DDA4230 Reinforcement learning Greedy algorithms and ETC II

Lecture 6

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

To analyze the regret of greedy algorithms and ETC.
Suggested reading: Chapter 6 of *Bandit algorithms*;

2 Greedy algorithms and ETC

2.1 The greedy algorithm

Algorithm 1: The greedy algorithm

Output: $\pi(t), t \in \{0, 1, \ldots, T\}$
while $0 \leq t \leq m-1$ **do**

$$\pi(t) = t + 1$$

while $m \leq t \leq T$ **do**

$$\pi(t) = \arg\max_{i \in [m]} \left\{ \frac{1}{N_{t-1,i}} \sum_{t'=0}^{t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} \right\}$$

The worst-case regret of the greedy algorithm is $O(T)$.

2.2 The ε-greedy algorithm

If $\varepsilon_t > \varepsilon$ holds for some constant $\varepsilon > 0$, then the regret of the ε-greedy algorithm is $O(T)$.

By carefully choosing $\varepsilon_t = O(1/t)$, we can obtain an algorithm with its regret at most $O(\log T)$.

Theorem 1 *Assume that $r(i)$ is 1-subgaussian for each i. By choose $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2} m\}$ for some sufficiently large absolute constant C, the regret under the ε-greedy algorithm satisfies*

$$\overline{R}_T \leq C' \sum_{i \geq 2} \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max \left\{ e, \frac{T\Delta_{\min}^2}{m} \right\} \right), \tag{1}$$

where C' is an absolute constant.

Algorithm 2: The ε-greedy algorithm

Input: $\varepsilon_t, t \in \{0, 1, \ldots, T\}$ the exploration parameters
Output: $\pi(t), t \in \{0, 1, \ldots, T\}$
while $0 \leq t \leq m-1$ **do**

$$\pi(t) = t + 1$$

while $m \leq t \leq T$ **do**

$$\pi(t) \sim \begin{cases} \arg\max_{i \in [m]} \left\{ \frac{1}{N_{t-1,i}} \sum_{t'=0}^{t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} \right\} & \text{with probability } 1 - \varepsilon_t \\ i \quad \text{with probability } \varepsilon_t/m, \text{ for each } i \in [m] \end{cases}$$

Proof: Let $x = \lfloor \frac{1}{2m} \sum_{t'=1}^{t} \varepsilon_{t'} \rfloor$.

For an suboptimal arm i, at time t,

$$\begin{aligned} \mathbb{P}(a_t = i) &\leq \frac{\varepsilon_t}{m} + (1 - \varepsilon_t)\mathbb{P}(\hat{\mu}_{t,i} \geq \hat{\mu}_{t,1}) \\ &\leq \frac{\varepsilon_t}{m} + (1 - \varepsilon_t)(\mathbb{P}(\hat{\mu}_{t,i} \geq \mu_i + \frac{\Delta_i}{2}) + \mathbb{P}(\hat{\mu}_{t,1} \leq \mu_1 - \frac{\Delta_i}{2})) \end{aligned}$$

We then desire to bound $\mathbb{P}(\hat{\mu}_{t,i} \geq \mu_i + \frac{\Delta_i}{2})$ and $\mathbb{P}(\hat{\mu}_{1,i} \leq \mu_i - \frac{\Delta_i}{2})$. Let $\eta_{t',i}$ to be the empirical mean of arm i after t' pulls and $\mathrm{NR}_{t,i}$ to be the number of pulls of arm i caused by random exploration up to time t.

$$\begin{aligned} \mathbb{P}(\hat{\mu}_{t,i} \geq \mu_i + \frac{\Delta_i}{2}) &= \sum_{t'=0}^{t} \mathbb{P}(N_{t,i} = t', \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) \\ &= \sum_{t'=0}^{t} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2})\mathbb{P}(\hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) \\ &\leq \sum_{t'=0}^{t} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) \exp(-\Delta_i^2 t'/2) \\ &= \sum_{t'=0}^{x} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) \exp(-\Delta_i^2 t'/2) \\ &\quad + \sum_{t'=x+1}^{\infty} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) \exp(-\Delta_i^2 t'/2) \\ &\leq \sum_{t'=0}^{x} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \sum_{t'=x+1}^{\infty} \exp(-\Delta_i^2 t'/2) \\ &\leq \sum_{t'=0}^{x} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \frac{2}{\Delta_i^2} \exp(-\Delta_i^2 x/2) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t'=0}^{x} \mathbb{P}(\mathrm{NR}_{t,i} \leq t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \frac{2}{\Delta_i^2} \exp(-\Delta_i^2 x/2) \\
&\leq \sum_{t'=0}^{x} \mathbb{P}(\mathrm{NR}_{t,i} \leq t') + \frac{2}{\Delta_i^2} \exp(-\Delta_i^2 x/2) \\
&\leq (x+1)\mathbb{P}(\mathrm{NR}_{t,i} \leq x) + \frac{2}{\Delta_i^2} \exp(-\Delta_i^2 x/2) \\
&\leq (x+1) \exp(-x/5) + \frac{2}{\Delta_i^2} \exp(-\Delta_i^2 x/2).
\end{aligned}
$$

