Unifying the Stochastic and the Adversarial Bandits with Knapsack

Anshuka Rangi [1], **Massimo Franceschetti**[1] and **Long Tran-Thanh** [2]
[1] University of California, San Diego,
[2] University of Southampton
arangi@ucsd.edu, massimo@ece.ucsd.edu, L.Tran-Thanh@soton.ac.uk

Abstract

This work investigates the adversarial Bandits with Knapsack (BwK) learning problem, where a player repeatedly chooses to perform an action, pays the corresponding cost of the action, and receives a reward associated with the action. The player is constrained by the maximum budget that can be spent to perform the actions, and the rewards and the costs of these actions are assigned by an adversary. This setting is studied in terms of expected regret, defined as the difference between the total expected rewards per unit cost corresponding the best fixed action and the total expected rewards per unit cost of the learning algorithm. We propose a novel algorithm EXP3.BwK and show that the expected regret of the algorithm is order optimal in the budget. We then propose another algorithm EXP3++.BwK, which is order optimal in the adversarial BwK setting, and incurs an almost optimal expected regret in the stochastic BwK setting where the rewards and the costs are drawn from unknown underlying distributions. These results are then extended to a more general online learning setting, by designing another algorithm EXP3++.LwK and providing its performance guarantees. Finally, we investigate the scenario where the costs of the actions are large and comparable to the budget. We show that for the adversarial setting, the achievable regret bounds scale at least linearly with the maximum cost for any learning algorithm, and are significantly worse in comparison to the case of having costs bounded by a constant, which is a common assumption in the BwK literature.

1 Introduction

Sequential decision making is a fundamental task faced by any agent interacting with the environment, and the optimal (or near optimal) behavior has been studied in in a number of settings [Auer *et al.*, 2002; Puterman, 2014; Rangi *et al.*, 2018c; 2018b; 2018a]. Multi-Armed Bandit (MAB) is a sequential decision making problem under uncertainty that is based on balancing the trade-off between exploration and exploitation, i.e. "the conflict between taking actions whose benefits will be seen later and taking actions which yield immediate rewards." A common feature in various applications of MAB is that the resources consumed during the decision making process are limited. For instance, scientists experimenting with alternative medical treatments may be limited by the number of patients participating in the study as well as by the cost of the material used in the treatments. Similarly, in web advertisements, a website experimenting with displaying advertisements is constrained by the number of users who visit the site as well as by the advertisers' budgets. A retailer engaging in price experimentation faces inventory limits along with a limited number of consumers. A model which incorporates a budget constraint on these supply limits is Bandits with Knapsack (BwK). This can be seen as a game between a player and an adversary (or environment) that evolves for T rounds, and the player is constrained by a budget B on the resources consumed during the decision making process. At each round $t \leq T$, the player performs an action i from a set of K actions, pays a cost for this action i from the budget B and receives a reward in $[0, 1]$ for this action. The game terminates when the player runs out of the budget, therefore T is dependent on B and the player's strategy. The reward and the cost can vary from application to application. For example, in web advertisement, the reward is the click through rate and the cost is the space occupied by the advertisement on the web page. In medical trials, the reward is the success rate of the medicine and the cost corresponds to the cost of the material used.

The BwK problem can be classified into two categories: stochastic BwK and adversarial BwK. In stochastic BwK, the reward and the cost of the actions are independent and identically distributed (i.i.d) sequences over T rounds drawn from fixed unknown distributions. In adversarial BwK, the sequence of the rewards and the costs associated with each action over T rounds is assigned by an oblivious adversary before the game starts. The objective of the player is to minimize the expected regret in these settings.

The stochastic BwK setting has been extensively studied in the literature [Tran-Thanh *et al.*, 2010; 2012a; Ding *et al.*, 2013; Badanidiyuru *et al.*, 2013; Agrawal and Devanur, 2014; Tran-Thanh *et al.*, 2012b; Agrawal and Devanur, 2016; Xia *et al.*, 2016; Sankararaman and Slivkins, 2018; Rangi and Franceschetti, 2018a]. However, limited attention has been received by the adversarial BwK setting [Immorlica *et*

al., 2018; Zhou and Tomlin, 2018]. It has been shown that the competitive ratio, defined as the ratio of the expected rewards corresponding to best fixed distribution over actions and the expected rewards of any learning algorithm, is at least $\Omega(\log(B))$ [Immorlica *et al.*, 2018]. Thus, the difference between the expected rewards corresponding to the best fixed distribution over the actions and the expected rewards of the learning algorithm, is at least $\Omega(B)$. The work in [Zhou and Tomlin, 2018], assumes that the rewards are greater than the costs at each round t for every action i, and reduces the problem to an adversarial MAB setting [Auer *et al.*, 2002]. This assumption is somewhat uncommon in the literature and does not have physical meaning in many applications. For example, in a medical trial, the success rate (i.e., the reward) and the cost of the material (i.e., the cost) cannot be compared with each other. Likewise, in web advertisement, the comparison of the click through rate (i.e., the reward) and the space occupied by the advertisement on the web page (i.e., the cost) has no meaning. Thus, determining the feasibility of sub-linear regret bounds in the adversarial BwK setting, and the designing an algorithm without constraints on the rewards and the costs remain important open problems.

