Fundamentals of Reinforcement Learning
Master IASD

Olivier Cappé (CNRS)

DI ENS, Université PSL

February 2022

Roadmap

Markov Decision Process (MDP)

Recall from previous courses.

Markov Decision Process (MDP)

An MDP transitions from (S_t, A_t) to (S_{t+1}, R_{t+1}) according to

- $\mathbb{P}(S_{t+1} = s|\mathcal{H}_t) = \mathbb{P}(S_{t+1} = s|S_t, A_t) = p(s|S_t, A_t)$
- $\mathbb{E}(R_{t+1}|\mathcal{H}_t) = \mathbb{P}(R_{t+1}|S_t, A_t) = r(S_t, A_t)$

where $\mathcal{H}_t = \sigma\left((S_i, A_i)_{i=0}^t, (R_i)_{i=1}^t\right)$ and S_{t+1} and R_{t+1} are conditionally independent given $\mathcal{H}_t$.

⚠ Many RL texts assume that (S_t, A_t) trigers reward R_t instead of R_{t+1} (check indices...); in Control theory, rewards are usually replaced by costs (to be minimized)

Policy

The agent's goal is to design a policy π.

Policy

Given S_t, a policy π defines the action A_t according to

$$\mathbb{P}_\pi\left(A_t = a | (S_i)_{i=0}^t, (A_i)_{i=0}^{t-1}, (R_i)_{i=1}^t\right) = \mathbb{P}_\pi(A_t = a | S_t) = \pi(a|S_t)$$

The choice of the policy induces a Markov reward model on $(S_i, R_i)_{i=1}^t$:

- $\mathbb{P}_\pi(S_{t+1} = s|\mathcal{H}_t) = \mathbb{P}_\pi(S_{t+1} = s|S_t) = \sum_a p(s|S_t, a)\pi(a|S_t)$
- $\mathbb{E}_\pi(R_{t+1}|\mathcal{H}_t) = \mathbb{E}_\pi(R_{t+1}|S_t) = \sum_a r(S_t, a)\pi(a|S_t)$

Value Functions

Infinite Horizon with Discount

$$v_\pi(s) = \mathbb{E}_\pi \left(\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| S_0 = s \right)$$

v_π is horizon-free in the sense that $v_\pi(s) = \mathbb{E}_\pi \left(\sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \middle| S_t = s \right)$ (however it depends on γ and involves an infinite horizon).

Value Functions (Contd.)

Episodic

$$v_\pi^\tau(s) = \mathbb{E}_\pi \left(\sum_{t=1}^{\infty} \mathbb{1}\{\tau \geq t\} R_t \middle| S_0 = s \right)$$

where τ is a $(\mathcal{H}_t)$-stopping time, such that $\mathbb{P}(\tau < \infty) = 1$ for all policies.

- Typically τ is associated with a goal or termination state s_G, i.e., $\tau = \inf\{t \geq 1 : S_t = S_G\}$.
- Tossing independent Bernoullis with probability of continuation γ at each timestep yields a geometrically distributed τ with $\mathbb{P}(\tau \geq t) = \gamma^{t-1}$ and $\mathbb{E}[\tau] = 1/(1-\gamma)$, and is such that $v_\pi^\tau(s) = v_\pi(s)$; however with larger variance trajectories due to Rao-Blackwell Theorem.

Value Functions (Contd.)

Finite Horizon

$$v_\pi^T(s) = \mathbb{E}_\pi \left(\sum_{t=1}^{T} R_t \middle| S_0 = s \right)$$

In some models, one can define an average-reward limit

$$\lim_{T \to \infty} \frac{1}{T} v_\pi^T(s)$$

but this does not always exists.

Bellman Equations

Value and Q-Value (or State–Action value) Functions

$$v_\pi(s) = \sum_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right] \pi(a|s)$$

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} q_\pi(s',a') \pi(a'|s')$$

Optimal Value and Q-Value

$$v^*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v^*(s') \right]$$

$$q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q^*(s',a')$$

Temporal Difference (TD) learning algorithms differ in goal (estimating the value of a policy/estimating the value of the optimal policy) and learning conditions (off policy/on policy) but use a common principle shared by TD(0), Q-Learning and SARSA.

