22. Introduction

PDF Version

Online learning in reinforcement learning refers to the idea that a learner is placed in an (initially) *unknown* MDP. By interacting with the MDP, the learner collects data about the unknown transition and reward function. The learner's goal is to collect as much reward as possible, or output a near-optimal policy. The difference to planning is that the learner does *not* have access to the true MDP. Unlike in batch RL, the learner gets to decide what actions to play. Importantly, this means the learner's action affect the data that is available to the learner (sometimes refered to as "closed loop").

The fact that the learner needs to create its own data leads to an important decision: Should the learner sacrifice reward to collect more data that will improve decision making in the future? Or should it act according to what seems currently best? Clearly, too much exploration will be costly if the learner chooses actions with low reward too often. On the other hand, playing actions that appear optimal with limited data comes at the risk of missing out on even better rewards. In the literature, this is commonly known as *exploration-exploitation dilemma*.

The exploration-exploitation dilemma is not specific to the MDP setting. It already arises in the simpler (multi-armed) bandit setting (i.e. an MDP with only one state and stochastic reward).

In the following, we focus on finite-horizon episodic (undiscounted) MDPs . The learner interacts with the MDP for episodes of length . At the beginning of each episode , an initial state is sampled from the initial distribution . The data collected during the episode is

where is the action chosen by the learner at step , is the next state and is the (possibly stochastic) reward.

This model contains some important settings as a special case. Most notably,

- recovers the contextual bandit setting, where the "context" is sampled from the distribution
- and is the finite multi-armed bandit setting.

Sample complexity and regret: How good is the learner?

The goal of the learner is to collect as much reward as possible. We denote as the reward collected by the learner in episode . The total reward is . For the analysis it will be useful to introduce a normalization: Instead of directly arguing about the total reward, we compare the learner to the value of the best policy in the MDP. This leads to the notation of *regret* defined as follows:

A learner has sublinear expected regret if as . Sublinear regret means that the average reward of the learner approaches the optimal value as the number of episodes increases. Certainly that is a desirable property!

Before we go on to construct learners with small regret, we briefly note that there are also other objectives. The most common alternative is PAC - which stands for *probably approximately correct*. A learner is said to be -PAC if upon termination in episode , it outputs a policy such that with probability at least . We have discussed PAC bounds already in the context of planning.

The difference to bounding regret is that in the first episodes, the learner does not 'pay' for choosing suboptimal actions. This is sometimes called a *pure exploration problem*. Note that a learner that achieves sublinear regret can be converted into a PAC learner (discussed in the notes). However, this may lead to a suboptimal (large) in the PAC framework.

-greedy

There exist many ideas on how to design algorithms with small regret. We first note that a "greedy" agent can easily fail: Following the best actions according to some empirical

estimate can easily get you trapped in a supoptimal policy (think of some examples where this can happen!).

A simple remedy is to add a small amount of "forced" exploration: With (small) probability , we choose an action uniformly at random. Thereby we eventually collect samples from all actions to improve our estimates. With probabilty we follow the "greedy" choice, that is the action that appears best under our current estimates. This gives raise to the name -greedy.

It is often possible to show that -greedy converges. By carefully choosing the exploration probability , we may show that in finite MDPs, the regret is at most . As we will discuss later, there are multiple algorithms that achieve a regret of only . Thus, -greedy is not the best algorithm to minimize regret.

Not unexpectedly, this type of exploration can be quite sub-optimal. It is easy to construct examples, where -greedy takes exponential time (in the number of states) to reach an optimal policy. Can you find an example (Hint: construct the MDP such that each time the agent explores a suboptimal action, the agent is reset to the starting state)?

On the upside, -greedy is very simple and can easily used in more complex scenarios. In fact, it is a popular choice when using neural network function approximations, where theoretically grounded exploration schemes are much harder to obtain.

