CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 7: Multi-Armed Bandits (April 19)

Lecturer: Yu-Xiang Wang *Scribes: Ari Polakof*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

7.1 Multi-arm bandits: Problem Setup

- No state or equivalently there's only one state and k-actions $a \in A = \{1, 2, ..., k\}$
- Decide which arm to pull in every iteration, where we can think of the horizon to be 1
- Get reward $\sum_{t=1}^{T} R_t$
- $E[R_t|A_t = a] = \mu_a$ and $R_t = \mu_a + Noise$, where $E[Noise] = 0$
- Define regret as $T \max_{a\in[k]} \mathbb{E}[R_t|a] - \sum_{t=1}^{t=T} \mathbb{E}_{a\sim\pi}[\mathbb{E}[R_t|a]]$
- No regret means sublinear scaling in T.

$$\lim_{T\to\infty} \frac{1}{T} Regret_T = 0$$

- The regret (upper) bound needs to apply to all problem instances

7.1.1 Exploration first

- Spend first N steps exploring, picking each action $\frac{N}{k}$ times, where k is the number of actions.
- Define

$$\hat{Q}_t(a) = \frac{\sum_{t=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- For t = N+1, N+2, ..., T:

$$A_t = \operatorname*{argmax}_a Q_t(a)$$

Recall the concentration inequalities:

Hoeffding's inequality: Assume $X_1, ..., X_m$ are independent and $P(a_i \le x_i \le b_i) = 1$

$$S_n = X_1 + ... + X_n$$

$$P(S_n - \mathbb{E}[S_n] \ge t) \le e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

Easier version, if $0 < X_i < B$, with probability $1 - \delta$

$$|\overline{X} - \mathbb{E}[\overline{X}]| \leq \sqrt{\frac{B^2 log(2/\delta)}{2n}}$$

Regret analysis of Exploration First

Since we take each action $\frac{N}{K}$ times, by Hoeffding's, with probability $\geq 1 - \frac{\delta}{k}$

$$|\hat{Q}(a) - Q(a) \leq \sqrt{\frac{klog(2k/\delta)}{2N}}$$

for all $a \in A$, using union bound

$$\sup_{a \in A} |\hat{Q}(a) - Q(a)| \leq \sqrt{\frac{K}{2N} log \frac{2k}{\delta}} = \epsilon$$

Regret for Exploration Phase:

$$\frac{N}{K} \sum_a \max_{a'} Q(a') - Q(a) \leq N$$

since $0 \leq Q(a) \leq 1$

Regret for Exploitation Phase:

Define $\hat{a}^* = \text{argmax}_a \hat{Q}(a)$

$$(T - N)(Q(a^*) - Q(\hat{a^*})$$

$$= (T - N)[Q(a^*) - \hat{Q}(a^*) + \hat{Q}(a^*) - \hat{Q}(\hat{a^*}) + \hat{Q}(\hat{a^*}) - Q(\hat{a^*})]$$

$$\leq (T - N) \cdot 2\epsilon$$

since $Q(a^*) - \hat{Q}(a^*) \leq \epsilon$, $\hat{Q}(a^*) - \hat{Q}(\hat{a^*}) \leq 0$, $\hat{Q}(\hat{a^*}) - Q(\hat{a^*}) \leq \epsilon$

$$\leq 2T \sqrt{\frac{K}{2N} \log \frac{2K}{\delta}}$$

Total Regret:

$$Regret = N + 2T \sqrt{\frac{K}{2N} \log \frac{2k}{\delta}} = O(T^{\frac{2}{3}} K^{\frac{1}{3}} (\log \frac{2k}{\delta})^{\frac{1}{3}}))$$

where we chose $N = T^{\frac{2}{3}} k^{\frac{1}{3}} (\log \frac{2k}{\delta})^{\frac{1}{3}}$

7.1.2 ϵ-greedy strategy

- With probability 1-ϵ choose

$$A_t = \underset{a}{\text{argmax}}\, Q_t(a)$$

- With probability ϵ choose an action uniformly at random.