By the choice of ε_t, we have $x \geq \frac{C}{\Delta_i^2} \log \frac{t\Delta_i^2\sqrt{e}}{Cm}$, which upper bounds the probability of pulling arm i by $O(\log t)/t^{(1+\varepsilon)}$ at time t for some ε. We then have $\sum_t \frac{\varepsilon_t}{m} + (1-\varepsilon_t)(\mathbb{P}(\hat{\mu}_{t,i} \geq \mu_i + \frac{\Delta_i}{2}) + \mathbb{P}(\hat{\mu}_{t,1} \leq \mu_1 - \frac{\Delta_i}{2})) = O(\log T)$, as desired. □

2.3 Explore-then-commit algorithms

Algorithm 3: The explore-then-commit algorithm

Input: k: number of exploration on each arm
Output: $\pi(t), t \in \{0, 1, \ldots, T\}$
while $0 \leq t \leq km - 1$ **do**

$$a_t = (t \bmod m) + 1$$

while $km \leq t \leq T - 1$ **do**

$$a_t = \arg\max_{i \in [m]} \frac{1}{k} \sum_{t'=0}^{mk} r_{t'} \mathbb{1}\{a_{t'} = i\}$$

Theorem 2 *Assume that $r(i)$ is 1-subgaussian for each i. The regret under ETC satisfies*

$$\overline{R}_T \leq k \sum_{i \in [m]} \Delta_i + (T - mk) \sum_{i \in [m]} \Delta_i e^{-k\Delta_i^2/4}. \tag{2}$$

Particularly, for two-armed bandits ($m = 2$), taking $k = \lceil \max\left\{1, 4\Delta_2^{-2} \log(T\Delta_2^2/4)\right\} \rceil$ yields

$$\overline{R}_T \leq \Delta_2 + (4 + e^{-2})\sqrt{T}. \tag{3}$$

We refer the proof to Section 6 and Exercise 6.1 of *Bandit algorithms.*

In fact, if the rewards are Gaussian with variance 1, the gap-dependent regret bound under $m = 2$ can be further improved by a more careful choice of k. Denote $\Delta = \Delta_2$ and the π below denotes the Archimedes' constant instead of a policy.

Theorem 3 *Assume that $r(i)$ is 1-subgaussian for each i and $T \geq 4\sqrt{2\pi e}/\Delta^2$. By choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret of ETC satisfies*

$$O(\frac{1}{\Delta}\log T\Delta^2) + o(\log T) + \Delta, \tag{4}$$

where $W(y)\exp(W(y)) = y$ denotes the Lambert function.

Proof: Let $A = r_0 - r_1 + r_2 - \cdots - r_{2k-1}$. The regret is composed of a deterministic exploration regret of $k\Delta$ and a regret $(T-2k)\Delta$ of exploitation which happens when $A \leq 0$. As $A \sim N(k\Delta, 2k)$,

$$\begin{aligned}
\overline{R}_T &= \Delta(k + (T-2k)\mathbb{P}(A \leq 0)) \\
&\leq \Delta(k + T\mathbb{P}(N(0,1) \leq -\Delta\sqrt{\frac{k}{2}})) \\
&\leq \Delta(\frac{2}{\Delta^2}W(\frac{T^2\Delta^4}{32\pi}) + 1 + T\mathbb{P}(N(0,1) \leq -\sqrt{W(\frac{T^2\Delta^4}{32\pi})})) \\
&\leq \Delta(\frac{2}{\Delta^2}W(\frac{T^2\Delta^4}{32\pi}) + 1 + T\frac{\frac{1}{\sqrt{2\pi}}\exp(-W(\frac{T^2\Delta^4}{32\pi}))}{\sqrt{W(\frac{T^2\Delta^4}{32\pi})}}) \\
&= \Delta(\frac{2}{\Delta^2}W(\frac{T^2\Delta^4}{32\pi}) + 1 + \frac{4}{\Delta^2}) \\
&\leq \Delta(\frac{2}{\Delta^2}(\log\frac{T^2\Delta^4}{32\pi} - \log\log\frac{T^2\Delta^4}{32\pi} + \log(1+\frac{1}{e})) + 1 + \frac{4}{\Delta^2}),
\end{aligned}$$

which achieves the desired order of bound. □

The choice of k is determined by minimizing $(k + T\mathbb{P}(N(0,1) \leq -\Delta\sqrt{\frac{k}{2}})$. Taking derivative with respect to k, we have

$$T\Delta\frac{1}{\sqrt{8k}}\frac{1}{\sqrt{2\pi}}\exp(-\Delta^2\frac{k}{4}) = 1$$

or equivalently $k\frac{\Delta^2}{2}\exp(k\frac{\Delta^2}{2}) = \frac{T^2\Delta^4}{32\pi}$, which hints us about the optimum $k_* = \frac{2}{\Delta^2}W(\frac{T^2\Delta^4}{32\pi})$ up to its rounding.

Acknowledgement

This lecture notes partially use material from *Reinforcement learning: An introduction*, and *Bandit algorithms*. For the proofs, we also referred to *On explore-then-commit strategies* by Garivier, Kaufmann, and Lattimore and *Finite-time analysis of the multiarmed bandit problem* by Auer, Cesa-bianchi, and Fischer.