MVA: Reinforcement Learning (2021/2022) **Assignment 3**

Exploration in Reinforcement Learning (theory)

Lecturers: *M. Pirotta* *(December 16, 2021)*

Solution by SIMON Hugo

Instructions

- The deadline is **January 16, 2022. 23h59**
- By doing this homework you agree to the late day policy, collaboration and misconduct rules reported on Piazza.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.
- Answers should be provided in **English**.

1 Best Arm Identification

In best arm identification (BAI), the goal is to identify the best arm in as few samples as possible. We will focus on the fixed-confidence setting where the goal is to identify the best arm with high probability $1-\delta$ in as few samples as possible. A player is given k arms with expected reward μ_i. At each timestep t, the player selects an arm to pull (I_t), and they observe some reward ($X_{I_t,t}$) for that sample. At any timestep, once the player is confident that they have identified the best arm, they may decide to stop.

δ-correctness and fixed-confidence objective. Denote by τ_δ the stopping time associated to the stopping rule, by $i^\star$ the best arm and by $\widehat{i}$ an estimate of the best arm. An algorithm is δ-correct if it predicts the correct answer with probability at least $1-\delta$. Formally, if $\mathbb{P}_{\mu_1,\ldots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ and $\tau_\delta < \infty$ almost surely for any $\mu_1, \ldots, \mu_k$. Our goal is to find a δ-correct algorithm that minimizes the sample complexity, that is, $\mathbb{E}[\tau_\delta]$ the expected number of sample needed to predict an answer. Assume that the best arm $i^\star$ is unique (i.e., there exists only one arm with maximum mean reward).

Notation

- I_t: the arm chosen at round t.
- $X_{i,t} \in [0,1]$: reward observed for arm i at round t.
- μ_i: the expected reward of arm i.
- $\mu^\star = \max_i \mu_i$.
- $\Delta_i = \mu^\star - \mu_i$: suboptimality gap.

Consider the following algorithm

Input: k arms, confidence δ
$S = \{1, \dots, k\}$
for $t = 1, \dots$ **do**
  Pull **all** arms in S
  $S = S \setminus \left\{ i \in S \ : \ \exists j \in S, \ \widehat{\mu}_{j,t} - U(t, \delta') \geq \widehat{\mu}_{i,t} + U(t, \delta') \right\}$
  if $|S| = 1$ **then**
    STOP
    return S
  end
end

The algorithm maintains an active set S and an estimate of the empirical reward of each arm $\widehat{\mu}_{i,t} = \frac{1}{t} \sum_{j=1}^{t} X_{i,j}$.

- Compute the function $U(t, \delta)$ that satisfy the any-time confidence bound. Let

$$\mathcal{E} = \bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} \{ |\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta') \} .$$

  Using Hoeffding's inequality and union bounds, shows that $\mathbb{P}(\mathcal{E}) \leq \delta$ for a particular choice of δ'. This is called "bad event" since it means that the confidence intervals do not hold.

  Proof. It suffices to find ε s.t. $\forall i, \mathbb{P}(|\widehat{\mu}_{i,t} - \mu_i| > \varepsilon) \leq \delta \frac{c}{t^2}$ where $c = \frac{6}{k\pi^2}$. Indeed, the union bound then gives $\mathbb{P}(\mathcal{E}) \leq \sum_{i,t=1}^{k,+\infty} \delta \frac{c}{t^2} = \delta k c \sum_{t=1}^{+\infty} \frac{1}{t^2} = \delta$. But via Hoeffding inequality

$$\forall i, \mathbb{P}(|\widehat{\mu}_{i,t} - \mu_i| > \varepsilon) \leq 2e^{-2t\varepsilon^2} := \delta \frac{c}{t^2} \iff \varepsilon = \sqrt{\frac{1}{2t} \log(\frac{2t^2}{\delta c})} := U(t, \delta \frac{c}{t^2})$$

  □

- Show that with probability at least $1 - \delta$, the optimal arm $i^\star = \arg\max_i \{\mu_i\}$ remains in the active set S. Use your definition of δ' and start from the condition for arm elimination. From this, use the definition of $\neg\mathcal{E}$.

