The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Budget-Constrained Multi-Armed Bandits with Multiple Plays

Datong P. Zhou,[1] **Claire J. Tomlin**[2]
[1]Dept. of Mechanical Engineering, [2]Dept. of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720
{datong.zhou, tomlin}@berkeley.edu

Abstract

We study the multi-armed bandit problem with multiple plays and a budget constraint for both the stochastic and the adversarial setting. At each round, exactly K out of N possible arms have to be played (with $1 \leq K \leq N$). In addition to observing the individual rewards for each arm played, the player also learns a vector of costs which has to be covered with an a-priori defined budget B. The game ends when the sum of current costs associated with the played arms exceeds the remaining budget.

Firstly, we analyze this setting for the stochastic case, for which we assume each arm to have an underlying cost and reward distribution with support $[c_{\min}, 1]$ and $[0, 1]$, respectively. We derive an Upper Confidence Bound (UCB) algorithm which achieves $O(NK^4 \log B)$ regret.

Secondly, for the adversarial case in which the entire sequence of rewards and costs is fixed in advance, we derive an upper bound on the regret of order $O(\sqrt{NB \log(N/K)})$ utilizing an extension of the well-known `Exp3` algorithm. We also provide upper bounds that hold with high probability and a lower bound of order $\Omega((1 - K/N)^2 \sqrt{NB/K})$.

1 Introduction

The multi-armed bandit (MAB) problem has been extensively studied in machine learning and statistics as a means to model online sequential decision making. In the classic setting popularized by (Auer, Cesa-Bianchi, and Fischer 2002), (Auer et al. 2002), the decision-maker selects exactly one arm at a given round t, given the observations of realized rewards from arms played in previous rounds $1, \ldots, t-1$. The goal is to maximize the cumulative reward over a fixed horizon T, or equivalently, to minimize regret, which is defined as the difference between the cumulative gain achieved, had the decision-maker always played the best arm, and the realized cumulative gain. The analysis of this setting reflects the fundamental tradeoff between the desire to learn better arms (exploration) and the possibility to play arms believed to have high payoff (exploitation).

A variety of practical applications of the MAB problem include placement of online advertising to maximize the click-through rate, in particular online sponsored search auctions (Rusmevichientong and Williamson 2005) and ad-exchange platforms (Chakraborty et al. 2010), channel selection in radio networks (Huang, Liu, and Ding 2008), or learning to rank web documents (Radlinski, Kleinberg, and Joachims 2008). As acknowledged by (Ding et al. 2013), taking an action (playing an arm) in practice is inherently costly, yet the vast majority of existing bandit-related work used to analyze such examples forgoes any notion of cost. Furthermore, the above-mentioned applications rarely proceed in a strictly sequential way. A more realistic scenario is a setting in which, at each round, *multiple* actions are taken among the set of all possible choices.

These two shortcomings motivate the theme of this paper, as we investigate the MAB problem under a budget constraint in a setting with time-varying rewards and costs and multiple plays. More precisely, given an a-priori defined budget B, at each round the decision maker selects a combination of K distinct arms from N available arms and observes the individual costs and rewards, which corresponds to the *semi-bandit* setting. The player pays for the materialized costs until the remaining budget is exhausted, at which point the algorithm terminates and the cumulative reward is compared to the theoretical optimum and defines the weak regret, which is the expected difference between the payout under the best fixed choice of arms for all rounds and the actual gain. In this paper, we investigate both the stochastic and the adversarial case. For the stochastic case, we derive an upper bound on the expected regret of order $O(NK^4 \log B)$, utilizing Algorithm `UCB-MB` inspired by the upper confidence bound algorithm `UCB1` first introduced by (Auer, Cesa-Bianchi, and Fischer 2002). For the adversarial case, Algorithm `Exp3.M.B` upper and lower-bounds the regret with $O(\sqrt{NB \log(N/K)})$ and $\Omega((1 - K/N)^2 \sqrt{NB/K})$, respectively. These findings extend existing results from (Uchiya, Nakamura, and Kudo 2010) and (Auer et al. 2002), as we also provide an upper bound that holds with high probability. To the best of our knowledge, this is the first case that addresses the *adversarial* budget-constrained case, which we therefore consider to be the main contribution of this paper.

Related Work

In the extant literature, attempts to make sense of a cost component in MAB problems occur in (Tran-Thanh et al. 2010)

and (Tran-Thanh et al. 2012), who assume *time-invariant* costs and cast the setting as a knapsack problem with only the rewards being stochastic. In contrast, (Ding et al. 2013) proposed algorithm UCB-BV, where per-round costs and rewards are sampled in an IID fashion from unknown distributions to derive an upper bound on the regret of order $O(\log B)$. The papers that are closest to our setting are (Badanidiyuru, Kleinberg, and Slivkins 2013) and (Xia et al. 2016). The former investigates the stochastic case with a resource consumption. Unlike our case, however, the authors allow for the existence of a "null arm", which is tantamount to skipping rounds, and obtain an upper bound of order $O(\sqrt{B})$ rather than $O(\log B)$ compared to our case. The latter paper focuses on the stochastic case, but does not address the adversarial setting at all.

The extension of the single play to the multiple plays case, where at each round $K \geq 1$ arms have to be played, was introduced in (Anantharam, Varaiya, and Walrand 1986) and (Agrawal, Hegde, and Teneketzis 1990). However, their analysis is based on the original bandit formulation introduced by (Lai and Robbins 1985), where the regret bounds only hold asymptotically (in particular not for a finite time), rely on hard-to-compute index policies, and are distribution-dependent. Influenced by the works of (Auer, Cesa-Bianchi, and Fischer 2002) and (Agrawal 2002), who popularized the usage of easy-to-compute upper confidence bounds (UCB), a recent line of work has further investigated the combinatorial bandit setting. For example, (Gai, Krishnamachari, and Jain 2012) derived an $O(NK^4 \log T)$ regret bound in the stochastic semi-bandit setting, utilizing a policy they termed "Learning with Linear Rewards" (LLR). Similarly, (Chen, Wang, and Yuan 2013) utilize a framework where the decision-maker queries an oracle that returns a fraction of the optimal reward. Other, less relevant settings to this paper are found in (Cesa-Bianchi and Lugosi 2009) and later (Combes et al. 2015), who consider the adversarial bandit setting, where only the sum of losses for the selected arms can be observed. Furthermore, (Kale, Reyzin, and Schapire 2010) investigate bandit slate problems to take into account the ordering of the arms selected at each round. Lastly, (Komiyama, Honda, and Nakagawa 2015) utilize Thompson Sampling to model the stochastic MAB problem.

