T/\delta)\rho^{-\frac{1}{3}})$.

The proof of the lemma is based on Lemma 3. The reward of Algorithm 2 in each batch j is at least the reward of distribution A in that batch scaled down by $\max_{i,\tau \in \mathcal{T}_j} \mathbb{E}_{a \sim A}\left[c_{\tau,i}(a)\right]$. This factor can be improved to $\mathbb{E}_{a \sim A}\left[c_{t,i}(a)\right]$ for any $t \in \mathcal{T}_j$ by introducing an additive error that depends on the variance of consumptions on that round. Using the condition on $\mathcal{E}$, this additive error over all batches is sublinear, which proves the lemma. The full proof can be found in Appendix C.

We now present the main result of this section. Using Lemma 7 we can get a strictly better bound. Our result is parametric and depends on a parameter $x \in [\rho, 1]$.

Theorem 8 *Against an adaptive $(\sigma_\mathtt{r}, \sigma_\mathtt{c})$-stationary adversary, for any $\delta > 0$ with probability at least $1 - \delta$, Algorithm 2 guarantees at least a $\alpha_\rho(\sigma_\mathtt{r}, \sigma_\mathtt{c})$ fraction of $\mathtt{OPT_{FD}}$, where*

$$\alpha_\rho(\sigma_\mathtt{r}, \sigma_\mathtt{c}) = \min_{x \in [\rho,1]} \left(\max\left\{ \rho, x\sigma_\mathtt{c}, \sigma_\mathtt{r}\frac{x}{d+x} \right\} + \max\left\{ \rho\sigma_\mathtt{r}\frac{1-x}{x}, \sigma_\mathtt{r}\sigma_\mathtt{c}(1-x) \right\} \right)$$

and regret $\mathtt{Reg}$ which is sublinear if $\mathcal{E}/\rho$ is sublinear, where $\mathtt{Reg}$ and $\mathcal{E}$ are defined in Lemma 7.

Remark 9 *The bound of Theorem 8 improves the bound of Theorem 4 significantly when there is only one resource $d = 1$ and $\sigma_\mathtt{r}$ is much larger than ρ and $\sigma_\mathtt{c}$. For example, if $\sigma_\mathtt{c} \leq \rho \ll \sigma_\mathtt{r}$ then the bound of Theorem 8 becomes*

$$\alpha_\rho(\sigma_\mathtt{r}, \sigma_\mathtt{c}) = \begin{cases} 2\sigma_\mathtt{r}(\sqrt{\rho} - \rho), & \text{if } \sigma_\mathtt{r}^2 \geq \rho \\ \sigma_\mathtt{r}^2 + \rho - 2\rho\sigma_\mathtt{r}, & \text{if } \sigma_\mathtt{r}^2 \leq \rho \end{cases}$$

in which case the bound of Theorem 4 is ρ. We also showcase this improvement in Figure 1 using some numerical examples, where we compare the bounds of Theorems 4, 6, and 8 for small values of ρ and $\sigma_\mathtt{c}$ and arbitrary values of $\sigma_\mathtt{r}$.

Acknowledgments

References

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In Moshe Babaioff, Vincent Conitzer, and David A. Easley, editors, *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*, pages 989–1006. ACM, 2014. doi: 10.1145/2600057.2602844. URL https://doi.org/10.1145/2600057.2602844.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 207–216. IEEE Computer Society, 2013. doi: 10.1109/FOCS.2013.30. URL https://doi.org/10.1109/FOCS.2013.30.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1109–1134. JMLR.org, 2014. URL http://proceedings.mlr.press/v35/badanidiyuru14.html.

Santiago R. Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. In Constantinos Daskalakis, Moshe Babaioff, and Hervé Moulin, editors, *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*, page 609. ACM, 2017. doi: 10.1145/3033274.3084088. URL https://doi.org/10.1145/3033274.3084088.

Santiago R. Balseiro, Haihao Lu, and Vahab S. Mirrokni. Dual mirror descent for online allocation problems. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 613–628. PMLR, 2020. URL http://proceedings.mlr.press/v119/balseiro20a.html.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Oper. Res.*, 63 (5):1227–1244, 2015. doi: 10.1287/opre.2015.1408. URL https://doi.org/10.1287/opre.2015.1408.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT 2012 - The*

25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland, volume 23 of *JMLR Proceedings*, pages 42.1–42.23. JMLR.org, 2012. URL http://proceedings.mlr.press/v23/bubeck12b/bubeck12b.pdf.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2767–2783. PMLR, 2022. URL https://proceedings.mlr.press/v162/castiglioni22a.html.

Nikhil R. Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A. Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *J. ACM*, 66 (1):7:1–7:41, 2019. doi: 10.1145/3284177. URL https://doi.org/10.1145/3284177.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Yonatan Gur, Assaf Zeevi, and Omar Besbes. Stochastic multi-armed-bandit problem with non-stationary rewards. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 199–207, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/903ce9225fca3e988c2af215d4e544d3-Abstract.html.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3-4):157–325, 2016. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

Nicole Immorlica, Karthik Abinav Sankararaman, Robert E. Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 202–219. IEEE Computer Society, 2019. doi: 10.1109/FOCS.2019.00022. URL https://doi.org/10.1109/FOCS.2019.00022.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2286–2305. PMLR, 2020. URL http://proceedings.mlr.press/v125/kesselheim20a.html.

