Reinforcement Learning

Greg Strabel

March 6, 2023

1 Key Concepts

Definition 1.1. A **Markov Decision Process** is a 4-tuple (S, A, P, R), where:

- S is a set of states
- A is a set of actions
- P is a probability measure such that $P(s_{t+1} = s'|s_t = s, a_t = a)$ is the probability of transitioning to state s' at time $t+1$ given that the state at time t is s and the agent has performed action a at time t.
- R is a reward function. The agent receives the immediate reward $R(s, a)$ for performing action a in state s.

Definition 1.2. A decision policy is a function $\pi : S \times A \to [0, 1]$ such that

$$\int_A \pi(s, a) da = 1 \qquad \forall s \in S \tag{1}$$

The space of all decision functions is denoted Π.

Given a discount rate $\gamma \in (0, 1)$ and an initial state s_0, the objective of the decision agent is:

$$\max_{\pi \in \Pi} \mathbb{E}_{s_{1:\infty}, a_{0:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi, s_0 = s \right] \tag{2}$$

Definition 1.3. The state-value of a policy π for a Markov Decision Process (S, A, P, R) is a function $V^\pi : S \to \mathbb{R}$ defined as

$$V^\pi(s) = \mathbb{E}_{s_{1:\infty}, a_{0:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi, s_0 = s \right] \tag{3}$$

Note that

$$\begin{aligned}V^\pi(s) &= \mathbb{E}_{s_{1:\infty},a_{0:\infty}}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|\pi,s_0=s\right]\\ &= \int_A R(s,a)\pi(s,a)da + \mathbb{E}_{s_1,a_0}\left[\mathbb{E}_{s_{2:\infty},a_{1:\infty}}\left[\sum_{t=1}^{\infty}\gamma^t R(s_t,a_t)|\pi,s_1=s'\right]|\pi,s_0=s\right]\\ &= \int_A R(s,a)\pi(s,a)da + \gamma\mathbb{E}_{s_1,a_0}\left[V^\pi(s_1)|\pi,s_0=s\right]\end{aligned} \tag{4}$$

Definition 1.4. The action-value of a policy π for a Markov Decision Process (S, A, P, R) is a function $Q^\pi : S \times A \to \mathbb{R}$ defined as

$$Q^\pi(s,a) = \mathbb{E}_{s_{1:\infty},a_{1:\infty}}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|\pi,s_0=s,a_0=a\right] \tag{5}$$

2 Policy Gradient Theorem (PGT)

Definition 2.1. For $s, s' \in S$, let $\rho^\pi(s \to s', 0) = \mathbb{1}\{s = s'\}$ and $\rho^\pi(s \to s', 1) = \int_A \pi(a,s)p(s'|s,a)da$. For $s, s' \in S$ and $k \geq 1$, define $\rho^\pi(s \to s', k+1)$ recursively by

$$\rho^\pi(s \to s', k+1) = \int_S \rho^\pi(s \to x, k)\rho^\pi(x \to s', 1)dx \tag{6}$$

Definition 2.2.

$$J(\theta) = \int_S p_0(s_0)V^\pi(s_0)ds_0 = \int_S p_0(s_0)\int_A \pi(s_0,a_0,\theta)Q^\pi(s_0,a_0)da_0 ds_0 \tag{7}$$

Theorem 1. *Policy Gradient Theorem*

$$\nabla_\theta J(\theta) = \int_S \rho^\pi(s)\int_A \nabla_\theta \pi(s,a,\theta)Q^\pi(s,a)dads \tag{8}$$

where

$$\rho^\pi(s) = \int_S \sum_{t=0}^{\infty}\gamma^t p_0(s_0)\rho^\pi(s_0 \to s, t)ds_0 \tag{9}$$

See Appendix for the proof of the PGT.