2 Main Results

In this section, we formally define the budgeted, multiple play multi-armed bandit setup and present the main theorems, whose results are provided in Table 1 together with a comparison to existing results in the literature. We first describe the stochastic setting (Section 2.1) and then proceed to the adversarial one (Section 2.2). Illuminating proofs for the theorems in this section are presented in Section 3. Technical proofs are relegated to the supplementary document (Zhou and Tomlin 2017).

2.1 Stochastic Setting

The definition of the stochastic setting is based on the classic setup introduced in (Auer, Cesa-Bianchi, and Fischer 2002), but is enriched by a cost component and a multiple play constraint. Specifically, given a bandit with N distinct arms, each arm indexed by $i \in [N]$ is associated with an unknown reward and cost distribution with unknown means $0 < \mu_r^i \leq 1$ and $0 < c_{\min} \leq \mu_c^i \leq 1$, respectively. Realizations of costs $c_{i,t} \in [c_{\min}, 1]$ and rewards $r_{i,t} \in [0, 1]$ are independently and identically distributed. At each round t, the decision maker plays exactly K arms ($1 \leq K \leq N$) and subsequently observes the individual costs and rewards only for the played arms, which corresponds to the *semi-bandit* setting. Before the game starts, the player is given a budget $0 < B \in \mathbb{R}_+$ to pay for the materialized costs $\{c_{i,t} \mid i \in a_t\}$, where a_t denotes the indexes of the K arms played at time t. The game terminates as soon as the sum of costs at round t, namely $\sum_{j \in a_t} c_{j,t}$ exceeds the remaining budget.

Notice the minimum $c_{\min}$ on the support of the cost distributions. This assumption is not only made for practical reasons, as many applications of bandits come with a minimum cost, but also to guarantee well-defined "bang-per-buck" ratios $\mu^i = \mu_r^i / \mu_c^i$, which our analysis in this paper relies on.

The goal is to design a deterministic algorithm $\mathcal{A}$ such that the expected payout $\mathbb{E}\left[G_{\mathcal{A}}(B)\right]$ is maximized, given the budget and multiple play constraints. Formally:

$$\begin{aligned} \underset{a_1,\ldots,a_{\tau_{\mathcal{A}}(B)}}{\text{maximize}} \quad & \mathbb{E}\left[\sum_{t=1}^{\tau_{\mathcal{A}}(B)} \sum_{i \in a_t} r_{i,t}\right] \\ \text{subject to} \quad & \mathbb{E}\left[\sum_{t=1}^{\tau_{\mathcal{A}}(B)} \sum_{i \in a_t} c_{i,t} \leq B\right] \\ & |a_t| = K,\ 1 \leq K \leq N\ \forall\, t \in [\tau_{\mathcal{A}}(B)] \end{aligned} \tag{1}$$

In (1), $\tau_{\mathcal{A}}(B)$ is the stopping time of algorithm $\mathcal{A}$ and indicates after how many steps the algorithm terminates, namely when the budget is exhausted. The expectation is taken over the randomness of the reward and cost distributions.

The performance of algorithm $\mathcal{A}$ is evaluated on its expected regret $\mathcal{R}_{\mathcal{A}}(B)$, which is defined as the difference between the expected payout (gain) $\mathbb{E}[G_{\mathcal{A}^*}]$ under the optimal strategy $\mathcal{A}^*$ (which in each round plays a^*, namely the set of K arms with the largest bang-per-buck ratios) and the expected payout $\mathbb{E}[G_{\mathcal{A}}]$ under algorithm $\mathcal{A}$:

$$\mathcal{R}_{\mathcal{A}}(B) = \mathbb{E}[G_{\mathcal{A}^*}(B)] - \mathbb{E}[G_{\mathcal{A}}(B)]. \tag{2}$$

Our main result in Theorem 1 upper bounds the regret achieved with Algorithm 1. Similar to (Auer, Cesa-Bianchi, and Fischer 2002) and (Ding et al. 2013), we maintain time-varying upper confidence bounds $U_{i,t}$ for each arm i

$$U_{i,t} = \bar{\mu}_t^i + e_{i,t}, \tag{3}$$

where $\bar{\mu}_t^i$ denotes the sample mean of the observed bang-per-buck ratios up to time t and $e_{i,t}$ the exploration term defined in Algorithm 1. At each round, the K arms associated with the K largest confidence bounds are played. For initialization purposes, we allow all N arms to be played exactly once prior to the while-loop.

Theorem 1. *There exist constants c_1, c_2, and c_3, which are functions of $N, K, c_{\min}, \Delta_{\min}, \mu_i, \mu_c$ only, such that Algorithm 1 (UCB-MB) achieves expected regret*

$$\mathcal{R}_{\mathcal{A}}(B) \leq c_1 + c_2 \log(B + c_3) = O(NK^4 \log B). \tag{4}$$

Algorithm	Upper Bound	Lower Bound	Authors
`Exp3`	$O\left(\sqrt{NT\log N}\right)$	$\Omega\left(\sqrt{NT}\right)$	(Auer et al. 2002)
`Exp3.M`	$O\left(\sqrt{NTK\log\frac{N}{K}}\right)$	$\Omega\left(\left(1-\frac{K}{N}\right)^2\sqrt{NT}\right)$	(Uchiya, Nakamura, and Kudo 2010)
`Exp3.M.B`	$O\left(\sqrt{NB\log\frac{N}{K}}\right)$	$\Omega\left(\left(1-\frac{K}{N}\right)^2\sqrt{NB/K}\right)$	This paper
`Exp3.P`	$O\left(\sqrt{NT\log\left(NT/\delta\right)}+\log(NT/\delta)\right)$		(Auer et al. 2002)
`Exp3.P.M`	$O\left(K^2\sqrt{NT\frac{N-K}{N-1}\log\left(NT/\delta\right)}+\frac{N-K}{N-1}\log(NT/\delta)\right)$		This paper
`Exp3.P.M.B`	$O\left(K^2\sqrt{\frac{NB}{K}\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)}+\frac{N-K}{N-1}\log\left(\frac{NB}{K\delta}\right)\right)$		This paper
`UCB1`	$O(N\log T)$		(Auer, Cesa-Bianchi, and Fischer 2002)
`LLR`	$O(NK^4\log T)$		(Gai, Krishnamachari, and Jain 2012)
`UCB-BV`	$O(N\log B)$		(Ding et al. 2013)
`UCB-MB`	$O(NK^4\log B)$		This paper

