Online RL / 23. Tabular MDPs

23. Tabular MDPs

In this lecture we will analize an online learning algorithm for the finite-horizon episodic MDP setting. Let $M = (\mathcal{S}, \mathcal{A}, P^*, r, \mu, H)$ be an MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, *unknown* transition matrix P^*, *known* reward function $r_a(s) \in [0,1]$, an initial state distribution μ, and length of each episode $H \geq 1$. The star-superscript in P^* is used to distinquish the true environment from other (e.g. estimated) environments that occur in the algorithm and the analysis. The assumption that the reward function r is known is for simplicity. In fact, most of the hardness (in terms of sample complexity and designing the algorithm) comes from unknown transition probabilities.

We will focus on the finite-horizon setting where the learner interacts with the MDP over $k = 1, \dots, K$ episodes of length $H \geq 1$. Most, but not all ideas translate to the infinite-horizon discounted or average reward settings.

Recall that the regret is defined as follows:

$$R_K = \sum_{k=1}^{K} v_0^*(S_0^{(k)}) - V_k$$

where $V_k = \sum_{h=0}^{H-1} r_{A_h^{(k)}}(S_h^{(k)})$.

UCRL: Upper Confidence Reinforcement Learning

The UCRL algorithm implements the optimism princple. For this we need to define a set of plausible models. First, we define the maximum likelihood estimates using data from rounds $1, \dots, k-1$:

$$P_a^{(k)}(s, s') = \frac{N_k(s, a, s')}{1 \vee N_k(s, a)}$$

The definition makes use of the notation $a \vee b = \max(a, b)$, and empirical counts:

$$\begin{aligned} N_k(s, a) &= \sum_{k'<k} \sum_{h<H} \mathbb{I}(S_h^{(k)} = s, A_h^{(k)} = a) \\ N_k(s, a, s') &= \sum_{k'<k} \sum_{h<H} \mathbb{I}(S_h^{(k)} = s, A_h^{(k)} = a, S_{h+1}^{(k)} = s') \end{aligned}$$

Define the confidence set

$$C_{k,\delta} = \{P_a(s) \;:\; \forall s,a \; \|P_a^{(k)}(s) - P_a(s)\|_1 \leq \beta_\delta(N_k(s,a))\}$$

where $\beta_\delta : \mathbb{N} \to (0,\infty)$ is a function that we will choose shortly. Our goal of choosing β_δ is to ensure that

1. $P^* \in C_{k,\delta}$ for all $k = 1,\dots,K$ with probability at least $1-\delta$
2. $C_{k,\delta}$ is “not too large”

The second point will appear formually in the proof, however note that from a statistical perspective, we want the confidence set to be as efficient as possible.

With the confidence set, we can now introduce the UCRL algorithm:

UCRL (Upper confidence reinforcement learning):

In episodes $k = 1,\dots,K$,

1. Compute confidence set $C_{k,\delta}$
2. Use policy $\tilde{\pi}_k = \arg\max_\pi \max_{P \in C_{k,\delta}} v_P^\pi$
3. Observe episode data $S_0^{(k)}, A_0^{(k)}, S_1^{(k)}, \dots, S_{H-1}^{(k)}, S_{H-1}^{(k)}, S_H^{(k)}$

Note that we omitted the rewards from the observation data. Since we made the assumption that the reward vector $r_a(s)$ is known, we can always recompute the rewards from the state and action sequence.

For now we we also glance over the point of how to compute the optimistic policy π_k efficently, but we will get back to this point later.

Step 1: Defining the confidence set

Lemma (L1-confidence set): Let $\beta_\delta(u) = 2\sqrt{\frac{S\log(2)+\log(u(u+1)SA/\delta)}{2u}}$ and define the confidence sets

$$C_{k,\delta} = \{P_a(s) \;:\; \forall s,a \; \|P_a^{(k)}(s) - P_a(s)\|_1 \leq \beta_\delta(N_k(s,a))\}$$

Then, with probability at least $1-\delta$,

$$\forall k \geq 1, \quad P^* \in C_{k,\delta}$$

Proof: Let s, a be fixed and denote by $X_v \in \mathcal{S}$ the next state observed upon visiting (s,a) the v^{th} time. Assume that (s,a) was visited in total u times. Then $P_{u,a}(s,s') = \frac{1}{u}\sum_{v=1}^{u} \mathbb{I}(X_v = s')$.

