Multi-Armed Bandits with Betting

Alexandru Niculescu-Mizil ANICULE@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA

Abstract

We study an extension to the stochastic multi-armed bandit problem where the learner has a budget of K "coins" it can use in each round. The learner can use the coins to play multiple arms in each round, having the option to "bet" multiple coins on an arm. At the end of the round, the arms generate a reward that is proportional to the amount of coins invested in them.

1. Introduction

The stochastic multi-armed bandit problem (Lai & Robbins, 1985) is motivated in terms of playing slot machines in a gambling casino. At each round, the gambler has *a single* coin he or she uses to play exactly one of N slot machines. The played machine yields a random payoff from a fixed distribution that is unknown to the gambler. The goal of the gambler is to maximize the expected payoff.

In this paper we consider an extension where the gambler has, at each round, K coins available for play, and the slot machines accept bets. If the player bets m coins on a machine, then the machine will return m times the payoff of that round. It is important to note that betting m coins on a machine results in obtaining a single sample from the rewards distribution of that machine (multiplied by m), not m independent samples. At each round, the gambler must divide *all* of his or hers K coins among the machines in such a way as to maximize the total expected payoff.

A different way to look at the same problem is in the context of collaborative learning: K honest and altruistic learners are collaborating to to maximize the cumulative total expected payoff. In each round, each learner selects exactly one of the N arms, and receives the payoff of the selected arm in that round. If two learners select the same arm, they both receive the same reward. At the end of the round the learners share information about the payoffs received, and coordinate on which arms to select in the next round.

As in the classical setting, performance is measured in terms of regret, i.e. the difference between the payoff obtained by the algorithm, and the payoff obtained by the optimal strategy of betting all K coins on the arm with the highest average payoff.

A naive approach would be to reduce this problem to a classical multi-armed bandit problem by betting all K coins on the machine selected, for instance, by the UCB1 algorithm (Auer et al., 2002a). This would yield a regret of $\mathcal{O}(K \cdot \ln T)$, where T is the number of rounds played, and the constant depends on the mean payoffs of the different machines. In this paper we present an algorithm that achieves a regret of $\mathcal{O}(K + \ln T)$ by splitting the coins among multiple machines in order to emphasize exploration.

2. Related work

The classical multi-armed bandit problem, the simplest example of an exploration-exploitation trade-off problem, has received significant attention in the literature. In the stochastic setting, Lai and Robbins (1985) gave asymptotically optimal strategies, and Auer et al. (2002a) introduced the UCB1 algorithm and showed that the optimal regret bounds of $\mathcal{O}(\ln T)$ can be achieved uniformly over time. For the non-stochastic version, Auer et al. (2002b) proposed the Exp3 algorithm that achieves a regret of $\mathcal{O}(\sqrt{T \cdot N \cdot \ln N})$.

In the context of Competitive Collaborative Learning, Awerbuch and Kleinberg (2008) studied, in the non-stochastic setting, a more general version of the problem studied in this paper, where the learners can be dishonest. They provide an algorithm with a regret bound of $\mathcal{O}((K+N) \cdot \ln^4(K+N) \cdot T^{3/4})$

A more restrictive version of the problem studied here, where, at each round, the gambler has to play exactly K distinct machines, with each machine accepting a single coin, has been studied by Anantharam et al. (1987) under the name of multi-armed bandit problems with multiple plays.

3. Main Result

We consider the stochastic rewards model, where the reward for arm i at time t, $r_i(t)$, is drawn independently from a fixed unknown distribution $P_i(\cdot)$ with mean μ_i. WLOG let the arms be ordered by the true average reward: $\mu_1 > ... > \mu_N$. Let $\Delta_i = \mu_1 - \mu_i$. We denote by $n_{i,t}$ the

number of times arm i has been played up to time t. Let

$$c_i^t = \sqrt{\frac{2\ln t}{n_{i,t}}}$$

Similar to UCB1 we keep a confidence interval of $\pm c_i^t$ around the observed mean reward for each arm. We say that an arm i is dominated if there exist another arm j such that the lower confidence bound of arm j is higher than the upper confidence bound of arm i.