Table 1: Regret Bounds in Adversarial and Stochastic Bandit Settings

In Theorem 1, $\Delta_{\min}$ denotes the smallest possible difference of bang-per-buck ratios among non-optimal selections $a \neq a^*$, i.e. the second best choice of arms:

$$\Delta_{\min} = \sum_{j\in a^*}\mu^j - \max_{a,a\neq a^*}\sum_{j\in a}\mu^j. \tag{5}$$

Similarly, the proof of Theorem 1 also relies on the largest such difference $\Delta_{\max}$, which corresponds to the worst possible choice of arms:

$$\Delta_{\max} = \sum_{j\in a^*}\mu^j - \min_{a,a\neq a^*}\sum_{j\in a}\mu^j. \tag{6}$$

Comparing the bound given in Theorem 1 to the results in Table 1, we recover the $O(N\log B)$ bound from (Ding et al. 2013) for the single-play case.

Algorithm 1 `UCB-MB` for Stochastic MAB

Initialize: $t = 1$. Play all arms together exactly once. Let $\bar{\mu}^i_{r,1} = r_{i,1}$, $\bar{\mu}^i_{c,1} = c_{i,1}$, $\bar{\mu}^i_1 = \frac{\bar{\mu}^i_{r,1}}{\bar{\mu}^i_{c,1}}\ \forall\ i \in [N]$, $n_{i,1} = 1$, $e_{i,1} = 0\ \forall\ i \in [N]$, $G_{\mathcal{A}} = 0$.

while true **do**
  $a_t \leftarrow$ Indexes of K arms with K largest $U_{i,t}$.
  if $\sum_{j\in a_t} c_{j,t} > B$ **then**
    return Gain $G_{\mathcal{A}}$, stopping time $\tau_{\mathcal{A}}(B) = t$
  end if
  $G_{\mathcal{A}} \leftarrow G_{\mathcal{A}} + \sum_{i\in a_t} r_{i,t}, \quad B \leftarrow B - \sum_{i\in a_t} c_{i,t}$
  $n_{i,t} \leftarrow n_{i,t} + 1 \quad \forall\ i \in a_t$
  $t \leftarrow t + 1$
  $e_{i,t} \leftarrow \frac{\sqrt{(K+1)\log t/n_{i,t}}(1+1/c_{\min})}{c_{\min}-\sqrt{(K+1)\log t/n_{i,t}}}$
end while

2.2 Adversarial Setting

We now consider the adversarial case that makes no assumptions on the reward and cost distributions whatsoever. The setup for this case was first proposed and analyzed by (Auer et al. 2002) for the single play case (i.e. $K = 1$), a fixed horizon T, and an oblivious adversary. That is, the entire seqence of rewards for all arms is fixed in advance and in particular cannot be adaptively changed during runtime. The proposed randomized algorithm `Exp3` enjoys $O(\sqrt{NT\log N})$ regret. Under *semi-bandit* feedback, where the rewards for a given round are observed for each arm played, (Uchiya, Nakamura, and Kudo 2010) derived a variation of the single-play `Exp3` algorithm, which they called `Exp3.M` and enjoys regret $O\left(\sqrt{NTK\log(N/K)}\right)$, where K is the number of plays per round.

We consider the extension of the classic setting as in (Uchiya, Nakamura, and Kudo 2010), where the decision maker has to play exactly $1 \leq K \leq N$ arms. For each arm i played at round t, the player observes the reward $r_i(t) \in [0,1]$ and, unlike in previous settings, additionally the cost $0 < c_{\min} < c_i(t) < 1$. As in the stochastic setting (Section 2.1), the player is given a budget $B > 0$ to pay for the costs incurred, and the algorithm terminates after $\tau_{\mathcal{A}}(B)$ rounds when the sum of materialized costs in round $\tau_{\mathcal{A}}(B)$ exceeds the remaining budget. The gain $G_{\mathcal{A}}(B)$ of algorithm $\mathcal{A}$ is the sum of observed rewards up to and including round $\tau_{\mathcal{A}}(B) - 1$. The expected regret $\mathcal{R}_{\mathcal{A}}(B)$ is defined as in (2), where the gain of algorithm $\mathcal{A}$ is compared against the best set of arms that an omniscient algorithm $\mathcal{A}^*$, which knows the reward and cost sequences in advance, would select, given the budget B. In contrast to the stochastic case, the expectation is now taken with respect to algorithm $\mathcal{A}$'s internal randomness.

Upper Bounds on the Regret We begin with upper bounds on the regret for the budget constrained MAB with multiple plays and later transition towards lower bounds and upper bounds that hold with high probability. Algorithm 2, which we call `Exp3.M.B`, provides a randomized algorithm to achieve sublinear regret. Similar to the original `Exp3` algorithm developed by (Auer et al. 2002), Algorithm `Exp3.M.B` maintains a set of time-varying weights $\{w_i(t)\}_{i=1}^N$ for all arms, from which the probabilities for

each arm being played at time t are calculated (line 10). As noted in (Uchiya, Nakamura, and Kudo 2010), the probabilities $\{p_i(t)\}_{i=1}^N$ sum to K (because exactly K arms need to be played), which requires the weights to be capped at a value $v_t > 0$ (line 3) such that the probabilities $\{p_i(t)\}_{i=1}^N$ are kept in the range $[0, 1]$. In each round, the player draws a set of distinct arms a_t of cardinality $|a_t| = K$, where each arm has probability $p_i(t)$ of being included in a_t (line 11). This is done by employing algorithm `DependentRounding` introduced by (Gandhi, Khuller, and Parthasarathy 2006), which runs in $O(K)$ time and $O(N)$ space. At the end of each round, the observed rewards and costs for the played arms are turned into estimates $\hat{r}_i(t)$ and $\hat{c}_i(t)$ such that $\mathbb{E}[\hat{r}_i(t) \mid a_t, \ldots, a_1] = r_i(t)$ and $\mathbb{E}[\hat{c}_i(t) \mid a_t, \ldots, a_1] = c_i(t)$ for $i \in a_t$ (line 16). Arms with $w_i(t) < v_t$ are updated according to $(\hat{r}_i(t) - \hat{c}_i(t))$, which assigns larger weights as $\hat{r}_i(t)$ increases and $\hat{c}_i(t)$ decreases, as one might expect.

