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

Lecture 19: UCB Algorithm and Adversarial Bandit Problem

Prof. Jacob Abernethy *Scribe: Hossein Keshavarz*

Announcements

- There is no class on Wednesday, November 27
- Good job on project proposals so far!

19.1 Review on stochastic multi-armed bandit problem

Consider a gambler playing against a $N-$armed slot machine and sequentially pulls up arms to minimize the expected regret. Each arm, $i \in \{1,\ldots,N\}$ has a probability distribution $\mathcal{D}_i$ supported on the closed interval $[0,1]$. Let μ_i be the expected value of the corresponding loss to arm i and define $\tilde{\mathcal{I}} = \operatorname{argmin}_{1\le j\le N} \mu_j$ as the index of the optimal expected value among arms. Suppose that the gambler selects arm I_t at round t, then the expected regret of the gambler up to round T is defined as

$$\mathbb{E} - \text{regret} = \mathbb{E}\left(\sum_{t=1}^{T}\left(X_{I_t,t} - X_{\tilde{\mathcal{I}},t}\right)\right)$$

where $X_{i,t} \sim \mathcal{D}_i$ denotes the loss associated with choosing arm i at round t. It's worthwhile to mention that the expected value is taken over the distributions $\{\mathcal{D}_i\}_{i=1}^{N}$ and the randomness of choosing arms.

19.2 Analysis of greedy algorithm

This section is devoted to state and prove the theorem regarding the upper bound on the expected regret of the greedy algorithm. The greedy algorithm is introduced in the last lecture.

Theorem 19.1. *Suppose that there exists a positive scalar Δ such that $\mu_j - \mu_{\tilde{\mathcal{I}}} \ge \Delta$ for any $j \neq \tilde{\mathcal{I}}$. Then the expected regret of the greedy algorithm has the following upper bound*

$$\mathbb{E} - \mathit{regret} \le 1 + \frac{2N\log(2NT)}{\Delta^2}$$

Proof. Let us to decompose the expected regret as sum of two terms in which the first term is the associated regret to the sampling phase and the second term corresponds to the exploitation phase.

$$\begin{aligned}\mathbb{E}-\text{regret} &= \mathbb{E}\left(\sum_{t=1}^{mN}\left(X_{I_t,t}-X_{\tilde{\mathcal{I}},t}\right)\right)+\mathbb{E}\left(\sum_{t=1+mN}^{T}\left(X_{I_t,t}-X_{\tilde{\mathcal{I}},t}\right)\right) \overset{(a)}{\le} Nm+\mathbb{E}\left(\sum_{t=1}^{mN}\left(X_{I_t,t}-X_{\tilde{\mathcal{I}},t}\right)\right)\\ &\overset{(b)}{\le} Nm+\mathbb{E}\left(\sum_{t=1}^{mN}\mathbb{I}_{I_t\neq\tilde{\mathcal{I}}}\right)=Nm+\sum_{t=1+mN}^{T}\mathbb{P}\left(I_t\neq\tilde{\mathcal{I}}\right)\le Nm+T\max_{1+Nm\le t\le T}\mathbb{P}\left(I_t\neq\tilde{\mathcal{I}}\right)\end{aligned} \tag{19.1}$$

Note that inequalities (a) and (b) are immediate consequence of the fact that $X_{k,t}\in[0,1]$ for all $1\le k\le N$ and $1\le t\le T$. Therefore, in order to complete the proof, we need to control the deterministic term $\mathbb{P}\left(I_t\neq\tilde{\mathcal{I}}\right)$ uniformly from above. If $I_t\neq\tilde{\mathcal{I}}$ then $\hat{\mu}_{I_t}\le\hat{\mu}_{\tilde{\mathcal{I}}}$, hence using some straightforward algebraic manipulation we obtain

$$\Delta\le\mu_{I_t}-\mu_{\tilde{\mathcal{I}}}=\left(\mu_{I_t}-\hat{\mu}_{I_t}\right)+\left(\hat{\mu}_{I_t}-\hat{\mu}_{\tilde{\mathcal{I}}}\right)+(\hat{\mu}_{\tilde{\mathcal{I}}}-\mu_{\tilde{\mathcal{I}}})\le\left|\mu_{I_t}-\hat{\mu}_{I_t}\right|+0+\left|\hat{\mu}_{\tilde{\mathcal{I}}}-\mu_{\tilde{\mathcal{I}}}\right|\le 2\max_{1\le j\le N}\left|\mu_j-\hat{\mu}_j\right|$$

In other words, $\mathbb{P}\left(I_t\neq\tilde{\mathcal{I}}\right)\le\mathbb{P}\left(\max_{1\le j\le N}\left|\mu_j-\hat{\mu}_j\right|\ge\frac{\Delta}{2}\right)$ for any t. Combination of union bound (inequality (a)) and Hoeffding's inequality (used to prove inequality (b)) give us the desired upper bound.