An additional open problem is to provide an algorithm that is satisfactory for both stochastic and adversarial BwK settings simultaneously. In many real-world situations, there is no information about the stochastic or the adversarial nature of the problem. Thus, the deployed algorithm has to be able to perform well in both the cases. Finally, the literature of the BwK problem typically assumes that the costs are bounded by a constant independent of the budget B, and it is unknown whether the state-of-the-art regret bounds hold for the case of large costs (i.e., when costs are comparable to the budget B).

In this framework, the contribution of our work is three fold. First, we study the adversarial BwK setting in terms of a new definition of expected regret defined as the difference between the total expected rewards per unit cost corresponding the best fixed action (instead of best fixed distribution over actions) and the total expected rewards per unit cost of the learning algorithm. We establish that, unlike in [Immorlica *et al.*, 2018], sub-linear regret bounds of $O(\sqrt{B})$ are feasible for this new notion of regret in BwK setting. We propose a new algorithm EXP3.BwK and show that its expected regret is $O(\sqrt{BK\log K})$. We also prove that the expected regret for any learning algorithm is at least $\Omega(\sqrt{KB})$ in the adversarial BwK setting, and establish that EXP3.BwK is order optimal. Additionally, this analysis is carried out without the assumption on the rewards and the costs previously used in [Zhou and Tomlin, 2018]. Second, we unify the stochastic and the adversarial settings by proposing EXP3++.BwK. This algorithm incurs an expected regret of $O(\sqrt{BK\log K})$ and $O(\log^2(B))$ in the adversarial and the stochastic BwK settings respectively. In the stochastic BwK setting, the regret bound has an additional factor of $\log(B)$ in comparison to the optimal expected regret i.e. $O(\log(B))$. Thus, EXP3++.BwK exhibits an almost optimal behavior in both the stochastic and the adversarial settings simultaneously. We also study the performance of another algorithm EXP3++.LwK for a generic online learning setting where the player's feedback can be viewed as a time-varying graph G_t at round t and a directed edge $k \to j$ in G_t indicates that choosing action k at round t also reveals the reward and the cost of action j at round t. Finally, we show that if the maximum cost is bounded above by B^α, where $\alpha \in [0,1]$, then the expected regret in the adversarial BwK setup scales at least linearly with the maximum cost, namely it is $\Omega(B^\alpha)$. Thus, for $\alpha > 1/2$, it is impossible to achieve a regret bound of $O(\sqrt{B})$, which is feasible for small costs. The proofs of all the results are available online [Rangi *et al.*, 2018d].

1.1 Related Work

In the MAB literature, the problem of finding one algorithm for both the stochastic and the adversarial setting has been referred as "Best of Both Worlds" [Bubeck and Slivkins, 2012; Auer and Chiang, 2016; Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Lykouris *et al.*, 2018]. Initial works on this problem reduced it to a detection problem, where the algorithms initially assume the rewards are stochastic, and perform an irreversible switch to EXP3.P if the beginning of the game is estimated to exhibit an adversarial or non-stochastic behavior [Bubeck and Slivkins, 2012; Auer and Chiang, 2016]. Unlike the algorithms in [Bubeck and Slivkins, 2012; Auer and Chiang, 2016], EXP3++ starts by assuming that the rewards exhibit a non-stochastic behavior and adapts itself as it encounters stochastic behavior on rewards [Seldin and Slivkins, 2014; Seldin and Lugosi, 2017]. The algorithm guarantees an expected regret of $O(\log^2(T))$ and $O(\sqrt{T})$ in the stochastic and the adversarial MAB settings respectively. The problem of stochastic bandits corrupted with adversarial samples has also been studied in the regime of small corruptions [Lykouris *et al.*, 2018]. The work provides the regret analysis of the algorithm in terms of the corruption in the rewards, and shows that the decay in performance is order optimal in this corruption.

The "Best of Both Worlds" problem has not been studied in the BwK setting. This setting is more complex in comparison to the MAB setting as the adversary can choose both the rewards and the costs associated with the actions. Additionally, the adversary can dynamically choose to corrupt the rewards or the costs at each step which can lead to stochastic environment based on the cumulative effects of the corruptions.

