Machine Learning Coms-4771

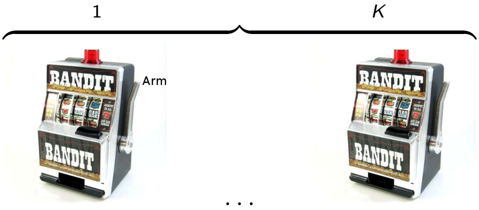
Multi-Armed Bandit Problems

Lecture 20

Multi-armed Bandit Problems

The Setting:

- K arms (or actions)
- Each time t, each arm i pays off a bounded real-valued reward $x_i(t)$, say in $[0, 1]$.
- Each time t, the learner chooses a single arm $i_t \in \{1, \ldots, K\}$ and receives reward $x_{i_t}(t)$. The goal is to maximize the return.



t=1	0.3	0.2	0.8	0.4	0.0
t=2	0.7	0.1	0.9	0.5	0.1
t=3	0.5	0.3	0.7	0.3	0.3

...

The simplest instance of the exploration-exploitation problem

Bandits for targeting content

- Choose the best content to display to the next visitor of your website
- Content options = slot machines
- Reward = user's response (e.g., click on a ad)
- A simplifying assumption: no context (no visitor profiles). In practice, we want to solve contextual bandit problems.

- Stochastic bandits: Each arm i is associated with some unknown probability distribution with expectation μ_i. Rewards are drawn iid.
  The largest expected reward: $\mu^* = \max_{i \in \{1,\ldots,K\}} \mathbf{E}[x_i]$
  Regret after T plays:
  $$\mu^* T - \sum_{t=1}^{T} \mathbf{E}[x_{i_t}(t)]$$
  expectation is over the draws of rewards and the randomness in player's strategy
- Adversarial (nonstochastic) bandits: No assumption is made about the reward sequence (other than it's bounded). Regret after T plays:
  $$\max_i \sum_{t=1}^{T} x_i(t) - \sum_{t=1}^{T} \mathbf{E}[x_{i_t}(t)]$$
  expectation is only over the randomness in the player's strategy

Stochastic Bandits: Upper Confidence Bounds Strategy

UCB

- Play each arm once
- At any time $t > K$ (deterministically) play machine i_t maximizing

$$\bar{x}_j(t) + \sqrt{\frac{2 \ln t}{T_{j,t}}},$$

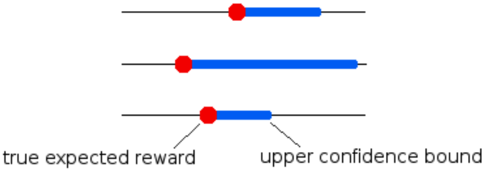
over $j \in \{1, \ldots, K\}$ where

- $\bar{x}_j$ is the average reward obtained from machine j
- $T_{j,t}$ is the number of times j has been played so far

UCB

Intuition:

The second term $\sqrt{2 \ln t / T_{j,t}}$ is the the size of the one-sided $(1 - 1/t)$-condifence interval for the average reward (using Chernoff-Hoeffding bounds).



Theorem

(Auer, Cesa-Bianchi, Fisher) At time T, the regret of the UCB policy is at most

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

where $\Delta^* = \mu^* - \max_{i:\mu_i<\mu^*} \mu_i$ (the gap between the best expected reward and the expected reward of the runner up).

Stochastic Bandits: ϵ-greedy

Randomized policy: ϵ_t-greedy

Parameter: schedule $\epsilon_1, \epsilon_2, \ldots$, where $0 \leq \epsilon_t \leq 1$.
At each time t

- (exploit) with probability $1 - \epsilon_t$, play the arm i_t with the highest current average return
- (explore) with probability ϵ, play a random arm

Is there a schedule of ϵ_t which guarantees logarithmic regret? Constant ϵ causes linear regret. Fix: let ϵ go to 0 as our estimates of the expected rewards become more accurate.

