The Multi-Armed Bandit Problem

Nicolò Cesa-Bianchi

Università degli Studi di Milano

The bandit problem

[Robbins, 1952]

$\dots$

K slot machines

- Rewards $X_{i,1}, X_{i,2}, \dots$ of machine i are i.i.d. $[0, 1]$-valued random variables
- An allocation policy prescribes which machine I_t to play at time t based on the realization of $X_{I_1,1}, \dots, X_{I_{t-1},t-1}$
- Want to play as often as possible the machine with largest reward expectation

$$\mu^* = \max_{i=1,\dots,K} \mathbb{E}\, X_{i,1}$$

Bandits for targeting content

- Choose the best content to display to the next visitor of your website
- Goal is to elicit a response from the visitor (e.g., click on a banner)
- Content options = slot machines
- Response rate = reward expectation
- Simplifying assumptions:
  1. fixed response rates
  2. no visitor profiles

Definition (Regret after n plays)

$$\mu^* n - \sum_{t=1}^{n} \mathbb{E}\, X_{I_t,t}$$

Theorem (Lai and Robbins, 1985)

There exist allocation policies satisfying

$$\mu^* n - \sum_{t=1}^{n} \mathbb{E}\, X_{I_t,t} \leqslant c\, K \ln n \qquad \textit{uniformly over } n$$

Constant c roughly equal to $1/\Delta^*$, where

$$\Delta^* = \mu^* - \max_{j\,:\,\mu_j < \mu^*} \mu_j$$

A simple policy

UCB [Agrawal, 1995]

1. At the beginning play each machine once
2. At each time $t > K$ play machine I_t maximizing

$$\overline{X}_{i,t} + \sqrt{\frac{2 \ln t}{T_{i,t}}} \qquad \text{over } i = 1, \dots, K$$

- $\overline{X}_{i,t}$ is the average reward obtained from machine i
- $T_{i,t}$ is number of times machine i has been played

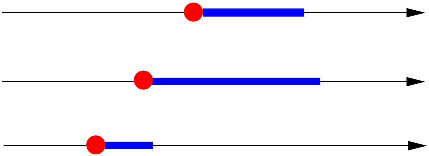
A finite-time regret bound

Theorem (Auer, C-B, and Fisher, 2002)

At any time n, the regret of the UCB *policy is at most*

$$\frac{8K}{\Delta^*}\ln n + 5K$$

Upper confidence bounds



$\sqrt{(2 \ln t)/T_{i,t}}$ is the size (using Chernoff-Hoeffding bounds) of the one-sided confidence interval for the average reward within which μ_i falls with probability $1 - \frac{1}{t}$

The epsilon-greedy policy

Input parameter: schedule $\varepsilon_1, \varepsilon_2, \dots$ where $0 \leqslant \varepsilon_t \leqslant 1$
At each time t:

1. with probability $1 - \varepsilon_t$ play the machine I_t with the highest average reward
2. with probability ε_t play a random machine

Is there a schedule of ε_t guaranteeing logarithmic regret?

The tuned epsilon-greedy policy

Theorem (Auer, C-B, and Fisher, 2002)

If $\varepsilon_t = 12/(d^2 t)$ *where* d *satisfies* $0 < d \leqslant \Delta^*$ *then the instantaneous regret at any time* n *of tuned* ε*-greedy is at most*

$$O\left(\frac{K}{dn}\right)$$

Practical performance

The UCB TUNED policy:

$$\sqrt{\frac{2\ln t}{T_{i,t}}} \qquad \text{is replaced by} \qquad \sqrt{\frac{\ln t}{T_{i,t}} \min\left\{\frac{1}{4}, V_{j,t}\right\}}$$

where $V_{j,t}$ is an upper confidence bound for the variance of machine j

Practical performance

- Optimally tuned ε-greedy performs almost always best unless there are several nonoptimal machines with wildly different response rates
- Performance of ε-greedy is quite sensitive to bad tuning
- UCB TUNED performs comparably to a well-tuned ε-greedy and is not very sensitive to large differences in the response rates

The nonstochastic bandit problem

[Auer, C-B, Freund, and Schapire, 2002]

What if probability is removed altogether?

Nonstochastic bandits

Bounded real rewards $x_{i,1}, x_{i,2}, \ldots$ are deterministically assigned to each machine i

- Analogies with repeated play of an unknown game [Baños, 1968; Megiddo, 1980]
- Allocation policies are allowed to randomize

0 1 0 0 7 9 9 8 9 0 0 1

5 7 9 6 0 0 2 2 0 0 0 1

0 2 0 1 0 1 0 0 8 9 8 7

Definition (Regret)

$$\max_{i=1,\dots,K} \left(\sum_{t=1}^{n} x_{i,t} \right) - \mathbb{E}\left[\sum_{t=1}^{n} x_{I_t,t} \right]$$

Competing against arbitrary policies

0 1 0 0 7 9 9 8 9 0 0 1

5 7 9 6 0 0 2 2 0 0 0 1

0 2 0 1 0 1 0 0 8 9 8 7

Tracking regret

Regret against an arbitrary and unknown policy $(j_1, j_2, \dots, j_n)$

$$\sum_{t=1}^{n} x_{j_t,t} - \mathbb{E}\left[\sum_{t=1}^{n} x_{I_t,t}\right]$$

Theorem (Auer, C-B, Freund, and Schapire, 2002)

For all fixed S, the regret of the weight sharing policy against any policy $\mathbf{j} = (j_1, j_2, \dots, j_n)$ is at most

$$\sqrt{S\, nK \ln K}$$

where S is the number of times $\mathbf{j}$ switches to a different machine