Optimism Principle

A popular technique to construct regret minimizing algorithms is based on *optimism in the face of uncertainty*. To formally define the idea, let be the set of possible environments (e.g. finite MDPs). We make the realizability assumption that the true environment is in this set. After obtaining data in rounds , the learner uses the observations to compute a set of plausible models . The plausible model set is such that it contains the true model with high probabilty. Although this is not always required, it is useful to think of a decreasing sequence of sets . This simply means that as more data arrives, the learner is able to exclude models that are statistically unlikely to produce the observation data.

The optimism principle is to act according to the policy that achieves the highest reward among all plausible models, i.e.

At this point it not be clear why this leads to an efficient learning algorithm (with small

regret). The idea is that the learner systematically obtains data about the environment. For example, if data contradicts the optimistic model , then is excluded from the set of plausible models in the future. Consequently, the learner chooses a different policy in the next round.

On the other hand, the learner ensures that with high probability. In this case, it is often possible to show that the gap is small (more specifically, behaves like a statistical estimation error of order with a leading constant that depends on the "size" of).

One should also ask if the optimization problem can be solved efficiently. This is far from always the case. Often one needs to rely on heuristics to implement the optimistic policy, or use other exploration techniques such as Thompson sampling (see below).

How much regret the learner has of course depends on the concrete setting at hand. In the next lecture we will see how we can make use of optimism to design (and analyize) an online learning algorithm for finite MDPs. The literature has produced a large amount of papers with algorithms that use the optimism principle in many settings. This however does not mean that optimism is a universal tool. More recent literature has also pointed out limitations of the optimsm principle, and in lieu proposed other design ideas.

Notes

Other Exploration Techniques

Some other notable exploration strategies are:

- Phased-Elimination and Experimental Design
- Thompson Sampling
- Information-Directed Sampling (IDS) and Estimation-To-Decisions (E2D)

References

The paper showing the details behind how to convert between Regret and PAC bounds.

- Dann, C., Lattimore, T., & Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. Advances in Neural Information Processing Systems, 30. [link]

23. Tabular MDPs

PDF Version

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:

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

Define the confidence set

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

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

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

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

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

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 $\|p\|_1 = \sup_{\|x\|_\infty \le 1}\langle p, x\rangle$. 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 $\mathbb{E}[\Delta_v] = 0$, $|\Delta_v| \le 1$ and $(\Delta_v)_{v=1}^u$ is an i.i.d. random variable. Therefore Hoeffding's inequality implies that with probability at least $1-\delta$,

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

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 $1-\delta$,

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

In a last step, we take a union bound over $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $u \ge 1$. 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 $1-3\delta$:

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

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

$$R_K \leq \tilde{\mathcal{O}}\left(H^{3/2}S\sqrt{AK\log(1/\delta)}\right)$$

Proof: Denote by π_k the UCRL policy defined as

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

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:

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

Note that we used that $v_{0,\tilde{P}_k}^*(S_0^{(k)}) = v_{0,\tilde{P}_k}^{\pi_k}(S_0^{(k)})$ 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 $\text{(I)} \leq 0$. 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 $\xi_k = \text{(III)}$. 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:

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

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

$$v^{\pi}_{h,P} = r^{\pi} + M_{\pi} P v^{\pi}_{h+1,P}\ , \quad 0 \le h \le H-1$$
$$v^{\pi}_{H,P} = 0$$

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

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

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

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

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 $m_1 + \cdots + m_k \geq 0$:

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

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 $N_k(s,a) = M_1 + \cdots + M_{k-1}$. Then

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

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

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

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, $\hat{r}_a^{(k)}(s) \ge r_a(s)$. Moreover, we have $\hat{r}^{(k)}_{A_h^{(k)}}(S_h^{(k)}) \ne r_{A_h^{(k)}}(S_h^{(k)})$ 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 $\pi_k = \arg\max_\pi v_{\tilde{P}_{1:H,k}}^\pi$ the optimal policy for the optimistic model $\tilde{P}_{1:H,k}$. Then $v_{\tilde{P}_{1:H,k}}^{\pi^k} = v_{\tilde{P}_{1:H,k}}^* = v^{(k)}$ 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:

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

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 $S_0^{(k)}, A_0^{(k)}, S_1^{(k)}, \ldots, S_{H-1}^{(k)}, S_{H-1}^{(k)}, S_H^{(k)}$

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]

24. Featurized MDPs

PDF Version

In tabular (finite-horizon) MDPs $M = (\mathcal{S}, \mathcal{A}, P, r, \mu, H)$, roughly speaking, the learner has to learn about reward and transition probabilities for *all* states and actions in the worst-case. This is reflected in lower bounds on the regret that scale with $R_K \geq \Omega(H^{3/2}\sqrt{ASK})$ (in the time in-homogeneous case).

In many applications the state space can be huge, and reinforcement learning is often used together with function approximation. In such settings, we want to avoid bounds that scale directly with the number of states S. The simplest parametric models often rely on state-action features and linearly parametrized transition and reward functions. The goal is to obtain bounds that scale with the complexity of the function class (e.g. the feature dimension in linear models), and are *independent* of S and A.

Historically, many ideas for online learning in linear MDP models are borrowed from the linear bandit model. Beyond what is written here, you may find it helpful to read about stochstic linear bandits and LinUCB (see chapters 19 and 20 of the Bandit Book).

Linear Mixture MDPs

We focus on the episodic, finite-horzion MDPs $M = (\mathcal{S}, \mathcal{A}, P_h, r_h, \mu, H)$ with time in-homogenous reward r_h and transition matrix P_h. We let $\mathcal{S}$ be a finite but possibly very large state space, and $\mathcal{A}$ be a finite action space. With care, most of the analysis can be extended to infinite state and action spaces. As before, we assume that the reward function $r_h(s,a) \in [0,1]$ is known.

We now impose additional (linear) structure on the transition kernel P_h. For this we assume the learner has access to features $\phi(s,a,s') \in \mathbb{R}^d$ that satisfy $\|\phi(s,a,s')\|_2 \leq 1$. In time-inhomogeneous *linear mixture MDPs*, the transition kernel is of the form

$$P_{h,a}(s,s') = \langle \phi(s,a,s'), \theta_h^* \rangle$$

for some unkown parameter $\theta_h^* \in \mathbb{R}^d$ with $\|\theta_h^*\|_2 \leq 1$. We remark that tabular MDPs are recovered using $\phi(s,a,s') = e_{s,a,s'}$, where $e_{s,a,s'}$ are the unit vectors in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$.

For any function $V : \mathcal{S} \to \mathbb{R}$, we define

$$\phi_V(s,a) = \sum_{s'} \phi(s,a,s')V(s') \in \mathbb{R}^d$$

Note that $\langle \phi_V(s,a), \theta^* \rangle$ predicts the expected value of $V(s')$ when s' is sampled from $P_{h,a}(s)$:

$$P_{h,a}(s)V = \sum_{s'} P_{h,a}(s,s')V(s') = \sum_{s'} \langle \phi(s,a,s'), \theta^*_h \rangle V(s') = \langle \phi_V(s,a), \theta^*_h \rangle$$

Value Targeted Regression (VTR)

Now that we have specified the parametrized model, the next step is to construct an estimator of the unknown paramter. An estimator of θ^* allows us to predict the value of any policy. For the algorithm, we are particularly interested in constructing optimistic estimates of the value function. Hence we will also need a confidence set.