2 Problem Formulation

A player has a total budget B to perform actions. At each round t, the player performs an action $i_t \in [K]$, pays the cost $c_t(i_t)$ and receives the reward $r_t(i_t)$, where $c_t(i) \in [c_{min}, c_{max}]$ is the cost of action i and $r_t(i) \in [0,1]$ is the reward of action i. Note that $c_{max} = 1$ which is also a common assumption in the literature, and we will investigate the case with larger costs in Section 5. The objective of a player is to design a learning algorithm $\mathcal{A}$ such that

$$\begin{aligned} &\max_{\{i_1,i_2,\ldots,i_{\tau(\mathcal{A})}\}} \mathbf{E}\Big[\sum_{t=1}^{\tau(\mathcal{A})} r_t(i_t)\Big] \\ &\text{s.t. } \mathbf{P}\Big(\sum_{t=1}^{\tau(\mathcal{A})} c_t(i_t) \le B\Big) = 1, \end{aligned} \tag{1}$$

where $\tau(\mathcal{A})$ is the number of rounds feasible in the budget B using algorithm $\mathcal{A}$. The optimization problem in (1) is a knapsack problem, and is known to be NP-hard [Kellerer *et al.*, 2004]. Given that at each round t, the reward and the cost of all the actions are known and fixed i.e. $r_t(i) = r(i)$ and $c_t(i) = c(i)$, the best fixed action i^* is the action with the highest efficiency, namely

$$i^* = \text{argmax}_{i\in[K]} e(i), \tag{2}$$

where efficiency $e(i) = r(i)/c(i)$ for an action $i \in [K]$. This choice of best fixed action is referred as a greedy algorithm $\mathcal{A}^G$ and satisfies the following [Kellerer *et al.*, 2004]

$$G(\mathcal{A}^G) \leq G(\mathcal{A}^*) \leq G(\mathcal{A}^G) + \max_{i\in[K]} e(i), \tag{3}$$

where $G(\mathcal{A}) = \mathbf{E}\big[\sum_{t=1}^{\tau(\mathcal{A})} r_t(i_t)\big]$ is the expected reward of $\mathcal{A}$, and $\mathcal{A}^*$ is the optimal solution of (1). Intuitively, the best fixed action i^* returns the maximum rewards per unit cost. Since the total cost spent is at most B, $G(\mathcal{A}^G)$ is at most $e(i^*) \cdot B$.

We define the expected regret in both stochastic and adversarial settings with respect to the best fixed optimal action based on the efficiency according to $\mathcal{A}^G$. In the stochastic setting, for all t and $i \in [K]$, the sequence of rewards $r_t(i)$ is i.i.d with mean $\mu(i)$ and the sequence of cost $c_t(i)$ is i.i.d with mean $\rho(i)$. However, the expected reward $\mu(i)$ and expected cost $\rho(i)$ can be correlated. Thus, the efficiency of an action i is $e(i) = \mu(i)/\rho(i)$ in stochastic setting, and the expected regret of an algorithm $\mathcal{A}$ with respect to best fixed action, given by $\mathcal{A}^G$, is defined as

$$R(\mathcal{A}) = T(i^*)\mu(i^*) - \mathbf{E}\Big[\sum_{t=1}^{\tau(\mathcal{A})} r_t(i_t)\Big], \tag{4}$$

where $i^* = \text{argmax}_{i\in[K]} e(i)$, and $T(i)$ is the number of rounds of action i feasible in the budget B. Thus, the best fixed action i^* returns maximum expected rewards per unit expected cost. The regret, defined in (4), is previously studied in the literature of stochastic BwK setting as well, and the performance guarantees are presented in terms of $\Delta(i) = e(i^*) - e(i)$ [Ding *et al.*, 2013; Tran-Thanh *et al.*, 2012b; Rangi and Franceschetti, 2018a].

In the adversarial setting, $r_t(i)$ and $c_t(i)$ are chosen by an adversary. Unlike the efficiency in the stochastic setting, the efficiency of an action i is time varying, and is denoted by $e_t(i) = r_t(i)/c_t(i)$ at round t. Since the reward per unit cost varies across t, the best fixed action in the hindsight is defined as

$$i^* = \text{argmax}_{i\in[K]} \sum_{t=1}^{T(i)} e_t(i). \tag{5}$$

The action i^* maximizes the total rewards per unit cost across $T(i^*)$ rounds, and $T(i^*)$ takes into account the total budget utilized while performing i^* repeatedly. Since $c_{max} = 1$, the maximum rewards achievable in adversarial setting is at most $\sum_{t=1}^{T(i^*)} e_t(i^*)$. Therefore, the expected regret in this setting is defined as

$$R(\mathcal{A}) = \mathbf{E}\Big[\sum_{t=1}^{T(i^*)} e_t(i^*) - \sum_{t=1}^{\tau(A)} e_t(i_t)\Big]. \tag{6}$$

For $c_t(i) = 1$, action i^* reduces to the best fixed action over B rounds in MAB setup, and the regret in (6) reduces to the regret in adversarial MAB setting [Auer *et al.*, 2002].

3 Adversarial Bandits with Knapsack

In this section, we propose an order optimal algorithm EXP3.BwK for the adversarial BwK setting, and show that sub-linear regret guarantees are feasible in this setting.

EXP3.BwK is presented in Algorithm 1. At each round t, EXP3.BwK selects an action $i_t = i$ with probability $p_t(i)$, pays the cost $c_t(i_t)$ from the remaining budget B_r, and receives the reward $r_t(i_t)$. If the budget B_r is insufficient, i.e. $c_t(i_t) > B_r$, then the algorithm terminates without attempting to find other feasible actions which can be performed using the remaining budget B_r. Since $c_{max} = 1$, B_r at the termination round is at most unity, and the number of rounds feasible with this remaining budget is at most $1/c_{min}$, which is independent of B.

