DDA4230 Reinforcement learning Greedy algorithms

Lecture 4

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

To understand the greedy and ε-greedy algorithms and complete our first analysis with a logarithmic regret.
Suggested reading: Chapter 6 of *Bandit algorithms*;

2 Recap: Multi-armed bandits

The problem of multi-armed bandits is a special case of the MDP we defined

- $\mathcal{A} = [m] = \{1, 2, \ldots, m\}$;
- $\mathcal{R}(s, a) = r(a)$ some unknown stochastic function $r(\cdot)$;
- The horizon T is finite.

The policy is aware of the problem structure but has no prior knowledge of the reward function. We mainly consider the asymptotic worst-case performance, namely, if the algorithm achieves a regret of either $O(\log T)$, $O(T)$, or some other orders.

The policy $\pi(\cdot, t)$ is a mapping from the historical information and the current time t to an action. We define the regret as

$$\overline{R}_t = \sum_{i=1}^{m} \mathbb{E}[N_{t,i}]\Delta_i,$$

where $N_{t,i} = \sum_{t'=0}^{t} \mathbb{1}\{a_{t'} = i\}$ and $\Delta_i = \mu^* - \mu_i$. Denote $\Delta_{\min} = \min_{i:\Delta_i>0} \Delta_i$ as the minimum non-zero gap between an arm and an optimal arm.

In the following analysis, without loss of generality we assume that arm 1 is optimal. From the form $\overline{R}_t$ written above, it is intuitive to bound $N_{t,i}$ for each i to analyze the performance of a given policy.

3 Greedy algorithms

3.1 The greedy algorithm

The idea of the greedy algorithm is to pull each arm once and then always pull the arm with the best empirical mean reward. This algorithm focuses purely on exploitation and does not consider exploration. The algorithm is described in the below chart.[1]

[1] When the context is clear we write $x = \arg\max\{\cdot\}$ to denote $x \in \arg\max\{\cdot\}$ with an arbitrary tiebreaker.

```
Algorithm 1: The greedy algorithm
  Output: $\pi(t), t \in \{0, 1, \ldots, T\}$
  while $0 \le t \le m-1$ do
      $$\pi(t) = t + 1$$
  while $m \le t \le T$ 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 $O(T)$. In fact, any algorithm achieves a regret at most $O(T)$. It suffices to showing that the greedy algorithm obtains this regret in some bandit instances. Consider a two-armed bandit instance where $r(1)$ and $r(2)$ follow Bernoulli distributions with mean p and q respectively (assume $p > q$, without loss of generality), then $\mathbb{P}(r_1 = 0, r_2 = 1) = q(1-p)$. When this event is true, the algorithm will pull arm 2 for the rest of the horizon, which induces a regret of at least $q(1-p)\Delta_2 T + o(T)$.

The regret order of $O(T)$ holds even when the algorithm explores each arm for k times in the beginning for a constant k.

3.2 The ε-greedy algorithm

The ε-greedy algorithm is a variant of the greedy algorithm, which is built upon the philosophy of *being optimistic is good.* The algorithm is derived to include exploration in the algorithm. The ε-greedy algorithm takes a non-deterministic policy that forces exploration on arms which look sub-optimal. The details are given below.

