Lecture 3a: RL Setting, Multi-Armed Bandit

Qi Zhang

Last Updated: September 2025

1 The RL Setting

In the previous lectures, we focused on two problem settings: the planning setting where the (finite-horizon) MDP of interest $M = (\mathcal{S}, \mathcal{A}, P, R, H)$ is fully known and we aim to compute an optimal policy for it, and the generative model setting where (P, R) is unknown but can be queried for any state-action pair multiple times, and we aim to find an (near-)optimal policy after querying as few times as possible.

Obviously, the generative model setting relaxes the requirements for planing. In this note, we introduce the reinforcement learning (RL) setting that is even more relaxed: (P, R) is unknown and cannot be queried as a generative model; instead, we can only access them through the agent-environment interactions, i.e., *actually* taking actions and observing the corresponding next states and rewards, which is described below:

Protocol 1 MDP interaction (finite-horizon)

```
1: for episode $k = 1, \dots, K$ do
2:     learner observes initial state $s_1$ sampled by environment;
3:     for timestep $h = 1, \dots, H$ do
4:         learner takes action $a_h$;                    ▷ by an RL algorithm
5:         learner observes next state $s_{h+1}$ and reward $r_h$ sampled by environment;
6:     end for
7: end for
```

In Protocol 1, the agent is referred to as the learner to emphasize the fact that we are in the RL setting. The interaction is repeated over episodes (indexed by k) in this finite-horizon case, and similar interaction protocols exist for the infinite-horizon case.

RL algorithm. A so-called *RL algorithm* essentially chooses the action to take at each timestep (line 4); all other lines are simply the learner observing information revealed by the environment. In general, the learner adaptively chooses the actions based on all previous information, i.e., the decision of a_h in episode k is based on

$$\underbrace{(s_{1:H}^{1:k-1}, a_{1:H}^{1:k-1}, r_{1:H}^{1:k-1})}_{\text{the previous } k-1 \text{ episodes}} \quad \text{and} \quad \underbrace{(s_1^k, a_1^k, r_1^k, \dots, s_{h-1}^k, a_{h-1}^k, r_{h-1}^k, s_h^k)}_{\text{the in-episode transitions up to the current state}}$$

where the superscript indexes the episodes.

RL evaluation. Protocol 1 induces two alternative evaluation criteria for RL algorithms, both of which have been extensively studied:

- *Exploration-exploitation.* The first measures the total rewards gathered *within* the K episodes. Under this criterion, the learner is asked to balance between doing state-action pairs that are less tried (exploration) and redoing the state-action pairs that have been most promising so far (exploitation).

- *Pure exploration.* The second ignores the rewards gathered in these episodes; instead, it further asks the learner to recommend a policy by the end of the interaction and measures the quality of that policy. Under this criterion, the learner is asked to gather high quality information within the interaction budget in order to recommend a best possible policy.

For both criteria, the performance of an RL algorithm hinges on how efficiently it explores the environment through its action selection.

2 Multi-Armed Bandits

Designing and analyzing RL algorithms for full-width MDPs is non-trivial. Many of the challenges, especially those related to exploration, also manifest in simplified MDP instances known as Multi-Armed Bandits (MABs). A MAB can be viewed as an MDP with a horizon of $H = 1$ and a single state, so the state transition function is out of the picture. To make the problem non-trivial, we consider the case where the reward upon taking each action is random.

Therefore, with an abuse of notation (to respect the convention), a MAB can be specified by some K real-valued distributions, $\{R_i\}_{i=1}^K$, i.e., the stochastic reward function for K actions or *arms*. Each R_i is assumed to have bounded support $[0, 1]$ and let μ_i denote its mean. Accordingly, Protocol 1 reduces to:

Protocol 2 MAB interaction

```
for round $t = 1, \dots, T$ do
    learner chooses an arm $i_t \in [K]$          ▷ chosen by an MAB algorithm;
    learner observes reward $r_t \sim R_{i_t}$    ▷ sampled independently;
end for
```

We use the following notations regarding an MAB:

K	number of arms/actions
R_i	reward distribution number of arm i
μ_i	mean of R_i
$\mu^* := \max_{i \in [K]} \mu_i$	mean of best arm
$\Delta_i := \mu^* - \mu_i$	suboptimality of arm i
T	total number of rounds
i_t	arm played at round t
$r_t \sim R_{i_t}$	reward received at round t
$N_{i,t} := \sum_{s=1}^t \mathbb{I}\{i_s = i\}$	number of times arm i has been played by end of round t
$\hat{\mu}_{i,t} := \frac{1}{N_{i,t}} \sum_{s=1}^t \mathbb{I}\{i_s = i\} r_s$	average reward received from arm i by end of round t

For MAB, the exploration-exploitation criterion reduces to maximizing the total reward within the T rounds, or equivalently, minimizing the *regret*:

$$\text{Regret}(T) := \textstyle\sum_{t=1}^T (\mu^* - r_t) = T\mu^* - \sum_{t=1}^T r_t \tag{1}$$

Intuitively, the regret measures difference between the learner's reward and the expected reward by playing the best arm per round, summed over T rounds for Regret(T).

