Lecture 11 - UCVI and PSRL

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

Policy Evaluation With a Known Model

Model-based Policy Evaluation:

- When at least one of $P_{\mathcal{T}}$ and r (or $P_{\mathcal{R}}$) is known, the problem is policy evaluation with a known model.
- When both $P_{\mathcal{T}}$ and r (or $P_{\mathcal{R}}$) are unknown we can make an effort to estimate a $\hat{P}_{\mathcal{T}}$ such that $P_{\mathcal{T}}$ and $\hat{P}_{\mathcal{T}}$ are close in some measure of discrepancy (or $\hat{r}$ and $\hat{P}_{\mathcal{R}}$, respectively).

If otherwise and we only utilize the access to the environment transition, the method is categorized as model-free policy evaluation.

Policy Evaluation With a Known Model

- When both $P_{\mathcal{T}}$ and r (or $P_{\mathcal{R}}$) are unknown we can make an effort to estimate a $\hat{P}_{\mathcal{T}}$ such that $P_{\mathcal{T}}$ and $\hat{P}_{\mathcal{T}}$ are close in some measure of discrepancy (or $\hat{r}$ and $\hat{P}_{\mathcal{R}}$, respectively).

$$\hat{P}_{\mathcal{T}}(s' \mid s, a) = \frac{N(s,a,s')}{N(s,a)}, \;\; \hat{P}_{\mathcal{R}}(r|s,a) = \frac{N(r)}{N(s,a)}, \;\; r(s,a) = \sum_n \frac{r_n}{N(s,a)}$$

where $r_n \sim P_{\mathcal{R}}(r|s,a)$.

Model-based Reinforcement Learning

Assume that $0 \leq r \leq 1$. Let $\varepsilon \in (0, \frac{1}{1-\gamma})$. There is an absolute constant c such that once one have collected at least

$$N \geq \frac{\gamma}{(1-\gamma)^4} \frac{n^2 m \log(cnm/\delta)}{\varepsilon^2}$$

samples for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ pair, then we could estimate $\hat{P}$ and $\hat{Q}^\pi$ such that with probability at least $1-\delta$,

$$\|P(\cdot \mid s,a) - \hat{P}(\cdot \mid s,a)\|_1 \leq (1-\gamma)^2 \varepsilon$$

for every (s,a) pair, and

$$\|Q^\pi - \hat{Q}^\pi\|_\infty \leq \varepsilon$$

Model-based Reinforcement Learning

The natural question remaining is that if we are able to obtain N samples for each (s, a) pair so as to fulfill the condition of the lemma.

The answer is, unfortunately, no, in general.

Motivations for Exploring in Discrete MDPs

- Having a good estimate of the transition kernel and the reward function requires the number N of samples to be large in every (s, a) pair.
- This is not possible in general, but we should increase the number of visits to those states with fewer samples. This is called exploration.
- Exploration helps RL to find a near-optimal policy when the MDPs are not known to the learner. The algorithm must focus on:
  - Sample Complexity: the number of samples required to find a near-optimal policy.
  - Regret achieved in the process of finding a near-optimal policy.

Episodic discrete MDPs

In the episodic setting, the learner acts for some finite number of steps, starting from a fixed starting state s_0, the learner observes the trajectory, and the state resets to s_0. This setting is applicable to the finite-horizon and infinite-horizon settings.

- Finite horizon MDPs. Here, each episode lasts for H steps, and then the state is reset to $s_0 \sim \rho_0$.

Episodic discrete MDPs

In the episodic setting, the learner acts for some finite number of steps, starting from a fixed starting state s_0, the learner observes the trajectory, and the state resets to s_0. This setting is applicable to the finite-horizon and infinite-horizon settings.

- Infinite horizon MDPs. Here, it is still natural to work in an episodic model for learning, where each episode terminates after a finite number of steps. Here, it is often natural to assume either the agent can terminate the episode at will or that the episode will terminate at each step with probability $1 - \gamma$.

Episodic discrete MDPs

In the episodic setting, the learner acts for some finite number of steps, starting from a fixed starting state s_0, the learner observes the trajectory, and the state resets to s_0. This setting is applicable to the finite-horizon and infinite-horizon settings.

- Finite horizon MDPs. Here, each episode lasts for H steps, and then the state is reset to $s_0 \sim \rho_0$.
- Infinite horizon MDPs. Here, it is still natural to work in an episodic model for learning, where each episode terminates after a finite number of steps.

We can study both sample complexity and regret under these settings.

Episodic discrete MDPs

The episodic setting is challenging in that

- The agent has to engage in some exploration in order to gain information at the relevant state, and therefore is a suitable environment for us to discuss exploration-based topics.
- This exploration must be strategic, in the sense that simply behaving randomly will not lead to information being gathered quickly enough.