The algorithm proceeds in two phases for each round. In the first phase, a single coin per arm is assigned for the exploration of the non-dominated arms. The dominated arms can be safely excluded as the best arm will, with high probability, not be among them. In the second phase, if there are any coins left over from the first phase, they are used to exploit the arm with the highest observed average reward. The algorithm is detailed in Algorithm 1.

Initialize $z_i = 0$ and $n_i = 0$ for all $i \in [N]$.
for $t = 1$ *to* T **do**
  Compute $ub_i = \left(\frac{z_i}{n_i} + \sqrt{\frac{2\ln t}{n_i}}\right)$ for all $i \in [N]$.
  Compute $lb_i = \left(\frac{z_i}{n_i} - \sqrt{\frac{2\ln t}{n_i}}\right)$ for all $i \in [N]$.
  PHASE 1:
  $A_t \leftarrow \{i | ub_i > lb_j \forall j \in [N]\}$
  $S_t \leftarrow \{$the up to K arms in A_t with highest $ub_i\}$
  Bet one coin on each arm in S_t
  PHASE 2:
  if $|A_t| < K$ **then**
    Bet $K - |A_t|$ coins on the arm with highest $\frac{z_i}{n_i}$
  end
  Observe payoffs $r_i(t)$ for all arms $i \in S_t$
  $z_i \leftarrow z_i + r_i(t)$ for all arms $i \in S_t$
  $n_i \leftarrow n_i + 1$ for all arms $i \in S_t$
end

Algorithm 1:

Theorem 1 *The regret of the algorithm is at most:*

$$\ln T \sum_{i=2}^{N} \frac{40}{\Delta_i} + \left(2 + \frac{2\pi^2}{3}\right) \sum_{i=2}^{N} \Delta_i + \\ + (K-1) \cdot \sum_{i=2}^{N} \left(\frac{16}{\Delta_i} + \frac{\pi^4}{90}\right)$$

Proof:

To prove the theorem we will separately bound the regret incurred in each phase of the algorithm.

Regret in Phase1. To play an arm i in the first phase of the algorithm, it must be that either its upper confidence bound exceeds that of arm 1, or its upper confidence bound is lower than that of arm 1, but its higher than the lower confidence bound of arm 1. Let $n_{i,t}^h$ be the number of times arm i has been played when its upper confidence bound was higher than that of arm 1, up to time t, and $n_{i,t}^l$ be the number of times arm i has been played when its upper confidence bound was lower than that of arm 1, up to time t. Let $n_{i,t} = n_{i,t}^h + n_{i,t}^l$ be the total number of times arm i has been played. The regret incurred in phase 1 will be:

$$R_1(T) = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_{i,T}] = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_{i,T}^h] + \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_{i,T}^l]$$

$\mathbb{E}[n_{i,T}^h]$ can be bounded in the same way as for UCB1 algorithm. Let S_t be the set of arms played at time t. We have

$$\begin{aligned} n_{i,T}^h &= 1 + \sum_{t=1}^{T} \{i \in S_t, ub_{i,t} \geq ub_{1,t}\} \\ &\leq l + \sum_{t=1}^{T} \{ub_{i,t} \geq ub_{1,t}, n_{i,t-1}^h \geq l\} \\ &\leq l + \sum_{t=1}^{T} \left\{\hat{\mu}_{i,t-1} + c_{t-1,n_{i,t-1}} \geq \right. \\ &\quad \left. \geq \hat{\mu}_{1,t-1} + c_{t-1,n_{1,t-1}}, n_{i,t-1}^h \geq l\right\} \\ &\leq l + \sum_{t=1}^{T} \left\{ \min_{0<s_1<t} \hat{\mu}_{1,s_1} + c_{t-1,s_1} \leq \right. \\ &\quad \left. \leq \min_{l<s_i<t} \hat{\mu}_{1,s_i} + c_{t-1,s_i} \right\} \\ &\leq l + \sum_{t=1}^{\infty} \sum_{s_1=1}^{t-1} \sum_{s_i=l}^{t-1} \{\hat{\mu}_{1,s_1} + c_{t,s_1} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\} \end{aligned}$$