Raunak Kumar and Robert Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/7a62d9a4c03377d1175b8859b4cc16d4-Abstract-Conference.html.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/69469da823348084ca8933368ecbf676-Abstract-Conference.html.

Thodoris Lykouris, Vahab S. Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 114–122. ACM, 2018. doi: 10.1145/3188745.3188918. URL https://doi.org/10.1145/3188745.3188918.

Anshuka Rangi, Massimo Franceschetti, and Long Tran-Thanh. Unifying the stochastic and the adversarial bandits with knapsack. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3311–3317. ijcai.org, 2019. doi: 10.24963/ijcai.2019/459. URL https://doi.org/10.24963/ijcai.2019/459.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In Amos J. Storkey and Fernando Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 1760–1770. PMLR, 2018. URL http://proceedings.mlr.press/v84/sankararaman18a.html.

Aleksandrs Slivkins. Book announcement: Introduction to multi-armed bandits. *SIGecom Exch.*, 18(1):28–30, 2020. doi: 10.1145/3440959.3440965. URL https://doi.org/10.1145/3440959.3440965.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression, 2023.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Trans. Inf. Theory*, 58(8):5588–5611, 2012. doi: 10.1109/TIT.2012.2198613. URL https://doi.org/10.1109/TIT.2012.2198613.

Siwei Wang, Longbo Huang, and John C. S. Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/89ae0fe22c47d374bc9350ef99e01685-Abstract.html.

Appendix A. Deferred proofs of Section 5

In this sections we present the deferred proofs for the guarantee of Theorem 4. We first present the proof of Lemma 3.

Proof of Lemma 3 We start the proof by reminding the two no-regret guarantees of Algorithm 1. (4) and (5), for $T' = T_{\mathtt{A}}$, the algorithm's stopping time imply that for any $\delta > 0$ it holds that

$$\mathbb{P}\left[\sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t) \geq \max_{a \in K} \sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(a, I_t) - \mathtt{Reg}_{\max}(T, \delta)\right] \geq 1 - \delta. \tag{7}$$

and

$$\mathbb{P}\left[\sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t) \leq \min_{i \in [d] \cup \{0\}} \sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, i) + \mathtt{Reg}_{\min}(T, \delta)\right] \geq 1 - \delta \tag{8}$$

We start by analyzing the second bound, (8). For any $\delta_1 > 0$, with probability at least $1 - \delta_1$,

$$
\begin{aligned}
\sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t) - \mathtt{Reg}_{\min}(T, \delta_1) \;&\leq\; \min_{i \in [d] \cup \{0\}} \sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, i) \\
&= \sum_{t=1}^{T_{\mathtt{A}}} R_t(A_t) + \min_{i \in [d] \cup \{0\}} \left\{ \frac{\mathbb{1}\left[i > 0\right]}{\rho} \sum_{t=1}^{T_{\mathtt{A}}} (\rho - C_{t,i}(A_t)) \right\} \\
&\leq\; \mathtt{REW} - (T - T_{\mathtt{A}}) + \frac{1}{\rho} \tag{9}
\end{aligned}
$$

where the last inequality holds by the following case analysis:

- If $T_{\mathtt{A}} = T$, then, by setting $i = 0$:

$$\min_{i \in [d] \cup \{0\}} \left\{ \frac{\mathbb{1}\left[i > 0\right]}{\rho} \sum_{t=1}^{T_{\mathtt{A}}} (\rho - C_{t,i}(A_t)) \right\} \leq 0 = -(T - T_{\mathtt{A}})$$

- If $T_{\mathtt{A}} < T$, then the algorithm runs out of some resource $i' \in [d]$ on round $T_{\mathtt{A}} + 1$, meaning that $\sum_{t=1}^{T_{\mathtt{A}}+1} C_{t,i'}(A_t) > \rho T$. By setting $i = i'$ we get

$$\min_{i \in [d] \cup \{0\}} \left\{ \frac{\mathbb{1}\left[i > 0\right]}{\rho} \sum_{t=1}^{T_{\mathtt{A}}} (\rho - C_{t,i}(A_t)) \right\} \leq \frac{1}{\rho} \sum_{t=1}^{T_{\mathtt{A}}} (\rho - C_{t,i'}(A_t)) = \frac{1}{\rho} (T_{\mathtt{A}}\rho - \rho T + 1)$$

Now, we analyze (7). For any $\delta_2 > 0$, with probability at least $1 - \delta_2$ it holds that

$$
\begin{aligned}
\sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(A_t, I_t) + \mathtt{Reg}_{\max}(T, \delta_2) \;&\geq\; \max_{a \in [K]} \sum_{t=1}^{T_{\mathtt{A}}} \mathcal{L}_t(a, I_t) \\
&\geq\; \sum_{t=1}^{T_{\mathtt{A}}} \mathop{\mathbb{E}}_{a \sim A} \left[\mathcal{L}_t(a, I_t) \right] \\
&= \sum_{t=1}^{T_{\mathtt{A}}} \mathop{\mathbb{E}}_{a \sim A} \left[R_t(a) \right] + \sum_{t=1}^{T_{\mathtt{A}}} \frac{\mathbb{1}\left[I_t > 0\right]}{\rho} (\rho - \mathop{\mathbb{E}}_{a \sim A} \left[C_{t,I_t}(a) \right]) \tag{10}
\end{aligned}
$$