In EXP3.BwK, the empirical estimate $\hat{e}_t(i)$ of efficiency (defined in Algorithm 1) of action i is the observed efficiency $e_t(i) = r_t(i)/c_t(i)$ scaled with the probability $p_t(i)$. This empirical estimate is used to maintain the set of time-varying weights $\{w_t(i)\}_{i=1}^K$ for all actions. The difference in $w_t(i)$ and $w_{t-1}(i)$ is controlled by $\exp(\gamma c_{min}\hat{e}_t(i)/K)$ as $\gamma c_{min}\hat{e}_t(i)/K$ less than unity. Thus, $w_t(i)$ is proportional to the cumulative estimated efficiency of action i observed until round t i.e $w_t(i) \propto \exp(\sum_{n=1}^{t-1} \hat{e}_n(i))$.

The sampling probability $p_t(i)$ is dependent on the time-varying weight $w_t(i)$ and the exploration constant γ/K. The weight $w_t(i)$ favors the selection of an action with higher cumulative estimated efficiencies until round t. In other words, it favors the action which may yield immediate rewards, and is responsible for the exploitation in the algorithm. On the other hand, the exploration constant γ/K ensures that the player is always exploring with a positive probability in search of the optimal action i^*. This balances the trade-off between exploration and exploitation.

In EXP3.BwK, we exploit the idea of efficiency measure $e_t(i)$ for tracking the contributions of each action $i \in [K]$. The use of this measure is motivated from the greedy algorithm $\mathcal{A}^G$, and its performance guarantees with respect to the optimal solution (see (3)). The advantages of using this measure are two folds. First, it eliminates the need of the assumption that at each round, the reward is greater than the cost for each action [Zhou and Tomlin, 2018]. Second, it can track rewards of the algorithm $\mathcal{A}$ irrespective of the measure or scale of the rewards and the costs of the actions. For example, in a recommendation system, the space (i.e. the cost) of the item and the click rate (i.e. the reward) of the item are not comparable on same scale. However, the returns can be tracked using the efficiency measure $e_t(i)$. An alternate choice of efficiency measure can be $r_t(i) - c_t(i)$ for tracking the contributions of different actions. However, this can only be used if both the rewards and the costs can be compared on a linear scale, which is not true in many practical applications.

The following theorem shows that the expected regret of EXP3.BwK is sub-linear in the budget B.

Algorithm 1 EXP3.BwK

Initialization: γ ; For all $i \in [K]$, $w_1(i) = 1$, and $\hat{e}_1(i) = 0$; $t = 1$; $B_r = B$
while $B_r > 0$ **do**
  $W_t = \sum_{i\in[K]} w_t(i)$
  Update $p_t(i) = (1-\gamma)w_t(i)/W_t + \gamma/K$
  Choose $i_t = i$ with probability $p_t(i)$.
  Observe $(r_t(i_t), c_t(i_t)$
  if $c_t(i_t) > B_r$ **then**
    Break;
  end if
  $B_r = B_r - c_t(i_t)$
  For all $i \in [K]$, $\hat{e}_t(i) = r_t(i)\mathbf{1}(i = i_t)/c_t(i_t)p_t(i)$, and $w_{t+1}(i) = w_t(i) \cdot \exp(\gamma \cdot c_{min} \cdot \hat{e}_t(i)/K)$
  $t = t + 1$
end while

Theorem 1. *For $\gamma = \sqrt{c_{min}K\log(K)/B(e-1)}$, the expected regret, as defined in (6), of EXP3.BwK is at most*

$$R(E) \leq \frac{2}{c_{min}}\sqrt{\frac{(e-1)BK\log(K)}{c_{min}}}, \tag{7}$$

where E denotes EXP3.BwK.

The following theorem provides the lower bound on the expected regret of any learning algorithm in the adversarial BwK setting.

Theorem 2. *There exists an adversary such that for any player's learning algorithm $\mathcal{A}$, the expected regret of the algorithm $\mathcal{A}$ is at least $\Omega(\sqrt{KB/c_{min}})$.*

Combining Theorem 1 and Theorem 2, it follows that the expected regret of EXP3.BwK is order optimal in the budget B and has an additional factor of $1/c_{min}$. Additionally, unlike [Immorlica *et al.*, 2018], Theorem 1 and Theorem 2 establish that it is feasible to attain sub-linear regret bounds with respect to the best fixed action based on the efficiency for BwK with single constraint (1). The order optimality of EXP3.BwK also highlights an important feature of an alternate class of algorithms in the adversarial BwK setup. Consider a new class of algorithms $\mathcal{G}$ which looks for an alternative action to perform after the algorithm is unable to pay the cost, i.e. $c_t(i_t) > B_r$, in order to utilize the remaining budget efficiently. Since EXP3.BwK terminates if it is unable to pay $c_t(i_t)$, EXP3.BwK does not belong to $\mathcal{G}$, and is still order optimal in the budget B. Therefore, the expected regret of the algorithms in this new class $\mathcal{G}$ will have same dependency on B as that of EXP3.BwK. Additionally, the difference between the expected regret of EXP3.BwK and the algorithms in $\mathcal{G}$ will be at most $1/c_{min}$, a constant independent of B. The algorithms in $\mathcal{G}$ faces an additional challenge of designing an appropriate criterion for the termination of the algorithm because the costs are assigned by the adversary.