Algorithm 2 `Exp3.M.B`: Budget Constrained Multi-Armed Bandit, Multiple Play, Adversarial

Initialize: $w_i = 1$ for $i \in [N]$, gain $G_{\mathcal{A}} = 0$.

1: **while** $B > 0$ **do**
2: **if** $\arg\max_{i \in [N]} w_i(t) \geq \left(\frac{1}{K} - \frac{\gamma}{N}\right) \sum_{j=1}^N \frac{w_i(t)}{1-\gamma}$ **then**
3: Determine v_t as follows: $1/K - \gamma/N =$

$$\frac{v_t(1-\gamma)}{\sum_{i=1}^N v_t \cdot \mathbb{1}(w_i(t) \geq v_t) + w_i(t) \cdot \mathbb{1}(w_i(t) < v_t)}$$

4: Define set $\tilde{S}(t) = \{i \in [N] \mid w_i(t) \geq v_t\}$.
5: Define weights $\tilde{w}_i(t) = v_t$ for $i \in \tilde{S}(t)$.
6: **else**
7: Define set $\tilde{S}(t) = \{\}$.
8: **end if**
9: Define weights $\tilde{w}_i(t) = w_i(t)$ for $i \in [N] \setminus \tilde{S}(t)$.
10: Calculate probabilities for each $i \in [N]$:

$$p_i(t) = K\left((1-\gamma)\frac{\tilde{w}_i(t)}{\sum_{j=1}^N \tilde{w}_j(t)} + \frac{\gamma}{N}\right).$$

11: Play arms $a_t \sim p_1, \ldots, p_N$.
12: **if** $\sum_{i \in a_t} c_i(t) > B$ **then**
13: **return** Gain $G_{\texttt{Exp3.M.B}}$, stopping time $\tau_{\mathcal{A}}(B) = t$
14: **end if**
15: $B \leftarrow B - \sum_{i \in a_t} c_i(t)$, $G_{\mathcal{A}} \leftarrow G_{\mathcal{A}} + \sum_{i \in a_t} r_i(t)$.
16: Calculate estimated rewards and costs to update weights for each $i \in [N]$:

$$\hat{r}_i(t) = r_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$\hat{c}_i(t) = c_i(t)/p_i(t) \cdot \mathbb{1}(i \in a_t)$$
$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \mathbb{1}_{i \in \tilde{S}(t)}\right]$$

17: **end while**

Theorem 2. *Algorithm* `Exp3.M.B` *achieves regret*

$$\mathcal{R} \leq 2.63\sqrt{1 + \frac{B}{gc_{\min}}}\sqrt{gN\log(N/K)} + K, \quad (7)$$

where g is an upper bound on $G_{\max}$, the maximal gain of the optimal algorithm. This bound is of order $O(\sqrt{BN\log(N/K)})$.

The runtime of Algorithm `Exp3.M.B` and its space complexity is linear in the number of arms, i.e. $O(N)$. If no bound g on $G_{\max}$ exists, we have to modify Algorithm 2. Specifically, the weights are now updated as follows:

$$w_i(t+1) = w_i(t) \exp\left[\frac{K\gamma}{N}\left[\hat{r}_i(t) - \hat{c}_i(t)\right] \cdot \mathbb{1}_{i \in a_t}\right]. \quad (8)$$

This replaces the original update step in line 16 of Algorithm 2. As in Algorithm `Exp3.1` in (Auer et al. 2002), we use an adaptation of Algorithm 2, which we call `Exp3.1.M.B`, see Algorithm 3. In Algorithm 3, we define cumulative expected gains and losses

$$\hat{G}_i(t) = \sum_{s=1}^t \hat{r}_i(s), \quad (9a)$$

$$\hat{L}_i(t) = \sum_{s=1}^t \hat{c}_i(s). \quad (9b)$$

and make the following, necessary assumption:

Assumption 1. $\sum_{i \in a} r_i(t) \geq \sum_{i \in a} c_i(t)$ *for all* $a \in \mathcal{S}$ *possible* K*-combinations and* $t \geq 1$.

Assumption 1 is a natural assumption, which is motivated by "individual rationality" reasons. In other words, a user will only play the bandit algorithm if the reward at any given round, for any possible choice of arms, is at least as large as the cost that incurs for playing. Under the caveat of this assumption, Algorithm `Exp3.1.M.B` utilizes Algorithm `Exp3.1.M` as a subroutine in each epoch until termination.

Algorithm 3 Algorithm `Exp3.1.M.B` with Budget B

Initialize: $t = 1$, $w_i = 1$ for $i \in [N]$, $r = 0$.

1: **while** $\sum_{t=1}^T \sum_{i \in a_t} c_i(t) \leq B$ **do**
2: Define $g_r = \frac{N\log(N/K)}{(e-1)-(e-2)c_{\min}} 4^r$
3: Restart `Exp3.M.B` with $\gamma_r = \min(1, 2^{-r})$
4: **while** $\max_{a \in \mathcal{S}} \sum_{i \in a} (\hat{G}_i(t) - \hat{L}_i(t)) \leq g_r - \frac{N(1-c_{\min})}{K\gamma_r}$ **do**
5: Draw $a_t \sim p_1, \ldots, p_N$, observe $r_i(t)$ and $c_i(t)$ for $i \in a_t$, calculate $\hat{r}_i(t)$ and $\hat{c}_i(t)$.
6: $\hat{G}_i(t+1) \leftarrow \hat{G}_i(t) + \hat{r}_i(t)$ for $i \in [N]$
7: $\hat{L}_i(t+1) \leftarrow \hat{L}_i(t) + \hat{c}_i(t)$ for $i \in [N]$
8: $t \leftarrow t + 1$
9: **end while**
10: **end while**
11: **return** Gain $G_{\texttt{Exp3.1.M.B}}$

Proposition 1. *For the multiple plays case with budget, the regret of Algorithm* `Exp3.1.M.B` *is upper bounded by*

$$\mathcal{R} \leq 8\left[(e-1)-(e-2)c_{\min}\right]\frac{N}{K}+2N\log\frac{N}{K}+K+ 8\sqrt{\left[(e-1)-(e-2)c_{\min}\right](G_{\max}-B+K)N\log(N/K)} \tag{10}$$

This bound is of order $O((G_{\max}-B)N\log(N/K))$ and, due to Assumption 1, not directly comparable to the bound in Theorem 2. One case in which (10) outperforms (7) occurs whenever only a loose upper bound of g on $G_{\max}$ exists or whenever $G_{\max}$, the return of the best selection of arms, is "small".