Note that

$$\rho^\pi(s) = \int_S \sum_{t=0}^{\infty}\gamma^t p_0(s_0)\rho^\pi(s_0 \to s, t)ds_0 = \sum_{t=0}^{\infty}\gamma^t P\left(s_t = s|\pi\right) \tag{10}$$

so that

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \int_S \rho^\pi(s) \int_A \nabla_\theta \pi(s,a,\theta) Q^\pi(s,a) dads \\
&= \sum_{t=0}^{\infty} \gamma^t \int_S \int_A P\left(s_t = s | \pi\right) \pi(s,a,\theta) \nabla_\theta \ln \pi(s,a,\theta) Q^\pi(s,a) dads \\
&= \sum_{t=0}^{\infty} \gamma^t \int_S \int_A P\left(s_t = s, a_t = a | \pi\right) \nabla_\theta \ln \pi(s,a,\theta) Q^\pi(s,a) dads
\end{aligned} \tag{11}
$$

Additionally, as note in [SML+15], for any function $f(s_{0:t}, a_{0:t-1})$,

$$
\begin{aligned}
&\mathbb{E}_{s_{0:\infty} a_{0:\infty}} \left[\nabla_\theta \ln \pi(s_t, a_t, \theta) f(s_{0:t}, a_{0:t-1}) | \pi \right] \\
&= \mathbb{E}_{s_{0:t} a_{0:t-1}} \left[\mathbb{E}_{s_{t+1:\infty} a_{t:\infty}} \left[\nabla_\theta \ln \pi(s_t, a_t, \theta) f(s_{0:t}, a_{0:t-1}) | \pi, s_{0:t} a_{0:t-1} \right] | \pi \right] \\
&= \mathbb{E}_{s_{0:t} a_{0:t-1}} \left[f(s_{0:t}, a_{0:t-1}) \mathbb{E}_{s_{t+1:\infty} a_{t:\infty}} \left[\nabla_\theta \ln \pi(s_t, a_t, \theta) | \pi, s_{0:t} a_{0:t-1} \right] | \pi \right] \\
&= \mathbb{E}_{s_{0:t} a_{0:t-1}} \left[f(s_{0:t}, a_{0:t-1}) \mathbb{E}_{a_t} \left[\nabla_\theta \ln \pi(s_t, a_t, \theta) | \pi, s_t \right] | \pi \right] \\
&= \mathbb{E}_{s_{0:t} a_{0:t-1}} \left[f(s_{0:t}, a_{0:t-1}) \cdot \int_A \frac{\nabla_\theta \pi(s_t, a, \theta)}{\pi(s_t, a, \theta)} \pi(s_t, a, \theta) da | \pi \right] \\
&= \mathbb{E}_{s_{0:t} a_{0:t-1}} \left[f(s_{0:t}, a_{0:t-1}) \cdot 0 | \pi \right] \\
&= 0
\end{aligned} \tag{12}
$$

3 Q-learning

Q-learning is an approach to reinforcement learning that estimates the action-value function of the optimal policy. Q-learning represents the action-value function as a function, $Q(s, a, \theta)$, typically a deep neural network, with a vector of parameters θ. [MKS+15]

References

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[SML+15] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2015.

Algorithm 1: Q-learning with experience replay

Input: Replay memory D with capacity N
Action-value function Q with random weights θ
Target action-value function $\hat{Q}$ with weights $\theta^{-} = \theta$

Output: Weights θ

for *episode = 0 to M* **do**
    Sample s_0 from emulator
    for t *= 0 to* T **do**
        $a_t = \begin{cases} \text{select random action} & \text{with probability } \epsilon \\ \text{argmax}_a \, Q(s_t, a_t, \theta) & \text{with probability } 1 - \epsilon \end{cases}$
        Execute action a_t in emulator, observe reward r_t and state s_{t+a}
        Store (s_t, a_t, r_t, s_{t+1}) in replay memory D
        Sample random minibatch of transitions (s_j, a_j, r_j, a_{j+1}) from D
        Set $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma \text{max}_a \hat{Q}(s_{j+1}, a, \theta^{-}) & \text{otherwise} \end{cases}$
        Perform a gradient descent step on $(y_j - Q(s_j, a_j, \theta))^2$ with respect to θ
        Every C steps reset $\theta^{-} \leftarrow \theta$