In contrast, the pure exploration criterion is formalized by minimizing

$$\Delta_{\hat{i}} = \mu^* - \mu_{\hat{i}}$$

where $\hat{i}$ is the arm that the learner *recommends* after T or a certain number of rounds.

Both Regret(T) and $\Delta_{\hat{i}}$ are random variables because the arms i_t are chosen with randomness and the individual rewards are sampled from distributions R_{i_t}. We typically consider their expectation ($\mathbb{E}[\text{Regret}(T)]$) or provide high probability guarantees. The theoretical analyses of MAB algorithms often care about the asymptotical behavior of the regret as a function of T, i.e., how quickly the regret increases as T increases.

We next describe some classical algorithms and analyze their regret.

2.1 ϵ-greedy

The ϵ-*greedy* algorithm is a simple way to balance exploration vs exploitation. At round t, w.p. (with probability) ϵ the learner chooses a random arm; w.p. $1 - \epsilon$ the learner chooses the empirically best arm:

$$i_t \sim \text{Uniform}([K]) \text{ w.p. } \epsilon \quad \text{and} \quad i_t = \arg\max_{i \in [K]} \ \hat{\mu}_{i,t-1} \text{ w.p. } 1 - \epsilon.$$

Here, ϵ is a constant. Because the learner picks the worst arm, $\underline{i} := \arg\min \mu_i$, w.p. at least ϵ/K every round, the expected regret is lower bounded as

$$\mathbb{E}[\text{Regret}(T)] \geq (\mu^* - \mu_{\underline{i}}) \frac{\epsilon}{K} T \tag{2}$$

which scales linearly with T.

2.2 Explore-then-commit

A simple idea is to try each arm multiple times and then favor the empirically best one.

Exploration by uniform sampling. Suppose we play each arm n times with n being a predefined positive integer. Let $\hat{\mu}_i$ be the average reward of arm i over the n plays. By Hoeffding's inequality (recall that the rewards are bounded in $[0, 1]$), we have, for any $\epsilon > 0$,

$$\Pr\Big(\underbrace{|\hat{\mu}_i - \mu_i| \geq \epsilon}_{=:B_i(\epsilon)}\Big) \leq 2e^{-2n\epsilon^2}. \tag{3}$$

We define the ϵ-*bad event* to be the event where $B_i(\epsilon)$ occurs for at least one arm i, i.e.,

$$\epsilon\text{-bad event} := \textstyle\bigcup_{i \in [K]} B_i(\epsilon).$$

By the union bound, we have

$$\Pr(\epsilon\text{-bad event}) = \Pr\left(\textstyle\bigcup_{i \in [K]} B_i(\epsilon)\right) \leq \textstyle\sum_{i \in [K]} \Pr(B_i(\epsilon)) \leq 2Ke^{-2n\epsilon^2}$$

where the last inequality is due to (3).

Let the ϵ-*clean event* to be the complement of the ϵ-bad event, i.e.,

$$\epsilon\text{-clean event: } |\hat{\mu}_i - \mu_i| < \epsilon \text{ for } \textit{all} \text{ arms } i \in [K]$$

so we have

$$\Pr(\epsilon\text{-clean event}) = 1 - \Pr(\epsilon\text{-bad event}) \geq 1 - 2Ke^{-2n\epsilon^2}. \tag{4}$$

Exploitation by committing. Let $\hat{i} := \arg\max_{i\in[K]} \hat{\mu}_i$ be the empirically best arm after trying each arm n times. Intuitively, when n is large, the suboptimality of $\hat{i}$, $\Delta_{\hat{i}} = \mu^* - \mu_{\hat{i}}$, should be small with high probability. We can make this argument formal. Letting $i^* := \arg\max_{i\in[K]} \mu_i$ be the actual best arm, decompose the suboptimality as

$$\Delta_{\hat{i}} = \mu_{i^*} - \mu_{\hat{i}} = \underbrace{\mu_{i^*} - \hat{\mu}_{i^*}}_{\text{(i)}} + \underbrace{\hat{\mu}_{i^*} - \hat{\mu}_{\hat{i}}}_{\leq 0} + \underbrace{\hat{\mu}_{\hat{i}} - \mu_{\hat{i}}}_{\text{(ii)}} \leq \text{(i)} + \text{(ii)}.$$

Therefore, we have

$$\Pr(\Delta_{\hat{i}} < \epsilon) \geq \Pr\left(\tfrac{\epsilon}{2}\text{-clean event}\right) \geq 1 - 2Ke^{-\frac{n\epsilon^2}{2}} \tag{5}$$

For Homework 3's 1a, provide a justification for (5).