We now prove a high probability lower bound on the last quantity. Towards that end, we define for every $\tau \in \{0\} \cup [T]$ and $a \in [K]$

$$Z_\tau(a) = \sum_{t=1}^\tau (R_t(a) - r_t(a)) + \sum_{t=1}^\tau \frac{\mathbb{1}\left[I_t > 0\right]}{\rho} (c_{t,I_t}(a) - C_{t,I_t}(A)).$$

We notice that for every a, $Z_\tau(a)$ is a martingale since

$$\mathbb{E}\left[Z_\tau(a) - Z_{\tau-1}(a)\big|\mathcal{H}_{\tau-1}\right] = \mathbb{E}\left[R_t(a) - r_t(a) + \frac{\mathbb{1}\left[I_\tau > 0\right]}{\rho}\left(c_{\tau,I_\tau}(a) - C_{\tau,I_\tau}(a)\right)\big|\mathcal{H}_{\tau-1}\right] = 0$$

where the last equality follows from the definitions of r_t and $c_{t,i}$: for all a and all i it holds that $\mathbb{E}\left[R_t(a) - r_t(a)\big|\mathcal{H}_{\tau-1}\right] = 0$ as well as $\mathbb{E}\left[C_{t,i}(a) - c_{t,i}(a)\big|\mathcal{H}_{\tau-1}\right] = 0$.

Since $Z_\tau(a)$ is a martingale and $|Z_\tau(a) - Z_{\tau-1}(a)| \leq 1 + 1/\rho$ we can use the Azuma-Hoeffding inequality and the union bound to get that for any $\delta_3 > 0$ with probability at least $1 - \delta_3$ it holds for all $T' \in [T]$ and all $a \in [K]$ that

$$Z_{T'}(a) - Z_0(a) \geq -\left(1 + \frac{1}{\rho}\right)\sqrt{2T\log(TK/\delta_3)} \geq -\texttt{Reg}_{\max}(T, \delta_3)$$

Combining the above for $T' = T_\mathtt{A}$, we can make every $R_t(a)$ and $C_{t,I_t}(a)$ in (10) into $r_t(a)$ and $c_{t,I_t}(a)$, with an additive error of $\texttt{Reg}_{\max}(T, \delta_3)$. This yields that with probability at least $1 - \delta_2 - \delta_3$

$$
\begin{aligned}
\sum_{t=1}^{T_\mathtt{A}} \mathcal{L}_t(A_t, I_t) + \texttt{Reg}_{\max}(T, \delta_2) &\geq \sum_{t=1}^{T_\mathtt{A}} \mathbb{E}_{a \sim A}\left[r_t(a)\right] + \sum_{t=1}^{T_\mathtt{A}} \frac{\mathbb{1}\left[I_t > 0\right]}{\rho}(\rho - \mathbb{E}_{a \sim A}\left[c_{t,I_t}(a)\right]) - \texttt{Reg}_{\max}(T, \delta_3) \\
&\geq \sum_{t=1}^{T_\mathtt{A}} \mathbb{E}_{a \sim A}\left[r_t(a)\right] - \texttt{Reg}_{\max}(T, \delta_3) \\
&\geq \sum_{t=1}^{T} \mathbb{E}_{a \sim A}\left[r_t(a)\right] - (T - T_\mathtt{A}) - \texttt{Reg}_{\max}(T, \delta_3) \qquad (11)
\end{aligned}
$$

where in the second inequality we used that $\max_{t,i} \mathbb{E}_{a \sim A}\left[c_{t,i}(a)\right] \leq \rho$ and in the last one that $r_t(a) \leq 1$.

Combining (9) and (11), using the union bound, and setting $\delta_1 = \delta_2 = \delta_3 = \delta/3$ we get the desired result with an additive error of $2\texttt{Reg}_{\max}(T, \delta/3) + \texttt{Reg}_{\min}(T, \delta/3)$. This is the same as the one in the lemma's statement because asymptotically they are the same and by the definition of $\texttt{Reg}_{\max}$ and $\texttt{Reg}_{\min}$ (which are defined using $O(\cdot)$ notation). ∎

Before proving Theorem 4, we prove two lemmas that show the effects of a $(\sigma_\mathtt{r}, \sigma_\mathtt{c})$-stationary adversary to the optimal solution.

The first lemma lower bounds the reward of the optimal distribution of distribution A^* after its stopping time, T^*. We need such a bound because we cannot guarantee that the algorithm has gained a large part of $\texttt{OPT}_{\texttt{FD}}$ before round T^*, especially when T^* is much smaller than T. We prove a bound that depends on $\sigma_\mathtt{r}$ and T^*/T.

Lemma 10 *Let (A^*, T^*) be the optimal solution to (1) and let $x = T^*/T$. Then it holds that*

$$\sum_{t=T^*+1}^{T} \mathbb{E}_{a \sim A^*}\left[r_t(a)\right] \geq \sigma_\mathtt{r} \frac{1-x}{x} \texttt{OPT}_{\texttt{FD}}.$$

Definition 1 only states that the rewards of each action are stationary but we need this for a distribution of actions.