Lower Bound on the Regret Theorem 3 provides a lower bound of order $\Omega((1-K/N)^2\sqrt{NB/K})$ on the weak regret of algorithm `Exp3.M.B`.

Theorem 3. *For $1 \leq K \leq N$, the weak regret $\mathcal{R}$ of Algorithm* `Exp3.M.B` *is lower bounded as follows:*

$$\mathcal{R} \geq \varepsilon\left(B-\frac{BK}{N}-2Bc_{\min}^{-3/2}\varepsilon\sqrt{\frac{BK\log(4/3)}{N}}\right), \tag{11}$$

where $\varepsilon \in (0,1/4]$. Choosing ε as

$$\varepsilon=\min\left(\frac{1}{4},\ \frac{(1-K/N)c_{\min}^{3/2}}{4\sqrt{\log(4/3)}}\sqrt{\frac{N}{BK}}\right)$$

yields the bound

$$\mathcal{R} \geq \min\left(\frac{c_{\min}^{3/2}(1-K/N)^2}{8\sqrt{\log(4/3)}}\sqrt{\frac{NB}{K}},\ \frac{B(1-K/N)}{8}\right). \tag{12}$$

This lower bound differs from the upper bound in Theorem 1 by a factor of $\sqrt{K\log(N/K)}(N/(N-K))^2$. For the single-play case $K=1$, this factor is $\sqrt{\log N}$, which recovers the gap from (Auer et al. 2002).

High Probability Upper Bounds on the Regret For a fixed number of rounds (no budget considerations) and single play per round ($K=1$), (Auer et al. 2002) proposed Algorithm `Exp3.P` to derive the following upper bound on the regret that holds with probability at least $1-\delta$:

$$G_{\max}-G_{\texttt{Exp3.P}} \leq 4\sqrt{NT\log(NT/\delta)} + 4\sqrt{\frac{5}{3}NT\log N}+8\log\left(\frac{NT}{\delta}\right). \tag{13}$$

Theorem 4 extends the non-budgeted case to the multiple play case.

Theorem 4. *For the multiple play algorithm ($1 \leq K \leq N$) and a fixed number of rounds T, the following bound on the regret holds with probability at least $1-\delta$:*

$$\begin{aligned}\mathcal{R} &= G_{\max}-G_{\textit{Exp3.P.M}}\\ &\leq 2\sqrt{5}\sqrt{NKT\log(N/K)}+8\frac{N-K}{N-1}\log\left(\frac{NT}{\delta}\right)\\ &\quad+2(1+K^2)\sqrt{NT\frac{N-K}{N-1}\log\left(\frac{NT}{\delta}\right)}.\end{aligned} \tag{14}$$

For $K=1$, (14) recovers (13) save for the constants, which is due to a better ε-tuning in this paper compared to (Auer et al. 2002). Agreeing with intuition, this upper bound becomes zero for the edge case $K \equiv N$.

Theorem 4 can be derived by using a modified version of Algorithm 2, which we name `Exp3.P.M`. The necessary modifications to `Exp3.M.B` are motivated by Algorithm `Exp3.P` in (Auer et al. 2002) and are provided in the following:

- Replace the outer while loop with **for** $t=1,\dots,T$ **do**
- Initialize parameter α:
$$\alpha = 2\sqrt{(N-K)/(N-1)\log(NT/\delta)}.$$
- Initialize weights w_i for $i \in [N]$:
$$w_i(1)=\exp\left(\alpha\gamma K^2\sqrt{T/N}/3\right).$$
- Update weights for $i \in [N]$ as follows:
$$w_i(t+1)=w_i(t)\times\exp\left[\mathbb{1}_{i\notin\tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t)+\frac{\alpha}{p_i(t)\sqrt{NT}}\right)\right]. \tag{15}$$

Since there is no notion of cost in Theorem 4, we do not need to update any cost terms.

Lastly, Theorem 5 extends Theorem 4 to the budget constrained setting using algorithm `Exp3.P.M.B`.

Theorem 5. *For the multiple play algorithm ($1 \leq K \leq N$) and the budget $B > 0$, the following bound on the regret holds with probability at least $1-\delta$:*

$$\begin{aligned}\mathcal{R} &= G_{\max}-G_{\textit{Exp3.P.M.B}}\\ &\leq 2\sqrt{3}\sqrt{\frac{NB(1-c_{\min})}{c_{\min}}\log\frac{N}{K}}\\ &\quad+4\sqrt{6}\frac{N-K}{N-1}\log\left(\frac{NB}{Kc_{\min}\delta}\right)\\ &\quad+2\sqrt{6}(1+K^2)\sqrt{\frac{N-K}{N-1}\frac{NB}{Kc_{\min}}\log\left(\frac{NB}{Kc_{\min}\delta}\right)}.\end{aligned} \tag{16}$$

To derive bound (16), we again modify the following update rules in Algorithm 2 to obtain Algorithm `Exp3.P.M.B`:

- Initialize parameter α:
$$\alpha = 2\sqrt{6}\sqrt{(N-K)/(N-1)\log(NB/(Kc_{\min}\delta))}.$$
- Initialize weights w_i for $i \in [N]$:
$$w_i(1)=\exp\left(\alpha\gamma K^2\sqrt{B/(NKc_{\min})}/3\right).$$
- Update weights for $i \in [N]$ as follows:
$$w_i(t+1)=w_i(t)\times\exp\left[\mathbb{1}_{i\notin\tilde{S}(t)}\frac{\gamma K}{3N}\left(\hat{r}_i(t)-\hat{c}_i(t)+\frac{\alpha\sqrt{Kc_{\min}}}{p_i(t)\sqrt{NB}}\right)\right].$$

The estimated costs $\hat{c}_i(t)$ are computed as $\hat{c}_i(t) = c_i(t)/p_i(t)$ whenever arm i is played at time t, as is done in Algorithm 2.