Regret analysis. In the *explore-then-commit* algorithm, the learner chooses some integer n; it plays each arm n times in the first nK rounds and then commits to playing the empirically best arm $\hat{i}$ in the rest $(T - nK)$ rounds. The analysis above facilitates the following way to decompose the expected regret based on whether event $\Delta_{\hat{i}} < \epsilon$ occurs or not. The rational is that, conditioned on $\Delta_{\hat{i}} < \epsilon$, committing to $\hat{i}$ for the last $(T - nK)$ rounds incurs small regret; otherwise, the regret will be large but it happens with the small probability of $\Pr(\Delta_{\hat{i}} \geq \epsilon)$. Formalizing this, we have

$$\mathbb{E}[\text{Regret}(T)] = \mathbb{E}[\text{Regret}(T) \mid \Delta_{\hat{i}} < \epsilon] \Pr(\Delta_{\hat{i}} < \epsilon) + \mathbb{E}[\text{Regret}(T) \mid \Delta_{\hat{i}} \geq \epsilon] \Pr(\Delta_{\hat{i}} \geq \epsilon) \tag{6}$$

where

$$\begin{aligned}
&\mathbb{E}[\text{Regret}(T) \mid \Delta_{\hat{i}} < \epsilon] \leq 1 \cdot nK + \epsilon(T - nK) && \text{(regret} \leq 1 \text{ per round during exploration)}\\
&\Pr(\Delta_{\hat{i}} < \epsilon) \leq 1 &&\\
&\mathbb{E}[\text{Regret}(T) \mid \Delta_{\hat{i}} \geq \epsilon] \leq 1 \cdot T && \text{(regret} \leq 1 \text{ per round)}\\
&\Pr(\Delta_{\hat{i}} \geq \epsilon) \leq 2Ke^{-\frac{n\epsilon^2}{2}} && \text{(due to (5))}
\end{aligned}$$

and therefore

$$\mathbb{E}[\text{Regret}(T)] \leq (1 - \epsilon)nK + T\epsilon + T \cdot 2Ke^{-\frac{n\epsilon^2}{2}} =: g(n, \epsilon).$$

Here, the bound above holds for any $\epsilon > 0$ and any integer $n > 0$ such that $nK \leq T$. Ideally, we hope to find a realization of (n, ϵ) such that $g(n, \epsilon)$ is minimized. However, because a closed-form minimizer is hard to obtain, we choose a realization of (n, ϵ) (which would depend on T) so $g(n, \epsilon)$ scales with T moderately, e.g., *sublinearly* if $g(n, \epsilon) = O(T^\alpha)$ for some $0 < \alpha < 1$. To achieve sublinear scaling, let's try making the last term of $g(n, \epsilon)$ a $O(1)$ term with respect to T, i.e., by setting

$$-\frac{n\epsilon^2}{2} = -\ln T \quad \text{so that} \quad T \cdot 2Ke^{-\frac{n\epsilon^2}{2}} = T \cdot 2Ke^{-\ln T} = T \cdot 2K \cdot T^{-1} = 2K. \tag{7}$$

Equivalently, we should set $\epsilon = \sqrt{\frac{2\ln T}{n}}$, with n to be chosen later. With this, we have the first two terms of $g(n, \epsilon)$ as

$$(1-\epsilon)nK + T\epsilon \le nK + T\epsilon = nK + T\sqrt{\frac{2\ln T}{n}} =: g_1(n)$$

where the first inequality amplifies $(1-\epsilon)$ to 1, which turns out to be tolerable. Minimizing $g_1(n)$, e.g., by setting derivative $g_1'(n) = 0$, we choose n as

$$n = \left(\frac{T^2 \ln T}{2K^2}\right)^{\frac{1}{3}} \tag{8}$$

which gives $g_1(n) = O\left(K^{\frac{1}{3}} T^{\frac{2}{3}} (\ln T)^{\frac{1}{3}}\right)$, and therefore we have the sublinear bound

$$\mathbb{E}[\text{Regret}(T)] \le g_1(n) + 2K = O\left(K^{\frac{1}{3}} T^{\frac{2}{3}} (\ln T)^{\frac{1}{3}}\right) \tag{9}$$

for the explore-then-commit algorithm.