Let $(V_h^{(j)})_{h \le H}^{j<k}$ be a sequence of value functions constructed up to episode $k-1$. Let $\phi_{h,j} = \phi_{V_{h+1}^{(j)}}(S_h^{(j)}, A_h^{(j)})$ and $y_{h,j} = V_{h+1}^{(j)}(S_{h+1}^{(j)})$. By constrution, we have that $\mathbb{E}[y_{h,j}] = \langle \phi_{h,j}, \theta^* \rangle$ and $|y_{h,j}| \le H$. Define the *regularized least-squares estimator*

$$\hat{\theta}_{h,k} = \arg\min_{\theta} \sum_{j=0}^{k-1} \left(\langle \phi_{h,j}, \theta \rangle - y_{h,j} \right)^2 + \lambda \|\theta\|^2$$

Let $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ be the idenity matrix. We have the following closed form for $\hat{\theta}_{h,k}$:

$$\hat{\theta}_{h,k} = \Sigma_{h,k}^{-1} \sum_{j=0}^{k-1} \phi_{h,j} y_{h,j} \qquad \text{where} \quad \Sigma_{h,k} = \sum_{j=0}^{k-1} \phi_{h,j}\phi_{h,j}^\top + \lambda \mathbf{I}_d$$

The next step is to quantify the uncertainy in the estimation. Mirroring the steps in the tabular setting, we construct a confidence set for $\hat{\theta}_{h,k}$.

For a positive (semi-)definite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and vector $v \in \mathbb{R}^d$, define the (semi-)norm $\|a\|_\Sigma = \sqrt{\langle v, \Sigma v \rangle}$. We make use of the following elliptical confidence set for $\hat{\theta}_{h,k}$

$$C_{h,\delta}^{(k)} = \{\theta : \|\theta - \hat{\theta}_{h,k}\|_{\Sigma_{h,k}}^2 \le \beta_{h,k,\delta}\}$$

where

$$\beta_{h,k,\delta}^{1/2} = H\sqrt{\log\det(\Sigma_{h,k}) - \log\det(\Sigma_{h,0}) + 2\log(1/\delta)} + \sqrt{\lambda}$$

The log determinant of $\Sigma_{h,k}$ can be computed online by the algorithm. For the analysis, it is useful to further upper bound $\beta_{h,k,\delta}$. It is possible to show the following upper bound on $\beta_{h,k,\delta}$ that holds independent of the data sequence:

$$\beta_{h,k,\delta}^{1/2} \leq H\sqrt{d\log(1+k/(d\lambda)) + 2\log(1/\delta)} + \sqrt{\lambda}$$

For a derivation of the above inequality see Lemma 19.4 of the Bandit Book. The next lemma formally specifies the confidence probabilty.

Lemma (Online Least-Squares Confidence) Fix some $0 \leq h < H$. Then

$$\mathbb{P}[\theta_h^* \in \cap_{k\geq 1} C_{h,k,\delta}] \geq 1-\delta$$

Proof: The above result is presented as Theorem 2 in Abbasi-Yadkori et al (2011), where the proof can also be found. ■

The confidence set can be used to derive bounds on the estimation error with probability at least $1-\delta$ as follows:

$$|\langle \phi_V(s,a), \hat{\theta}_{h,k} - \theta^* \rangle| \leq \|\phi_V(s,a)\|_{\Sigma_{h,k}^{-1}} \|\hat{\theta}_{h,k} - \theta^*\|_{\Sigma_{h,k}} \leq \beta_{h,k,\delta}^{1/2} \|\phi_V(s,a)\|_{\Sigma_{h,k}^{-1}}$$

The first inequality is by Cauchy-Schwarz and the second inequality uses the confidence bound from the previous lemma.

UCRL-VTR

Similar to the tabular UCRL and UCBVI algorithms, UCRL-VTR uses the estimates $\hat{\theta}_{h,k}$ to compute an optimistic policy. One way of obtaining an optimistic policy is from optimistic Q-estimates $Q_h^{(k)}(s,a)$ defined via backwards induction. Then UCRL-VTR follows the greedy policy w.r.t. the optimistic Q-values.

