DDA4230 Reinforcement learning UCB algorithms

Lecture 7

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

To understand algorithms based on upper confidence bounds (UCB), in terms of the regret analysis and its advantages over the approaches we previously discussed.
Suggested reading: Chapter 7, 8, and 10 of *Bandit algorithms*; Five miracles of mirror descent (Lecture 5, lecture notes);

2 Recap: The ε-greedy and ETC algorithms

For ε-greedy, by choosing $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2} m\}$ for some absolute constant C, the regret satisfies

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

where C' is an absolute constant.

For ETC under 2-armed bandits, when $T \geq 4\sqrt{2\pi e}/\Delta^2$, By choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret satisfies

$$\overline{R}_T \leq O(\frac{1}{\Delta} \log T\Delta^2) + o(\log T) + \Delta, \qquad (2)$$

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

3 The UCB algorithms

3.1 Motivation

Despite the fact that the ETC algorithm with $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$ rounds of exploration gives an optimal bound on regret (the regret bound is $\log T$, where no algorithm can achieve a better order), the algorithm has a few limitations. Executing the algorithm requires the knowledge of Δ, which is usually not available in real applications. For example, the algorithm chooses a news article to be displayed and Δ denotes the difference between the popularity of articles that has not been published before. The algorithm also uses T, the horizon of the problem. This can be unknown in real applications as well (it is fixed, but not revealed to the agent in advance). With these constraints, the theoretical result obtained by ETC is applying to 2-armed bandits only.

The UCB algorithm we discuss today addresses all these problems and is one (out of two) of the most popular approaches in practice. The advantages of UCB are listed as follows:

1. The algorithm does not use Δ_i. Note that the regret bound can still depend on Δ_i, though.
2. The algorithm does not use T. Note that the version we present today still uses T for simplicity and we refer the variant without T to Chapter 8 of the book.
3. UCB is work on multi-armed bandits with any finite number of arms.

Algorithm 1: The UCB algorithm.

```
Input: $\delta$: confidence level.
Output: $\pi(t), t \in \{0, 1, \dots, T\}$
while $t \leq T - 1$ do
    $$\pi_t = \arg\max_{i \in [m]} \mathrm{UCB}_i(t-1, \delta),$$
    where ties break arbitrarily and for $i \in [m]$,
    $$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty\,, & N_{i,t-1} = 0\,, \\ \frac{1}{N_{i,t-1}} \sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} + \sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}\,, & N_{i,t-1} > 0\,; \end{cases}$$
```

3.2 The optimism principle

The UCB algorithm is based on the principle of optimism in the face of uncertainty, which states that

one should act as if the environment is as nice as plausibly possible.

In fact, this principle is applicable to other bandit algorithms as well and is beyond the finite-armed stochastic bandit problem. In general, taking an optimistic view of the unknown leads to exploration while taking a pessimistic view of new options discourages exploration

For UCB, the optimism principle means using the data observed so far to assign to each arm a value, called the upper confidence bound. The first term,

$$\hat{\mu}_{i,t-1} = \frac{1}{N_{i,t-1}} \sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'}\}\,,$$

is the empirical mean of the rewards collected from arm i, where $N_{i,t-1} = \sum_{t' \leq t-1} \mathbb{1}\{a_{t'}\}$ is the number of times arm i has been pulled up to time $t-1$. Recall the Chernoff-Hoeffding bound for n independent 1-sub-Gaussian random variables

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \leq z) \geq 1 - \exp(-2nz^2)\,.$$

The second term, $\sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}$, is an at least $(1-\delta)$-order statistics of μ_i (which is $\mathbb{E}[r(i)]$, the true mean of arm i's reward). Then with high probability the UCB term is an overestimate of the unknown mean, if $N_{i,t-1}$ is a constant

$$\mathbb{P}(\mu_i \geq \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}) \leq \delta\,.$$

While $N_{i,t-1}$ is also a random variable which is not independent of $\hat{\mu}_{i,t-1}$, the claim holds up to constant factors (Exercise 7.1 on the book).