It is not unclear if we can obtain bounds that scale better with T, for example, by setting (n, ϵ) differently in (7) (e.g., $-\frac{n\epsilon^2}{2} = \ln T^{-\beta}$).

2.3 UCB

We now introduce an algorithm that employs the principle of *optimism in the face of uncertainty* for strategic exploration, which favors actions/arms that we are less certain about. As we do so, the less certain actions/arms will be better explored and their values/rewards will become more and more certain. For MABs, the Upper Confidence Bound (UCB) algorithm implements this idea by, again, the Hoefflindg's inequality, which also enjoys a regret bound better than that of explore-then-commit (9).

Assume that the rewards are bounded in $[0, 1]$. If $\hat{\mu}_i$ is the average reward of arm i over n plays of it, we have

$$\Pr\left(\hat{\mu}_i - \mu_i \le -\epsilon\right) \le e^{-2n\epsilon^2} \tag{10}$$

For Homework 3's 1b, provide a justification for (10).

Here, we omit the other one-side inequality, because UCB uses the one above to form a high-probability upper bound of the true mean: rewriting (10) we have $\Pr\left(\hat{\mu}_i + \epsilon \le -\mu_i\right) \le e^{-2n\epsilon^2}$, or equivalently, by setting $\delta = e^{-2n\epsilon^2}$ we have

$$\Pr\left(\hat{\mu}_i + \epsilon(n, \delta) \le \mu_i\right) \le \delta \quad \text{where } \epsilon(n, \delta) := \sqrt{\frac{\ln(1/\delta)}{2n}} \tag{11}$$

In words, w.p $\ge 1 - \delta$ (i.e., with high confidence), $\hat{\mu}_i + \epsilon(n, \delta)$ is an upper bound of the true mean μ_i. The UCB algorithm selects the arm with the highest upper confidence bound given above: for round $t = 1, 2, \ldots, T$,

$$i_t := \arg\max_{i \in [K]} \underbrace{\hat{\mu}_{i,t-1} + B_i(\delta_t)}_{=:\text{UCB}_{i,t}} \text{ where } B_i(\delta_t) := \epsilon(N_{i,t-1}, \delta_t) = \sqrt{\frac{\ln(1/\delta_t)}{2N_{i,t-1}}}. \tag{12}$$

Recall that $N_{i,t-1}$ is number of times arm i has been played by end of round $t - 1$ and $\hat{\mu}_{i,t-1}$ is the corresponding empirical mean. Here, we let the confidence level δ_t depend on round number t

and potentially T and K as well. We next detail how to choose δ_t to obtain a $O(\sqrt{T})$ regret bound (which is better than (9)).

To analyze the regret of the UCB algorithm (12), we use the similar recipe as (6) to decompose the expected regret conditioned on some good/bad event. The bad event is when the empirical means are far off the true mean, which happens with small probability due to Hoeffding's. However, as a difference from explore-then-commitment, in UCB the number of plays $N_{i,t-1}$ is a random variable that depends on previous plays, whereas explore-then-commitment uses a predefined constant n for exploration. This difference is crucial because Hoeffding's only applies when the number of samples n is a constant:

$$\Pr\big(\underbrace{|\hat{\mu}_{i,t-1} - \mu_i| \geq B_i(\delta_t) \text{ where } N_{i,t-1} = n \text{ for constant } n}_{\text{bad event } E^n_{i,t}}\big) \leq 2\delta_t$$

Here, the bad event $E^n_{i,t}$ is when arm i has been played exactly n times by end of round $t-1$ (i.e., $N_{i,t-1} = n$) and the empirical mean is far off. We can then bound the probability of the general bad event:

$$\Pr\big(\underbrace{|\hat{\mu}_{i,t-1} - \mu_i| \geq B_i(\delta_t)}_{=:E_{i,t}}\big) \leq 2t\delta_t. \tag{13}$$

For Homework 3's 1c, provide a justification for (13) (Hint: Use the union bound over all possible values that $N_{i,t-1} = n$ can take).