The ideas developed in EXP3.BwK form the basis for designing an algorithm that achieves almost optimal performance guarantees in both the stochastic and the adversarial BwK settings simultaneously.

4 One Practical Algorithm for Both Stochastic and Adversarial BwK

In this section, we propose EXP3++.BwK, and show that it achieves almost optimal performance guarantees in both the stochastic and the adversarial BwK settings simultaneously.

Before discussing EXP3++.BwK, let us briefly focus on the fundamental difference between the optimal algorithms in the stochastic and the adversarial settings. In the stochastic setting, the algorithms focus on exploration in the initial stage until reliable estimates of the expected rewards $\mu(i)$ and expected costs $\rho(i)$ are achieved. Then, the algorithms shift their focus on exploitation by choosing actions which yield immediate rewards, and perform exploration only with a small probability. For instance, in UCB type of algorithms, the probability of exploration or choosing sub-optimal actions decays as $O(1/t^2)$ with round t [Tran-Thanh *et al.*, 2012a; Ding *et al.*, 2013; Rangi and Franceschetti, 2018a]. In greedy algorithms, the probability of exploration is exactly zero after a fixed round (or time instance), and the algorithm chooses the best action based on its knowledge at this time instance [Tran-Thanh *et al.*, 2010; 2012b]. On the contrary, in the adversarial setting, the algorithms are always exploring, and looking for the actions with high returns [Auer *et al.*, 2002; Rangi and Franceschetti, 2018b]. For instance, in EXP3.BwK, the exploration constant γ/K does not change with round t, and is dependent only on the total number of rounds i.e. $\Theta(B)$ in the BwK setup. Intuitively, an algorithm that is optimal in both these settings needs to efficiently adapt its exploration phase based on the nature of the observations while balancing the exploration and exploitation trade-off.

EXP3++.BwK is presented in Algorithm 2. It is built upon the ideas of the efficiencies in the stochastic and the adversarial BwK settings. Like EXP3.BwK, at each round t, EXP3++.BwK selects an action $i_t = i$ with sampling probability $\tilde{p}_t(i)$, pays the cost $c_t(i_t)$ from the remaining budget B_r, and receives reward $r_t(i_t)$. The algorithm terminates if it is unable to pay the cost $c_t(i_t)$ at any round t, namely $c_t(i_t) > B_r$.

The sampling probability $\tilde{p}_t(i)$ is dependent on two time varying parameters: the exploration parameter $\epsilon_t(i)$ and the exploitation parameter $p_t(i)$. Unlike EXP3.BwK, exploration parameter $\epsilon_t(i)$ is time-varying, and helps to efficiently adapt the exploration phase of the algorithm based on the stochastic or non-stochastic nature of the past observations. It is a function of $\text{UCB}_t(i)$, $\text{LCB}_t(i)$ and $\hat{\Delta}_t(i)$ (see Algorithm 2). At each round t, EXP3++.BwK maintains an Upper Confidence Bound $\text{UCB}_t(i)$ and a Lower Confidence Bound $\text{LCB}_t(i)$ on the stochastic efficiency $e(i) = \mu(i)/\rho(i)$ of action i, where

$$\text{UCB}_t(i) = \min\left\{\frac{1}{c_{min}}, \bar{e}_t(i) + \frac{(1+1/\lambda)\eta_t(i)}{\lambda - \eta_t(i)}\right\}, \tag{8}$$

$$\text{LCB}_t(i) = \max\left\{0, \bar{e}_t(i) - \frac{(1+1/\lambda)\eta_t(i)}{\lambda - \eta_t(i)}\right\}, \tag{9}$$