$$\begin{aligned}\eta=\max_{1+Nm\le t\le T}\mathbb{P}\left(I_t\neq\tilde{\mathcal{I}}\right) &\le \mathbb{P}\left(\max_{1\le j\le N}\left|\mu_j-\hat{\mu}_j\right|\ge\frac{\Delta}{2}\right)\overset{(a)}{\le} N\max_{1\le j\le N}\mathbb{P}\left(\left|\mu_j-\hat{\mu}_j\right|\ge\frac{\Delta}{2}\right)\\ &\overset{(b)}{\le} 2N\exp\left(\frac{m\Delta^2}{2}\right)\end{aligned} \tag{19.2}$$

Letting the optional parameter $m=\frac{2}{\Delta^2}\log\left(\frac{2N}{\eta}\right)$ and substituting inequality (19.2) into inequality (19.1) yield

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

Choosing $\eta=\frac{1}{T}$ terminates the proof. □

Note that the reason that Δ^2 appeared in the denumerator of the expected regret is that the inequality $\mathbb{E}\left(\sum_{t=1}^{mN}\left(X_{I_t,t}-X_{\tilde{\mathcal{I}},t}\right)\right)\le Nm$ is not tight enough.

Lemma 19.2 (Hoeffding's inequality)**.** *Let $Z_1,\ldots,Z_n$ are independent and identically distributed random variables such that $Z_1\in[0,1]$ almost sure and $\mathbb{E}(Z_1)=\mu$. Then,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}Z_i-\mu\right|\ge\epsilon\right)\le 2\exp\left(-2n\epsilon^2\right)$$

19.3 Upper Confidence Bound (UCB) algorithm

In order to introduce UCB algorithm let us to introduce some notation. Define, $T_j(t) := \sum_{s=1}^{t} \mathbb{I}_{(I_s=j)}$ as the number of times that the gambler played the arm j up to time t and let the empirical expected regret of arm j as the following

$$\hat{\mu}_{j,t} = \frac{1}{\left|T_j(t)\right|} \sum_{s=1}^{t} X_{j,s} \mathbb{I}_{(I_s=j)} \tag{19.3}$$

```
Algorithm 1. UCB algorithm to solve multi armed bandit problem
Initialization Play each arm once
For $t = 1$ to $T$
    For $j = 1$ to $N$
        Update $\hat{\mu}_{j,t}$ according to the identity (19.3)
    end
    Play the arm $I_t = \operatorname{argmin}_{1\le j\le N}\left(\hat{\mu}_{j,t} - \sqrt{\frac{3\log t}{|T_j(t)|}}\right)$
end
```

The arm selection criterion of UCB algorithm encourages to choose the rarely chosen arms. The next theorem characterizes the upper bound on the expected regret of UCB algorithm. The interested reader is referred to [1] for further details and proof of Theorem 19.3.

Theorem 19.3. *Let $\Delta_j = \mu_j - \mu_{\tilde{\mathcal{I}}}$ such that $\Delta_j > 0$ for any $j \neq \tilde{\mathcal{I}}$. Then the expected regret of UCB algorithm can be upper bounded by the following inequality.*

$$\mathbb{E} - \mathit{regret} \le \mathcal{O}\left(\sum_{j\neq\tilde{\mathcal{I}}} \frac{\log T}{\Delta_j}\right)$$

Theorem 19.3 shows that the expected regret of UCB algorithm is far better than greedy method for small Δ_j's. Strictly speaking, if $\Delta_{\min} = \min_{j\neq\tilde{\mathcal{I}}} \Delta_j$, then

$$\mathbb{E} - \text{regret} \le \mathcal{O}\left(\frac{N \log T}{\Delta_{\min}}\right)$$

19.4 Adversarial Bandit

Unlike the multi armed bandit setting where the loss of each action has a stationary distribution over time, in the adversarial bandit problem there is no statistical assumption about the form of the generating process of losses. In this new formulation, the associated regret to each arm is determined at each round by an adversary and the player only knows the reward of previously chosen actions. The only assumption about the loss vector is that, $\ell^t \in [0,1]^N$ for each round t. Since the adversary can assign low reward to the previously selected actions and high rewards to the unseen arms, hence, a deterministic policy of arm selection can not optimize the expected regret function. Finally, we need to mention that the distribution of action I_t only depends on the loss of previous actions, $\left\{\ell^s_{I_s}\right\}_{s=1}^{t-1}$.

Due to the lack of full information about the associated losses of choosing arms, the player can't run exponential weighted algorithm to optimize the regret function. One doubtful solution is to estimate of loss vector, ℓ^t, based on the observation of a single component $\ell^t_{I_t}$. Taking advantage of the conditional expectation property, we can show that if $\tilde{l}^t$ is an unbiased estimator of ℓ^t, i.e. $\mathbb{E}\left(\tilde{\ell}^t|\mathcal{F}_{t-1}\right) = \ell^t$ where $\mathcal{F}_{t-1}$ is the generated $\sigma-$field by observations up to round $t-1$, then the expected regrets with respect to ℓ^t and $\tilde{\ell}^t$ are the same.

$$\mathbb{E}\left(\sum_{t=1}^{T}\tilde{\ell}^t.\left(p^t-p^*\right)\right)=\mathbb{E}\left\{\mathbb{E}\left(\sum_{t=1}^{T}\tilde{\ell}^t.\left(p^t-p^*\right)|\mathcal{F}_{t-1}\right)\right\}=\mathbb{E}\left\{\sum_{t=1}^{T}\mathbb{E}\left(\tilde{\ell}^t|\mathcal{F}_{t-1}\right).\left(p^t-p^*\right)\right\}=\mathbb{E}\left(\sum_{t=1}^{T}\ell^t.\left(p^t-p^*\right)\right)$$

19.4.1 Unbiased estimation of ℓ^t

We claim that the following procedure which is called exponential weighted algorithm with $\epsilon-$exploration generates an unbiased estimator of ℓ^t.