Proof For every $a \in [K]$ let $p(a) = \mathbb{P}[A^* = a]$. We have that

$$
\begin{aligned}
\sum_{t=T^*+1}^{T} \mathop{\mathbb{E}}_{a \sim A^*} [r_t(a)] &= \sum_{a \in [K]} p(a) \sum_{t=T^*+1}^{T} r_t(a) \\
&\geq \sum_{a \in [K]} p(a)(T - T^*) \min_t r_t(a) \\
&\geq \sigma_{\mathtt{r}} \sum_{a \in [K]} p(a)(T - T^*) \max_t r_t(a) \qquad (\sigma_{\mathtt{r}}\text{-stationarity}) \\
&\geq \sigma_{\mathtt{r}} \sum_{a \in [K]} p(a) \frac{T - T^*}{T^*} \sum_{t=1}^{T^*} r_t(a) \\
&= \sigma_{\mathtt{r}} \frac{1 - x}{x} \mathtt{OPT}_{\mathtt{FD}} \qquad (T^* = xT)
\end{aligned}
$$

which proves the lemma. ∎

We proceed to prove the second lemma, for the consumption of the optimal action. In contrast to the previous lemma, we prove an upper bound on the maximum consumption across all resources and rounds.

Lemma 11 *Let (A^*, T^*) be the optimal solution to (1) and let $x = T^*/T$. Then, for all $i \in [d]$,*

$$
\max_t \mathop{\mathbb{E}}_{a \sim A^*} [c_{t,i}(a)] \leq \frac{\rho}{x \sigma_{\mathtt{c}}}.
$$

To prove this lemma we use the fact that by round T^* the total consumption of any resource under action A^* is at most $T\rho$. We then bound the maximum consumption using $\sigma_{\mathtt{c}}$-stationarity.
Proof For every $a \in [K]$ let $p(a) = \mathbb{P}[A^* = a]$. Fix a resource $i \in [d]$. We have that

$$
\begin{aligned}
\max_t \mathop{\mathbb{E}}_{a \sim A^*} [c_{t,i}(a)] &= \max_t \sum_{a \in [K]} p(a) c_{t,i}(a) \\
&\leq \sum_{a \in [K]} p(a) \max_t c_{t,i}(a) \qquad (\text{Jensen's inequality}) \\
&\leq \frac{1}{\sigma_{\mathtt{c}}} \sum_{a \in [K]} p(a) \min_t c_{t,i}(a) \qquad (\sigma_{\mathtt{c}}\text{-stationary}) \\
&\leq \frac{1}{\sigma_{\mathtt{c}}} \sum_{a \in [K]} p(a) \frac{1}{T^*} \sum_{t=1}^{T^*} c_{t,i}(a) \qquad (\text{minimum less than average}) \\
&= \frac{1}{\sigma_{\mathtt{c}}} \frac{1}{T^*} \sum_{t=1}^{T^*} \mathop{\mathbb{E}}_{a \sim A^*} [c_{t,i}(a)] \\
&\leq \frac{1}{\sigma_{\mathtt{c}}} \frac{1}{T^*} T\rho = \frac{\rho}{x \sigma_{\mathtt{c}}} \qquad (\text{using (1)})
\end{aligned}
$$

This proves the lemma. ∎

We know prove Theorem 4. The proof combines Lemmas 3, 10, and 11 and guarantees a bound based on x. By taking the worst case of over x we get the theorem.

Proof of Theorem 4 The theorem trivially holds if $\sigma_c \leq \rho$, by the competitive ratio guarantee of $1/\rho$ of the algorithm in purely adversarial settings.

We are going to use Lemma 3 with a carefully chosen distribution of actions. Let (A^*, T^*) be the optimal solution to (1). Let $x = T^*/T$. We set A to be the arm distribution that plays the optimal distribution A^* with probability $\max\{\rho, \sigma_c x\}$ and otherwise the null arm. We notice that for every resource i and round t,

$$\mathbb{E}_{a \sim A}[c_{t,i}(a)] = \max\{\rho, \sigma_c x\} \mathbb{E}_{a \sim A^*}[c_{t,i}(a)] \leq \max\{\rho, \sigma_c x\} \min\left\{1, \frac{\rho}{\sigma_c x}\right\} = \rho$$

where in the inequality we used Lemma 11 and $c_{t,i}(a) \leq 1$. The above allows us to use Lemma 3: for any $\delta > 0$ with probability at least $1 - \delta$ we have

$$\texttt{REW} + \texttt{Reg}_{\max}(T, \delta) + \texttt{Reg}_{\min}(T, \delta) \geq \sum_{t=1}^{T} \mathbb{E}_{a \sim A}[r_t(a)] = \max\{\rho, \sigma_c x\} \sum_{t=1}^{T} \mathbb{E}_{a \sim A^*}[r_t(a)]$$

Combining the above with Lemma 10 we get that, up to the additive term $\texttt{Reg}_{\max}(T, \delta) + \texttt{Reg}_{\min}(T, \delta)$, the algorithm guarantees with high probability the following fraction of $\texttt{OPT}_{\texttt{FD}}$:

$$\max\{\rho, \sigma_c x\}\left(1 + \sigma_r \frac{1-x}{x}\right) = \max\left\{\rho + \rho\sigma_r \frac{1-x}{x}, \sigma_r \sigma_c + \sigma_c x(1 - \sigma_r)\right\}.$$

The above term is decreasing in x while $x \leq \frac{\rho}{\sigma_c}$ and increasing afterwards. This means that the minimizing x is $\frac{\rho}{\sigma_c}$, which proves the desired competitive ratio. ∎

Appendix B. Deferred proofs of Section 6

In this section, we present the deferred proof of Theorem 6.