```
Algorithm 2: The $\varepsilon$-greedy algorithm
  Input: $\varepsilon_t, t \in \{0, 1, \ldots, T\}$ the exploration parameters
  Output: $\pi(t), t \in \{0, 1, \ldots, T\}$
  while $0 \le t \le m-1$ do
      $$\pi(t) = t + 1$$
  while $m \le t \le T$ do
      $$\pi(t) \sim \begin{cases} \arg\max_{i \in [m]} \left\{ \frac{1}{N_{t-1,i}} \sum_{t'=0}^{t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} \right\} & \text{with probability } 1 - \varepsilon_t \\ i \quad \text{with probability } \varepsilon_t/m, & \text{for each } i \in [m] \end{cases}$$
```

The algorithm amounts to the choice of the exploration parameters ε_t.

We first establish a negative result when ε_t does not diminish with t. In fact, if $\varepsilon_t > \varepsilon$ holds for some constant $\varepsilon > 0$, then for $T - m$ rounds, the algorithm has a probability at least ε to pull a random arm. As pulling a random arm induces an expected regret of $\frac{1}{m}(\Delta_2 + \cdots + \Delta_m)$ per step, the regret of the algorithm is at least

$$\overline{R}_t \geq \frac{1}{m}(\Delta_2 + \cdots + \Delta_m)\varepsilon(T - m).$$

Again, a regret in order $O(T)$ is the worst possible regret and is not desired.

By carefully choosing ε_t as a decreasing function of t, we can obtain an algorithm with its regret at most $O(\log T)$.

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

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

where C' is an absolute constant.

The proof of the theorem is two-fold. First, the cost of exploration, being $\overline{R}_t = \frac{1}{m}(\Delta_2 + \cdots + \Delta_m)\varepsilon$ for $\varepsilon_t = O(1)$, reduces to $\overline{R}_t = \frac{1}{m}(\Delta_2 + \cdots + \Delta_m)O(1 + \frac{1}{2} + \cdots + \frac{1}{T}) = \frac{1}{m}(\Delta_2 + \cdots + \Delta_m)O(\log T)$ with the annealing of ε_t. Second, we show that the probability of pulling a suboptimal arm in a round after $\log T$ explorations is very thin (as thin as at most $O(\log T/T)$). This can be done by showing that the empirical mean of a suboptimal gap has a small enough probability to deviate by at least Δ_i, compared to the empirical mean of the optimal arm established by at least $\log T$ pulls on each arm.

ε-greedy, with Theorem 1, is the first algorithm we introduce to obtain a logarithmic regret. Despite this, the choice for ε requires information on the gap of suboptimality. This is called gap-dependent (also known as problem-dependent, instance-dependent, and distribution-dependent) algorithms in bandits. When prior knowledge on such a gap is not available, one will have to pull each arm for a few times to get an estimation of this gap and plug in the estimation (known as bootstrap). This can cause the performance of the algorithm to be uncertain in practice.

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

For a suboptimal arm i, at time t,