  Proof. Under the event $\neg\mathcal{E}$, we have $\forall i, t \quad |\widehat{\mu}_{i,t} - \mu_i| \leq U(t, \delta')$. Suppose then i^* is dropped i.e. $\exists j \quad \widehat{\mu}_{j,t} - U(t, \delta') \geq \widehat{\mu}_{i^*,t} + U(t, \delta') \implies \mu_j \geq \mu_{i^*}$, **contradiction**. Then under $\neg\mathcal{E}$, the optimal arm i^* remains in the active set S, which appends $1 - \delta$-surely. □

- Under event $\neg\mathcal{E}$, show that an arm $i \neq i^\star$ will be removed from the active set when $\Delta_i \geq C_1 U(t, \delta')$ for some constant $C_1 \in \mathbb{N}$. Compute the time required to have such condition for each non-optimal arm. Use the condition of arm elimination applied to arm $i^\star$.[1]

  Proof. If $\neg\mathcal{E}$ and $\Delta_i \geq 4U(t, \delta')$ then

$$\widehat{\mu}_{i^*,t} - U(t, \delta') \geq \mu_{i^*} - 2U(t, \delta') \geq \mu_i + 2U(t, \delta') \geq \widehat{\mu}_{i,t} + U(t, \delta')$$

  So arm i get rejected. Moreover, $\Delta_i \geq 4U(t, \delta') \iff \log(bt) \leq at$ where $a = \frac{\Delta_i^2}{16}$ and $b = \sqrt{\frac{2}{\delta c}}$. Then by [1] $T_i \geq \frac{\log b/a + \sqrt{2 \log b/a - 1}}{a}$ meaning T_i is in the order of $\Delta_i^{-2} \log(\frac{\delta}{k\Delta_i^4})$ □

[1] Note that $at \geq \log(bt)$ can be solved using Lambert W function. We thus have $t \geq \frac{-W_{-1}(-a/b)}{a}$ since, given $a = \Delta_i^2$ and $b = 2k/\delta$, $-a/b \in (-1/e, 0)$. We can make the bound more explicit by noticing that $-1 - \sqrt{2u} - u \leq W_{-1}(-e^{-u-1}) \leq -1 - \sqrt{2u} - 2u/3$ for $u > 0$ [Chatzigeorgiou, 2016]. Then $t \geq \frac{1+\sqrt{2u}+u}{a}$ with $u = \log(b/a) - 1$.
[1],

- Compute a bound on the sample complexity (after how many pulls the algorithm stops) for identifying the optimal arm w.p. $1-\delta$.

  Proof. Under event $\neg\mathcal{E}$, the previous bound implies all sub-optimal arms get rejected after $T = \max_{i\neq i^*} T_i$ in the order of $\Delta_{min}^{-2}\log(\frac{\delta}{k\Delta_{min}^4})$ where $\Delta_{min} = \min_{i\neq i^*}\Delta_i$. $\square$

- We assumed that the optimal arm $i^\star$ is unique. Would the algorithm still work if there exist multiple best arms? Why?

  Answer. As $\Delta_{min}\to 0$, $T\to+\infty$. We have no guarantees the algorithm terminates if the optimal arm is not unique. $\square$

Note that also a variations of UCB are effective in pure exploration.

2 Regret Minimization in RL

Consider a finite-horizon MDP $M^\star = (S, A, p_h, r_h)$ with stage-dependent transitions and rewards. Assume rewards are bounded in $[0,1]$. We want to prove a regret upper-bound for UCBVI. We will aim for the suboptimal regret bound ($T = KH$)

$$R(T) = \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \widetilde{O}(H^2 S\sqrt{AK})$$

Define the set of plausible MDPs as

$$\mathcal{M}_k = \{M = (S, A, p_{h,k}, r_{h,k})\ :\ r_{h,k}(s,a)\in B_{h,k}^r(s,a), p_{h,k}(\cdot|s,a)\in B_{h,k}^p(s,a)\}$$

Confidence intervals can be anytime or not.

- Define the event $\mathcal{E} = \{\forall k, M^\star\in\mathcal{M}_k\}$. Prove that $\mathbb{P}(\neg\mathcal{E})\le\delta/2$. First step, construct a confidence interval for rewards and transitions for each (s,a) using Hoeffding and Weissmain inequality (see appendix), respectively. So, we want that

  $$\mathbb{P}\Big(\forall k,h,s,a : |\widehat{r}_{hk}(s,a) - r_h(s,a)| \le \beta_{hk}^r(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \le \beta_{hk}^p(s,a)\Big) \ge 1-\delta/2$$

  Proof. We want

  $$\begin{aligned}
  \forall k,h,s,a \quad \mathbb{P}\left(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \varepsilon\right) &\le 2e^{-2N_{hk}(s,a)\varepsilon^2} = \frac{\delta}{2KHSA}\\
  \iff \varepsilon &= \sqrt{\frac{1}{2N_{hk}(s,a)}\log(\frac{4KHSA}{\delta})} := \beta_{hk}^r(s,a)\\
  \forall k,h,s,a \quad \mathbb{P}\left(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \varepsilon\right) &\le 2^S e^{-\frac{N_{hk}(s,a)}{2}\varepsilon^2} = \frac{\delta}{2KHSA}\\
  \iff \varepsilon &= \sqrt{\frac{2}{N_{hk}(s,a)}\log(\frac{2^{S+1}KHSA}{\delta})} := \beta_{hk}^p(s,a)
  \end{aligned}$$