The Markov property implies that $(X_v)_{v=1}^u$ is i.i.d. Note that for any vector $p \in \mathbb{R}^S$ we can write the 1-norm as $\|p\|_1 = \sup_{\|x\|_\infty \leq 1} \langle p, x\rangle$. Therefore

$$\|P_{u,a}(s) - P_a^*(s)\|_1 = \max_{x \in \{\pm 1\}^S} \langle P_{u,a}(s) - P_a^*(s), x\rangle$$

Fix some $x \in \{\pm 1\}^S$.

$$\begin{aligned}\langle P_{u,a}(s) - P_a^*(s), x\rangle &= \frac{1}{u}\sum_{v=1}^{u}\sum_{s'} x_{s'}\big(\mathbb{I}(X_v = s') - P_a^*(s,s')\big)\\ &= \frac{1}{u}\sum_{v=1}^{u}\Delta_v\end{aligned}$$

where in the last line we defined $\Delta_v = \sum_{s' \in \mathcal{S}} x_{s'}\big(\mathbb{I}(X_v = s') - P_a^*(s,s')\big)$. Note that $\mathbb{E}[\Delta_v] = 0$, $|\Delta_v| \leq 1$ and $(\Delta_v)_{v=1}^u$ is an i.i.d. random variable. Therefore Hoeffding's inequality implies that with probability at least $1-\delta$,

$$\frac{1}{u}\sum_{v=1}^{u}\Delta_v \leq 2\sqrt{\frac{\log(1/\delta)}{2u}}$$

Next note that $|\{\pm 1\}^S| = 2^S$, therefore taking the union bound over all $x \in \{\pm 1\}^S$, we get that with probability at least $1-\delta$,

$$\|P_{u,a}(s) - P_a^*(s)\|_1 \leq 2\sqrt{\frac{S\log(2) + \log(1/\delta)}{2u}}$$

In a last step, we take a union bound over $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $u \geq 1$. For taking the union bound over the infinite set of natural numbers, we can use the following simple trick. Note that

$$\sum_{u=1}^{\infty}\frac{\delta}{u(u+1)} = \delta$$

This follows from the simple obseration that $\frac{1}{u(u+1)} = \frac{1}{u} - \frac{1}{u+1}$ and using a telescoping sum argument. Therefore, with probability at least $1-\delta$, for all $u \geq 1$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$

$$\|P_{u,a}(s) - P^*_a(s)\|_1 \leq 2\sqrt{\frac{S\log(2) + \log(u(u+1)SA/\delta)}{2u}}$$

Lastly, the claim follows by noting that $P_a^{(k)}(s) = P_{N_k(s,a),a}(s)$. ■

Step 2: Bounding the regret

Theorem (UCRL Regret): The regret of UCRL defined with confidence sets $C_{k,\delta}$ satisfies with probability at least $1-3\delta$:

$$R_K \leq 4c_\delta H^2\sqrt{SAHK} + c_\delta H^2 SA + 3H\sqrt{\frac{HK}{2}\log(1/\delta)}$$

where $c_\delta = \sqrt{2S\log(2) + \log(HK(HK+1)SA/\delta)}$. In particular, for large enough K, surpressing constants and logarithmic factors, we get

$$R_K \leq \tilde{\mathcal{O}}\big(H^2 S\sqrt{AK\log(1/\delta)}\big)$$

Proof: Denote by π_k the UCRL policy defined as

$$\pi_k = \arg\max_{\pi} \max_{P\in C_{k,\delta}} v^\pi_{0,P}(S_0^{(k)})$$

Further, let $\tilde{P}^{(k)} = \arg\max_{P\in C_{k,\delta}} v^*_{0,P}(S_0^{(k)})$ be the optimistic model.

In what follows we assume that we are on the event $\mathcal{E} = \cap_{k\geq 1} C_{k,\delta}$. By the previous lemma, $\mathbb{P}(\mathcal{E}) \geq 1-\delta$.