1. With probability ϵ, choose $p^t = \frac{1}{N}\langle 1,\ldots,1\rangle$ and select $I_t \sim p^t$ (uniformly at random). Let $\tilde{\ell}^t = \langle 0,\ldots,0,\frac{N\ell^t_{I_t}}{\epsilon},0,\ldots,0\rangle$.
2. With probability $1-\epsilon$ choose p^t by exponential weighted algorithm on the loss vectors $\{\tilde{\ell}^s\}_{s=1}^{t-1}$ and let $\tilde{\ell}^t = 0$.

proof of claim.

$$\mathbb{E}\left(\tilde{\ell}^t\right)=(1-\epsilon).0+\epsilon\sum_{j=1}^{N}\frac{1}{N}\langle 0,\ldots,0,\frac{N\ell^t_j}{\epsilon},0,\ldots,0\rangle=\ell^t$$

□

Since $\tilde{\ell}^t$ is an unbiased estimator of ℓ^t, so at the first glance, it seems that the expected regret of above algorithm is exactly equal to the expected regret of EWA. However, a contingent reader notices that $\tilde{\ell}^t$ is no longer in the closed cube $[0,1]^N$. Recalling the proof of Theorem 3.2. in the lecture notes, one can easily show that

$$\mathbb{E}\text{−regret of EWA with } \epsilon \text{ exploration} \leq T\epsilon+(1-\epsilon)\left(\frac{\log N}{\eta}+\eta\sum_{t=1}^{T}\|\tilde{\ell}^t\|_\infty^2\right)=\mathcal{O}\left(T\epsilon+\frac{\log N}{\eta}+\eta T\frac{N^2}{\epsilon^2}\right)$$

Now choosing the regularization parameters $\epsilon^2 = \frac{N}{T}\sqrt{2T\log N}$ and $\eta = \frac{\epsilon}{N}\sqrt{\frac{2\log N}{T}}$, the optimal upper bound is given by

$$\mathbb{E}-\text{regret of EWA with } \epsilon \text{ exploration} \leq \mathcal{O}\left(\sqrt{N}T^{\frac{3}{4}}(\log N)^{\frac{1}{4}}\right) \tag{19.4}$$

It's worthwhile to mention that although $\max_{1\le t\le T} \| \tilde{\ell}^t \|_\infty \le \frac{N}{\epsilon}$, but most of the times (with probability 1ϵ), we have $\tilde{\ell}^t = 0 \in [0,1]$. Hence, the proof can be slightly modified in a smart way to obtain the following upper bound.

$$\mathbb{E} - \text{regret of EWA with } \epsilon \text{ exploration} \le \mathcal{O}\left(T\epsilon + \frac{\log N}{\eta} + \eta T \frac{N^2}{\epsilon}\right)$$

which leads to $\mathbb{E} - \text{regret of EWA with } \epsilon \text{ exploration} \le \mathcal{O}\left(\sqrt{N} T^{\frac{2}{3}}\right)$.

Question: Is there any algorithm with the expected regret $\mathcal{O}\left(\sqrt{NT}\right)$? Yes! EXP3 algorithm

19.5 EXP3 Algorithm

Let $\tilde{L}^t$ be the cumulative loss up to round t.

Algorithm 2. EXP3 algorithm to for adversarial multi-armed bandit problem

```
Input Regularization parameter $\eta$
Initialization St initial value for $p^1$
For $t = 1$ to $T$
  For $j = 1$ to $N$
    Sample $I_t$ absed on the distribution $p^t$
    Observe $\ell^t_{I_t}$
    Let $\tilde{\ell}^t = \langle 0, \ldots, 0, \frac{\ell^t_{I_t}}{p^t_{I_t}}, 0, \ldots, 0 \rangle$
    Update $p^{t+1}$ by $p_j^{t+1} = \frac{\exp\left(-\eta \tilde{L}_j^t\right)}{\sum_{j=1}^{N} \exp\left(-\eta \tilde{L}_j^t\right)}$
end
```

The EXP3 algorithm will be analyzed in the next class.

References

[1] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time analysis of the multi-armed bandit problem", *Machine learning* 47, no. 2-3 (2002): 235-256.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem", *In Foundations of Computer Science*, 1995. Proceedings., 36th Annual Symposium on, pp. 322-331. IEEE, 1995.