DDA4230 Reinforcement learning UCVI and PSRL

Lecture 11

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

In this lecture we will introduce exploration in discrete Markov decision processes and several algorithms with exploration techniques. The idea of upper confident bounds is extended to upper confident value iteration (UCVI, also known as UCRL). The idea of Thompson sampling is extended to posterior sampling for reinforcement learning (PSRL).

Suggested reading: Chapter 3 and 8 of *Reinforcement learning: An introduction*; Chapter 7 of *Reinforcement learning: Theory and algorithms*; *(More) Efficient Reinforcement Learning via Posterior Sampling* by Osband, Van Roy, and Russo; *Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds* by Agrawal and Jia.

2 Recap: Model-based reinforcement learning

A model-based approach does not know the model, but can maintain an estimation of it and use the estimation when calculating the value function. It is then very straightforward to consider replacing P with $\hat{P}$ for some estimation $\hat{P}$, where the most simple way to obtain $\hat{P}$ is to use the empirical distribution of the state transitions collected from the trajectories.

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

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

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 $1-\delta$,

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

for every (s,a) pair, and

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

for every policy π.

The number N of samples needed could be reduced to

$$N \ge \frac{c}{(1-\gamma)^3} \frac{nm \log(cnm/\delta)}{\varepsilon^2},$$

if we only desire the accurate estimate on the optimal policy and value, i.e., $\|Q^* - \hat{Q}^*\|_\infty \le \varepsilon$ and $\|Q^* - Q^{\hat{\pi}^*}\|_\infty \le \varepsilon$. This improvement is minimax optimal for estimating the optimal policy and value in this way.

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**.

3 Exploration in discrete MDPs

While the optimal policy could be directly computed once we have a good estimate of the transition kernel and the reward function, estimating these variables using the empirical average requires the number N of samples to be large in every (s, a) pair. This is not possible in general, but it motivates us to incorporate exploration into our algorithm, to increase the number of visits towards those states with less samples. In fact, much of reinforcement learning is concerned with finding a near optimal policy (or obtaining near optimal reward) in settings where the MDPs is not known to the learner. We will study these questions in a few different models of how the agent obtains information about the unknown underlying MDP.

In each of these settings, we are interested in understanding two things: the number of samples required to find a near optimal policy, i.e., the *sample complexity*; and the cumulative (sublinear) *regret* achieved in the process of finding a near optimal policy.

Ultimately, we are interested to extend the results to cases where number of states and actions is large, or, possibly, countably or uncountably infinite. This could be achieved in a variety of ways, including approximation through a deep neural network.

3.1 Episodic discrete MDPs

In the episodic setting, in every episode, the learner acts for some finite number of steps, starting from a fixed starting state $s_0 \sim \rho_0$, the learner observes the trajectory, and the state resets to $s_0 \sim \rho_0$. This episodic model of feedback is applicable to both 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. Even for infinite horizon MDPs it is 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 $1 - \gamma$. After termination, we again assume that the state is reset to $s_0 \sim \rho_0$. Note that, if each step in an episode is terminated with probability $1 - \gamma$, then the observed cumulative reward in an episode of a policy provides an unbiased estimate of the infinite-horizon, discounted value of that policy. In this setting, we are often interested in either the number of episodes it takes to find a near optimal policy, which is a probably approximately correct (PAC) guarantee on the sample complexity, or we are interested in a regret guarantee.

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.

In this lecture notes we assume the MDP to be with a finite horizon and a fixed start state s_0. We discuss the problem in the episodic setting. 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. It is straightforward to extend this setting where the starting state is sampled from a distribution, i.e., $s_0 \sim \rho_0$.

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

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

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.

3.2 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. As usual, the former is estimated by the empirical average of the reward and the latter is estimated by the frequency of the transition.

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. A regret of $\sqrt{K}$ could be obtained with standard arguments in probabilities.

Theorem 1 *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*

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

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.

The proofs could be found in Chapter 7 of *Reinforcement learning: Theory and algorithms* and the referred papers thereof. We leave the proof of the theorems as an extended reading to the readers.

3.3 Posterior sampling for reinforcement learning

In discrete RL, most theoretical results are induced by the optimism principal and some variants of the upper confident bounds. 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 (PSRL) was proposed in 2013 and was improved in 2017 and thereafter. See suggested reading for more references. In similar settings, PSRL achieves a regret at most $O(\sqrt{n^2mH^3K\log(nmHK)})$.

Algorithm 1: UCVI

Input: δ: confidence level

while $k \leq K-1$ **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

$$\begin{cases} \infty\,, & N_h^k(s,a) = 0\,, \\ \dfrac{1}{N_h^k(s,a)} \displaystyle\sum_{k' \leq k-1} r_h^{k'} \mathbb{1}\{(s_h^{k'}, a_h^{k'}) = (s,a)\} + \sqrt{\dfrac{4H^2 \log(nmHK/\delta)}{N_h^k(s,a)}}\,, & N_h^k(s,a) > 0\,; \end{cases}$$

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 $h \in [H]$

Theorem 2 *The regret of PSRL*

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

Note that this regret is a minimax regret (which is the one we usually encounter). It is distinguished from the Bayesian regret.

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.

3.4 Stationary v.s. non-stationary MDPs

In analysis of discrete MDPs (in and out of this course), it is natural for us to study both the stationary and the non-stationary models, where we typically assume stationary dynamics in the infinite-horizon setting and time-dependent dynamics in the finite-horizon

Algorithm 2: PSRL

Input: Prior $p(\theta_0)$ on the distribution of P_h and r_h
Initialize $\theta = \theta_0$
while $k \leq K - 1$ **do**
  Sample P_h, r_h from θ
  Run value iteration on P_h, r_h
  Update the posterior probability distribution of θ_{k+1} by

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

setting. From a theoretical perspective, the finite-horizon, time-dependent setting is often more amenable to analysis, where optimal statistical rates often require simpler arguments. However, we should note that from a practical perspective, time-dependent MDPs are rarely utilized because they lead to policies and value functions that consume $O(H)$ larger memory to be stored in a computer than those in the stationary setting. In practice, we often incorporate temporal information t directly into the definition of the state, which leads to more compact value functions and policies (when coupled with function approximation methods, which attempt to represent both the values and policies in a more compact form).

Acknowledgement

This lecture notes partially use material from *Reinforcement learning: An introduction* and *Reinforcement learning: Theory and algorithms*.