Fix $k \geq 1$ and decompose the (instantenous) regret in round k as follows:

$$\begin{aligned} v_0^*(S_0^{(k)}) - V_k = &\underbrace{v^*_{0,P^*}(S_0^{(k)}) - v^*_{0,\tilde{P}_k}(S_0^{(k)})}_{\text{(I)}} \\ &+ \underbrace{v^{\pi_k}_{0,\tilde{P}_k}(S_0^{(k)}) - v^{\pi_k}_{0,P^*}(S_0^{(k)})}_{\text{(II)}} \\ &+ \underbrace{v^{\pi_k}_{0,P^*}(S_0^{(k)}) - V_k}_{\text{(III)}} \end{aligned}$$

Note that we used that $v^*_{0,\tilde{P}_k}(S_0^{(k)}) = v^{\pi_k}_{0,\tilde{P}_k}(S_0^{(k)})$ which holds because by definition π_k is an optimal policy for $\tilde{P}_k$.

The first term is easily bounded. This is the crucial step that makes use of the optimism principle. By $P^* \in C_{k,\delta}$ and the choice of $\tilde{P}_k$ it follows that $(\mathrm{I}) \leq 0$. In particular, we already eliminated the dependence on the (unknown) optimal policy from the regret bound!

The last term is also relatively easy to control. Denote $\xi_k = (\mathrm{III})$. Note that by the definition of the value function we have $\mathbb{E}[\xi_k | S_0^{(k)}] = 0$ and $|\xi_k| \leq H$. Hence ξ_k behaves like noise! If ξ_k was an i.i.d variable we could directly apply Hoeffding's inequality to bound $\sum_{k=1}^K \xi_k$.

The sequence ξ_k has a property that allows us to obtain a similar bound. Let

$$\mathcal{F}_k = \{S_0^{(l)}, A_0^{(l)}, S_1^{(l)}, \dots, S_{H-1}^{(l)}, S_{H-1}^{(l)}, S_H^{(l)}\}_{l=1}^{k-1}$$

be the data available to the learner at the beginning of the episode k. Then by definition of the value function, $\mathbb{E}[\xi_k | \mathcal{F}_k, S_0^{(k)}] = 0$.

A sequence of random variables $(\xi_k)_{k\geq 1}$ with this property is called a martingale difference sequence. Lucky for us, most properties that hold for (zero-mean) i.i.d. sequences can also be shown for martingale difference sequences. The analogue result to Hoeffding's inequality is called the Azuma-Hoeffding's inequalty. Applied to the sequence ξ_k, Azuma-Hoeffdings inequality implies that

$$\sum_{k=1}^K \xi_k \leq H\sqrt{\frac{K}{2}\log(1/\delta)}$$

It remains to bound term (II) in the regret decomposition:

$$(\mathrm{II}) = v_{0,P^*}^{\pi_k}(S_0^{(k)}) - v_{0,\tilde{P}^{(k)}}^{\pi_k}(S_0^{(k)})$$

Using the Bellman equation, we can recursively compute the value function for any policy π:

$$\begin{aligned} v_{h,P}^\pi &= r^\pi + M_\pi P v_{h+1,P}^\pi \;, \quad 0 \leq h \leq H-1 \\ v_{H,P}^\pi &= 0 \end{aligned}$$

We introduce the following shorthand for the value difference of policy π_k under models P^* and $\tilde{P}^{(k)}$:

$$\delta_h^{(k)} = v_{h,\tilde{P}^{(k)}}^{\pi_k}(S_h^{(k)}) - v_{h,P^*}^{\pi_k}(S_h^{(k)})$$