$$\eta_t(i) = \sqrt{\frac{\alpha\log(K^{1/\alpha}t)}{2N_t(i)}}, \tag{10}$$

Algorithm 2 EXP3++.BwK

Initialization: For all $i \in [K]$, $w_1(i) = 1$, $\hat{e}_1(i) = 0$, $\bar{e}_1(i) = 0$, $N_1(i) = 1$ $\delta_1(i) > 0$; $t = 1$, $\gamma_t = 0.5c_{min}\sqrt{\log(K)/tK}$; $B_r = B$;
Perform each action once and update for all $i \in [K]$, $\bar{e}_1(i) = r_1(i)/c_1(i)$, $B_r = B_r - \sum_{i\in[K]} c_1(i)$ and $t = K + 1$.
while $B_r > 0$ **do**
  For all $i \in [K]$, update:
    $\text{UCB}_t(i)$ (see (8))
    $\text{LCB}_t(i)$ (see (9))
    $\hat{\Delta}_t(i)$ (see (11))
    $\delta_t(i) = \beta \log(t)/(t\hat{\Delta}_t(i)^2)$
    $\epsilon_t(i) = \min\{1/2K, 0.5\sqrt{\log(K)/t}, \delta_t(i)\}$
    $p_t(i) = \frac{\exp(-\gamma_t \hat{L}_{t-1}(i))}{\sum_{j\in[K]} \exp(-\gamma_t \hat{L}_{t-1}(j))}$
    $\tilde{p}_t(i) = (1 - \sum_{j\neq i} \epsilon_t(j))p_t(i) + \epsilon_t(i)$
  Choose $i_t = i$ with probability $\tilde{p}_t(i)$.
  Observe $(r_t(i_t), c_t(i_t))$
  if $c_t(i_t) > B_r$ **then**
    exit;
  end if
  $B_r = B_r - c_t(i_t)$
  For all $i \in [K]$, update:
    $\hat{e}_t(i) = r_t(i)\mathbf{1}(i = i_t)/\tilde{p}_t(i)c_t(i)$.
    $\hat{\ell}_t(i) = \mathbf{1}(i = i_t)/c_{min}\tilde{p}_t(i) - \hat{e}_t(i)$.
    $\hat{L}_t(i) = \sum_{n=1}^{t} \hat{\ell}_n(i)$
    $N_t(i) = N_{t-1}(i) + \mathbf{1}(i = i_t)$.
    $\bar{r}_t(i) = \sum_{n=1}^{t} r_n(i)\mathbf{1}(i = i_n)/N_t(i)$
    $\bar{c}_t(i) = \sum_{n=1}^{t} c_n(i)\mathbf{1}(i = i_n)/N_t(i)$
    $\bar{e}_t(i) = \bar{r}_t(i)/\bar{c}_t(i)$
  t=t+1
end while

$\lambda \leq c_{min}$ and $N_t(i)$ is the number of times action i has been selected until round t. These UCB and LCB on the stochastic efficiency are used to estimate the gap $\Delta(i)$. At round t, the estimate $\hat{\Delta}_t(i)$ of $\Delta(i)$ is defined as

$$\hat{\Delta}_t(i) = \max\{0, \max_{j\neq i} \text{LCB}_t(j) - \text{UCB}_t(i)\}. \tag{11}$$

It can been shown that for all $i \in [K]$, we have

$$\frac{\Delta(i)}{2} \leq \hat{\Delta}_t(i) \leq \Delta(i), \tag{12}$$

with high probability as $t \to \infty$ in the stochastic BwK setting [Rangi *et al.*, 2018d]. Thus, $\hat{\Delta}_t(i)$ is a reliable estimate of $\Delta(i)$, and helps to adapt the exploration of the algorithm via $\epsilon_t(i)$. In the stochastic BwK setup, combining the fact that $\Delta(i^*) = 0$ and (12) holds, the exploration parameter $\epsilon_t(i^*)$ of the optimal action i^* tends to zero, and favors its selection. On the other hand, the exploitation parameter $p_t(i)$ is computed based on the empirical estimate $\hat{e}_t(i)$ of efficiency until round t. Similar to EXP3.BwK, this favors the selection of an action with higher cumulative estimated efficiency $\sum_{n=1}^{t-1} \hat{e}_{t-1}(i)$ or lower cumulative estimated inefficiency $\sum_{n=1}^{t-1} \hat{\ell}_{t-1}(i)$. In conclusion, the sampling probability $\tilde{p}_t(i)$ is dependent on both the estimates of the efficiencies $\bar{e}_t(i)$ and $\hat{e}_t(i)$ where $\bar{e}_t(i)$ and $\hat{e}_t(i)$ are crucial in the stochastic BwK and the adversarial BwK settings respectively. $\bar{e}_t(i)$ controls the exploration through $\epsilon_t(i)$, and $\hat{e}_t(i)$ controls the exploitation through $p_t(i)$.

The following theorem provides the performance guarantees of EXP3++.BwK in the stochastic BwK setting.

Theorem 3. *In the stochastic BwK setting, for $\alpha = 3$ and $\beta = 256/c_{min}^2$, the expected regret of EXP3++.BwK is at most*

$$R(F) \leq M\left(\sum_{i:\Delta(i)>0} \frac{\log^2(B/c_{min})}{c_{min}^2\Delta(i)}\right), \tag{13}$$

where F denotes the algorithm EXP3++.BwK, and M is a constant.

The optimal regret guarantees in the stochastic BwK setting are $O(\log(B/c_{min}))$[Tran-Thanh *et al.*, 2012a; Ding *et al.*, 2013; Rangi and Franceschetti, 2018a]. Using Theorem 3, EXP3++.BwK has an additional factor of $\log(B/c_{min})$ in comparison to the optimal regret bounds in the literature. This additional factor is also common in the literature of the "Best of Both Worlds" problem in MAB setting [Seldin and Slivkins, 2014; Lykouris *et al.*, 2018].