The intuitive reason why this leads to sublinear regret is simple. Assuming the upper confidence bound assigned to the optimal arm is indeed an overestimate, then another arm can only be played if its upper confidence bound is larger than that of the optimal arm, which in turn is larger than the mean of the optimal arm. And yet this cannot happen too often because the additional data provided by playing a suboptimal arm means that the upper confidence bound for this arm will eventually fall below that of the optimal arm.

3.3 Analysis

Algorithm 1 first explores all arms exactly once and then estimates each arm using the (sample-mean based) upper bound of its δ-confidence interval obtained from the Chernoff-Hoeffding bound. Intuitively, the arm chosen in round t either has a large sample mean or is underexplored compared to other arms. A suboptimal arm is unlikely to be played long since its optimism bonus is decreasing to zero. The key ingredient lies in choosing a good confidence level δ, which again balances the trade-off between exploration and exploitation.

Theorem 1 *Assume the rewards of arms are* 1*-sub-Gaussian. Let* $\delta = T^{-2}$*. The regret under UCB is at most*

$$\overline{R}_T \leq 3\sum_{i=1}^{m} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i}.$$

Proof: Let $c \in (0,1)$ be a constant. Consider the intersection of two events: 1) UCB overestimates arm 1 in all T rounds; 2) UCB for each suboptimal arm i falls below μ_1 at time t_i, where t_i is chosen as the smallest integer such that $(1-c)\Delta_i \geq \sqrt{2\log(1/\delta)/t_i}$. This intersected event occurs with probability at least $1 - T\delta - \sum_{i:\Delta_i>0}\exp(-t_i c^2\Delta_i^2/2)$. Applying the total probability formula, one obtains an upper bound for $\overline{R}_T$ using $\overline{R}_t = \sum_{i=1}^{m}\mathbb{E}\left[N_{i,t}\right]\Delta_i$ from LN4 that

$$\overline{R}_T \leq \sum_{i:\Delta_i>0} T\left(T\delta + \exp(-t_i c^2\Delta_i^2/2)\right)\Delta_i + \sum_{i\in[m]} t_i\Delta_i\,.$$

Letting $\delta = T^{-2}$ and $c = 1/2$ yields the desired result. □

Theorem 1 may seem loose when Δ_i are small. This can be fixed by separating the arms into two parts: those with sub-optimality gap less than $\sqrt{16m\log T/T}$ and greater than

$\sqrt{16m \log T/T}$. Bounding $\mathbb{E}[N_{i,T}]$ by T in the first part and by Theorem 1 in the second part gives

$$\overline{R}_T \leq 3 \sum_{i \in [m]} \Delta_i + 8\sqrt{mT \log T}\,. \tag{3}$$

The regret in Theorem 1 differs only by a log factor compared to the optimal bound by ETC. Yet it does not require knowledge on the suboptimality gaps. This gains UCB popularity in many practical situations.

There are a few things we could consider for extension. First of all, the confidence level in Theorem 1 depends on horizon. This can be removed by choosing δ in a decreasing format, say, $\delta_t = (1 + t \log^2 t)^{-1}$, and the resulting bound on regret remains unchanged (even better from simulation results). Secondly, when calculating the UCB, the Hoeffding inequality is used, which can be rather loose sometimes. For example, consider the Bernoulli bandits whose means are close to 0 or 1. In such situations, one could apply the Chernoff bound instead, which gives a confidence interval based on relative entropy. The corresponding regret will be improved by a taming factor (which usually depends on variance).

The $\sqrt{\log T}$ factor in (3) can be removed by making the confidence level arm-dependent. This is achieved in an algorithm termed mini-max optimal strategy (MOSS) in stochastic bandits (see Chapter 9 of the book). This algorithm achieves asymptotic optimality which matches the lower bound for minimax stochastic bandits.

3.4 Alternative proof

A relatively wider range of UCB algorithms can be proved by techniques in mirror descent. This alternative proof is fun and can be useful in many research problems. We leave it to the reader as suggested reading materials.

Acknowledgement

This lecture notes partially use material from *Bandit algorithms*.