Proof of Theorem 6 There is only one resource, $d = 1$. Let y be a real number such that $\rho \leq y < 1$. We create an instance with the null action and another K actions, where $K = 1 + \lceil \frac{1-y}{\rho} \rceil$. Additionally, let $z = \frac{1-y}{K-1}$. We are going to pick y that does not depend on T, which entails that K and z also do not depend on T. We split the rounds into K batches of rounds. The first batch has Ty rounds and the rest have $\frac{T(1-y)}{K-1} = Tz \leq T\rho$ rounds each. We examine $K + 1$ different outcomes, indexed by $q \in [K + 1]$. The adversary decides the value of q. We are going to introduce outcome $q = K + 1$ later in the proof and initially focus on the first K outcomes.

We now present the reward and consumption of each arm in each batch and outcome, which can also be found in Table 1. During the rounds of batch $i \in [K]$, only the i-th arm can have positive reward and consumption. Let $\varepsilon < 1$ and let $r_t^{(q)}(a), c_t^{(q)}(a)$ denote the reward and consumption of arm a in outcome q. The reward and consumption of arm $i = 1$ are

$$\left(r_t^{(q)}(1), c_t^{(q)}(1)\right) = \begin{cases} \left(\varepsilon^{K-1}, \frac{\rho}{y}\right), & \text{if } t \in \text{1st batch and for any outcome } q \\ (0, 0), & \text{otherwise} \end{cases} \tag{12}$$

20

and the rewards and costs for every arm $2 \geq i \geq K$

$$\left(r_t^{(q)}(i), c_t^{(q)}(i)\right) = \begin{cases} \left(\varepsilon^{K-i}, 1\right), & \text{if } t \in i\text{-th batch and outcome is } i \leq q \leq K \\ (0,0), & \text{otherwise} \end{cases}$$

Outcome q	Action $i = 1$, batch $i = 1$	$i = 2$	$i = 3$	$3 < i \leq K$
$q = 1$	$\left(\varepsilon^{K-1}, \frac{\rho}{y}\right)$	$(0,0)$	$(0,0)$	$(0,0)$
$q = 2$	$\left(\varepsilon^{K-1}, \frac{\rho}{y}\right)$	$(\varepsilon^{K-2}, 1)$	$(0,0)$	$(0,0)$
$q = 3$	$\left(\varepsilon^{K-1}, \frac{\rho}{y}\right)$	$(\varepsilon^{K-2}, 1)$	$(\varepsilon^{K-3}, 1)$	$(0,0)$
$3 < q \leq K$	$\left(\varepsilon^{K-1}, \frac{\rho}{y}\right)$	$(\varepsilon^{K-2}, 1)$	$(\varepsilon^{K-3}, 1)$	$(\varepsilon^{K-i}, 1)$ if $q \geq i$ else $(0,0)$

Table 1: Rewards and consumptions in the first K outcomes. Each cell shows the tuple of reward and consumption of an action i during batch i in the q-th outcome. The reward and consumption of action $a \neq i$ in batch i are both always 0.

The optimal solution for outcome $2 \leq q \leq K$ is at least the reward of playing the q-th arm with probability 1, resulting in its reward being

$$\mathtt{OPT}_q \geq \varepsilon^{K-q} T z$$

Now fix any algorithm which plays the i-th arm on batch i, $n_i T$ times. Because of the budget constraints, it must hold that

$$\frac{\rho}{y} n_1 T + \sum_{i=2}^{K} n_i T \leq T \rho \tag{13}$$

The reward of the algorithm on the q-th outcome is

$$\mathtt{REW}_q = \sum_{j=1}^{q} \varepsilon^{K-j} n_j T \leq \varepsilon^{K-q} (n_q + \varepsilon) T$$

This means that the fraction of the optimal solution the algorithm achieves in the q-th outcome for $q \geq 2$ is

$$\frac{\mathtt{REW}_q}{\mathtt{OPT}_q} \leq \frac{n_q}{z} + \frac{\varepsilon}{z}$$

Since ε, ρ, K are constants w.r.t. T, we notice that for every q it must hold that $n_q = \Theta(1)$; otherwise, if $n_q = o(1)$, in outcome q the algorithm would get a 0 fraction of the optimal solution as $T \to \infty$. This means that the above two relations are true when $T \to \infty$, in which case any regret terms disappear. Now we can take $\varepsilon \to 0$. Since the algorithm must guarantee a ρ fraction of the optimal solution this yields for $q \geq 2$

$$n_q \geq \rho z$$

which makes (13)

$$n_1 \leq y(1 - z(K - 1)) = y^2 \tag{14}$$

where in the last inequality we used the definition of $z = \frac{1-y}{K-1}$.