Let $\mathcal{F}_{h,k}$ contain all observation data up to episode k and step h including S_h^k. Using the Bellman equation, we can write

$$\begin{aligned}
\delta_h^{(k)} &= M_{\pi_k}\tilde{P}^{(k)} v^{\pi_k}_{h+1,\tilde{P}^{(k)}}(S_h^{(k)}) - M_{\pi_k}P^* v^{\pi_k}_{h+1,P^*}(S_h^{(k)}) \pm M_{\pi_k}P^* V_{h+1,\tilde{P}^{(k)}}(S_h^{(k)}) \\
&= (M_{\pi^k}(\tilde{P}^{(k)} - P^*) v^{\pi_k}_{h+1,\tilde{P}^{(k)}})(S_h^{(k)}) + (M_{\pi_k}P^*(v^{\pi_k}_{h+1,\tilde{P}^{(k)}} - v^{\pi_k}_{h+1,P^*})(S_h^{(k)}) \\
&\le \|P^*_{A_h^{(k)}}(S_h^{(k)}) - \tilde{P}^{(k)}_{A_h^{(k)}}(S_h^{(k)})\|_1 H + \delta_{h+1}^{(k)} + \underbrace{\left(\mathbb{E}[\delta_{h+1}^{(k)}|\mathcal{F}_{h,k}] - \delta_{h+1}^{(k)}\right)}_{=:\eta_{h+1}^{(k)}} \\
&\le 2H\beta_\delta(N_k(S_h^{(k)}, A_h^{(k)})) + \delta_{h+1}^{(k)} + \eta_{h+1}^{(k)}
\end{aligned}$$

The first inequality uses that for any two vectors w, v, we have $\langle w, v\rangle \le \|w\|_1\|v\|_\infty$ and $\|v^{\pi_k}_{h+1,\tilde{P}^{(k)}}\|_\infty \le H$. Further we use that π_k is a deterministic policy, therefore $M_{\pi_k}P(S_h^{(k)}) = P_{A_h^{(k)}}(S_h^{(k)})$. The second follows from the definition of the confidence set in the previous lemma:

$$\begin{aligned}
&\|P^*_{A_h^{(k)}}(S_h^{(k)}) - \tilde{P}^{(k)}_{A_h^{(k)}}(S_h^{(k)})\|_1 \\
&\le \|P^*_{A_h^{(k)}}(S_h^{(k)}) - P^{(k)}_{A_h^{(k)}}(S_h^{(k)})\|_1 + \|P^{(k)}_{A_h^{(k)}}(S_h^{(k)}) - \tilde{P}^{(k)}_{A_h^{(k)}}(S_h^{(k)})\|_1 \\
&\le 2\beta_\delta(N_k(S_h^{(k)}, A_h^{(k)}))
\end{aligned}$$

Telescoping and using that $\delta_H^{(k)} = 0$ yields

$$\delta_0^{(k)} \le \eta_1^{(k)} + \dots + \eta_{H-1}^{(k)} + 2H\underbrace{\sum_{h=0}^{H-1} \beta_\delta(N_k(S_h^{(k)}, A_h^{(k)}))}_{(\mathrm{IV})}$$

Note that $(\eta_h^{(k)})_{h=1}^{H-1}$ is another martingale difference sequence (with $\eta_h^{(k)}| \le H$) that can be bounded by Azuma-Hoeffding:

$$\sum_{k=1}^K \sum_{h=1}^{H-1} \eta_h^{(k)} \le 2H\sqrt{\frac{HK}{2}\log(1/\delta)}$$

It remains to bound term (IV). For this we make use of the following algebraic lemma:

Lemma:

For any sequence $m_1, \dots, m_k$ that satisfies $m_1 + \dots + m_k \ge 0$:

$$\sum_{k=1}^K \frac{m_k}{\sqrt{1 \vee (m_1 + \dots + m_k)}} \le 2\sqrt{m_1 + \dots + m_k}$$

Proof of Lemma: Let $f(x) = 1/\sqrt{x}$. $f(x)$ is a concave function on $(0, \infty)$. Therefore $f(A + x) \leq f(A) + xf'(A)$ for all $A, A + x, > 0$. This translates to:

$$\sqrt{A + x} \leq \sqrt{A} + \frac{x}{2\sqrt{A}}$$

The claim follows from telescoping. ■

Continuing the proof of the theorem where we need to bound (IV). Denote $c_\delta = \sqrt{2S\log(2) + \log(HK(HK + 1)SA/\delta)}$. Further let $M_k(s, a) = \sum_{h=1}^{H-1} \mathbb{I}(S_h^{(k)} = s, A_h^{(k)} = a)$ and note that $N_k(s, a) = M_1 + \dots + M_{k-1}$. Then