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

Now we investigate the first term, ε_t/m, in the upper bound of $\mathbb{P}(a_t = i)$. Recall that

$$\varepsilon_t = \begin{cases} 1, & \text{if } t \leq \frac{Cm}{\Delta_{\min}^2} \\ \frac{Cm}{t\Delta_{\min}^2}, & \text{if } t > \frac{Cm}{\Delta_{\min}^2}\,. \end{cases}$$

We have

$$\begin{aligned}\sum_{t=1}^{T}\frac{\varepsilon_t}{m} &= \sum_{t=1}^{\lfloor \frac{Cm}{\Delta_{\min}^2}\rfloor}\frac{1}{m} + \frac{C}{\Delta_{\min}^2}\cdot\frac{1}{m}\sum_{t=\lfloor \frac{Cm}{\Delta_{\min}^2}\rfloor+1}^{T}\frac{1}{t}\\ &\le \frac{Cm}{\Delta_{\min}^2}\cdot\frac{1}{m} + \frac{Cm}{\Delta_{\min}^2}\cdot\frac{1}{m}\int_{\frac{Cm}{\Delta_{\min}^2}}^{T}\frac{1}{t}dt\\ &= \frac{C}{\Delta_{\min}^2}(1+\log\frac{T\Delta_{\min}^2}{Cm})\,.\end{aligned}$$

Similarly,

$$2x_t \le \frac{1}{m}\sum_{t'=1}^{t}\varepsilon_{t'} \le \frac{C}{\Delta_{\min}^2}(1+\log\frac{t\Delta_{\min}^2}{Cm})\,.$$

A casual bound of the harmonic number states that $\log t + \frac{1}{2} < \sum_{t'=1}^{t}\frac{1}{t'} < \log t + 1$. As such,

$$\begin{aligned}2x_t &\ge \frac{1}{m}\sum_{t'=1}^{t}\varepsilon_{t'} - 1\\ &\ge \sum_{t=1}^{\lfloor \frac{Cm}{\Delta_{\min}^2}\rfloor}\frac{1}{m} + \frac{Cm}{\Delta_{\min}^2}\cdot\frac{1}{m}(\sum_{t=1}^{t}\frac{1}{t} - \sum_{t=1}^{\lfloor \frac{Cm}{\Delta_{\min}^2}\rfloor}\frac{1}{t}) - 1\\ &\ge (\frac{Cm}{\Delta_{\min}^2} - 1)\frac{1}{m} + \frac{Cm}{\Delta_{\min}^2}\cdot\frac{1}{m}(\log t + \frac{1}{2} - (\log\frac{Cm}{\Delta_{\min}^2} + 1)) - 1\\ &\ge \frac{C}{\Delta_{\min}^2}(1+\log\frac{t\Delta_{min}^2}{Cme^{\frac{1}{2}}}) - (1+\frac{1}{m})\,.\end{aligned}$$

We then desire to bound $\mathbb{P}(\hat{\mu}_{t,i} \ge \mu_i + \frac{\Delta_i}{2})$ and $\mathbb{P}(\hat{\mu}_{1,i} \le \mu_i - \frac{\Delta_i}{2})$. 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.

$$\begin{aligned}\mathbb{P}(\hat{\mu}_{t,i} \ge \mu_i + \frac{\Delta_i}{2}) &= \sum_{t'=0}^{t}\mathbb{P}(N_{t,i} = t', \hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\\ &= \sum_{t'=0}^{t}\mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\mathbb{P}(\hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\\ &\le \sum_{t'=0}^{t}\mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\exp(-\Delta_i^2 t'/8)\\ &= \sum_{t'=0}^{x_t}\mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\exp(-\Delta_i^2 t'/8)\\ &\quad + \sum_{t'=x_t+1}^{T}\mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \ge \mu_i + \frac{\Delta_i}{2})\exp(-\Delta_i^2 t'/8)\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{t'=0}^{x_t} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \sum_{t'=x_t+1}^{\infty} \exp(-\Delta_i^2 t'/8) \\
&\leq \sum_{t'=0}^{x_t} \mathbb{P}(N_{t,i} = t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \frac{8}{\Delta_i^2} \exp(-\Delta_i^2 x_t/8) \\
&\leq \sum_{t'=0}^{x_t} \mathbb{P}(\mathrm{NR}_{t,i} \leq t' \mid \hat{\eta}_{t',i} \geq \mu_i + \frac{\Delta_i}{2}) + \frac{8}{\Delta_i^2} \exp(-\Delta_i^2 x_t/8) \\