Now we introduce the $(K+1)$-th outcome, which is $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-stationary. In this outcome, during the first batch, the rewards and consumptions of the actions are identical to the other outcomes: only action 1 has positive reward and cost, as defined in (12). All actions have zero rewards and consumptions on the remaining batches, except for action 1. More specifically, for $t > Ty$, action 1 has

$$\left(r_t^{(K+1)}(1), c_t^{(K+1)}(1) \right) = \left(\sigma_{\mathbf{r}} \epsilon^{K-1}, \min\left\{ 1, \frac{\rho}{y} \frac{1}{\sigma_{\mathbf{c}}} \right\} \right)$$

which we notice is $(\sigma_{\mathbf{r}}, \sigma_{\mathbf{c}})$-stationary. Let $1/a = \min\left\{ 1, \frac{\rho}{y} \frac{1}{\sigma_{\mathbf{c}}} \right\}$.

In this new outcome, the optimal solution is the same as the 1st outcome, i.e., play the 1st arm with probability 1:

$$\mathtt{OPT}_{K+1} = \mathtt{OPT}_1 = \varepsilon^{K-1} Ty$$

Because all the outcomes are the same during the first batch, the algorithm cannot distinguish which batch it is in. This means it must play the first arm during the first batch $n_1 T$ times, entailing that the remaining budget after the first batch is

$$T\rho - n_1 T \frac{\rho}{y}$$

This entails that in the new outcome the number of times the first arm is played after the first batch is at most

$$\left(T\rho - n_1 T \frac{\rho}{y} \right) a$$

which proves that the reward for this outcome is at most

$$
\begin{aligned}
\mathtt{REW}_{K+1} \;&\leq\; \varepsilon^{K-1} n_1 T + \sigma_{\mathbf{r}} \varepsilon^{K-1} \left(T\rho - n_1 T \frac{\rho}{y} \right) a \\
&=\; \varepsilon^{K-1} T \left(\rho \sigma_{\mathbf{r}} a + n_1 \left(1 - \frac{\rho \sigma_{\mathbf{r}} a}{y} \right) \right) \\
&\leq\; \varepsilon^{K-1} T \left(\rho \sigma_{\mathbf{r}} a + y^2 \left(1 - \frac{\rho \sigma_{\mathbf{r}} a}{y} \right) \right) \qquad \left(\text{using (14) and } \frac{\rho \sigma_{\mathbf{r}} a}{y} \leq 1 \right) \\
&=\; \varepsilon^{K-1} T \left(\rho \sigma_{\mathbf{r}} a + y^2 - \rho y \sigma_{\mathbf{r}} a \right)
\end{aligned}
$$

making the fraction of the optimal solution that the algorithm gets at most

$$
\begin{aligned}
\frac{\mathtt{REW}_{K+1}}{\mathtt{OPT}_{K+1}} \;&\leq\; \frac{\rho \sigma_{\mathbf{r}} a + y^2 - \rho y \sigma_{\mathbf{r}} a}{y} \\
&=\; y + \frac{\rho \sigma_{\mathbf{r}}(1 - y)}{y} \alpha \\
&=\; y + \frac{\rho \sigma_{\mathbf{r}}(1 - y)}{y} \max\left\{ 1, \frac{y \sigma_{\mathbf{c}}}{\rho} \right\}
\end{aligned}
$$

Similar to before, because the above expression does not depend on T, it also holds as $T \to \infty$. This makes any regret terms disappear, making the above the inverse of the competitive ratio. Minimizing it over y we set

$$y = \begin{cases} \rho, & \text{if } \sigma_{\mathbf{r}} \leq \rho \\ \sqrt{\rho \sigma_{\mathbf{r}}}, & \text{if } \rho \leq \sigma_{\mathbf{r}} \leq \frac{\rho}{\sigma_{\mathbf{c}}^2} \\ \frac{\rho}{\sigma_{\mathbf{c}}}, & \text{if } \sigma_{\mathbf{r}} \geq \frac{\rho}{\sigma_{\mathbf{c}}^2} \end{cases}$$

which proves the desired bound simply by substituting. $\blacksquare$

Appendix C. Deferred proofs of Section 7

In this section, we present the deferred proofs of Section 7. We start with Lemma 7.

Proof of Lemma 7 For ease of notation, for the distribution mentioned in the lemma's statement A, we use $r_t(A) = \mathbb{E}_{a \sim A}\left[r_t(a)\right]$ and $c_{t,i}(A) = \mathbb{E}_{a \sim A}\left[c_{t,i}(a)\right]$.