Episodic discrete MDPs

Let's assume, in every episode $k \in [K]$, the learner acts for H step starting from a fixed starting state s_0 and, at the end of the H-length episode, the state is reset to s_0.

The goal of the agent is to minimize the expected cumulative regret over K episodes

$$\overline{R}_K = \mathbb{E}\left[K V^*(s_0) - \sum_{k=0}^{K-1} \sum_{h=0}^{H-1} r(s_h^k, a_h^k) \right],$$

where the expectation is with respect to the randomness of the MDP environment and any randomness of the agent's policy and (s_h^k, a_h^k) denotes the state-action pair in the h-th step of the k-th episode.

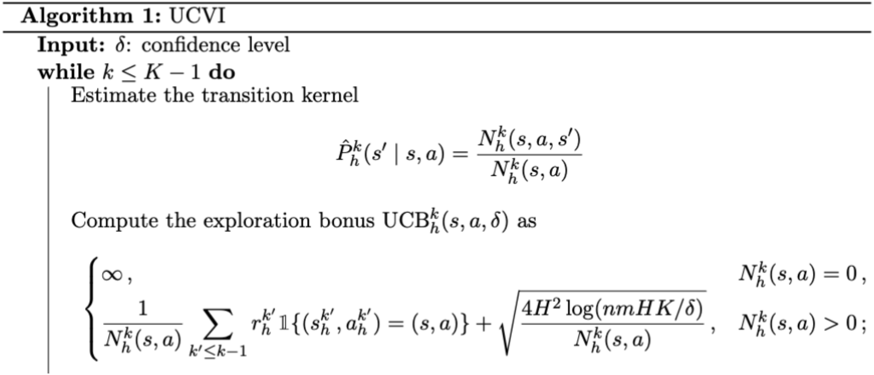
UCB Value Iteration

Without loss of generality, we present the UCB value iteration algorithm (UCVI) on the non-stationary setting.

- The reward function r_h and the probability transition kernel P_h are assumed to change over the horizon $[H]$.
- The estimation of r_h and P_h up to the collection of the first $k-1$ episodes are denoted by $\hat{r}_h^k$ and $\hat{P}_h^k$, respectively.

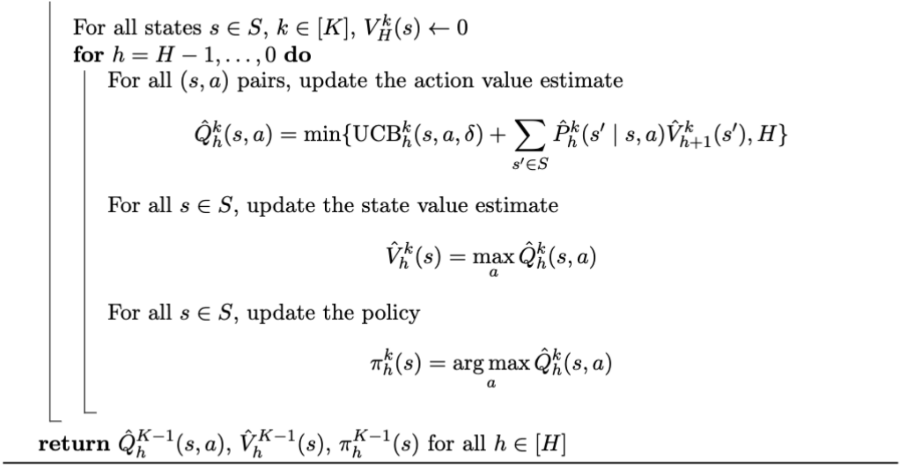
The exploration is encouraged by a UCB exploration bonus term $\sqrt{\frac{4H^2\log(nmHK/\delta)}{N_h^k(s,a)}}$, which is similar to the UCB algorithm in multi-armed bandits.

UCB Value Iteration



```
Algorithm 1: UCVI
Input: $\delta$: confidence level
while $k \le K - 1$ do
    Estimate the transition kernel
        $\hat{P}_h^k(s' \mid s, a) = \frac{N_h^k(s, a, s')}{N_h^k(s, a)}$
    Compute the exploration bonus $\mathrm{UCB}_h^k(s, a, \delta)$ as
        $\begin{cases} \infty\,, & N_h^k(s,a) = 0\,, \\ \frac{1}{N_h^k(s,a)} \sum_{k' \le k-1} r_h^{k'} \mathbb{1}\{(s_h^{k'}, a_h^{k'}) = (s,a)\} + \sqrt{\frac{4H^2 \log(nmHK/\delta)}{N_h^k(s,a)}}\,, & N_h^k(s,a) > 0\,; \end{cases}$
```