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=0}^{H-1} \beta_\delta(N_k(S_h^{(k)}, A_h^{(k)})) &\leq c_\delta \sum_{s,a} \sum_{k=1}^{K} \sum_{h=0}^{H-1} \frac{\mathbb{I}(S_h^{(k)} = s, A_h^{(k)} = a)}{\sqrt{1 \vee N_k(s, a)}} \\
&= c_\delta \sum_{s,a} \sum_{k=1}^{K} \frac{M_k}{\sqrt{1 \vee (M_1 + \dots + M_{k-1})}}
\end{aligned}$$

Next, using the algebraic lemma above and the fact that $M_k(s, a) \leq H$, we find

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=0}^{H-1} \beta_\delta(N_k(S_h^{(k)}, A_h^{(k)})) &\leq c_\delta \sum_{s,a} \sum_{k=1}^{K} \frac{M_k(s, a)}{\sqrt{1 \vee (M_1(s, a) + \dots + M_{k-1}(s, a))}} \\
&\leq c_\delta \sum_{s,a} \sum_{k=1}^{K} \frac{M_k(s, a)}{\sqrt{1 \vee (M_1(s, a) + \dots + M_k(s, a) - H)}} \\
&\leq c_\delta \sum_{s,a} \sum_{k=1}^{K} \frac{M_k(s, a)\mathbb{I}(M_1(s, a) + \dots + M_k(s, a) > H)}{\sqrt{M_1(s, a) + \dots + M_k(s, a) - H}} + c_\delta HSA \\
&\leq 2c_\delta \sum_{s,a} \sqrt{N_k(s, a)} + c_\delta HSA \\
&\leq 2c_\delta SA \sqrt{\sum_{s,a} N_k(s, a)/SA} + c_\delta HSA \\
&= 2c_\delta \sqrt{SAHK} + c_\delta HSA
\end{aligned}$$

The last inequality uses Jensen's inequality.

Collecting all terms and taking the union bound over two applications of Azuma-Hoeffdings and the event $\mathcal{E}$ completes the proof. ■

Unknown reward functions

In our analysis of UCRL we assumed that the reward function is known. While this is quite a common assumption in the literature, it is mainly for simplicity. We also don't expect the bounds to change by much: Estimating the rewards is not harder than estimating the transition kernels.

To modify the analysis and account for unkown rewards, we first consider the case with deterministic reward function $r_a(s) \in [0, R_{\max}]$, where $R_{\max}$ is some known upper bound on the reward per step.

Embracing the idea of optimism, we define reward estimates

$$\hat{r}_a^{(k)}(s) = \begin{cases} r_{A_h^{(k')}}(S_h^{(k')}) & \text{(s,a) was visited in a round } k' < k \text{ and step } h \\ R_{\max} & \text{else.} \end{cases}$$

Clearly this defines an optimistic estimate, $\hat{r}_a^{(k)}(s) \geq r_a(s)$. Moreover, we have $\hat{r}^{(k)}_{A_h^{(k)}}(S_h^{(k)}) \neq r_{A_h^{(k)}}(S_h^{(k)})$ at most SA times. Therefore the regret in the previous analysis is increased by at most $R_{\max} SA$.

When the reward is stochastic, we can use a maximum likelihood estimate of the reward and construct confidence bounds around the estimate. This way we can define an optimistic reward. Still not much changes, as the reward estimates concentrate at the same rate as the estimates of P.

UCBVI: Upper Confidence Bound Value Iteration

Computing the UCRL policy can be quite challenging. However, we can relax the construction so that we can use backward induction. We define a time-inhomogenous relaxation of the confidence set:

$$C_{k,\delta}^H = \underbrace{C_{k,\delta} \times \cdots \times C_{k,\delta}}_{H \text{ times}}$$