Fix a batch j. Let $\mathcal{T}_j$ be the rounds of that batch and $\texttt{REW}_j$ be the total reward of the algorithm during it. The per-round budget during batch j is $\frac{\rho T_{\texttt{res}} - 1}{T_{\texttt{res}}} = \rho - 1/T_{\texttt{res}}$. We use Lemma 3, in which we use the distribution of actions that plays A (as defined in the description of Lemma 7) with probability $\min_i \min\left\{1, \frac{\rho}{\max_{t \in T_j} c_{t,i}(A)}\right\} - 1/T_{\texttt{res}}$ and the null arm otherwise. This proves that with probability $1 - \delta$,

$$\texttt{REW}_j \geq \left(\min_i \min\left\{1, \frac{\rho}{\max_{t \in \mathcal{T}_j} c_{t,i}(A)}\right\} - 1\right)\sum_{t \in T_j} r_t(A) - \texttt{Reg}_{\max}(T_{\texttt{res}}, \delta) - \texttt{Reg}_{\min}(T_{\texttt{res}}, \delta)$$

$$\geq \min_i \min\left\{1, \frac{\rho}{\max_{t \in \mathcal{T}_j} c_{t,i}(A)}\right\}\sum_{t \in T_j} r_t(A) - 1 - \texttt{Reg}_{\max}(T_{\texttt{res}}, \delta) - \texttt{Reg}_{\min}(T_{\texttt{res}}, \delta)$$

$$= \min_i \min\left\{1, \frac{\rho}{\max_{t \in \mathcal{T}_j} c_{t,i}(A)}\right\}\sum_{t \in T_j} r_t(A) - \texttt{Reg}_{\max}(T_{\texttt{res}}, \delta) - \texttt{Reg}_{\min}(T_{\texttt{res}}, \delta) \quad (15)$$

where in the last equality we used the fact that $\texttt{Reg}_{\min}(T_{\texttt{res}}, \delta)$ is asymptotically bigger than 1.

Let for every $i \in [d]$

$$\mathcal{E}_{ji} = \max_{t \in \mathcal{T}_j} c_{t,i}(A) - \min_{t \in \mathcal{T}_j} c_{t,i}(A).$$

We are going to prove that for any $t \in \mathcal{T}_j$ and $i \in [d]$ it holds that

$$\min\left\{1, \frac{\rho}{\max_{\tau \in \mathcal{T}_j} c_{\tau,i}(A)}\right\} \geq \min\left\{1, \frac{\rho}{c_{t,i}(A)}\right\} - \frac{1}{\rho}\mathcal{E}_{ji} \quad (16)$$

Combining (15) and (16) we get that with probability at least $1 - \delta$

$$\texttt{REW}_j + \texttt{Reg}_{\max}(T_{\texttt{res}}, \delta) + \texttt{Reg}_{\min}(T_{\texttt{res}}, \delta) \geq \sum_{t \in T_j} r_t(A) \min_i \min\left\{1, \frac{\rho}{c_{t,i}(A)}\right\} - \frac{T_{\texttt{res}}}{\rho}\max_i \mathcal{E}_{ji}$$

23

With probability $1 - \delta T_{\texttt{res}}/T$ the above inequality holds for all j; summing over all j gets us

$$\texttt{REW} + \frac{T}{T_{\texttt{res}}}\big(\texttt{Reg}_{\max}(T_{\texttt{res}}, \delta) + \texttt{Reg}_{\min}(T_{\texttt{res}}, \delta)\big) \geq \sum_{t=1}^{T} r_t(A) \min_i \min\left\{1, \frac{\rho}{c_{t,i}(A)}\right\} - \frac{T_{\texttt{res}}}{\rho}\sum_j \max_i \mathcal{E}_{ji}$$

which proves what we want by noticing that $\mathcal{E} \geq \sum_j \max_i \mathcal{E}_{ji}$ and substituting δ.

Now we need to prove (16). Fix i and j. We have that

$$\min\left\{1, \frac{\rho}{\max_{\tau \in \mathcal{T}_j} c_{\tau,i}(A)}\right\} + \frac{\mathcal{E}_{ji}}{\rho} = \min\left\{1, \frac{\rho}{\min_{\tau \in \mathcal{T}_j} c_{\tau,i}(A) + \mathcal{E}_{ji}}\right\} + \frac{\mathcal{E}_{ji}}{\rho}$$

Let $c = \min_{\tau \in \mathcal{T}_j} c_{\tau,i}(A)$. We notice that the above r.h.s. is increasing in $\mathcal{E}_{ji}$; this is obvious for $\rho > c + \mathcal{E}_{ji}$ and if $c + \mathcal{E}_{ji} \geq \rho$ the derivative is $\frac{(c+\mathcal{E}_{ji})^2 - \rho^2}{\rho(c+\mathcal{E}_{ji})} \geq 0$. This means that, because $\mathcal{E}_{ji} \geq 1 - c$, it holds that

$$\min\left\{1, \frac{\rho}{c + \mathcal{E}_{ji}}\right\} + \frac{\mathcal{E}_{ji}}{\rho} \geq \min\{1, \rho\} + \frac{1-c}{\rho} \geq \rho + 1 - c = (1-c) + c\frac{\rho}{c} \geq \min\left\{1, \frac{\rho}{c}\right\}$$

where in the last inequality we used the fact that the min is less than the weighted average. (16) follows by noticing that for any $t \in \mathcal{T}_j$, $c \leq c_{t,i}(A)$. ∎

Now we prove Theorem 8.