return θ

A Proof of the Policy Gradient Theorem

Proof. Let $\phi(s) = \int_A \nabla_\theta \pi(s,a,\theta) Q^\pi(s,a) da$. Then

$$\begin{aligned}
\nabla_\theta V^\pi(s) &= \int_A \left[\nabla_\theta \pi(s,a,\theta) Q^\pi(s,a) + \pi(s,a,\theta)\nabla_\theta Q^\pi(s,a)\right] da \\
&= \phi(s) + \int_A \pi(s,a,\theta)\nabla_\theta Q^\pi(s,a) da \\
&= \phi(s) + \int_A \pi(s,a,\theta)\nabla_\theta \left[r(s,a) + \gamma \int_S p(s'|s,a) V^\pi(s') ds'\right] da \\
&= \phi(s) + \gamma \int_S \int_A \pi(s,a,\theta) p(s'|s,a) da \nabla_\theta V^\pi(s') ds' \\
&= \int_S \gamma^0 \phi(s') \rho^\pi(s \to s', 0) ds' + \gamma \int_S \rho^\pi(s \to s', 1) \nabla_\theta V^\pi(s') ds'
\end{aligned} \tag{13}$$

Now suppose that for some $k \geq 0$,

$$\nabla_\theta V^\pi(s) = \sum_{i=0}^{k} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' + \gamma^{k+1} \int_S \rho^\pi(s \to s', k+1) \nabla_\theta V^\pi(s') ds' \tag{14}$$

Then

$$\begin{aligned}
\nabla_\theta V^\pi(s) &= \sum_{i=0}^{k} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' + \gamma^{k+1} \int_S \rho^\pi(s \to s', k+1) \nabla_\theta V^\pi(s') ds' \\
&= \sum_{i=0}^{k} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' \\
&+ \gamma^{k+1} \int_S \rho^\pi(s \to s', k+1) \int_S \gamma^0 \phi(s'') \rho^\pi(s' \to s'', 0) ds'' ds' \\
&+ \gamma^{k+1} \int_S \rho^\pi(s \to s', k+1) \gamma \int_S \rho^\pi(s' \to s'', 1) \nabla_\theta V^\pi(s'') ds'' ds' \\
&= \sum_{i=0}^{k+1} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' \\
&+ \gamma^{k+2} \int_S \int_S \rho^\pi(s \to s', k+1) \rho^\pi(s' \to s'', 1) ds' \nabla_\theta V^\pi(s'') ds'' \\
&= \sum_{i=0}^{k+1} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' + \gamma^{k+2} \int_S \rho^\pi(s \to s'', k+2) \nabla_\theta V^\pi(s'') ds''
\end{aligned} \tag{15}$$

Hence, by induction, for all $k \geq 0$,

$$\nabla_\theta V^\pi(s) = \sum_{i=0}^{k} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' + \gamma^{k+1} \int_S \rho^\pi(s \to s', k+1) \nabla_\theta V^\pi(s') ds' \tag{16}$$

Taking the limit as $k \to \infty$, we have

$$\begin{aligned}
\nabla_\theta V^\pi(s) &= \sum_{i=0}^{\infty} \int_S \gamma^i \phi(s') \rho^\pi(s \to s', i) ds' \\
&= \int_S \sum_{i=0}^{\infty} \gamma^i \rho^\pi(s \to s', i) \int_A \nabla_\theta \pi(s', a, \theta) Q^\pi(s', a) da ds'
\end{aligned} \tag{17}$$

Plugging this into the definition of $J(\theta)$ yields

$$\begin{aligned}
\nabla_\theta J(\theta) &= \int_S p_0(s_0) \nabla_\theta V^\pi(s_0) ds_0 \\
&= \int_S p_0(s_0) \int_S \sum_{i=0}^{\infty} \gamma^i \rho^\pi(s_0 \to s, i) \int_A \nabla_\theta \pi(s, a, \theta) Q^\pi(s, a) da ds ds_0 \\
&= \int_S \rho^\pi(s) \int_A \nabla_\theta \pi(s, a, \theta) Q^\pi(s, a) da ds
\end{aligned} \tag{18}$$

$\square$