UCRL-VTR

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

1 Set $V_H^{(k)}(s) = 0$. Compute $\hat{\theta}_{h,k}$ and $\Sigma_{h,k}$. Recursively define optimistic value functions

For $h = H-1, \ldots, 0$:

$$\hat{\theta}_{h,k} = \arg\min_{\theta} \sum_{j=1}^{k-1} \big(\langle \phi_{h,j}, \theta \rangle - y_{h,j}\big)^2 + \lambda \|\theta\|_2^2$$

$$\Sigma_{h,k} = \sum_{j=1}^{k} \phi_{h,j} \phi_{h,j}^\top + \lambda \mathbf{I}_d$$

$$Q_h^{(k)}(s,a) = \big(r_h(s,a) + \langle \phi_{V_{h+1}^{(k)}}(s,a), \theta_{h,k} \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(s,a)\|_{\Sigma_{h,k}^{-1}}\big) \wedge H$$

$$V_h^{(k)}(s) = \max_a Q_h^{(k)}(s,a)$$

2 Follow greedy policy w.r.t. $Q_h^{(k)}(s,a)$.

For $h = 0, \dots, H-1$:

$$A_h^{(k)} = \arg\max_{a \in \mathcal{A}} Q_h^{(k)}(S_h^{(k)}, a)$$

Let $\phi_{h,k} = \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})$ and $y_{h,k} = V_{h+1}^{(k)}(S_{h+1}^{(k)})$.

We are now in the position to state a regret bound for UCRL-VTR.

Theorem (UCRL-VTR Regret) The regret of UCRL-VTR satisfies with probability at least $1-2\delta$:

$$R_K \le \mathcal{O}\big(dH^2 \log(K) \sqrt{K \log(KH/\delta)}\big)$$

Note that the bound scales with the feature dimension d, but not the size of the state space or action space. The lower bound for this setting is $R_K \ge \Omega(dH^{3/2}\sqrt{K})$, therefore our upper bound is tight except for a factor $\sqrt{H}$.

Proof:

Our proof strategy follows the same steps as in the proof of UCRL.

Step 1 (Optimism):

Taking the union bound over $h = 0, \dots, H-1$, the previous lemma implies that with probability at least $1-\delta$, for all $h \in [H-1]$ and all $k \ge 0$, $\theta_h^* \in C_{h,\delta/H}^{(k)}$. In the following, we condition on this event. Using induction over $h = H, H-1, \dots, 0$, we can show that

$$V_0^*(S_h^{(k)}) \le V_0^{(k)}(S_h^{(k)})$$

Step 2 (Bellman recursion and estimation error):

For any $h = 0, \dots, H-1$, we find

$$\begin{aligned}
&V_h^{(k)}(S_h^{(k)}) - V_h^{\pi_k}(S_h^{(k)}) \\
&\le \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} - P^*_{h,A_h^{(k)}}(S_h^{(k)}) V_{h+1}^{\pi_k} \\
&= \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} - \theta^* \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k})
\end{aligned}$$

The inequality is by the definition of $V_h^{(k)}$ and dropping the truncation, and in the last line we add and subtract $P^*_{h,A_h^{(k)}}(S_h^{(k)}) V_{h+1}^{(k)} = \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \theta^* \rangle$. Further, by Cauchy-Schwarz on the event $\theta^* \in C_{k,\delta/H}$ we get

$$\langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} - \theta^* \rangle \le \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}}$$

Continuing the previous display, we find

$$\begin{aligned}
&V_h^{(k)}(S_h^{(k)}) - V_h^{\pi_k}(S_h^{(k)}) \\
&\le 2\beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k}) \\
&= 2\beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + V_{h+1}^{(k)}(S_{h+1}^{(k)}) - V_{h+1}^{\pi_k}(S_{h+1}^{(k)}) + \xi_{h,k}
\end{aligned}$$

where we defined

$$\xi_{h,k} = \big(P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k})\big) - \big(V_{h+1}^{(k)}(S_{h+1}^{(k)}) - V_{h+1}^{\pi_k}(S_{h+1}^{(k)})\big)$$