Then, by the union bound over all $i \in [K]$ and $t \in [T]$, we can bound the probability of our *unclean* event:

$$\Pr\big(\underbrace{|\hat{\mu}_{i,t-1} - \mu_i| \geq B_i(\delta_t) \text{ for any } i \in [K], t \in [T]}_{=:\text{unclean event}}\big) \leq \sum_{i=1}^{K}\sum_{t=1}^{T} 2t\delta_t = 2K\sum_{t=1}^{T} t\delta_t.$$

The corresponding *clean* event is the complement of unclean:

$$\text{clean event: } |\hat{\mu}_i - \mu_i| < B_i(\delta_t) \text{ for } \textit{all } i \in [K], t \in [T].$$

From now on, we try setting $\delta_t = \delta$ which, to be chosen carefully, does not depend on round number t but can potentially depend on T and/or K. The probability of unclean is then bounded as

$$\Pr(\text{unclean}) \leq 2K \textstyle\sum_{t=1}^{T} t\delta_t = 2K\delta \sum_{t=1}^{T} t \leq 2KT^2\delta.$$

For the unclean event, its contribution the expected regret can therefore be bounded as

$$\mathbb{E}[\text{Regret}(T)|\text{unclean}] \cdot \Pr(\text{unclean}) \leq (1 \cdot T) \cdot 2KT^2\delta = 2KT^3\delta.$$

We would like the bound above to scale as $O(T^{-\alpha})$ for some $\alpha \geq 0$. We here choose $\alpha = 1$ by setting

$$\delta = \frac{1}{T^4}, \text{ so that } \mathbb{E}[\text{Regret}(T)|\text{unclean}] \cdot \Pr(\text{unclean}) \leq 2KT^{-1} \tag{14}$$

while other choices such as $\alpha = 0$ like in (7) should also work.

For the clean event, we simply bound Pr(clean) ≤ 1 and decompose the regret per round as

$$\mathbb{E}[\text{Regret}(T)|\text{clean}] = \mathbb{E}\left[\textstyle\sum_{t=1}^{T}(\mu^* - r_t) \mid \text{clean}\right] = \mathbb{E}\left[\textstyle\sum_{t=1}^{T}(\mu^* - \mu_{i_t}) \mid \text{clean}\right]$$

Here, the second equality is intuitively straightforward but requires a rigorous proof using the law of total expectation which we omit here. Conditioned on the clean event, we bound the per step regret as

$$\mu^* - \mu_{i_t} = \mu_{i^*} - \mu_{i_t} \leq \text{UCB}_{i^*,t} - \mu_{i_t} \leq \text{UCB}_{i_t,t} - \mu_{i_t} = B_{i_t}(\delta_t) = \sqrt{\tfrac{2\ln T}{N_{i_t,t-1}}} \tag{15}$$

where in the last equality we set $\delta_t = \delta = \frac{1}{T^4}$.

For Homework 3's 1d, provide a justification for the two inequalities in (15).

Note the bound is vacuous when $N_{i_t,t-1} = 0$. To fix this, note that by the definition, UCB algorithm in (12) will choose each arm once in the first K rounds, as a never-chosen arm has a UCB value of infinity. Thus, we can proceed as

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)|\text{clean}] =& \mathbb{E}\left[\textstyle\sum_{t=1}^{K}(\mu^* - \mu_{i_t}) \mid \text{clean}\right] + \mathbb{E}\left[\textstyle\sum_{t=K+1}^{T}(\mu^* - \mu_{i_t}) \mid \text{clean}\right] \\ \leq& 1 \cdot K + \mathbb{E}\left[\textstyle\sum_{t=K+1}^{T}\sqrt{\tfrac{2\ln T}{N_{i_t,t-1}}} \mid \text{clean}\right]\end{aligned}$$

The second term above can be bounded by the following tricks

$$\sum_{t=K+1}^{T}\sqrt{\frac{1}{N_{i_t,t-1}}} = \sum_{i=1}^{K}\sum_{m=1}^{N_{i,T-1}}\sqrt{\frac{1}{m}} \leq \sum_{i=1}^{K} 2\sqrt{N_{i,T-1}} \leq 2 \cdot \sqrt{K} \cdot \sqrt{\textstyle\sum_{i=1}^{K} N_{i,T-1}} \leq 2\sqrt{KT}$$

where the first inequality is due to the "integral trick", $\sum_{m=1}^{M}\sqrt{\frac{1}{m}} \leq \int_{x=0}^{M} x^{-\frac{1}{2}}dx = 2\sqrt{M}$; the second inequality is due to Cauchy–Schwarz, $\sum_{i=1}^{K} a_i = \langle \mathbf{1}, \mathbf{a}\rangle \leq \|\mathbf{1}\|_2 \cdot \|\mathbf{a}\|_2$.

Putting it all together, for UCB with $\delta = \frac{1}{T^4}$, its expected regret can be bounded as

$$\mathbb{E}[\text{Regret}(T)] \leq 2KT^{-1} + K + 2\sqrt{2KT\ln T} = O(\sqrt{2KT\ln T}).$$