&\leq \sum_{t'=0}^{x_t} \mathbb{P}(\mathrm{NR}_{t,i} \leq t') + \frac{8}{\Delta_i^2} \exp(-\Delta_i^2 x_t/8) \\
&\leq (x_t + 1)\mathbb{P}(\mathrm{NR}_{t,i} \leq x_t) + \frac{8}{\Delta_i^2} \exp(-\Delta_i^2 x_t/8)\,.
\end{aligned}$$

The random variable $\mathrm{NR}_{t,i}$ is the summation of $\mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_t\}$, where the event $A_{t'}$ denotes a Bernoulli variable that at time t' arm i is pulled by random exploration, which happens with probability $\varepsilon_{t'}/m$ independent of other $A_{t'}$ events. The total variance $\sum_{i=1}^{t} \mathbb{V}[\mathbb{1}\{A_i\}] \leq \sum_{t'=1}^{t} \frac{\varepsilon_{t'}}{m}(1 - \frac{\varepsilon_{t'}}{m})$. By Bernstein's inequality

$$\begin{aligned}
\mathbb{P}(\frac{1}{t}\mathrm{NR}_{t,i} \geq \frac{1}{t}\mathbb{E}[\mathrm{NR}_{t,i}] - z) \geq& 1 - \exp(-\frac{t^2 z^2}{2\sum_{t'=1}^{t} \frac{\varepsilon_{t'}}{m}(1 - \frac{\varepsilon_{t'}}{m}) + tz}) \\
\geq& 1 - \exp(-\frac{t^2 z^2}{2\sum_{t'=1}^{t} \frac{\varepsilon_{t'}}{m} + tz}) \\
=& 1 - \exp(-\frac{t^2 z^2}{4\hat{x}_t + tz})\,.
\end{aligned}$$

As such,

$$\begin{aligned}
\mathbb{P}(\mathrm{NR}_{t,i} \leq x_t) &\leq \mathbb{P}(\mathrm{NR}_{t,i} \leq \hat{x}_t) \\
&= \mathbb{P}(\mathrm{NR}_{t,i} - \mathbb{E}[\mathrm{NR}_{t,i}] \leq \hat{x}_t - \mathbb{E}[\mathrm{NR}_{t,i}]) \\
&= \mathbb{P}(\mathrm{NR}_{t,i} - \mathbb{E}[\mathrm{NR}_{t,i}] \leq \hat{x}_t - \frac{1}{m}\sum_{t'=1}^{t} \varepsilon_{t'}) \\
&\leq \mathbb{P}(\mathrm{NR}_{t,i} - \mathbb{E}[\mathrm{NR}_{t,i}] \leq -\hat{x}_t) \\
&\leq \mathbb{P}(\frac{1}{t}\mathrm{NR}_{t,i} - \frac{1}{t}\mathbb{E}[\mathrm{NR}_{t,i}] \leq -\frac{\hat{x}_t}{t}) \\
&\leq \exp(-\frac{t^2(\frac{\hat{x}_t}{t})^2}{4\hat{x}_t + t(\frac{\hat{x}_t}{t})}) \\
&\leq \exp(-\frac{\hat{x}_t}{5}) \leq \exp(-\frac{x_t}{5})\,.
\end{aligned}$$

Therefore,

$$\mathbb{P}(\hat{\mu}_{t,i} \geq \mu_i + \frac{\Delta_i}{2}) \leq (x_t + 1)\exp(-\frac{x_t}{5}) + \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 x_t}{8})\,,$$

and by the same arguments,

$$\mathbb{P}(\hat{\mu}_{t,1} \leq \mu_1 - \frac{\Delta_i}{2}) \leq (x_t + 1)\exp(-\frac{x_t}{5}) + \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 x_t}{8})\,.$$

The regret

$$
\begin{aligned}
\overline{R}_T &= \sum_{i=1}^{m} \mathbb{E}[N_{T,i}]\Delta_i \\
&= \sum_{\Delta_i>0} \Delta_i + \sum_{\Delta_i>0} \Delta_i \sum_{t=m}^{T} \mathbb{P}(a_t = i) \\
&\le \sum_{\Delta_i>0} \Delta_i + \sum_{\Delta_i>0} \Delta_i \sum_{t=m}^{T} (\frac{\varepsilon_t}{m} + (1-\varepsilon_t)(\mathbb{P}(\hat{\mu}_{t,i} \ge \mu_1 + \frac{\Delta_i}{2}) + \mathbb{P}(\hat{\mu}_{t,1} \le \mu_1 - \frac{\Delta_i}{2}))) \\
&\le \sum_{\Delta_i>0} \Delta_i + \sum_{\Delta_i>0} \Delta_i \frac{C}{\Delta_{\min}^2}(1 + \log \frac{T\Delta_{\min}^2}{Cm}) \\
