DDA4230 Reinforcement learning	Explore-then-commit algorithms
Lecture 5	
Lecturer: Guiliang Liu	*Scribe: Baoxiang Wang, Jing Dong*

1 Goal of this lecture

To introduce and analyze explore-then-commit (ETC) algorithms.
Suggested reading: Chapter 6 of *Bandit algorithms*;

2 The explore-then-commit algorithm

Algorithm 1: The explore-then-commit algorithm

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

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

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

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

In the first km rounds (the *explore* part), then algorithm pulls each arm for k times. The algorithm then calculates the empirical mean $\frac{1}{k}\sum_{t'=0}^{mk-1} r_{t'} \mathbb{1}\{a_{t'} = i\}$ of the reward of each arm. After that (the *commit* part), the arm with the best empirical mean will be selected and will be pulled **for the rest of the horizon**, regardless of the reward it generates in the commit part of the algorithm.

We now show a general regret bound of ETC.

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

$$\overline{R}_T \le k \sum_{i \in [m]} \Delta_i + (T - mk) \sum_{i \in [m]} \Delta_i \exp\left(-\frac{k\Delta_i^2}{4}\right) . \tag{1}$$

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

$$\overline{R}_T \le \Delta_2 + \frac{4}{\Delta_2} + \frac{4}{\Delta_2} \log\left(\frac{T\Delta_2^2}{4}\right) . \tag{2}$$

Proof: Arm i is pulled for exactly k times in the first mk rounds. It is pulled for $T-mk$ times in the rest $T-mk$ rounds if the empirical mean at time $mk-1$ is optimal for arm i among all arms. Therefore, the expected number of pulls of arm i through the horizon is

$$\begin{aligned}\mathbb{E}\left[N_{T,i}\right] &= k + (T-mk)\ \mathbb{P}\big(i = \arg\max_{i'} \hat{\mu}_{mk-1,i'}\big) \\ &\leq k + (T-mk)\ \mathbb{P}\big(\hat{\mu}_{mk-1,i} \geq \hat{\mu}_{mk-1,1}\big) \\ &= k + (T-mk)\ \mathbb{P}\big(\hat{\mu}_{mk-1,i} - \mu_i - (\hat{\mu}_{mk-1,1} - \mu_1) \geq \Delta_i\big)\,.\end{aligned}$$

By the property of sub-Gaussian random variables, $\hat{\mu}_{mk-1,i} - \mu_i - (\hat{\mu}_{mk-1,1} - \mu_1)$ is $\sqrt{2/k}$ -sub-Gaussian. By the tail bound,

$$\mathbb{P}\left(\hat{\mu}_{mk-1,i} - \mu_i - (\hat{\mu}_{mk-1,1} - \mu_1) \geq \Delta_i\right) \leq \exp\left(-\frac{k\Delta_i^2}{4}\right)\,.$$

Therefore,

$$\begin{aligned}\overline{R}_T &= \sum_{i=1}^{m} \mathbb{E}\left[N_{T,i}\right] \Delta_i \\ &\leq \sum_{i=1}^{m} \Delta_i \left(k + (T-mk)\ \mathbb{P}\big(\hat{\mu}_{mk-1,i} - \mu_i - (\hat{\mu}_{mk-1,1} - \mu_1) \geq \Delta_i\big)\right) \\ &\leq \sum_{i=1}^{m} \Delta_i \left(k + (T-mk) \exp\left(-\frac{k\Delta_i^2}{4}\right)\right)\,.\end{aligned}$$

as we desired.

We then prove (2) when $m=2$. In fact, (1) reduces to

$$\begin{aligned}\overline{R}_T \leq& \Delta_2 \left(k + (T-mk)\exp\left(-\frac{k\Delta_2^2}{4}\right)\right) \\ \leq& \Delta_2 \left(k + T\exp\left(-\frac{k\Delta_2^2}{4}\right)\right)\,.\end{aligned}$$

Taking derivative against k helps us get $k_0 = 4\Delta_2^{-2}\log(T\Delta_2^2/4)$. Taking the maximum with 1 and ceiling make $k = \lceil \max\left\{1, 4\Delta_2^{-2}\log(T\Delta_2^2/4)\right\}\rceil$ a positive integer, where $k_0 \leq k \leq k_0 + 1$. Substituting this choice of k gives us

