Bandit-Based Methods

Jan-Hendrik Lange

November 18, 2014

Outline

We present the most important result from the paper “Finite-time Analysis of the Multiarmed Bandit Problem” by Auer et al., which is the foundation of bandit-based methods in Monte Carlo Tree Search.
Outline:

1. Introduction
2. Multiarmed Bandit Problem & Regret
3. UCB1 & Main Result
4. Proof

Introduction

- Consider a sequential decision problem where we are given K options each associated with a stochastically distributed *reward* (or *cost*).
- We seek to maximize rewards (or minimze costs) by figuring out the best option and taking that option as often as possible in a growing number of turns. $\longrightarrow$ exploration vs. exploitation
- Different interpretations are imagineable, most well-known is the model called *multiarmed bandit problem*.

Multiarmed Bandit Problem

- We are given random variables $X_{i,n}$ for $1 \leq i \leq K$, $n \in \mathbb{N}$, where $X_{i,n}$ describes the reward obtained by playing the i-th bandit for the n-th time.
- We assume independence of all random variables and identical (unknown) distributions for fixed i with $\mu_i = \mathrm{E}[X_{i,n}]$ being the expectation of playing machine i at any time n.

Regret

- Goal: Choose a *policy* that maximizes the expected rewards when successively playing the bandits.
- Equivalently, we can minimize the *regret* after n plays, which is defined as

$$R(n) = \mu^* n - \sum_{j=1}^{K} \mu_j \, \mathrm{E}[T_j(n)],$$

where $\mu^* = \max_j \mu_j$ and $T_j(n)$ denotes the number of times bandit j has been played during the first n plays in total.

Regret

It will be helpful to rewrite the regret as follows. Put $\Delta_j = \mu^* - \mu_j$, then we use $n = \sum_{j=1}^K T_j(n)$ to get

$$\begin{aligned} R(n) &= \mu^* n - \sum_{j=1}^{K} \mu_j \, \mathrm{E}[T_j(n)] = \mathrm{E}\left[\mu^* n - \sum_{j=1}^{K} \mu_j T_j(n)\right] \\ &= \mathrm{E}\left[\sum_{j=1}^{K} (\mu^* - \mu_j) T_j(n)\right] = \mathrm{E}\left[\sum_{j=1}^{K} \Delta_j T_j(n)\right] \\ &= \sum_{j:\mu_j<\mu^*} \Delta_j \, \mathrm{E}[T_j(n)]. \end{aligned}$$

Previous Results

- In a paper from 1985, Lai and Robbins described policies which ensured for any suboptimal machine j that

$$\mathrm{E}[T_j(n)] \leq c_j(n) \cdot \ln n, \quad \text{where } c_j(n) \to c_j \in \mathbb{R} \text{ as } n \to \infty,$$

  given the reward distributions are in a certain class.

- Moreover, they showed that under some mild assumptions any arbitrary policy satisfies

$$\mathrm{E}[T_j(n)] \geq c_j \cdot \ln n$$

  for large n, leaving the former policies (asymptotically) optimal.

Main Result

In a nutshell, the main result of Auer et al. is to give a very simple and efficient policy, called UCB1, which achieves

$$\mathrm{E}[T_j(n)] \leq c \cdot \ln n + c', \quad \text{where } 0 \leq c, c' \in \mathbb{R},$$

for all n, under very little assumptions on the underlying reward distributions.
This yields a bound on the regret $R(n)$ within a constant factor of $\ln n$ uniformly for all n.

UCB1

We define $\bar{X}_{j,n} = \frac{1}{n}\sum_{t=1}^{n} X_{j,t}$, i.e. the average reward of machine j in n successive plays.

The (deterministic) policy UCB1 proceeds as follows:

1. For $n = 1, \dots, K$, play bandit n. (Initialize by playing each machine once.)
2. After $n \geq K$ plays, select machine

$$i = \arg\max_{j} \; \bar{X}_{j,T_j(n)} + \sqrt{\frac{2\ln n}{T_j(n)}},$$

The name UCB1 (*Upper Confidence Bound*) relates to the second summand and will become clearer considering the proof.