Recursively appliying the previous inequality and summing over all episodes yields

$$\sum_{k=1}^{K} V_0^{(k)}(S_0^{(k)}) - V_0^{\pi_k}(S_0^{(k)}) \le \sum_{k=1}^{K} \sum_{h=0}^{H-1} 2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} + \xi_{h,k}$$

Note that $\xi_{h,k}$ is a martingale difference sequence, hence by Azuma-Hoeffdings inequality we have with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \sum_{h=0}^{H-1} \xi_{h,k} \le H\sqrt{\frac{HK}{2}\log(1/\delta)}$$

Step 3 (Cauchy-Schwarz):

Note that $\beta_{h,k,\delta}$ is non-decreasing in both h and k. Very little is lost by bounding $\beta_{h,k,\delta} \le \beta_{H,K,\delta}$. From the previous step, we are left to bound the sum over uncertainties $\|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}$. We start with an application of the Cauchy-Schwarz inequality. Applied to

sequences $(a_i)_{i=1}^n$, $(b_i)_{i=1}^n$, we have that $|\sum_{i=1}^n a_i b_i| \leq \sqrt{\sum_{i=1}^n a_i^2 \sum_{j=1}^n b_i^2}$. Applied to the regret, we get:

$$\sum_{k=1}^K \sum_{h=0}^{H-1} 2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} \leq \sum_{h=0}^{H-1} 2\beta_{h,K,\delta}^{1/2} \sqrt{K \sum_{k=1}^K \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2}$$

Step 4 (Elliptic potential lemma):

The penultima step is to control the sum over squared uncertainties $\|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2$. This classical result is sometimes refered to as the elliptic potential lemma:

$$\sum_{k=1}^K \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2 \leq \mathcal{O}(d \log(K))$$

The proof, as mentioned earlier, can be found as Lemma 19.4 in the Bandit Book.

Step 5 (Summing up):

It remains to chain the previous steps and take the union bound over the event where the confidence set contains the true parameter and the application of Azuma-Hoeffdings.

$$\begin{aligned} R_K &= \sum_{k=1}^K V_0^{(k)}(S_0^{(k)}) - V_0^{\pi_k}(S_0^{(k)}) \\ &\leq \sum_{k=1}^K \sum_{h=0}^{H-1} \left(2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} + \xi_{h,k}\right) \\ &\leq C \cdot H \beta_{H,K,\delta}^{1/2} \sqrt{d \log(K) K} + H^{3/2} \sqrt{2K \log(1/\delta)} \end{aligned}$$

For some universal constant C. This completes the proof. ■

Linear MDPs

So far we have seen the linear mixture MDP model. This is not the only way one can parameterize the transition matrix. An alternative is the *linear MDP* model, defined as follows for features $\phi(s,a) \in \mathbb{R}^d$ and parameters $\psi_h^* \in \mathbb{R}^{d \times S}$ and $\theta_h^* \in \mathbb{R}^d$:

$$\begin{aligned} P_h^*(s,s') &= \langle \phi(s,a), \psi_h^*(s') \rangle \\ r_h(s,a) &= \langle \phi(s,a), \theta_h^* \rangle \end{aligned}$$

Note that tabular MDPs are recovered using $\phi(s,a) = e_{s,a}$, where $e_{s,a}$ are the unit vectors in $\mathbb{R}^{\mathcal{S} \times \mathcal{A}}$.

Compared to the linear mixture model, an immediate observation is that the dependence on the the next state s' is pushed into the parameter $\psi_h(s') \in \mathbb{R}^d$. Consequently, the dimension of the parameter space scales with the number of states, and it is not immediately clear how we can avoid the S dependence in the regret bounds.

Another consequence of this model is that the Q-function for *any* policy is linear in the features $\phi(s,a)$.