$$\begin{aligned}\overline{R}_T \leq& \Delta_2 \left(k + T\exp\left(-\frac{k\Delta_2^2}{4}\right)\right) \\ \leq& \Delta_2 \left(k_0 + 1 + T\exp\left(-\frac{k_0\Delta_2^2}{4}\right)\right) \\ \leq& \Delta_2 \left(\frac{4}{\Delta_2^2}\log\left(\frac{T\Delta_2^2}{4}\right) + 1 + T\exp\left(-\frac{\Delta_2^2}{4}\cdot\frac{4}{\Delta_2^2}\cdot\log\left(\frac{T\Delta_2^2}{4}\right)\right)\right) \\ \leq& \Delta_2 \left(\frac{4}{\Delta_2^2}\log\left(\frac{T\Delta_2^2}{4}\right) + 1 + T\cdot\frac{4}{T\Delta_2^2}\right)\end{aligned}$$

$$\leq \Delta_2 + \frac{4}{\Delta_2} + \frac{4}{\Delta_2}\log\left(\frac{T\Delta_2^2}{4}\right),$$

as we desired. $\square$

Despite the fact that (2) gives an sublinear bound on regret, obtaining this regret bound depends on the knowledge of both the suboptimality gaps Δ_2 and the horizon T. These quantities are usually fixed but may not be revealed to the agent in advance. We call an algorithm that does not require the knowledge of T *any time*. Thus the ETC algorithm is not an any time algorithm.

It is possible to show that $\overline{R}_t \leq (\Delta_2 + e^{-2})\sqrt{T}$ when $m = 2$ (we leave it as an exercise). This will remove the dependency on $\frac{1}{\Delta_2}$ at a cost of a larger order of T. The dependence of Δ_2 could be removed while obtaining a regret bound of $O(T^{2/3})$, and the dependence on T can be resolved by a doubling trick without increasing the regret by too much.

In fact, if the rewards are Gaussian with variance at most 1, the gap-dependent regret bound under $m = 2$ can be further improved by $O(\log\log T)$ by a more careful choice of k. Denote $\Delta = \Delta_2$ and π as the Archimedes' constant.

Theorem 2 *Assume that $r(i)$ is Gaussian with variance at most 1 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*

$$\overline{R}_T \leq \Delta + \frac{2}{\Delta}(\log\frac{T^2\Delta^4}{32\pi} - \log\log\frac{T^2\Delta^4}{32\pi} + \log(1+\frac{1}{e}) + 2)\,, \tag{3}$$

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 $A \sim N(k\Delta, 2k)$,

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

where the last inequality is by the inequality $W(y) \leq \log((1+e^{-1})y/\log y)$ when $y \geq e$. $\square$

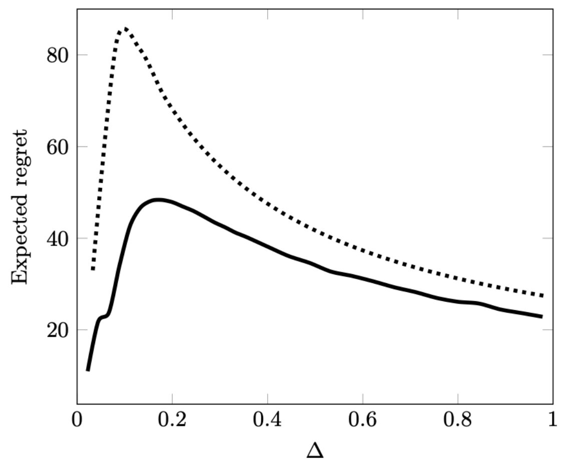

Figure 1: Regret (solid line) and regret upper bound (dashed line) of ETC with 2-armed bandit with underlying distribution being Gaussian.

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.

Some empirical results In the following figure we shall see that our upper bound is indeed not bad when the suboptimality gap Δ is large.

A Elimination algorithm

A simple way to avoid tuning the commitment time of ETC is to use elimination algorithm instead, which is a more generalized version of ETC. The intuition behind the algorithm is simple: we try to estimate the Δ_i and eliminate an arm (does not play this arm anymore) when its Δ_i is too large.

Theorem 3 *Assume that $r(i)$ is 1-sub-Gaussian for each i. The regret under the elimination algorithm with $m_\ell = 2^{4+2\ell}\log(\ell/\delta)$ and $\delta = T^{-1}\left(1 + m\pi^2/6\right)^{-1}$ is*

$$\overline{R}_T \leq \sum_{\Delta_i \neq 0} \Delta_i + \frac{16C}{\Delta_i}\log(Tm)$$

for some absolute constant C.

Algorithm 2: The elimination algorithm

Input: Sequence m_ℓ: number of exploration pulls on each arm at phase ℓ

Output: $\pi(t), t \in \{0, 1, \ldots, T\}$

Initialize active set $A_1 = \{1, \ldots, m\}$.

while $\ell = 1, 2, 3, \ldots$ **do**

  Choose each arm $i \in A_\ell$ exactly m_ℓ times.