Proof of Theorem 8 Fix the solution of (1), (A^*, T^*) and let $x = T^*/T$. We are going to use Lemma 7 by setting $A = A^*$. To get a bound on the competitive ratio, we need to lower bound

$$\sum_{t=1}^{T} \mathbb{E}_{a \sim A^*}[r_t(a)] \min\left\{1, \frac{\rho}{\max_i \mathbb{E}_{a \sim A^*}[c_{t,i}(a)]}\right\} \tag{17}$$

First we prove a bound on $\min_t \mathbb{E}_{a \sim A^*}[r_t(a)]$. The bound we prove is implicitly proven inside the proof of Lemma 10, but we include it for completeness:

$$\min_t \mathbb{E}_{a \sim A^*}[r_t(a)] = \min_t \sum_{a=1}^{K} \mathbb{P}[A^* = a]\, r_t(a)$$

$$\geq \sum_{a=1}^{k} \mathbb{P}[A^* = a] \min_t r_t(a)$$

$$\geq \sigma_{\mathbf{r}} \sum_{a=1}^{K} \mathbb{P}[A^* = a] \max_t r_t(a)$$

$$\geq \frac{\sigma_{\mathbf{r}}}{T^*} \sum_{a=1}^{K} \mathbb{P}[A^* = a] \sum_{t=1}^{T^*} r_t(a)$$

$$= \frac{\sigma_{\mathbf{r}}}{xT} \texttt{OPT}_{\texttt{FD}} \tag{18}$$

Let $c_t = \max_i \mathbb{E}_{a \sim A^*} [c_{t,i}(a)]$ and $\bar{c} = \max_t c_t$. We focus on the case where $\bar{c} \geq \rho$, as otherwise (17) is at least $\mathrm{OPT}_{\mathrm{FD}}$. We first focus on the sum of (17) for $t > T^*$:

$$\sum_{t=T^*+1}^{T} \mathbb{E}_{a \sim A^*} [r_t(a)] \min\left\{1, \frac{\rho}{c_t}\right\} \geq \frac{\rho}{\bar{c}} T(1-x) \min_t \mathbb{E}_{a \sim A^*} [r_t(a)] \geq \frac{\rho}{\bar{c}} \sigma_{\mathrm{r}} \frac{1-x}{x} \mathrm{OPT}_{\mathrm{FD}} \qquad (19)$$

where in the last inequality we used (18).

To analyze the sum in (17) for $t \leq T^*$ we take two cases. First, we have that

$$\sum_{t=1}^{T^*} \mathbb{E}_{a \sim A^*} [r_t(a)] \min\left\{1, \frac{\rho}{c_t}\right\} \geq \frac{\rho}{\bar{c}} \sum_{t=1}^{T^*} \mathbb{E}_{a \sim A^*} [r_t(a)] = \frac{\rho}{\bar{c}} \mathrm{OPT}_{\mathrm{FD}} \qquad (20)$$

Second, we have that

$$\begin{aligned}
\sum_{t=1}^{T^*} \mathbb{E}_{a \sim A^*} [r_t(a)] \min\left\{1, \frac{\rho}{c_t}\right\} &\geq \left(\min_t \mathbb{E}_{a \sim A^*} [r_t(a)]\right) \sum_{t=1}^{T^*} \min\left\{1, \frac{\rho}{c_t}\right\} \\
&\geq \left(\min_t \mathbb{E}_{a \sim A^*} [r_t(a)]\right) \frac{\rho T^{*2}}{\sum_{t=1}^{T^*} \max\{\rho, c_t\}} \qquad \text{(AM-HM inequality)} \\
&\geq \left(\min_t \mathbb{E}_{a \sim A^*} [r_t(a)]\right) \frac{\rho T^{*2}}{\sum_{t=1}^{T^*}(c_t + \rho)} \\
&\geq \left(\min_t \mathbb{E}_{a \sim A^*} [r_t(a)]\right) \frac{\rho T^{*2}}{d\rho T + \rho T^*} \qquad \left(\forall i : \sum_{t=1}^{T^*} c_{t,i}(A^*) \leq \rho T\right) \\
&= \min_t \mathbb{E}_{a \sim A^*} [r_t(a)] \frac{x^2 T}{d+x} \qquad (T^* = xT) \\
&= \sigma_{\mathrm{r}} \frac{x}{d+x} \mathrm{OPT}_{\mathrm{FD}} \qquad \text{(using (18))} \qquad (21)
\end{aligned}$$

Now we combine (19), (20), and (21) to get

$$\sum_{t=1}^{T} \mathbb{E}_{a \sim A^*} [r_t(a)] \min\left\{1, \frac{\rho}{\max_i c_{t,i}(A^*)}\right\} \geq \max\left\{\frac{\rho}{\bar{c}}, \sigma_{\mathrm{r}} \frac{x}{d+x}\right\} \mathrm{OPT}_{\mathrm{FD}} + \frac{\rho}{\bar{c}} \sigma_{\mathrm{r}} \frac{1-x}{x} \mathrm{OPT}_{\mathrm{FD}}$$

using Lemma 11 and that $\bar{c} \leq 1$ it holds that $\bar{c} \leq \min\{1, \rho/x\sigma_{\mathrm{c}}\}$ which we use to get the desired bound. ∎