Theorem

(Auer, Cesa-Bianchi, Fisher) If $\epsilon_t = 12/(d^2 t)$ where $0 < d \leq \Delta^*$, then the instantaneous regret (i.e., probability of choosing a suboptimal arm) at any time t of ϵ-greedy is at most

$$O(\frac{K}{dt}).$$

The regret of ϵ-greedy at time T (summing over the steps) is thus at most

$$O(\frac{\Delta^*}{d} K \log T)$$

(using $\sum_{t=1}^{T} \frac{1}{t} \approx lnT + \gamma$ where $\gamma \approx 0.5772$ is the Euler constant).

Practical performance (from Auer, Cesa-Bianchi and Fisher):

- Tuning the UCB: replace $\sqrt{2 \ln t / T_{i,t}}$ with

$$\sqrt{\frac{\ln t}{T_{i,t}} \min\{1/4, V_{i,t}\}},$$

where $V_{i,t}$ is an upper confidence bound for the variance of arm i. (The factor $1/4$ is an upper bound on the variance of any $[0, 1]$ bounded variable.) Performs significantly better in practice.

- ϵ-greedy is quite sensitive to bad parameter tuning and large differences in response rates. Otherwise an optimally tuned ϵ-greedy performs very well.

- UCB tuned performs comparably to a well-tuned ϵ-greedy and is not very sensitive to large differences in response rates.

Nonstochastic Bandits: Recap

- No assumptions are made about the generation of rewards.
- Modeled by an *arbitrary* sequence of reward vectors $x_1(t), \ldots, x_K(t)$, where $x_i(t) \in [0,1]$ is the reward obtained if action i is chosen at time t.
- At step t, the player chooses arm i_t and receives x_{i_t}.
- Regret after T plays (with respect to the best single action):

$$\underbrace{\max_j \sum_{t=1}^{T} x_j(t)}_{G_{\max}=\text{reward of the best action in hindsight}} \quad - \quad \underbrace{\sum_{t=1}^{T} \mathbf{E}[x_{i_t}(t)]}_{\text{expected reward of the player}}$$

Exp3 Algorithm (Auer, Cesa-Bianchi, Freund, and Schapire)

- Initialization: $w_i(1) = 1$ for $i \in \{1, \ldots, K\}$
- Set $\gamma = \min\{1, \sqrt{\frac{K \ln K}{(e-1)g}}\}$, where $g \geq G_{\max}$.
- For each $t = 1, 2, \ldots$
  - Set
$$p_i(t) = (1-\gamma)\frac{w_i}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$$
  - Draw i_t randomly according to $p_1(t), \ldots, p_K(t)$.
  - Receive reward $x_{i_t}(t) \in [0,1]$
  - For $j = 1, \ldots, K$ set the estimated rewards and update the weights:
$$\hat{x}_j(t) = \begin{cases} x_j(t)/p_j(t) & \text{if } j = i_t \\ 0 & \text{otherwise} \end{cases}$$
$$w_j(t+1) = w_j(t)\exp(\gamma \hat{x}_j(t)/K)$$

Exp3 Algorithm (Auer, Cesa-Bianchi, Freund, and Schapire)

Theorem: For any $T > 0$ and for any sequence of rewards, regret of the player is bounded by

$$2\sqrt{e-1}\sqrt{gK \ln K} \leq 2.63\sqrt{TK \ln K}$$

Observation: Setting $\hat{x}_{i_t}$ to $x_{i_t}(t)/p_{i_t}(t)$ guarantees that the expectations are equal to the actual rewards for each action:

$$\mathbf{E}[\hat{x}_j \mid i_1, \ldots, i_{t-1}] = p_j(t)x_j(t)/p_j(t) = x_j(t),$$

where the expectation is with respect to the random choice of i_t at time t (given the choices in the previous rounds). So dividing by p_{i_t} compensates for the reward of actions with small probability of being drawn.