Sketch of regret analysis for ϵ greedy:

- In expectation, each arm is chosen for at least ϵt times: By Hoeffding's, at time t:

$$N_t(a) \geq \frac{\epsilon t}{k} - O(\sqrt{\frac{k}{t}}) \geq \frac{\epsilon t}{2k}$$

- Condition on the number of times, and then apply Hoeffding's inequality/union bound for all t and a

$$\sup_a |\hat{Q}_t(a) - Q(a)| \leq O(\sqrt{\frac{k}{\epsilon t}})$$

- The regret bound is then:

$$\epsilon T + \sum_{t=1}^{T} C\sqrt{\frac{k}{\epsilon t}}$$

  where the first term comes from the exploration part and the second from the exploitation part. Note that we can bound the second term by observing that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}}$ is less than $\int_1^T \frac{1}{\sqrt{x}} dx = 2\sqrt{t} - 2$

7.1.3 Upper Confidence Bound algorithm

- Play each action $a \in A$ once. Given that we have k actions this corresponds to k steps
- for t = k+1, ..., T

$$A_t = \underset{a}{\text{argmax}}\, \hat{Q}_t(a) + \sqrt{\frac{\log(\frac{2TK}{\delta})}{2N_t(a)}}$$

  where

$$N_t(a) = \sum_{t=1}^{t-1} \mathbb{1}_{A_t=a}$$

$$\hat{Q}_t(a) = \frac{1}{N_t(a)}(R_a + \sum_{i=k+1}^{t-1} \mathbb{1}_{A_i=a} R_i)$$

Introduce Martingale

- A sequence of random variables $X_1, ..., X_n$ is a Martingale if for any n

$$\mathbb{E}[|X_n|] < \infty$$
$$\mathbb{E}[X_{n+1}|X_1, ..., X_n] = X_n$$

Introduce Azuma-Hoeffding's inequality

- Assume $X_1, ..., X_n$ are Martingale differences, then S_n is Martingale, where

$$S_n = X_1 + ... + X_n$$
$$\mathbb{P}[S_n \geq \epsilon] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

Regret analysis of UCB

Recall that we want to bound

$$\hat{Q}_t(a) = \frac{1}{N_t(a)}(R_a + \sum_{i=k+1}^{t-1} \mathbb{1}_{A_i=a} R_i)$$

Let

$$S_t = (R_a + \sum_{i=k+1}^{t-1} \mathbb{1}_{A_i=a} R_i)$$

Subtract the mean to make it zero mean

$$R_a - \mu_a + \sum_{i=k+1}^{t-1} \mathbb{1}_{A_i=a} R_i - \mathbb{E}[\mathbb{1}(A_i = a) R_i | History_{i-1}]$$

We know from UCB that $\mathbb{1}(A_i = a)$ is fixed

Let $X_i = \mathbb{1}(A_i = a) R_i$ conditioned on $X_1 ... X_{i-1}$

For those i where $A_i = a$ we set $b_i = 1$ $a_i = 0$.

$$R_a - Q(a) + \sum_{i=k+1}^{t-1} \mathbb{1}(A_i = a)(R_i - Q(a)$$

is martingale and so with probability $1 - \frac{\delta}{kT}$

$$|R_a - Q(a) + \sum_{i=k+1}^{t-1} \mathbb{1}(A_i = a)(R_i - Q(a)| \leq \sqrt{2N_t(a) \log \frac{kt}{\delta}}$$

Take union bound over all $a \in A$, all t, $k + 1 \leq t \leq T$, with probability $1 - \delta$

$$\sup_{t,a} \frac{1}{N_t(a)} |R_a - Q(a) + \sum_{i=k+1}^{t-1} \mathbb{1}(A_i = a)(R_i - Q(a)| \leq \sqrt{2 \log \frac{kT}{\delta}}$$