&\quad + \sum_{\Delta_i>0} \Delta_i \sum_{t=m}^{T} (1 - \min\{1, \frac{Cm}{t\Delta_{\min}^2}\}) \cdot 2((x_t+1)\exp(-\frac{x_t}{5}) + \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 x_t}{8}))\,.
\end{aligned}
$$

By $\frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{min}^2}{Cme^{\frac{1}{2}}}) - 1 \le x_t \le \frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{\min}^2}{Cm})$, we have

$$
\begin{aligned}
&(1 - \min\{1, \frac{Cm}{t\Delta_{\min}^2}\}) \cdot 2((x_t+1)\exp(-\frac{x_t}{5}) + \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 x_t}{8})) \\
\le\ & 2((\frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{\min}^2}{Cm}) + 1)\exp(-\frac{\frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{min}^2}{Cme^{\frac{1}{2}}}) - 1}{5}) \\
&+ \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 \cdot \frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{min}^2}{Cme^{\frac{1}{2}}}) - 1}{8})) \\
\le\ & 2((\frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{\min}^2}{Cm}) + 1)e^{\frac{1}{5}}(\frac{t\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-\frac{C}{10\Delta_{\min}^2}} + \frac{8}{\Delta_i^2}e^{\frac{1}{8}}(\frac{t\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-\frac{C\Delta_i^2}{16\Delta_{\min}^2}}) \\
\le\ & 2((\frac{C}{2\Delta_{\min}^2}(1+\log\frac{t\Delta_{\min}^2}{Cm}) + 1)e^{\frac{1}{5}}(\frac{t\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-2} + \frac{8}{\Delta_i^2}e^{\frac{1}{8}}(\frac{t\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-2})\,,
\end{aligned}
$$

where the last inequality is by letting C be sufficiently large such that $\frac{C}{10\Delta_{\min}^2} > 2$ and $\frac{C\Delta_i^2}{16\Delta_{\min}^2} > 2$. With $\sum_{t=1}^{\infty}\frac{1}{t^2} < \sum_{t=1}^{\infty}\frac{\log t}{t^2} < 1$, we have

$$
\begin{aligned}
\overline{R}_T &\le \sum_{\Delta_i>0} \Delta_i + \sum_{\Delta_i>0} \Delta_i \frac{C}{\Delta_{\min}^2}(1 + \log \frac{T\Delta_{\min}^2}{Cm}) \\
&\quad + \sum_{\Delta_i>0} \Delta_i \sum_{t=m}^{T} (1 - \min\{1, \frac{Cm}{t\Delta_{\min}^2}\}) \cdot 2((x_t+1)\exp(-\frac{x_t}{5}) + \frac{8}{\Delta_i^2}\exp(-\frac{\Delta_i^2 x_t}{8})) \\
&< \sum_{\Delta_i>0} \Delta_i + \sum_{\Delta_i>0} \Delta_i \frac{C}{\Delta_{\min}^2}(1 + \log \frac{T\Delta_{\min}^2}{Cm}) \\
&\quad + \sum_{\Delta_i>0} \Delta_i 2((\frac{C}{2\Delta_{\min}^2}(1 + 1 + \log\frac{\Delta_{\min}^2}{Cm}) + 1)e^{\frac{1}{5}}(\frac{\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-2} + \frac{8}{\Delta_i^2}e^{\frac{1}{8}}(\frac{\Delta_{\min}^2 e^{\frac{1}{2}}}{Cm})^{-2})\,,
\end{aligned}
$$

as we desired □

Acknowledgement

This lecture notes partially use material from *Reinforcement learning: An introduction*, and *Bandit algorithms*. We also used some scribe from *Stochastic bandits* by user Yiminithere on GitHub. The proofs is partially borrowed from *Finite-time analysis of the multiarmed bandit problem* by Auer, Cesa-bianchi, and Fischer.