Lemma:

Under the linear MDP assumption, for any policy π the Q-function $Q_h^\pi(s,a)$ is linear in the features $\phi(s,a)$. That is, there exist parameters $w_h^\pi \in \mathbb{R}^d$ such that

$$Q_h^\pi(s,a) = \langle \phi(s,a), w_h^\pi \rangle$$

Proof: The claim follows directly from the definition of Q_h^π and the assumptions on $r_h(s,a)$ and $P_{h,a}(s)$.

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + P_{h,a}(s) V_{h+1}^\pi \\ &= \langle \phi(s,a), \theta_h^* \rangle + \sum_{s'} V_{h+1}^\pi(s') \langle \phi(s,a), \psi_h^*(s') \rangle \\ &= \langle \phi(s,a), w_h^\pi \rangle \end{aligned}$$

where we defined $w_h^\pi = \theta_h^* + \sum_{s'} \psi_h^*(s') V_{h+1}^\pi(s')$ for the last equation. ■

In light of this lemma, our goal is to estimate $w_h^{\pi^*}$. This can be done using *least-squares value iteration* (LSVI). Let $\{S_1^{(j)}, A_1^{(j)}, \dots, S_{H-1}^{(j)}, A_{H-1}^{(j)}, S_H^{(j)}\}_{j=1}^{k-1}$ be the data available at the beginning of episode k. Denote $\phi_{h,j} = \phi(S_h^{(j)}, A_h^{(j)})$ and define targets $y_{h,j} = r_h(S_h^{(j)}, A_h^{(j)}) + \max_{a \in \mathcal{A}} Q_{h+1}^{(j)}(S_h^{(j)}, a)$ based on $Q_{h+1}^{(j)}(s,a)$ estimates obtained in episodes $j = 1, \dots, k-1$.

Least-squares value iteration solves the following problem:

$$\hat{w}_{h,k} = \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left(\langle \phi_{j,h}, w \rangle - y_{j,h} \right)^2 + \lambda \|w\|_2^2$$

The closed form solution is $w_{h,k} = \Sigma_{h,k}^{-1} \sum_{j=1}^{k-1} \phi_{h,j} y_{h,j}$ where $\Sigma_{h,k} = \sum_{j=1}^{k-1} \phi_{j,h} \phi_{j,h}^\top + \lambda \mathbf{I}_d$.

Based on the estimate $\hat{w}_{h,k}$, we can define optimistic Q- and V-estimates:

$$Q_h^{(k)}(s,a) = (\langle \phi(s,a), \hat{w}_{h,k} \rangle + \tilde{\beta}_{k,\delta}^{1/2} \|\phi(s,a)\|_{\Sigma_{h,k}^{-1}}) \wedge H$$
$$V_h^{(k)}(s) = \max_{a \in \mathcal{A}} Q_h^{(k)}$$

Assuming that the features satisfy $\|\phi(s,a)\|_2 \le 1$ and the true parameters satisfy $\|\theta_h^*\|_2 \le 1$ and $\|\psi_h^* v\|_2 \le \sqrt{d}$ for all $v \in \mathbb{R}^S$ with $\|v\|_\infty \le 1$, one can choose the confidence parameter as follows:

$$\tilde{\beta}_{k,h,\delta} = \mathcal{O}\left(d^2 \log(\frac{HK}{\delta}) \right)$$

This result is the key to unlock a regret bound that is independent of the size of the state space S. The proof requires a delicate covering argument. For details refer to chapter 8 of the RL Theory Book

LSVI-UCB

Algorithm: LSVI-UCB

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

1. Initialize $V_H^{(j)}(s) = 0$ for $j = 1, \dots, k-1$.

   For $h = H-1, \dots, 0$, compute optimistic Q estimates:

$$y_{h,j} = r_h(S_h^{(j)}, A_h^{(j)}) + V_{h+1}^{(j)}(S_h^{(j)}) \quad \forall\, j = 1, \dots, k-1$$
$$\phi_{h,j} = \phi(S_h^{(j)}, A_h^{(j)}) \quad \forall\, j = 1, \dots, k-1$$
$$\hat{w}_{h,k} = \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left(\langle \phi_{j,h}, w \rangle - y_{j,h} \right)^2 + \lambda \|w\|_2^2$$
$$\Sigma_{h,k} = \sum_{j=1}^{k-1} \phi_{j,h} \phi_{j,h}^\top + \lambda \mathbf{I}_d$$
$$Q_h^{(k)}(s,a) = (\langle \phi(s,a), \hat{w}_{h,k} \rangle + \tilde{\beta}_{k,\delta}^{1/2} \|\phi(s,a)\|_{\Sigma_{h,k}^{-1}}) \wedge H$$

2. For $h = 0, \dots, H-1$, follow greedy policy

$$A_h^{(k)} = \arg\max_{a \in \mathcal{A}} Q_h^{(k)}(S_h^{(k)}, a)$$

Note that computing the optimistic policy in episode k can be done in time $\mathcal{O}(Hd^2 + HAd)$ by incrementally updating the least-square estimates $\hat{w}_{h,k}$ using the Sherman-Morrison formula. Compared to UCRL-VTR, this avoids iteration over the state space S, which is a big advantage!

Theorem (LSVI-UCB Regret)

The regret of LSVI-UCB is bounded up to logarihmic factors and with probability at least $1-\delta$ as follows:

$$R_K \leq \tilde{\mathcal{O}}(d^{3/2}H^2\sqrt{K})$$

Proof: The proof idea follows a similar strategy as the proof we presented for UCRL-VTR. As mentioned before, the crux is to show a confidence bound for LSVI that is indepenent of the size of the state space. For details, we again refer you to chapter 8 of the RL Theory Book. ■

Notes

Bernstein-type bounds for VTR (UCRL-VTR$^+$)

The UCRL-VTR$^+$ algorithm is computationally efficient and able to obtain a regret upper bound of $\mathcal{O}(dH\sqrt{K})$, and $\mathcal{O}(d\sqrt{T}(1-\gamma)^{-1.5})$ in the episodic and discounted, infinite horizon setting respectively. These results rely on using bernstein-type bounds.

Better regret bounds for Linear MDPs (Eleanor)?

A careful reader might have noticed that the regret bound for LSVI-UCB, $\tilde{\mathcal{O}}(d^{3/2}H^2\sqrt{K})$, is not tight with the tabular lower bound, $\Omega(d\sqrt{K})$. The difference is in a factor of $\sqrt{d}$. The Eleanor algorithm (Algorithm 1 in Zanette et al (2020)) is able to shave of the factor of $\sqrt{d}$, obtaining a regret upper bound of $\tilde{\mathcal{O}}(dH^2\sqrt{K})$. However, it is not currently known if the alogrithm can be implemented in a computationally efficient way. The Eleanor algorithm operates under the assumption of low inherent Bellman error (Definition 1 in Zanette et al (2020)), which means the function class is approximately closed under the Bellman optimality operator. It is interesting to note that this assumption is more general than the Linear MDP, thus Eleanor is also able to operate under the Linear MDP assumption.

References

The UCRL-VTR paper.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., & Yang, L. (2020, November). Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning (pp. 463-474). PMLR. [link]

The UCRL-VTR$^+$ paper. It also shows the regret lower bound for linear mixture MDPs $\Omega(dH^{3/2}\sqrt{K})$.

Zhou, D., Gu, Q., & Szepesvari, C. (2021, July). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Conference on Learning Theory (pp. 4532-4576). PMLR. [link]

The LSVI-UCB paper.

Jin, C., Yang, Z., Wang, Z., & Jordan, M. I. (2020, July). Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory (pp. 2137-2143). PMLR. [link]

The Eleanor paper.

Zanette, A., Lazaric, A., Kochenderfer, M., & Brunskill, E. (2020, November). Learning near optimal policies with low inherent bellman error. In International Conference on Machine Learning (pp. 10978-10989). PMLR. Link