  Definitions of $\beta_{hk}^r(s,a)$ and $\beta_{hk}^p(s,a)$ provide $\mathbb{P}(\neg\mathcal{E})\le\delta/2$ via the union bound. $\square$

- Define the bonus function and consider the Q-function computed at episode k

  $$Q_{h,k}(s,a) = \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'}\widehat{p}_{h,k}(s'|s,a)V_{h+1,k}(s')$$

with $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$. Recall that $V_{H+1,k}(s) = V^\star_{H+1}(s) = 0$. Prove that under event $\mathcal{E}$, Q_k is optimistic, i.e.,

$$Q_{h,k}(s,a) \geq Q^\star_h(s,a), \forall s,a \tag{1}$$

where $Q^\star$ is the optimal Q-function of the unknown MDP $M^\star$. Note that $\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \geq r_{H,k}(s,a)$ and thus $Q_{H,k}(s,a) \geq Q^\star_H(s,a)$ (for a properly defined bonus). Then use induction to prove that this holds for all the stages h.

Proof. Under event $\mathcal{E}$

- *Init.* $Q_{H,k}(s,a) \geq Q^\star_H(s,a)$
- *Heredity*

$$\begin{aligned}
&Q^\star_h(s,a) - Q_{h,k}(s,a) \\
&\leq r_h(s,a) - \widehat{r}_{hk}(s,a) - b_{hk}(s,a) + \sum_{s'} p_{h,k}(s'|s,a) \max_a Q^*_{h+1}(s',a) - \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s') \\
&\leq \beta^r_{hk}(s,a) - b_{hk}(s,a) + H\beta^p_{hk}(s,a) := 0 \qquad \text{(this defines the bonus function)}
\end{aligned}$$

- *Conclusion* $\forall h,k,s,a \quad Q_{h,k}(s,a) \geq Q^\star_h(s,a)$

□

- In class we have seen that

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{2}$$

where $\delta_{hk}(s) = V_{hk}(s) - V^{\pi_k}_h(s)$ and $m_{hk} = \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})$. We now want to prove this result. Denote by a_{hk} the action played by the algorithm (you will have to use the greedy property).

1. Show that $V^{\pi_k}_h(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}$
2. Show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
3. Putting everything together prove Eq. 2.

Proof. Indexation by k is dropped to simplify notation.

1. Using definitions and the greedy property

$$\begin{aligned}
r(s_h,a_h) + \mathbb{E}_{p(s_h,a_h)}[V_{h+1}(s')] - \delta_{h+1}(s_{h+1}) - m_h &= r(s_h,a_h) + \mathbb{E}_{p(s_h,a_h)}[V_{h+1,k}(s') - \delta_{h+1,k}(s')] \\
&= r(s_h,a_h) + \mathbb{E}_{p(s_h,a_h)}[V^\pi_{h+1}(s')] \\
&= \max_a r(s_h,a) + \mathbb{E}_{p(s_h,a)}[V^\pi_{h+1}(s')] \\
&= V^\pi_h(s_h)
\end{aligned}$$

2. $V_h(s_h) = \min(H, \max_a Q_h(s_h,a)) \leq \max_a Q_h(s_h,a) = Q_h(s_h,a_h)$ by the greedy property.
3. From previous questions, we have that

$$\begin{aligned}
\forall k, \quad \delta_k(s_k) &= V_k(s_k) - V^\pi_k(s_k) \leq Q_h(s_h,a_h) - V^\pi_k(s_k) \\
&= Q_h(s_h,a_h) - r(s_h,a_h) - \mathbb{E}_{p(s_h,a_h)}[V_{h+1}(s')] + m_h + \delta_{k+1}(s_{k+1}) \\
\iff \delta_1(s_1) &= \delta_1(s_1) - \delta_{H+1}(s_{H+1}) = \sum_{h=1}^{H} \delta_k(s_k) - \delta_{k+1}(s_{k+1}) \\
&\leq \sum_{h=1}^{H} Q_h(s_h,a_h) - r(s_h,a_h) - \mathbb{E}_{p(s_h,a_h)}[V_{h+1}(s')] + m_h
\end{aligned}$$

□

- Since $(m_{hk})_{hk}$ is an MDS, using Azuma-Hoeffding we show that with probability at least $1-\delta/2$

$$\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log(2/\delta)} \tag{3}$$

Show that the regret is upper bounded with probability $1-\delta$ by

$$R(T) \leq 2\sum_{kh} b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH\log(2/\delta)}$$