3 Proofs

Proof of Theorem 1

The proof of Theorem 1 is divided into two technical lemmas introduced in the following. Due to space constraints, the proofs are relegated to the supplementary document.

First, we bound the number of times a non-optimal selection of arms is made up to stopping time $\tau_{\mathcal{A}}(B)$. For this purpose, let us define a counter $C_{i,t}$ for each arm i, initialized to zero for $t = 1$. Each time a non-optimal vector of arms is played, that is, $a_t \neq a^*$, we increment the smallest counter in the set a_t:

$$C_{j,t} \leftarrow C_{j,t} + 1, \quad j = \arg\min_{i \in a_t} C_{i,t}. \tag{17}$$

Ties are broken randomly. By definition, the number of times arm i has been played until time t is greater than or equal to its counter $C_{i,t}$. Further, the sum of all counters is exactly the number of suboptimal choices made so far:

$$n_{i,t} \geq C_{i,t} \quad \forall\, i \in [N],\ t \in [\tau_{\mathcal{A}}(B)].$$
$$\sum_{i=1}^{N} C_{i,t} = \sum_{\tau=1}^{t} \mathbb{1}(a_\tau \neq a^*) \quad \forall\, t \in [\tau_{\mathcal{A}}(B)].$$

Lemma 1 bounds the value of $C_{i,t}$ from above.

Lemma 1. *Upon termination of algorithm $\mathcal{A}$, there have been at most $O\left(NK^3 \log \tau_{\mathcal{A}}(B)\right)$ suboptimal actions. Specifically, for each $i \in [N]$:*

$$\mathbb{E}\left[C_{i,\tau_{\mathcal{A}}(B)}\right] \leq 1 + K\frac{\pi^2}{3} + (K+1)\left(\frac{\Delta_{\min} + 2K(1 + 1/c_{\min})}{c_{\min}\Delta_{\min}}\right)^2 \log \tau_{\mathcal{A}}(B).$$

Secondly, we relate the stopping time of algorithm $\mathcal{A}$ to the optimal action a^*:

Lemma 2. *The stopping time $\tau_{\mathcal{A}}$ is bounded as follows:*

$$\frac{B}{\sum_{i\in a^*}\mu_c^i} - c_2 - c_3 \log\left(c_1 + \frac{2B}{\sum_{i\in a^*}\mu_c^i}\right) \leq \tau_{\mathcal{A}} \leq \frac{2B}{\sum_{i\in a^*}\mu_c^i}\mu_c^i + c_1,$$

where c_1, c_2, and c_3 are the same positive constants as in Theorem 1 that depend only on $N, K, c_{\min}, \Delta_{\min}, \mu_c^i, \mu_r^i$.

Utilizing Lemmas 1 and 2 in conjunction with the definition of the weak regret (2) yields Theorem 1. See the supplementary document for further technicalities.

Proof of Theorem 2

The proof of Theorem 2 in influenced by the proof methods for Algorithms `Exp3` by (Auer et al. 2002) and `Exp3.M` by (Uchiya, Nakamura, and Kudo 2010). The main challenge is the absence of a well-defined time horizon T due to the time-varying costs. To remedy this problem, we define $T = \max\left(\tau_{\mathcal{A}}(B), \tau_{\mathcal{A}^*}(B)\right)$, which allows us to first express the regret as a function of T. In a second step, we relate T to the budget B.

Proof of Proposition 1

The proof of Proposition 1 is divided into the following two lemmas:

Lemma 3. *For any subset $a \in \mathcal{S}$ of K unique elements from $[N]$, $1 \leq K \leq N$:*

$$\sum_{t=S_r}^{T_r}\sum_{i\in a_t}(r_i(t) - c_i(t)) \geq \sum_{i\in a}\sum_{t=S_r}^{T_r}(\hat{r}_j(t) - \hat{c}_j(t)) - 2\sqrt{(e-1) - (e-2)c_{\min}}\sqrt{g_r N \log(N/K)}, \tag{18}$$

where S_r and T_r denote the first and last time step at epoch r, respectively.

Lemma 4. *The total number of epochs R is bounded by*

$$2^{R-1} \leq \frac{N(1-c_{\min})}{Kc} + \sqrt{\frac{\hat{G}_{\max} - \hat{L}_{\max}}{c}} + \frac{1}{2}, \tag{19}$$

where $c = \frac{N\log(N/K)}{(e-1)-(e-2)c_{\min}}$.

To derive Proposition 1, we combine Lemmas 3 and 4 and utilize the fact that algorithm `Exp3.1.M.B` terminates after $\tau_{\mathcal{A}}(B)$ rounds. See supplementary document for details.

Proof of Theorem 3

The proof follows existing procedures for deriving lower bounds in adversarial bandit settings, see (Auer et al. 2002), (Cesa-Bianchi and Lugosi 2006). The main challenges are found in generalizing the single play setting to the multiple play setting ($K > 1$) as well as incorporating a notion of cost associated with bandits.