Let $\tilde{P}_{1:H,k} := (\tilde{P}_{1,k}, \dots, \tilde{P}_{H,k}) = \arg\max_{P \in C_{k,\delta}^H} v_P^*(s_0^{(k)})$ be the optimistic (time-inhomogenous) transition matrices and $\pi_k = \arg\max_\pi v^\pi_{\tilde{P}_{1:H,k}}$ the optimal policy for the optimistic model $\tilde{P}_{1:H,k}$. Then $v^{\pi^k}_{\tilde{P}_{1:H,k}} = v^*_{\tilde{P}_{1:H,k}} = v^{(k)}$ is defined by the following backwards induction:

$$\begin{aligned} v_H^{(k)}(s) &= 0 \qquad \forall s \in [S] \\ Q_h^{(k)}(s,a) &= r(s,a) + \max_{P \in C_{k,\delta}} P_a(s) v_{h+1}^{(k)} \\ v_h^{(k)}(s) &= \max_a Q_h^{(k)}(s,a) \end{aligned}$$

Note that the maximum in the second line is a linear optimization with convex constraints that can be solved efficiently. Further, the proof of the UCRL regret still applies, because we used the same (step-wise) relaxation in the analysis.

We can further relax the backward induction to avoid the optimization over $C_{k,\delta}$ completely:

$$\begin{aligned}\max_{P\in C_{k,\delta}} P_a(s)v_{h+1}^{(k)} &\leq P_a^{(k)}(s)v_{h+1}^{(k)} + \max_{P\in C_{k,\delta}}(P_a(s) - P_a^{(k)}(s))v_{h+1}^{(k)}\\ &\leq P_a^{(k)}(s)v_{h+1}^{(k)} + \max_{P\in C_{k,\delta}} \|P_a(s) - P_a^{(k)}(s))\|_1 \|v_{h+1}^{(k)}\|_\infty\\ &\leq P_a^{(k)}(s)v_{h+1}^{(k)} + \beta_\delta(N_k(s,a))H\end{aligned}$$

This leads us to the the UCBVI (upper confidence bound value iteration) algorithm. In episode k, UCBVI uses value iteration for the estimated transition kernel $P_a^{(k)}(s)$ and optimistic reward function $r_a(s) + H\beta_\delta(N_k(s,a))$ to compute the policy.

UCBVI (Upper confidence bound value iteration):

In episodes $k = 1, \dots, K$,

1 Compute optimistic value function:

$$\begin{aligned}v_H^{(k)}(s) &= 0 \qquad \forall s \in [S]\\ b_k(s,a) &= H\beta_\delta(N_k(s,a))\\ Q_h^{(k)}(s,a) &= \min\left(r(s,a) + b_k(s,a) + P_a^{(k)}(s)v_{h+1}^{(k)}, H\right)\\ v_h^{(k)}(s) &= \max_a Q_h^{(k)}(s,a)\end{aligned}$$

1 Follow greedy policy $A_h^{(k)} = \arg\max_A Q_h^{(k)}(S_h^{(k)}, A)$

2 Observe episode data $S_0^{(k)}, A_0^{(k)}, S_1^{(k)}, \dots, S_{H-1}^{(k)}, S_{H-1}^{(k)}, S_H^{(k)}$

Note that we truncate the $Q_h^{(k)}$-function to be at most H, this avoids a blow up by a factor of H in the regret bound. Carefully checking that the previous analysis still applies shows that UCBVI has regret at most $R_K \leq \mathcal{O}(H^2 S\sqrt{AK})$.

By more carefully designing the reward bonuses for UCBVI, it is possible to achieve $R_K \leq \tilde{\mathcal{O}}(H^{3/2}\sqrt{SAK})$ which matches the lower bound up to logarithmic factors in the time in-homogeneous setting.

Notes

References

The original UCRL paper. Notice that they consider the infinite horizon average reward setting, which is different from the episodic setting we present.

Auer, P., & Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. Advances in neural information processing systems, 19. [link]

The UCBVI paper. Notice that they consider the homogeneous setting, which is different from the in-homogeneous setting we present.

Azar, M. G., Osband, I., & Munos, R. (2017, July). Minimax regret bounds for reinforcement learning. In International Conference on Machine Learning (pp. 263-272). PMLR. [link]

The paper that presents the lower bound. Notice the they consider the infinite horizon average reward setting. Thus, there results contains a diameter term D instead of a horizon term of H.

Auer, P., Jaksch, T., & Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. Advances in neural information processing systems, 21. [link]