Proof: Let $W_t = \sum_j w_j(t)$. We have

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{K} \frac{\overbrace{w_i(t+1)}^{w_i(t)exp(\gamma \hat{x}_i(t)/K)}}{W_t} = \sum_{i=1}^{K} \frac{p_i(t) - (\gamma/K)}{1-\gamma} exp(\gamma \hat{x}_i(t)/K)$$

$$\leq \sum_{i=1}^{K} \frac{p_i(t) - (\gamma/K)}{1-\gamma} (1 + \frac{\gamma}{k}\hat{x}_i(t) + (e-2)\frac{\gamma^2}{k^2}\hat{x}_i^2(t))$$

$$(using\ e^x \leq 1 + x + (e-2)x^2 \text{ for } x \in [0,1])$$

$$\leq \overbrace{\sum_{i=1}^{K} \frac{p_i(t) - (\gamma/k)}{1-\gamma}}^{=1} + \sum_{i=1}^{K} \frac{p_i(t)\gamma\hat{x}_i(t)}{(1-\gamma)K} + \frac{(e-2)\gamma^2}{(1-\gamma)K^2} \sum_i p_i(t)\hat{x}_i^2(t)$$

$$\leq 1 + \left[\frac{\gamma}{(1-\gamma)K} \underbrace{x_{i_t}(t)}_{\text{the only non-zero term}} + (e-2)\frac{(\gamma/K)^2}{1-\gamma} \sum_{i=1}^{K} \hat{x}_i(t) \right]$$

use approximation $1 + z \leq e^z$

Take logs:

$$\ln \frac{W_{T+1}}{W_t} \le \frac{\gamma}{(1-\gamma)K} x_{i_t}(t) + (e-2)\frac{(\gamma/K)^2}{1-\gamma}\sum_{i=1}^{K}\hat{x}_i(t)$$

Summing over t,

$$\ln \frac{W_{T+1}}{W_1} \le \frac{\gamma/K}{(1-\gamma)} \overbrace{G_{\exp 3}}^{\text{reward of Exp3}} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{t=1}^{T}\sum_{i=1}^{K}\hat{x}_i(t)$$

Now, for any fixed arm j

$$\ln \frac{W_{T+1}}{W_1} \ge \ln \frac{w_j(T+1)}{W_1} = \ln \frac{w_j(1)}{W_1} + (\gamma/K)\sum_{t=1}^{T}\hat{x}_j(t).$$

Combine with the upper bound,

$$\frac{\gamma}{K}\sum_{t=1}^{T}\hat{x}_j(t) - \ln K \le \frac{\gamma/K}{1-\gamma}G_{\exp 3} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{t=1}^{T}\sum_{i=1}^{K}\hat{x}_i(t)$$

Solve for $G_{\exp 3}$:

$$G_{\exp 3} \ge (1-\gamma)\sum_{t=1}^{T}x_j(t) - \frac{K}{\gamma}\ln K \cdot (1-\gamma) - (e-2)(\gamma/K)\sum_{t=1}^{T}\sum_{i=1}^{K}\hat{x}_i(t)$$

Take expectaion of both sides wrt distribution of $i_1, \ldots, i_T$:

$$\mathbf{E}[G_{\exp 3}] \geq (1-\gamma) \sum_{t=1}^{T} x_j(t) - \frac{K}{\gamma} \ln K - (e-2) \frac{\gamma}{K} K G_{\max}.$$

Since j was chosen arbitrarily, it holds for $j = \max$:

$$\mathbf{E}[G_{\exp 3}] \geq (1-\gamma) G_{\max} - \frac{K}{\gamma} \ln K - (e-2) \gamma G_{\max}$$

Thus

$$G_{\max} - \mathbf{E}[G_{\exp 3}] \leq \frac{K \ln K}{\gamma} + (e-1) \gamma G_{\max}$$

The value of γ in the algorithm is chosen to minimize the regret.

Comments:

- Don't need to know T in advance (guess and double)
- Possible to get high probability bounds (with a modified version of Exp3 that uses upper confidence bounds)
- Stronger notions of regret. Compete with the best in a class of strategies.
- The difference between $\sqrt{T}$ bounds and $\log T$ bounds is a bit misleading. The difference is not due to the adversarial nature of rewards but in the asymptotic quantification! $\log T$ bounds hold for any fixed set of reward distributions (so Δ^* is fixed **before** T, not after).