Theorem

Let $K > 1$ and $X_{i,n}$ be random rewards with support in $[0, 1]$. Suppose we play the bandits successively following policy UCB1. Then it holds that

$$R(n) \leq \left[8 \sum_{j:\mu_j < \mu^*} \left(\frac{\ln n}{\Delta_j} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^{K} \Delta_j \right).$$

Hoeffding's Inequality

For the proof, we need the following fact from probability theory. It is a special case of Hoeffding's inequality.

Fact

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with common range $[0,1]$ and mean μ. Denote their average by $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for all $a \geq 0$, we have

$$\mathrm{P}[\bar{X}_n \geq \mu + a] \leq e^{-2na^2}$$

$$\text{and} \quad \mathrm{P}[\bar{X}_n \leq \mu - a] \leq e^{-2na^2}.$$

Proof

For better presentation, we put $c_{t,s} = \sqrt{\frac{2\ln t}{s}}$. Also, for all expressions referring to some optimal bandit we add $*$ as a superscript.
Moreover, let the random variables I_t indicate the index of the machine played at time t. We use $[A]$ to denote the indicator function of some event A.
Thus, according to UCB1 we can write

$$[I_t = i] = 1 \iff i = \arg\max_j \ \bar{X}_{j,T_j(t-1)} + c_{t-1,T_j(t-1)}$$

to express bandit i has been picked at time t.

Proof

Let $i \in \{1, \ldots, K\}$ be any index and $\ell \in \mathbb{N}$. Then we have

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=K+1}^{n} [I_t = i] \leq \ell + \sum_{t=K+1}^{n} [I_t = i,\ T_i(t-1) \geq \ell] \\ &\leq \ell + \sum_{t=K+1}^{n} [\bar{X}^*_{T^*(t-1)} + c_{t-1,T^*(t-1)} \leq \bar{X}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}, \\ &\qquad\qquad T_i(t-1) \geq \ell] \\ &\leq \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} [\bar{X}^*_s + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i}]. \end{aligned}$$

Proof

Observe that, if $\bar{X}^*_s + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i}$, then at least one of the following must hold

$$\begin{aligned}
\bar{X}^*_s &\leq \mu^* - c_{t,s} \\
\bar{X}_{i,s_i} &\geq \mu_i + c_{t,s_i} \\
\mu^* &< \mu_i + 2c_{t,s_i}.
\end{aligned}$$

This is true, since if we assume the contrary, then

$$\bar{X}^*_s + c_{t,s} > \mu^* \geq \mu_i + 2c_{t,s_i} > \bar{X}_{i,s_i} + c_{t,s_i},$$

a contradiction.

Proof

Using Hoeffding's inequality, we can bound the first two events by

$$\mathrm{P}[\bar{X}^*_s \leq \mu^* - c_{t,s}] \leq e^{-2s\left(\sqrt{2\ln t/s}\right)^2} = e^{-4\ln t} = t^{-4}$$

and similarly $\mathrm{P}[\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i}] \leq t^{-4}$.
The third relation does not hold anymore as soon as s_i gets large enough. More precisely, for $s_i \geq (8\ln n)/\Delta_i^2$ we have

$$\mu^* - \mu_i - 2c_{t,s_i} = \mu^* - \mu_i - 2\sqrt{(2\ln t)/s_i} \geq \mu^* - \mu_i - \Delta_i = 0.$$

Proof

Hence, choosing $\ell = \lceil (8 \ln n)/\Delta_i^2 \rceil$, we finally obtain

$$\begin{aligned}
\mathrm{E}[T_i(n)] &\leq \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} \big(\mathrm{P}[\bar{X}_s^* \leq \mu^* - c_{t,s}] + \mathrm{P}[\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i}] \big) \\
&\leq \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t} \sum_{s_i=1}^{t} 2t^{-4} \\
&\leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}.
\end{aligned}$$

With $R(n) = \sum_{j : \mu_j < \mu^*} \Delta_j \, \mathrm{E}[T_j(n)]$, this is the assertion.

Conclusion

- The multiarmed bandit problem is "solved optimally" by policies that bound the regret asymptotically by $\ln n$.
- We have examined UCB1 as a very simple policy achieving this bound uniformly over n.
- This policy is widely used and leads to the UCT algorithm in Monte Carlo Tree Search.