  Let $\hat{\mu}_{i,\ell}$ be the average reward for arm i from this phase l only

  Update active set $A_{\ell+1} = \{i : \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \max_{j \in A_\ell} \hat{\mu}_{j,\ell}\}$

Proof: We first desire to show that the probability of eliminating the optimal arm decreases as the algorithm proceeds. By the sub-Gaussian tail bound, we have

$$\begin{aligned}
\mathbb{P}\left(1 \notin A_{\ell+1}, 1 \in A_\ell\right) &\leq \mathbb{P}\left(1 \in A_\ell,\ \text{exists } i \in A_\ell \backslash \{1\} : \hat{\mu}_{i,\ell} \geq \hat{\mu}_{1,\ell} + 2^{-\ell}\right) \\
&= \mathbb{P}\left(1 \in A_\ell,\ \text{exists } i \in A_\ell \backslash \{1\} : \hat{\mu}_{i,\ell} - \hat{\mu}_{1,\ell} \geq 2^{-\ell}\right) \\
&\leq m \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right).
\end{aligned}$$

Similarly, we have the probability of the optimal arm 1 and some suboptimal arm both in the active set bounded by

$$\begin{aligned}
\mathbb{P}(i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell) &\leq \mathbb{P}\left(1 \in A_\ell, i \in A_\ell, \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \hat{\mu}_{1,\ell}\right) \\
&= \mathbb{P}\left(1 \in A_\ell, i \in A_\ell, (\hat{\mu}_{i,\ell} - \mu_i) - (\hat{\mu}_{1,\ell} - \mu_1) \geq \Delta_i - 2^{-\ell}\right) \\
&\leq \exp\left(-\frac{m_\ell\left(\Delta_i - 2^{-\ell}\right)^2}{4}\right).
\end{aligned}$$

Let $\delta \in (0, 1)$ be some constant to be chosen later and $m_\ell = 2^{4+2\ell} \log(\ell/\delta)$ Then,

$$\begin{aligned}
\mathbb{P}\left(\text{ exists } \ell : 1 \notin A_\ell\right) &\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(1 \notin A_{\ell+1}, 1 \in A_\ell\right) \\
&\leq m \sum_{\ell=1}^{\infty} \exp\left(-\frac{m_\ell 2^{2\ell}}{4}\right) \\
&\leq m\delta \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} = \frac{m\pi^2\delta}{6}.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}\left(i \in A_{\ell_i+1}\right) &\leq \mathbb{P}\left(i \in A_{\ell_i+1}, i \in A_{\ell_i}, 1 \in A_{\ell_i}\right) + \mathbb{P}\left(1 \notin A_{\ell_i}\right) \\
&\leq \exp\left(-\frac{m_\ell\left(\Delta_i - 2^{-\ell_i}\right)^2}{4}\right) + \frac{m\pi^2\delta}{6}
\end{aligned}$$

$$
\begin{aligned}
&\le \exp\left(-\frac{m_\ell 2^{-2\ell_i}}{16}\right) + \frac{m\pi^2\delta}{6} \\
&\le \delta\left(1 + \frac{m\pi^2}{6}\right).
\end{aligned}
$$

Notice that if we choose $\delta = T^{-1}\left(1 + m\pi^2/6\right)^{-1}$ then $\mathbb{P}(\text{ exists } \ell : 1 \notin A_\ell) \le 1/T$ and $\mathbb{P}(i \in A_{\ell_i+1}) \le 1/T$.

To finish the proof, let i be a suboptimal action and notice that $2^{-\ell_i} \ge \Delta_i/4$, $2^{2\ell_i} \le 16/\Delta_i^2$. Furthermore, $m_\ell \ge m_1 \ge 1$ for $\ell \ge 1$. Hence,

$$
\begin{aligned}
\mathbb{E}[N_{T,i}] &\le T\mathbb{P}(i \in A_{\ell_i+1}) + \sum_{\ell=1}^{\ell_i \wedge T} m_\ell \\
&\le 1 + \sum_{\ell=1}^{\ell_i \wedge T} 2^{4+2\ell} \log\left(\frac{T}{\delta}\right) \\
&\le 1 + C 2^{2\ell_i} \log(Tm) \\
&\le 1 + \frac{16C}{\Delta_i^2} \log(Tm).
\end{aligned}
$$

where $x \wedge y$ denotes $\min\{x, y\}$ and $C > 1$ is a sufficiently large absolute constant derived by naively bounding the logarithmic term and the geometric series. The regret follows from summing this times each Δ_i. □

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.