In order for $\hat{\mu}_{1,s_1} + c_{t,s_1} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$ must be the case that at least one of the following is true:

$$\begin{aligned} \hat{\mu}_{1,s_1} &\leq \mu_1 - c_{t,s_1} \\ \hat{\mu}_{i,s_1} &\geq \mu_i + c_{t,s_i} \\ \mu_1 &\leq \mu_i + 2 \cdot c_{t,s_i} \end{aligned}$$

Applying Chernoff-Hoeffding bounds we have:

$$\begin{aligned} \mathbb{P}\left[\hat{\mu}_{1,s_1} \leq \mu_1 - c_{t,s_1}\right] &\leq e^{-4\ln t} = t^{-4} \\ \mathbb{P}\left[\hat{\mu}_{i,s_i} \leq \mu_i + c_{t,s_i}\right] &\leq e^{-4\ln t} = t^{-4} \end{aligned}$$

and for $s_i > (8\ln T)/\Delta_i^2$ we have:

$$\mu_1 - \mu_i - 2 \cdot c_{t,s_i}^i = \mu_1 - \mu_i - 2\sqrt{2\ln t / s_i} \geq$$

$$\geq \mu_1 - \mu_i - \Delta_i = 0$$

So we get

$$\begin{aligned}
\mathbb{E}[n^h_{i,T}] &\leq \frac{8\ln T}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty}\sum_{s_1=1}^{t-1}\sum_{s_i=\lceil 8\ln T/\Delta_i^2\rceil}^{t-1} \\
&\quad \times \left(\mathbb{P}\left[\hat{\mu}_{1,s_1} \leq \mu_1 - c_{t,s_1}\right] + \mathbb{P}\left[\hat{\mu}_{i,s_i} \leq \mu_i + c_{t,s_1}\right]\right) \\
&\leq \frac{8\ln T}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty}\sum_{s_1=1}^{t-1}\sum_{s_i=1}^{t-1} 2t^{-4} \\
&\leq \frac{8\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}
\end{aligned}$$

To bound $\mathbb{E}[n^l_{i,T}]$ we first note that $n_{1,t} \geq n^l_{i,t}$ for any t. Indeed, whenever arm i is selected and its upper confidence bound is lower than that of arm 1, arm 1 must be selected too because the algorithm selects arms in the order of their upper confidence bound. So the $n^l_{i,T}$ can be bounded by

$$\begin{aligned}
n^l_{i,T} &= 1 + \sum_{t=1}^{T}\{i \in S_t, ub_{i,t} \leq ub_{1,t}\} \\
&\leq l + \sum_{t=1}^{T}\{i \in S_t, ub_{i,t} \leq ub_{1,t}, n^l_{i,t-1} \geq l\} \\
&\leq l + \sum_{t=1}^{T}\Big\{\hat{\mu}_{i,t-1} + c_{t-1,n_{i,t-1}} \geq \\
&\qquad \geq \hat{\mu}_{1,t-1} - c_{t-1,n_{1,t-1}}, n_{1,t-1} \geq n^l_{i,t-1} \geq l\Big\} \\
&\leq l + \sum_{t=1}^{T}\left\{\min_{l<s_1<t}\hat{\mu}_{1,s_1} - c_{t-1,s_1} \leq \right. \\
&\qquad \left. \leq \min_{l<s_i<t}\hat{\mu}_{1,s_i} + c_{t-1,s_i}\right\} \\
&\leq l + \sum_{t=1}^{\infty}\sum_{s_1=l}^{t-1}\sum_{s_i=l}^{t-1}\{\hat{\mu}_{1,s_1} - c_{t,s_1} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\}
\end{aligned}$$