Select exactly K out of N arms at random to be the arms in the "good" subset a^*. For these arms, let $r_i(t)$ at each round t be Bernoulli distributed with bias $\frac{1}{2} + \varepsilon$, and the cost $c_i(t)$ attain $c_{\min}$ and 1 with probability $\frac{1}{2} + \varepsilon$ and $\frac{1}{2} - \varepsilon$, respectively, for some $0 < \varepsilon < 1/2$ to be specified later. All other $N - K$ arms are assigned rewards 0 and 1 and costs $c_{\min}$ and 1 independently at random. Let $\mathbb{E}_{a^*}[\,\cdot\,]$ denote the expectation of a random variable conditional on a^* as the set of good arms. Let $\mathbb{E}_u[\,\cdot\,]$ denote the expectation with respect to a uniform assignment of costs $\{c_{\min}, 1\}$ and rewards $\{0, 1\}$ to all arms. Lemma 5 is an extension of Lemma A.1 in (Auer et al. 2002) to the multiple-play case with cost considerations:

Lemma 5. *Let $f : \{\{0,1\},\{c_{\min},1\}\}^{\tau_{\max}} \to [0, M]$ be any function defined on reward and cost sequences $\{\mathbf{r}, \mathbf{c}\}$ of length less than or equal $\tau_{\max} = \frac{B}{Kc_{\min}}$. Then, for the best action set a^*:*

$$\mathbb{E}_{a^*}\left[f(\mathbf{r},\mathbf{c})\right] \leq \mathbb{E}_u[f(\mathbf{r},\mathbf{c})] + \frac{Bc_{\min}^{-3/2}}{2}\sqrt{-\mathbb{E}_u[N_{a^*}]\log(1-4\varepsilon^2)},$$

where N_{a^*} denotes the total number of plays of arms in a^* during rounds $t = 1$ through $t = \tau_{\mathcal{A}}(B)$, that is:

$$N_{a^*} = \sum_{t=1}^{\tau_{\mathcal{A}}(B)}\sum_{i\in a^*}\mathbb{1}\left(i \in a_t\right).$$

Lemma 5, whose proof is relegated to the supplementary document, allows us to bound the gain under the existence of K optimal arms by treating the problem as a uniform assignment of costs and rewards to arms. The technical parts of the proof can also be found in the supplementary document.

Proof of Theorem 4

The proof strategy is to acknowledge that Algorithm `Exp3.P.M` uses upper confidence bounds $\hat{r}_i(t) + \frac{\alpha}{p_i(t)\sqrt{NT}}$ to update weights (15). Lemma 6 asserts that these confidence bounds are valid, namely that they upper bound the actual gain with probability at least $1-\delta$, where $0 < \delta \ll 1$.

Lemma 6. *For* $2\sqrt{\frac{N-K}{N-1}\log\left(\frac{NT}{\delta}\right)} \leq \alpha \leq 2\sqrt{NT}$,

$$\mathbb{P}\left(\hat{U}^* > G_{\max}\right) \geq \mathbb{P}\left(\bigcap_{a\subset\mathcal{S}} \sum_{i\in a} \hat{G}_i + \alpha\hat{\sigma}_i > \sum_{i\in a} G_i\right) \geq 1-\delta,$$

where $a \subset \mathcal{S}$ denotes an arbitrary subset of K unique elements from $[N]$. $\hat{U}^*$ denotes the upper confidence bound for the optimal gain.

Next, Lemma 7 provides a lower bound on the gain of Algorithm `Exp3.P.M` as a function of the maximal upper confidence bound.

Lemma 7. *For $\alpha \leq 2\sqrt{NT}$, the gain of Algorithm `Exp3.P.M` is bounded below as follows:*

$$G_{\texttt{Exp3.P.M}} \geq \left(1 - \frac{5}{3}\gamma\right)\hat{U}^* - \frac{3N}{\gamma}\log(N/K) - 2\alpha^2 - \alpha(1+K^2)\sqrt{NT}, \tag{20}$$

where $\hat{U}^* = \sum_{j\in a^*} \hat{G}_j + \alpha\hat{\sigma}_j$ denotes the upper confidence bound of the optimal gain achieved with optimal set a^*.

Therefore, combining Lemmas 6 and 7 upper bounds the actual gain of Algorithm `Exp3.P.M` with high probability. See the supplementary document for technical details.

Proof of Theorem 5

The proof of Theorem 5 proceeds in the same fashion as in Theorem 4. Importantly, the upper confidence bounds now include a cost term. Lemma 8 is the equivalent to Lemma 6 for the budget constrained case:

Lemma 8. *For* $2\sqrt{6}\sqrt{\frac{N-K}{N-1}\log\frac{NB}{Kc_{\min}\delta}} \leq \alpha \leq 12\sqrt{\frac{NB}{Kc_{\min}}}$,

$$\mathbb{P}\left(\hat{U}^* > G_{\max} - B\right) \geq \mathbb{P}\left(\bigcap_{a\subset\mathcal{S}} \sum_{i\in a} \hat{G}_i - \hat{L}_i + \alpha\hat{\sigma}_i > \sum_{i\in a} G_i - L_i\right) \geq 1-\delta,$$

where $a \subset \mathcal{S}$ denotes an arbitrary time-invariant subset of K unique elements from $[N]$. $\hat{U}^*$ denotes the upper confidence bound for the cumulative optimal gain minus the cumulative cost incurred after $\tau_a(B)$ rounds (the stopping time when the budget is exhausted):

$$a^* = \max_{a\in\mathcal{S}} \sum_{t=1}^{\tau_a(B)} (r_i(t) - c_i(t)),$$

$$\hat{U}^* = \sum_{i\in a^*}\left(\alpha\hat{\sigma}_i + \sum_{t=1}^{\tau_{a^*}(B)} (\hat{r}_i(t) - \hat{c}_i(t))\right). \tag{21}$$

In Lemma 8, $G_{\max}$ denotes the optimal cumulative reward under the optimal set a^* chosen in (21). $\hat{G}_i$ and $\hat{L}_i$ denote the cumulative expected reward and cost of arm i after exhaustion of the budget (that is, after $\tau_a(B)$ rounds), respectively.

Lastly, Lemma 9 lower bounds the actual gain of Algorithm `Exp3.P.M.B` as a function of the upper confidence bound (21).

Lemma 9. *For $\alpha \leq 2\sqrt{\frac{NB}{Kc_{\min}}}$, the gain of Algorithm `Exp3.P.M.B` is bounded below as follows:*

$$G_{\texttt{Exp3.P.M.B}} \geq \left(1 - \gamma - \frac{2\gamma}{3}\frac{1-c_{\min}}{c_{\min}}\right)\hat{U}^* - \frac{3N}{\gamma}\log\frac{N}{K} - 2\alpha^2 - \alpha(1+K^2)\frac{BN}{Kc_{\min}}.$$

Combining Lemmas 8 and 9 completes the proof, see the supplementary document.