Q-Learning [Watkins, 1989]

Q-Learning performs off-policy learning of the optimal Q-Value by behaving according to π and recursively updating an estimate of q^*

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \alpha_t \left[R_{t+1} + \gamma \max_a Q_t(S_{t+1}, a) - Q_t(S_t, A_t) \right]$$

and $Q_{t+1}(s, a) = Q_t(s, a)$ for all other state–action pairs.

For any (s, a) and Q-Table q, define the mapping $h_{s,a} : q \mapsto h_{s,a}(q)$ by

$$h_{s,a}(q)(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q(s',a')$$

and $h_{s_a}(q)(s',a') = q(s',a')$ when $(s',a') \neq (s,a)$. Note that

1. $h_{s,a}(q^*) = q^*$ due to Bellman equations
2. $\|h_{s,a}(q) - q^*\|_\infty \leq \gamma \|q - q^*\|_\infty$

The Q-Learning recursion may be written as

$$Q_{t+1} = Q_t + \alpha_t \left[h_{S_t,A_t}(Q_t) - Q_t + \epsilon_{t+1}\right] \qquad \text{where } \mathbb{E}[\epsilon_{t+1}|\mathcal{H}_t] = 0$$

This form of update is known as a Stochastic Approximation (SA) scheme.

Stochastic Approximation

More generally,

Stochastic Approximation (AKA Robbins–Monro algorithm)

$$Q_{t+1} = Q_t + \alpha_t \left[h(Q_t) - Q_t + \epsilon_{t+1}\right] = (1-\alpha_t)Q_t + \alpha_t \left[h(Q_t) + \epsilon_{t+1}\right]$$

with $\mathbb{E}[\epsilon_{t+1}|\mathcal{H}_t] = 0$

is a general purpose scheme for finding the solutions of $h(q) = q$.

In the particular case where $h(q) - q$ may be interpreted as the gradient ∇f of a function f, one recovers the stochastic gradient algorithm (for *maximizing* f). This is not however the case for TD learning algorithms.

Convergence of SA

[Bertsekas & Tsitsiklis, 1996] study the SA scheme under the assumptions required for Q-learning and other TD algorithms.

[Bertsekas & Tsitsiklis, 1996] Proposition 4.4 (Simplified)

Assuming

- $\|h(q) - q^*\|_\infty \leq \gamma \|q - q^*\|_\infty$, with $\gamma < 1$
- $\mathbb{E}[\epsilon_{t+1}|\mathcal{H}_t] = 0$, $\mathbb{E}[\|\epsilon_{t+1}\|_\infty^2 |\mathcal{H}_t] \leq A + B\|Q_t - q^*\|_\infty^2$
- $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$

implies that $Q_t \to q^*$ (a. s.)

Assuming that the exploration policy π is such that each state–action pair (s, a) is visited infinitely often implies that Q_t (produced by Q-Learning) converges (a.s.) to q^*.

Using a Fixed Exploration Policiy Does Not Scale

The capacity of a fixed policy to visit "relevant regions" of the state–action space may decrease exponentially fast in the size of the state–action space.

The River Swim (Toy) Example [Strehl & Littman, 2008]

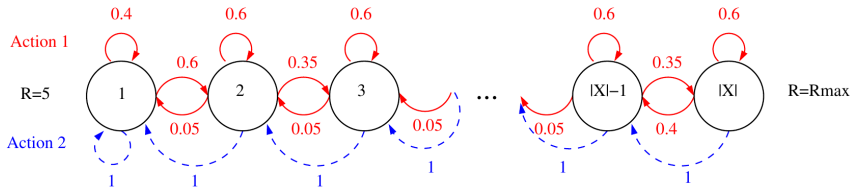

Figure: Riwer Swim Environment: The continuous red (resp. dotted blue) arrows represent the transitions when action 1 "R" (resp. 2 "L") is chosen.

Scalable Algorithms Require Some Form of Policy Update

To maintain some exploration, one typically use

- ϵ–greedy policy

$$\pi(a|s) = (1-\epsilon)\mathbb{1}\left(a = \arg\max_{a'} Q(s,a')\right) + \frac{\epsilon}{|\mathcal{A}|}$$

- Boltzmann (softmax) policy

$$\pi(a|s) = \frac{\exp(\beta Q(s,a))}{\sum_{a'}\exp(\beta Q(s,a'))}$$

with decaying ϵ or increasing β.

Analysis of convergence is much more involved, due to the difficulty of ensuring sufficient exploration when the policy is updated –see, e.g., [Singh *et al.*, 2000].