Let us define the UCB

$$\overline{Q}_t(a) = \hat{Q}_t(a) + \sqrt{\frac{2 \log \frac{kT}{\delta}}{N_t(a)}}$$

$$Q(a*) - Q(A_t) = Q(a^*) - \overline{Q}_t(a^*) + \overline{Q}_t(a*) - \overline{Q}(A_t) + \overline{Q}(A_t) - Q(A_t)$$

where the first term ≤ 0, the second ≤ 0 by UCB, and the last term, by concentration, $\leq 2 \cdot \epsilon$

Define

$$\Delta_a = Q(a^*) - Q(A_t)$$

$$
\begin{aligned}
Regret &= \sum_{a=1}^{k} \Delta_a + \sum_{t=k+1}^{T} Q(a^*) - Q(A_t) \\
&\leq K + \sum_{t=k+1}^{T} 2\sqrt{\frac{2\log\frac{2Tk}{\delta}}{N_t(A_t)}} \\
&= K + 2\sqrt{2\log\frac{2Tk}{\delta}} \sum_{a=1}^{k} \sum_{i=1}^{N_t(a)} \frac{1}{\sqrt{i}} \\
&\leq K + 4\sqrt{2\log\frac{2Tk}{\delta}} \sum_{a=N}^{k} \sqrt{N_t(a)} \\
&\leq k + 4\sqrt{2\log\frac{2Tk}{\delta}}\sqrt{KT}
\end{aligned}
$$

by Cauchy-Schwarz

$$= K + c\sqrt{KT\log\frac{2Tk}{\delta}}$$

Gap dependent analysis to obtain a tighter bound:

Claim: $N_t(a)\frac{\leq 2\sqrt{2}\log\frac{2Tk}{\delta}}{\Delta_a^2}$

Substitute above to get bound.

7.1.4 Summary of Regret Bounds in Multi-Armed Bandits

Let $\widetilde{O}$ hide constant log factors.

- Explore-First

$$\widetilde{O}(T^{\frac{2}{3}}k^{\frac{1}{3}})$$

- Epsilon greedy

$$\widetilde{O}(T^{\frac{2}{3}}k^{\frac{1}{3}})$$

- UCB

$$\widetilde{O}(\sqrt{TK})$$

7.2 Linear bandits: Multi-Armed Bandits with an infinite number of actions

- Each action is determined by a feature vector
- Action space is a compact set $A \subset \mathbb{R}^d$
- Reward is linear with noise: $R_t = \langle A_t, \mu_* \rangle + \eta_t$, where η_t independent and σ^2 subgaussian.
- Agent chooses a sequence of actions $A_1 ... A_T$
- Regret is defined as:

$$Regret_T = T \cdot \langle a^*, u_* \rangle - \sum_{t=1}^{T} \langle A_t, u_* \rangle$$

Note that in the textbook the notation is different: $A_i = D$, $a = x$

7.2.1 The LinUCB algorithm: Optimism in the Face of Uncertainty

- Consider the ridge regression at each time t

$$\hat{\mu}_t = \operatorname*{argmin}_{\mu \in W} \sum_{i=1}^{t-1} (r_i - \mu^T x_i)^2 + \lambda \|\mu\|^2$$

  Note that there is a closed form solution $\hat{\mu}_t = \Sigma_t^{-1} \sum_{i=1}^{t-1} x_i r_i$ where Σ_t is defined below

- Construct high probability confidence set of the parameter vector

$$Ball_t = \{\mu | (\mu - \hat{\mu}_t)^T \Sigma_t (\mu - \hat{\mu}_T) \leq B_t\}$$

  where $\Sigma_t = \sum_{i=1}^{t-1} x_i x_i^T + \lambda I_d$

- Choose actions that maximize the UCB

$$x_t = \operatorname*{argmax}_{x \in D} \max_{\mu \in Ball_t} \langle x, \mu \rangle$$