EECS598: Prediction and Learning: It's Only a Game Fall 2013

Lecture 18: The Bandit Setting

Prof. Jacob Abernethy *Scribe: Zhihao Chen*

Announcements

- Homework 2 grading nearly finished
- Homework 3 will be out within a week or so
- Class on Nov 27 (Wednesday before Thanksgiving) may be cancelled

18.1 Bandit setting

The player only finds out the loss associated with his choice, the losses on other choices are unknown to the player (as opposed to the "full information" setting, where all losses are known at the end of the round).

18.1.1 Basic setting

- N (slot machine) arms
- On round t, each arm i returns a loss $X_{i,t} \in [0,1]$ if selected
- Player selects an arm I_t (possibly at random) on round t, and consequently has a loss of $X_{I_t,t}$ in that round

Bandit restriction: The player only observes $X_{I_t,t}$ on round t.

18.1.2 Stochastic setting

Same as the basic setting, except the losses $X_{i,t}$ are i.i.d. with mean μ_i for all t.

Notion of regret We define the expected regret ($\mathbb{E}$-regret)/pseudo regret of making a sequence of choices $I_1, ..., I_T$:

$$\mathbb{E}\text{-regret} := \max_i \mathbb{E}\left[\sum_{t=1}^{T}(X_{I_t,t} - X_{i,t})\right]$$

Note that the expectation is taken with respect to

- the player's randomness in choosing an arm
- nature's randomness in assigning losses

18.2 A greedy algorithm for the bandit setting

Select some parameter $m \in \mathbb{Z}^+$
for $i = 1, ..., N$ **do**
  for $j = 1, ..., m$ **do**
    $t \leftarrow (i-1)N + m$
    Play $I_t = i$
    Observe the loss $X_{i,t} = Z_j$
  end for
  $\hat{\mu}_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} Z_j$
end for
for $t > Nm$ **do**
  Play $I_t = \hat{i} = \arg\min_i \hat{\mu}_i$
end for

Observation 18.1. *The algorithm has two phases*

1. *Exploration (sampling) phase ($t = 1, ..., Nm$)*
2. *Exploitation phase ($t > Nm$)*

Hoeffding's Inequality If $Z_1, ..., Z_m$ are i.i.d. with mean μ on $[0, 1]$, then

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{j=1}^{m} Z_j - \mu \right| > \epsilon \right) \leq 2e^{-2m\epsilon^2}$$

We use Hoeffding's inequality to prove the following theorem.

Theorem 18.2. *Assume there exists some $\Delta > 0$ such that $\mu_{i^*} < \mu_j - \Delta$ (the smallest mean and the second smallest mean have some positive difference) for all $j \neq i^*$. Then the expected regret of the greedy algorithm is*

$$\mathbb{E}\text{-}regret = O\left(\frac{N(\log N - \log T}{\Delta^2} \right)$$

for some appropriately chosen m.

Proof. Will be completed in the next class! □