The following theorem provides the performance guarantees of EXP3++.BwK in the adversarial BwK setting.

Theorem 4. *In the adversarial BwK setting, the expected regret of EXP3++.BwK is at most*

$$R(F) \leq \frac{2}{c_{min}}\sqrt{\frac{6BK\log(K)}{c_{min}}}. \tag{14}$$

Thus, like EXP3.BwK, EXP3++.BwK is order optimal in the adversarial BwK setting. In conclusion, using Theorem 3 and Theorem 4, EXP3++.BwK is order optimal in the adversarial BwK setting, and is almost optimal with an additional factor of $\log(B/c_{min})$ in the stochastic BwK setting.

Now, we extend these ideas beyond the MAB setting to a general online learning setting, where the player's feedback can be viewed as a time-varying graph G_t at round t [Alon *et al.*, 2015; Rangi and Franceschetti, 2018b]. At round t, a directed edge $k \to j$ in G_t indicates that choosing action k also reveals the reward and the cost of action j. Thus, $S_t(i) = \{j : i \to j \text{ is a directed edge in } G_t\}$ is the set of observable actions if i is performed. In MAB setting, $S_t(i) = \{i\}$. EXP3++.BwK can be modified to design a new algorithm EXP3++.LwK for this general online learning with Knapsack setting, which has not been addressed previously in the literature. The key difference is the estimation of the two efficiencies $\hat{e}_t(i)$ and $\bar{e}_t(i)$, and is presented in the following equations:

$$\hat{e}_t(i) = r_t(i)\mathbf{1}(i \in S_t(i_t))/(c_t(i)\sum_{j:j\to i} \tilde{p}_{j,t}), \tag{15}$$

$$\begin{aligned}\bar{e}_t(i) &= \bar{r}_t(i)/\bar{c}_t(i), \\ &= \sum_{n=1}^{t} r_n(i)\mathbf{1}(i \in S_n(i_n))/\sum_{n=1}^{t} c_n(i)\mathbf{1}(i \in S_n(i_n)).\end{aligned} \tag{16}$$

Likewise, $N_t(i) = \sum_{n=1}^{t} \mathbf{1}(i \in S_n(i_n))$ and $\hat{\ell}_t(i) = \mathbf{1}(i \in S_t(i_t))/c_{min} \sum_{j:j\to i} \tilde{p}_{j,t} - \hat{e}_t(i)$. The following theorem provides the performance guarantees of EXP3++.LwK in the adversarial online learning Knapsack setting.

Theorem 5. *In the adversarial online learning Knapsack setting, the expected regret of EXP3++.LwK is at most*

$$R(L) \leq b\sqrt{\sum_{n=1}^{T} mas(G_t)}, \tag{17}$$

where L denotes EXP3++.LwK, $T = B/c_{min}$, $mas(G_t)$ is the size of the maximal acyclic graph in G_t, and $b > 0$ is a constant dependent on K and c_{min}.

In the stochastic online learning Knapsack setting, the expected regret of EXP3++.LwK is $O(\log^2(B/c_{min}))$, and these guarantees are similar to the ones presented in Theorem 3. Now, the following theorem provides a lower bound on the expected regret in the online learning Knapsack setting.

Theorem 6. *For a sequence of feedback graphs $G_1, \ldots, G_T$ with independence sequence number $\beta(G_{1:T}) > 1$, there exists an adversary such that for any learning algorithm $\mathcal{A}$, the expected regret of the algorithm $\mathcal{A}$ is at least $\Omega(\sqrt{\beta(G_{1:T})B/c_{min}})$.*

Using Theorem 5 and 6, EXP3++.LwK is order optimal in two special cases of adversarial online learning: MAB and symmetric PI setting, i.e. the feedback graph $G_t = G$ is fixed and un-directed. In MAB setting $\beta(G_{1:T}) = \text{mas}(G_t) = K$, Theorem 6 recovers the lower bound in Theorem 2, and establishes that the algorithm is order optimal. Likewise, in symmetric PI setting $\beta(G_{1:T}) = \text{mas}(G_t)$ which implies the algorithm is order optimal. Note that any state-of-art algorithms in adversarial online learning setting are order optimal for these two special cases only [Alon *et al.*, 2015; Rangi and Franceschetti, 2018a], and the key challenges for closing this gap are highlighted in [Alon *et al.*, 2015].

5 BwK with Unbounded Cost

In the previous sections, we have shown that the sub-linear regret is achievable in the BwK and the online learning Knapsack settings for $c_{max} = 1$. We now explore the regime of varying c_{max}, and develop important insights about the achievable regret in this scenario. Following theorem presents the scaling of the lower bound on the expected regret with respect to c_{max} in the adversarial BwK setup.