For $\hat{\mu}_{1,s_1} - c_{t,s_1} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$ at least one of the following must hold:

$$\begin{aligned}
\hat{\mu}_{1,s_1} &\leq \mu_1 - c_{t,s_1} \\
\hat{\mu}_{i,s_1} &\geq \mu_i + c_{t,s_i} \\
\mu_1 - 2c_{t,s_1} &\leq \mu_i + 2c_{t,s_i}
\end{aligned}$$

But, for $s_1 > 32\ln T/\Delta_i^2$ and $s_i > 32\ln T/\Delta_i^2$ we get:

$$\begin{aligned}
&\mu_1 - 2c_{t,s_1} - \mu_i - 2\cdot c_{t,s_i} = \\
&= \mu_1 - \mu_i - 2\sqrt{2\ln t/s_1} - 2\sqrt{2\ln t/s_i} \\
&\geq \mu_1 - \mu_i - \Delta_i/2 - \Delta_i/2 = 0
\end{aligned}$$

Which leads to

$$\begin{aligned}
&\mathbb{E}[n^l_{i,T}] \leq \\
&\leq \frac{32\ln T}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty}\sum_{s_1=\lceil 32\ln T/\Delta_i^2\rceil}^{t-1}\sum_{s_i=\lceil 32\ln T/\Delta_i^2\rceil}^{t-1} \\
&\quad \times \left(\mathbb{P}\left[\hat{\mu}_{1,s_1} \leq \mu_1 - c_{t,s_1}\right] + \mathbb{P}\left[\hat{\mu}_{i,s_i} \leq \mu_i + c_{t,s_1}\right]\right) \\
&\leq \frac{32\ln T}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty}\sum_{s_1=1}^{t-1}\sum_{s_i=1}^{t-1} 2t^{-4} \\
&\leq \frac{32\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}
\end{aligned}$$

Putting all together we can bound the regret incurred in phase 1 of the algorithm

$$R_1(T) = \sum_{i=2}^{N}\Delta_i\mathbb{E}[n_{i,T}] \leq \ln T\sum_{i=2}^{N}\frac{40}{\Delta_i} + \left(2 + \frac{2\pi^2}{3}\right)\sum_{i=2}^{N}\Delta_i$$

Regret at phase 2 To bound the regret incurred in phase 2 we will use the Azuma-Hoeffding Inequality:

$$P[n_{i,t}\hat{\mu}_{i,t} > n_{i,t}\mu_i + n_{i,t}\Delta_i/2] \leq e^{-\frac{\Delta_i^2 n_{i,t}}{8}}$$

and

$$P[n_{1,t}\hat{\mu}_{1,t} < n_{1,t}\mu_1 + n_{1,t}\Delta_i/2] \leq e^{-\frac{\Delta_i^2 n_{1,t}}{8}}$$

Let A_t be the set of non-dominated arms at time t. Let us restrict our attention to only those time steps were $1 \in A_t$, and assume a maximum regret in the other time steps. In order to incur regret in phase 2, the algorithm must select an arm in A_t. Also it must still have leftover coins after phase 1, which means that all arms in A_t, and in particular arm 1, are selected this round ($n_{i,t} = n_{i,t-1} + 1$ for all $i \in A_t$). Let M_i be the number of times arm i is selected in phase 2, when arm 1 is not dominated. We can bound the expected value of M_i by:

$$\begin{aligned}
\mathbb{E}[M_i] &= \mathbb{E}\left[\sum_{t=1}^{T} P[\hat{\mu}_{i,t} > \hat{\mu}_{1,t}]\cdot 1(1, i \in A_t)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} P[n_{i,t}\hat{\mu}_{i,t} > n_{i,t}\mu_i + n_{i,t}\Delta_i/2]\cdot 1(i \in A_t)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} P[n_{1,t}\hat{\mu}_{1,t} < n_{1,t}\mu_1 + n_{1,t}\Delta_i/2]\cdot 1(1 \in A_t)\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{t=1}^{\infty} e^{-\frac{\Delta_i^2 n_{i,t}}{8}} \cdot 1(i \in A_t)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{\infty} e^{-\frac{\Delta_i^2 n_{1,t}}{8}} \cdot 1(1 \in A_t)\right] \\
&\leq \mathbb{E}\left[\sum_{n=1}^{\infty} e^{-\frac{\Delta_i^2 n}{8}}\right] + \mathbb{E}\left[\sum_{n=1}^{\infty} e^{-\frac{\Delta_i^2 n}{8}}\right] \\
&= \frac{1}{e^{\Delta_i^2/8-1}} + \frac{1}{e^{\Delta_i^2/8-1}} \leq \frac{16}{\Delta_i^2}
\end{aligned}
$$