Proof. Let be $C = 2H\sqrt{KH\log(2/\delta)}$

$$\begin{aligned}
&\mathbb{P}\left(R(T) > 2\sum_{kh} b_{hk} + C\right) = \mathbb{P}\left(\sum_{k} \delta_{1k}(s_{1k}) + V_1^\star - V_{1k} > 2\sum_{kh} b_{hk} + C\right) \\
&\leq \mathbb{P}\left(\sum_{hk} m_{hk} + (\widehat{r}_{hk} - r_h) + \sum_{s'} (\widehat{p}_{hk}(s') - p_{hk}(s'))V_{h+1k}(s') + \sum_{k} V_1^\star - V_{1k} > \sum_{kh} b_{hk} + C\right) \\
&\leq \mathbb{P}(\sum_{hk} m_{hk} > C) + \mathbb{P}(\bigcup_{kh} \{(\widehat{r}_{hk} - r_h) > \beta_{hk}^r\} \cup \{\|\widehat{p}_{hk} - p_{hk}\|_1 > \beta_{hk}^p\} \cup \{V_1^\star - V_{1k} > 0\}) \\
&\leq \mathbb{P}(\sum_{hk} m_{hk} > C) + \mathbb{P}(\neg\mathcal{E}) \leq \delta/2 + \delta/2 = \delta \qquad \text{(apply 3 and 1)}
\end{aligned}$$

□

- Finally, we have that [Domingues et al., 2021]

$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \lesssim \underset{\sim\sim\sim\sim}{H^2S^2A} + 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)}$$

Complete this by showing an upper-bound of $H\sqrt{SAK}$, which leads to $R(T) \lesssim H^2S\sqrt{AK}$

Proof. We first prove

$$\sum_{hk} \frac{1}{\sqrt{N_{hk}}} \lesssim 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)} \leq 2\sum_{h=1}^{H}\sqrt{SA\sum_{s,a} N_{hK}(s,a)} = 2H\sqrt{SAK} \qquad \text{(Jensen)}$$

Then

$$\sum_{hk} b_{hk} = \sum_{hk} \beta_{hk}^r(s,a) + H\beta_{hk}^p(s,a) \simeq \sum_{hk} H\sqrt{\frac{S}{N_{hk}}} \lesssim H^2S\sqrt{AK}$$

Finally

$$R(T) \leq 2\sum_{hk} b_{hk} + 2H\sqrt{KH\log(2/\delta)} \lesssim H^2S\sqrt{AK}$$

□

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in S \times A$ and $h = 1, \dots, H$

for $k = 1, \dots, K$ **do**
  Observe initial state s_{1k} *(arbitrary)*
  Estimate empirical MDP $\widehat{M}_k = (S, A, \widehat{p}_{hk}, \widehat{r}_{hk}, H)$ from $\mathcal{D}_k$

$$\widehat{p}_{hk}(s'|s,a) = \frac{\sum_{i=1}^{k-1} \mathbb{1}\{(s_{hi}, a_{hi}, s_{h+1,i}) = (s,a,s')\}}{N_{hk}(s,a)}, \quad \widehat{r}_{hk}(s,a) = \frac{\sum_{i=1}^{k-1} r_{hi} \cdot \mathbb{1}\{(s_{hi}, a_{hi}) = (s,a)\}}{N_{hk}(s,a)}$$

  Planning (by backward induction) for π_{hk} using $\widehat{M}_k$
  for $h = H, \dots, 1$ **do**
    $Q_{h,k}(s,a) = \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s')$
    $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$
  end
  Define $\pi_{h,k}(s) = \arg\max_a Q_{h,k}(s,a)$, $\forall s, h$
  for $h = 1, \dots, H$ **do**
    Execute $a_{hk} = \pi_{hk}(s_{hk})$
    Observe r_{hk} and $s_{h+1,k}$
    $N_{h,k+1}(s_{hk}, a_{hk}) = N_{h,k}(s_{hk}, a_{hk}) + 1$
  end
end

Algorithm 1: UCBVI

A Weissmain inequality

Denote by $\widehat{p}(\cdot|s,a)$ the estimated transition probability build using n samples drawn from $p(\cdot|s,a)$. Then we have that

$$\mathbb{P}(\|\widehat{p}_h(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \varepsilon) \leq (2^S - 2) \exp\Big(-\frac{n\varepsilon^2}{2}\Big)$$

References

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. CoRR, abs/1601.04895, 2016.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In ICML, volume 139 of Proceedings of Machine Learning Research, pages 2783–2792. PMLR, 2021.