Theorem 7. *Suppose that $c_{\max} = B^{\alpha}$. For any learning algorithm $\mathcal{A}$, there exists an adversary such that the expected regret of the algorithm is at least $\Omega(B^{\alpha})$.*

In the literature of BwK, the cost is always considered to be bounded above by a constant independent of B. Theorem 7 instead considers that the cost is bounded by a function of the budget B. It shows that the lower bound on the expected regret scales at least linearly with the maximum cost c_{max} in the adversarial BwK setup, and similar results hold for online learning Knapsack setting as well. If $\alpha > 1/2$, then it is impossible to achieve a regret bound of $O(\sqrt{B})$, which is order optimal in cases with small c_{max}.

In the adversarial BwK setup, the adversary can penalize the player in two ways. First, the adversary can control the reward of an action at any round. Second, the adversary can control the cost of an action, which is analogous to penalizing the player on the number of rounds T. For $\alpha > 1/2$, the penalty on the number of rounds T becomes significant, and the minimum achievable regret is no longer $\Omega(\sqrt{B})$. In this setting with $\alpha > 1/2$, the design of algorithms which achieve regret of $O(B^{\alpha})$ is left as a future work.

6 Conclusion

The study of BwK has been mostly focused on the stochastic regime. In this work, we considered the adversarial regime and proposed the order optimal algorithm EXP3.BwK for this setting. We also proposed the algorithm EXP3++.BwK, which achieves an expected regret of $O(\sqrt{KB\log(K)})$ and $O(\log^2(B))$ in the adversarial and stochastic settings respectively. Thus, the algorithm is order optimal in the adversarial regime, and has an additional factor of $\log(B)$ in the stochastic regime. It is the first algorithm that provides almost optimal performance guarantees in both stochastic and adversary BwK settings simultaneously. Using the ideas from EXP3++.BwK, a new algorithm EXP3++.LwK was designed for a general online learning setting, and its performance guarantees were provided.

All the results in the literature of BwK assume that the maximum cost is bounded by a constant independent of B. We have shown that if the cost is $O(B^{\alpha})$, then the expected regret is at least $\Omega(B^{\alpha})$. This setting is of particular interest when $\alpha > 1/2$ because the expected regret of $O(\sqrt{B})$, which is achievable in the setting where cost is bounded by a constant, becomes unachievable. Hence, there is a need to study this BwK setting, and design optimal algorithms whose expected regret is $O(B^{\alpha})$, which is left as a future work.

References

[Agrawal and Devanur, 2014] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

[Agrawal and Devanur, 2016] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*, pages 3450–3458, 2016.

[Alon *et al.*, 2015] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *JMLR WORKSHOP AND CONFERENCE PROCEEDINGS*, volume 40. Microtome Publishing, 2015.

[Auer and Chiang, 2016] Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic mul-

tiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

[Bubeck and Slivkins, 2012] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

[Ding *et al.*, 2013] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[Immorlica *et al.*, 2018] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *arXiv preprint arXiv:1811.11881*, 2018.

[Kellerer *et al.*, 2004] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer, Berlin, Germany, 2004.

[Lykouris *et al.*, 2018] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122. ACM, 2018.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Rangi and Franceschetti, 2018a] Anshuka Rangi and Massimo Franceschetti. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1345–1352. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[Rangi and Franceschetti, 2018b] Anshuka Rangi and Massimo Franceschetti. Online learning with feedback graphs and switching costs. *arXiv preprint arXiv:1810.09666*, 2018.

[Rangi *et al.*, 2018a] Anshuka Rangi, Massimo Franceschetti, and Stefano Marano. Consensus-based chernoff test in sensor networks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6773–6778. IEEE, 2018.

[Rangi *et al.*, 2018b] Anshuka Rangi, Massimo Franceschetti, and Stefano Marano. Decentralized chernoff test in sensor networks. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 501–505. IEEE, 2018.

[Rangi *et al.*, 2018c] Anshuka Rangi, Massimo Franceschetti, and Stefano Marano. Distributed chernoff test: Optimal decision systems over networks. *arXiv preprint arXiv:1809.04587*, 2018.

[Rangi *et al.*, 2018d] Anshuka Rangi, Massimo Franceschetti, and Long Tran-Thanh. Unifying the stochastic and the adversarial bandits with knapsack. *arXiv preprint arXiv:1811.12253*, 2018.

[Sankararaman and Slivkins, 2018] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pages 1760–1770, 2018.

[Seldin and Lugosi, 2017] Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759, 2017.

[Seldin and Slivkins, 2014] Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *ICML*, pages 1287–1295, 2014.

[Tran-Thanh *et al.*, 2010] Long Tran-Thanh, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[Tran-Thanh *et al.*, 2012a] Long Tran-Thanh, Archie C Chapman, Alex Rogers, and Nicholas R Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI*, pages 1134–1140, 2012.

[Tran-Thanh *et al.*, 2012b] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. Efficient crowdsourcing of unknown experts using multi-armed bandits. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 768–773, 2012.

[Xia *et al.*, 2016] Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *IJCAI*, pages 2210–2216, 2016.

[Zhou and Tomlin, 2018] Datong P Zhou and Claire J Tomlin. Budget-constrained multi-armed bandits with multiple plays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.