Now all that remains to do is to bound the number of times arm 1 is dominated by some other arm, and thus not in A_t. For arm 1 to be dominated at time t (i.e. its upper bound is lower than the lower bound of some other arm at time t) it must be the case that either

$$\hat{\mu}_{1,t} \leq \mu_1 - c_{s,t}$$

or

$$\hat{\mu}_{i,t} \leq \mu_i + c_{s,t}$$

for some arm $i \neq 1$.

Let D be the number of times arm 1 is dominated. We have

$$
\begin{aligned}
\mathbb{E}[D] &= \sum_{t=1}^{T} \mathbb{P}\left[1 \notin A_t\right] \\
&\leq \sum_{t=1}^{T} \left(\mathbb{P}\left[\hat{\mu}_{1,t} \leq \mu_1 - c_{s,t}\right] + \sum_{i=2}^{N} \mathbb{P}\left[\hat{\mu}_{i,t} \leq \mu_i + c_{s,t}\right] \right) \\
&\leq \sum_{t=1}^{\infty} N \cdot t^{-4} = N \cdot \frac{\pi^4}{90}
\end{aligned}
$$

Since there can be at most K-1 learners that incur regret in phase 2, the regret in phase 2 is less than:

$$
\begin{aligned}
R_2(T) &\leq (K-1)\left(\mathbb{E}[D] + \sum_{i=2}^{N} \mathbb{E}[M_i] \cdot \Delta_i\right) \\
&\leq (K-1) \cdot \sum_{i=2}^{N} \left(\frac{16}{\Delta_i} + \frac{\pi^4}{90}\right)
\end{aligned}
$$

Putting it all together we get the statement of the theorem.

■

4. Conclusions

We study an extension of the stochastic multi-armed bandit framework where the learner has a budget of K coins that it can use to play multiple arms at each round. The learner may chose to bet multiple coins on the same arm, in which case the reward received from that arm is proportional to the amount of coins invested. The learning algorithm we propose for this problem first allocates resources for the exploration of viable arms (arms that can not be ruled out with high probability as being suboptimal), then it uses the remaining resources, if any, for the exploitation of the most promising arm. We prove an upper bound of $\mathcal{O}(K + \ln T)$ on the regret of this algorithm.

In terms of a lower bound on regret, one can notice that, if the learner were to receive information on the rewards immediately after each of the K plays in a round, and also were to receive a fresh sample from the rewards distribution of an arm for every coin (as opposed to receiving a single sample per machine in each round) the problem would become a standard multi-armed bandit problem with $T \cdot K$ rounds. Hence, we get a lower bound on the regret of $\mathcal{O}(\ln T + \ln K)$ (Lai & Robbins, 1985). Bridging he gap between this lower bound and the linear in K upper bound obtained in this paper remains an open problem.

Acknowledgements

Thanks to Robert Kleinberg for helpful discussions.

References

Anantharam, V., Varaiya, P., , & Walrand, J. (1987). Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays part i: I.i.d. rewards. *IEEE Transactions on Automatic Control, 32*, 968976.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*, 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM J. Comput., 32*, 48–77.

Awerbuch, B., & Kleinberg, R. (2008). Competitive collaborative learning. *Journal of Computer and System Sciences, 74*, 1271–1288.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocations rules. *Adv. in Appl. Math., 6*, 4–22.