UCB Value Iteration



```
For all states $s \in S$, $k \in [K]$, $V_H^k(s) \leftarrow 0$
for $h = H - 1, \ldots, 0$ do
    For all $(s, a)$ pairs, update the action value estimate
        $$\hat{Q}_h^k(s,a) = \min\{\mathrm{UCB}_h^k(s,a,\delta) + \sum_{s' \in S} \hat{P}_h^k(s' \mid s,a)\hat{V}_{h+1}^k(s'), H\}$$
    For all $s \in S$, update the state value estimate
        $$\hat{V}_h^k(s) = \max_a \hat{Q}_h^k(s,a)$$
    For all $s \in S$, update the policy
        $$\pi_h^k(s) = \arg\max_a \hat{Q}_h^k(s,a)$$
return $\hat{Q}_h^{K-1}(s,a)$, $\hat{V}_h^{K-1}(s)$, $\pi_h^{K-1}(s)$ for all $h \in [H]$
```

Note that r is normalized to [0, 1]

UCB Value Iteration

Theorem

Without loss of generality assume that $r_h(s,a)$ is deterministic and known and is between 0 *to* 1. *Taking $\delta = 1/KH$, the regret of UCVI*

$$\overline{R}_T \leq 10\sqrt{n^2 m H^4 K \log(nmH^2K^2)}\,.$$

This regret bound could be improved to $\sqrt{nmH^4K} + n^2mH^3$, which is smaller than the above theorem by a factor of $\sqrt{n}$ when K is asymptotically large.

Posterior Sampling for Reinforcement Learning

In bandits, an alternative perspective to implement exploration is to use Thompson sampling, that is, to sample a bandit environment from a posterior distribution in every time step.

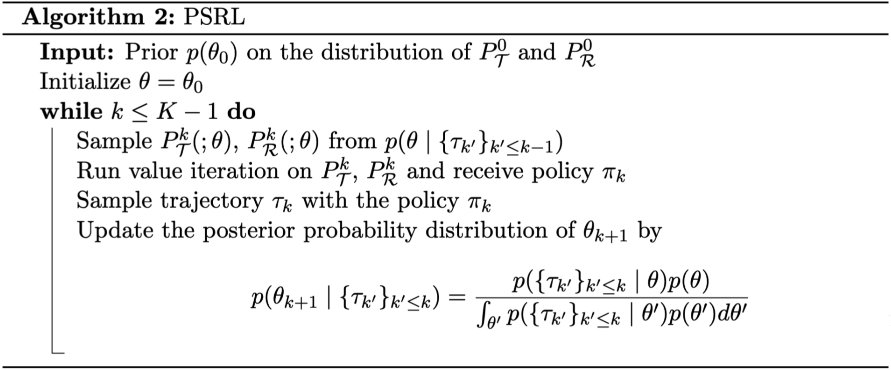
" *We wonder if a similar approach is possible in discrete RL, that is,*

To sample an MDP in every episode in episodic MDPs."

The answer is yes!.

Posterior Sampling for Reinforcement Learning

The likelihood $P_{\mathcal{T}}$ follows a categorical distribution, so their prior and posterior follow the Dirichlet distribution. The likelihood $P_{\mathcal{R}}$ follows the Bernoulli or Normal distribution, so their prior and posterior follow the Beta or Normal distribution.



```
Algorithm 2: PSRL
Input: Prior $p(\theta_0)$ on the distribution of $P^0_{\mathcal{T}}$ and $P^0_{\mathcal{R}}$
Initialize $\theta = \theta_0$
while $k \leq K - 1$ do
    Sample $P^k_{\mathcal{T}}(;\theta)$, $P^k_{\mathcal{R}}(;\theta)$ from $p(\theta \mid \{\tau_{k'}\}_{k' \leq k-1})$
    Run value iteration on $P^k_{\mathcal{T}}$, $P^k_{\mathcal{R}}$ and receive policy $\pi_k$
    Sample trajectory $\tau_k$ with the policy $\pi_k$
    Update the posterior probability distribution of $\theta_{k+1}$ by

        $p(\theta_{k+1} \mid \{\tau_{k'}\}_{k' \leq k}) = \frac{p(\{\tau_{k'}\}_{k' \leq k} \mid \theta) p(\theta)}{\int_{\theta'} p(\{\tau_{k'}\}_{k' \leq k} \mid \theta') p(\theta') d\theta'}$
```

Posterior Sampling for Reinforcement Learning

Theorem

The regret of PSRL

$$\overline{R}_T \leq \sqrt{30 n^2 m H^3 K \log(nmHK)} \, .$$

A point worth noting is that in practice, PSRL and TS are observed to outperform UCRL and UCB, respectively, in general, by a significant margin.

Stationary v.s. Non-Stationary MDPs

- Stationary dynamics in the infinite-horizon setting and time-dependent dynamics in the finite-horizon setting.
- From a theoretical perspective, the finite-horizon, time-dependent setting is often more amenable to analysis.
- From a practical perspective, time-dependent MDPs are rare because their value functions consume $O(H)$ larger memory than those in the stationary setting.

Question and Answering (Q&A)