4 Discussion and Conclusion

We discussed the budget-constrained multi-armed bandit problem with N arms, K multiple plays, and an a-priori defined budget B. We explored the stochastic as well as the adversarial case and provided algorithms to derive regret bounds in the budget B. For the stochastic setting, our algorithm `UCB-MB` enjoys regret $O(NK^4\log B)$. In the adversarial case, we showed that algorithm `Exp3.M.B` enjoys an upper bound on the regret of order $O(\sqrt{NB\log(N/K)})$ and a lower bound $\Omega((1-K/N)^2\sqrt{NB/K})$. Lastly, we derived upper bounds that hold with high probability.

Our work can be extended in several dimensions in future research. For example, the incorporation of a budget constraint in this paper leads us to believe that a logical extension is to integrate ideas from economics, in particular mechanism design, into the multiple plays setting (one might think about auctioning off multiple items simultaneously) (Babaioff, Sharma, and Slivkins 2009). A possible idea is to investigate to which extent the regret varies as the number of plays K increases. Further, we believe that in such settings, repeated interactions with customers (playing arms) give rise to strategic considerations, in which customers can misreport their preferences in the first few rounds to maximize their long-run surplus. While the works of (Amin, Rostamizadeh, and Syed 2013) and (Mohri and Munoz 2014) investigate repeated interactions with a single player only, we believe an extension to a pool of buyers is worth exploring. In this setting, we would expect that the extent of strategic behavior decreases as the number of plays K in each round increases, since the decision-maker could simply ignore users in future rounds who previously declined offers.

References

Agrawal, R.; Hegde, M. V.; and Teneketzis, D. 1990. Multi-Armed Bandits with Multiple Plays and Switching Cost. *Stochastics and Stochastic Reports* 29:437–459.

Agrawal, R. 2002. Sample Mean Based Index Policies with O(log n) Regret for the Multi-Armed Bandit Problem. *Machine Learning* 47:235–256.

Amin, K.; Rostamizadeh, A.; and Syed, U. 2013. Learning Prices for Repeated Auctions with Strategic Buyers. *Advances in Neural Information Processing Systems* 1169–1177.

Anantharam, V.; Varaiya, P.; and Walrand, J. 1986. Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem - Part I: IID Rewards. *IEEE Transactions on Automatic Control* 32:968–976.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The Nonstochastic Multi-Armed Bandit Problem. *SIAM Journal on Computing* 32:48–77.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47:235–256.

Babaioff, M.; Sharma, Y.; and Slivkins, A. 2009. Characterizing Truthful Multi-Armed Bandit Mechanisms. *Proceedings of the 10th ACM Conference on Electronic Commerce* 79–88.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with Knapsacks. *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science* 207–216.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Cesa-Bianchi, N., and Lugosi, G. 2009. Combinatorial Bandits. *Proceedings of the 22nd Annual Conference on Learning Theory*.

Chakraborty, T.; Even-Dar, E.; Guha, S.; Mansour, Y.; and Muthukrishnan, S. 2010. Selective Call Out and Real Time Bidding. *WINE* 6484:145–157.

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial Bandits: General Framework, Results and Applications. *International Conference on Machine Learning*.

Combes, R.; M. Sadegh Talebi; Proutiere, A.; and Lelarge, M. 2015. Combinatorial Bandits Revisited. *Advances in Neural Information Processing Systems* 2116 – 2124.

Ding, W.; Qin, T.; Zhang, X.-D.; and Liu, T.-Y. 2013. Multi-Armed Bandit with Budget Constraint and Variable Costs. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Gai, Y.; Krishnamachari, B.; and Jain, R. 2012. Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations. *IEEE/ACM Transactions on Networking* 20(5):1466–1478.

Gandhi, R.; Khuller, S.; and Parthasarathy, S. 2006. Dependent Rounding and its Applications to Approximation Algorithms. *Journal of the ACM (JACM)* 53(3):324–360.

Huang, S.; Liu, X.; and Ding, Z. 2008. Opportunistic Spectrum Access in Cognitive Radio Networks. *IEEE INFOCOM 2008 Proceedings* 2101–2109.

Kale, S.; Reyzin, L.; and Schapire, R. E. 2010. Non-Stochastic Bandit Slate Problems. *Advances in Neural Information Processing Systems* 1054–1062.

Komiyama, J.; Honda, J.; and Nakagawa, H. 2015. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-Armed Bandit Problems with Multiple Plays. *International Conference on Machine Learning* 1152–1161.

Lai, T. L., and Robbins, H. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6(1):4–22.

Mohri, M., and Munoz, A. 2014. Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers. *Advances in Neural Information Processing Systems* 1871–1879.

Radlinski, F.; Kleinberg, R.; and Joachims, T. 2008. Learning Diverse Rankings with Multi-Armed Bandits. *Proceedings of the 25th International Conference on Machine Learning* 784–791.

Rusmevichientong, P., and Williamson, D. P. 2005. An Adaptive Algorithm for Selecting Profitable Keywords for Search-Based Advertising Services. *Proceedings of the 7th ACM Conference on Electronic Commerce* 260–269.

Tran-Thanh, L.; Chapman, A.; F.L. Munoz De Cote; Jose, E.; Rogers, A.; and Jennings, N. R. 2010. Epsilon-First Policies for Budget-Limited Multi-Armed Bandits. *Twenty-Fourth AAAI Conference on Artificial Intelligence* 1211–1216.

Tran-Thanh, L.; Chapman, A.; Rogers, A.; and Jennings, N. R. 2012. Knapsack Based Optimal Policies for Budget–Limited Multi–Armed Bandits. *Twenty-Sixth AAAI Conference on Artificial Intelligence* 1134–1140.

Uchiya, T.; Nakamura, A.; and Kudo, M. 2010. Algorithms for Adversarial Bandit Problems with Multiple Plays. *International Conference on Algorithmic Learning Theory* 375–389.

Xia, Y.; Qin, T.; Ma, W.; Yu, N.; and Liu, T.-Y. 2016. Budgeted Multi-Armed Bandits with Multiple Plays. *Proceedings of the 25th International Joint Conference on Artificial Intelligence* 2210 – 2216.

Zhou, D., and Tomlin, C. 2017. Budget-Constrained Multi-Armed Bandits with Multiple Plays. "https://arxiv.org/pdf/1711.05928.pdf".