EECS 495: Randomized Algorithms **Lecture 13**
Multi-armed Bandits

Reading: none

Multi-armed bandits (Robbins '52):

- Slot machine with multiple levers
- levers give rewards according to distribution
- want to maximize sum of rewards
- no initial knowledge about payoffs

Problem: Tradoff between

- *exploit* lever with high expected payoff
- *explore* to get more info about expected payoffs of other levers

Example: Keyword Allocations

- n advertisers with CPC $v_1, \ldots, v_n$ and CTR $p_1, \ldots, p_n$
- one slot per keyword
- CTRs unknown, fixed over time
- which ad to show?

Problem:

- n arms
- reward $X_i \in [0, 1]$ of arm i
- X_i random variable with mean μ_i

Goal: Given finite horizon T, seek policy to minimize *regret*:

$$\max_{X_i} \left(T \times \mu_{i^*} - E[\sum_{t=1}^{T} R_t] \right)$$

where i^* is arm with highest μ_i.

Claim: There is a policy that obtains regret $O(\sqrt{nT \log T})$ (and hence per-turn regret vanishes for large T).

Question: Easy policies?

Algorithm: Play all arms for a while, then play best one.

- Play each arm for $T^{2/3}$ steps
- Choose arm with max sample ave and play for remaining $T - nT^{2/3}$ steps

Claim: $O(nT^{2/3}\sqrt{\log T})$ regret.

Claim: (Chernoff-Hoeffding's inequality). Let $X_1, \ldots, X_k$ be k independent draws from a distribution on $[0, 1]$ with mean μ. Let $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} X_i$ be sample average. Then:

$$\Pr[\hat{\mu} - \mu > \epsilon] \leq 2e^{-2k\epsilon^2}$$

and

$$\Pr[\hat{\mu} + \mu < \epsilon] \leq 2e^{-2k\epsilon^2}.$$

Proof: (of regret bound):

- $\hat{\mu_i}$ = sample ave of arm i. by hoeffding with $k = T^{2/3}$:

$$\Pr[|\mu_i - \hat{\mu_i}| > \frac{\sqrt{\log T}}{T^{1/3}}] \leq \frac{2}{T^2}$$

- Assume $n << T$. by union bound:

$$\Pr[\exists i : |\mu_i - \hat{\mu_i}| > \frac{\sqrt{\log T}}{T^{1/3}}] \leq \frac{2}{T}$$

- W/prob. $1 - \frac{2}{T}$, chosen arm i has $\mu_i \geq \mu_{i^*} - \frac{2\sqrt{\log T}}{T^{1/3}}$, so regret at most

$$nT^{2/3} + T \times \frac{2\sqrt{\log T}}{T^{1/3}} + \frac{2}{T} \times T$$

where

  - first term is regret due to initial explore
  - second term regret due to slightly sub-opt arm played at most T times
  - third term regret due to arm sub-opt by 1

Idea: Treating all arms equal wastes time. Play arm with highest upper confidence interval. Either

- also has higher mean, or
- narrow confidence interval

either way, we're happy.

Def: If $\hat{\mu_i}$ is sample ave and t_i is number of times played arm i, then the *index* Φ_i of i is $\hat{\mu_i} + \sqrt{\frac{\log T}{t_i}}$.

Algorithm: Index policy

- Play arm with highest index
- Update index

Claim: $O(\sqrt{nT \log T})$ regret.

Proof: Let

- i^* be arm with highest mean,
- $\Delta_i = \mu_{i^*} - \mu_i$ be per-turn regret for playing i,
- Q_i be exp. # times i is played in T steps.

For each arm $i \neq i^*$, $E(Q_i) \leq \frac{4 \log T}{\Delta_i^2} + 2$:

- $\Pr[\Phi_{i^*} < \mu_{i^*}] \leq 1/T$ no matter how long we play it.

  If i^* played continously, at each step t,

$$\Pr[\Phi_{i^*}(t) < \mu_{i^*}] = \Pr[\mu_{i^*} - \hat{\mu_{i^*}}(t) > \sqrt{\frac{\log T}{t}}] \leq \frac{1}{T^2}$$

  by Hoeffding. Dips below μ_{i^*} with prob. at most $\frac{1}{T}$ by union bound over steps.

- $\Pr[\Phi_i > \mu_i] \leq 1/T$ after enough trials.

  If i played for $t_i = \frac{4 \log T}{\Delta_i^2}$ steps (index is $\hat{\mu_i} + \Delta_i/2$), then

$$\Pr[\Phi_i > \mu_i] = \Pr[\hat{\mu_i} - \mu_i > \Delta_i/2] \leq 1/T$$

  by Hoeffding.

If neither event happens, play i at most t_i times, else w.p. at most $2/T$, play arm at most T times.

Regret is:

$$\sum_i \Delta_i E[Q_i] \approx \sum_i (\frac{4 \log T}{\Delta_i})$$

Define

- Arms with large regret $\Delta_i > \sqrt{\frac{4n \log T}{T}}$ incur total regret at most $n \frac{4 \log T}{\Delta_i} = 2\sqrt{nT \log T}$
- Arms with small regret $\Delta_i \leq \sqrt{\frac{4n \log T}{T}}$ incur total regret at most $T \max_i \Delta_i = 2\sqrt{nT \log T}$.

Lower bound:

Claim: Any bandit policy incurs regret $\Omega(\sqrt{nT})$.

Proof: $n-1$ arms with mean $1/2$; one arm with mean $1/2+\epsilon$ for $\epsilon = O(\sqrt{n/2})$. With t samples, variance becomes $\sqrt{1/t}$, so need $O(T/n)$ samples to decide if arm is *good* one with constant prob. Not enough samples to resolve all arms, so with constant probability fail and incur regret $\epsilon T = \Omega(\sqrt{nT})$.