Reinforcement learning: notes

Jerry Tworek

Contents

1 General notation

1.1 Markov decision processes

Definition 1.1.1. Markov Decision Process

A Markov Decision Process (**MDP**) is a tuple $(\mathcal{S}, \mathcal{A}, p, p_0, \mathcal{R}, \gamma)$, where:

- $\mathcal{S}$ is a set of states $s \in \mathcal{S}$.
- $\mathcal{A}$ is a set of actions. Abusing notation a bit, we define $\mathcal{A}(s) : \mathcal{S} \to \mathcal{A}$ as a set of actions available in state $s \in \mathcal{S}$.
- $p(s'|s, a)$ is a probability distribution of transition from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ due to action $a \in \mathcal{A}$.
- $p_0(s)$ is a probability distribution of initial state after action $a \in \mathcal{A}$.
- $\mathcal{R} : \mathcal{S} \to \mathbb{R}$ is a reward function.
- $\gamma \in [0, 1]$ is a discount factor.

Definition 1.1.2. Trajectory

An **MDP** trajectory is an alternating sequence of states, actions and rewards, a singular realization of an **MDP** process. More formally, let's define a *transition* as a tuple of $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$. Then, a trajectory h is a sequence of transitions $h_t = (s_t, a_t, r_t, s'_t)$ such, that $s'_t = s_{t+1}$, $r_t = \mathcal{R}(s'_t)$ for all n. Trajectories can be both finite and infinite. In various contexts we'll use notation of s'_t and s_{t+1} fully interchangeably when discussing trajectories. $T(h) \in \mathbb{N} \cup \{\infty\}$ denotes a length of a trajectory.

We consider various sets of trajectories for a given **MDP**:

- H_n is a set of trajectories of length n
- $H = \bigcup_{n \in \mathbb{N}} H_n$ is a set of all finite trajectories
- H^∞ is a set of all infinite trajectories
- $H^* = H \cup H^\infty$ is a set of all possible trajectories both finite and infinite
- $H(s)$ is a set of trajectories such, that $s_0 = s$
- $H(s, a)$ is a set of trajectories such, that $s_0 = s$ and $a_0 = a$.

Definition 1.1.3. Return

A return G_t following time t for given trajectory h is a sum of discounted rewards received from time t:

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_k = \sum_{k=0}^{T-t} \gamma^k r_{t+k} = r_t + \gamma G_{t+1}$$

1.2 Policies

Definition 1.2.1. Policy

A policy π is a probability distribution $\pi(a|s)$ for each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}(s)$.

Implicitly, for a given **MDP** policy π also defines a probability distribution on H^* where

$$\pi(h) = p_0(s_0) \cdot \pi(a_0|s_0) \cdot p(s_1|a_0, s_0) \cdot \pi(a_1|s_1) \cdot \ldots$$

Most of the time, when we write expectation $\mathbb{E}_\pi$ we mean expectation by this probability measure of π on H^*.

Definition 1.2.2. State value

A state-value $V : \mathcal{S} \to \mathbb{R}$ is an expected return we get following policy π from state s:

$$V_\pi(s) = \mathbb{E}_\pi \left[G_t | s_t = s \right]$$

We also define value of a policy $V(\pi)$ to be an expected return of a policy for a random trajectory sampled from a distribution of π on H^*.

$$V_\pi = \mathbb{E}_\pi \left[G_0 \right] = \mathbb{E}_{p_0(s_0)} V_\pi(s_0)$$

Definition 1.2.3. Action value

An action-value $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is an expected return we get following policy π from state s after choosing action a:

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[G_t | s_t = s, a_t = a \right]$$

Definition 1.2.4. Advantage

An advantage of a state-action pair is a difference:

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

Definition 1.2.5. Discounted visitation frequencies

For every state $s \in \mathcal{S}$ we can count how many times on average (discounted) we visit state in each trajectory:

$$\mathrm{dvf}_\pi(s) = \sum_{t=0}^{T} \gamma^t \mathbb{P}(s_t = s) = \sum_{t=0}^{T} \gamma^t \mathbb{1}_{\{s_t = s\}}$$

1.3 Policy identities

Below I'll prove a few useful identities we'll be referring through subsequent sections.

Lemma 1.3.1.

$$Q_\pi(s, a) = \mathbb{E}_{p(s'|s,a)} \left[\mathcal{R}(s') + \gamma V_\pi(s') \right]$$

Proof.

$$\begin{aligned}
Q_\pi(s,a) &= \mathbb{E}_\pi\left[G_t | s_t = s, a_t = a\right] \\
&= \mathbb{E}_\pi\left[\mathcal{R}(s'_t) + \gamma G_{t+1} | s_t = s, a_t = a\right] \\
&= \mathbb{E}_\pi\left[\mathcal{R}(s'_t) | s_t = s, a_t = a\right] + \mathbb{E}_\pi\left[\gamma G_{t+1} | s_t = s, a_t = a\right] \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s'_t)\right] + \mathbb{E}_{p(s'|s,a)}\mathbb{E}_\pi\left[\gamma G_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\right] \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s'_t)\right] + \mathbb{E}_{p(s'|s,a)}\gamma\mathbb{E}_\pi\left[G_{t+1} | s_{t+1} = s'\right] \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s'_t)\right] + \mathbb{E}_{p(s'|s,a)}\gamma V_\pi(s') \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s') + \gamma V_\pi(s')\right]
\end{aligned}$$

□

Lemma 1.3.2.

$$A_\pi(s,a) = \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s') + \gamma V_\pi(s') - V_\pi(s)\right]$$

Proof. Follows straight from lemma 1.3.1.

$$\begin{aligned}
A_\pi(s,a) &= Q_\pi(s,a) - V_\pi(s) \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s') + \gamma V_\pi(s')\right] - V_\pi(s) \\
&= \mathbb{E}_{p(s'|s,a)}\left[\mathcal{R}(s') + \gamma V_\pi(s') - V_\pi(s)\right]
\end{aligned}$$

□

Theorem 1.3.1. *For two policies π and π', we have the identity*

$$V_{\pi'} = V_\pi + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t A_\pi(s_t, a_t)\right]$$

Proof. Follows from lemma 1.3.2 and a telescoping sum of state-values.

$$\begin{aligned}
\mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t A_\pi(s_t, a_t)\right] &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t \left(r_t + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\right)\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t r_t\right] - \mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t \left(V_\pi(s_t) - \gamma V_\pi(s_{t+1})\right)\right] \\
&= \mathbb{E}_{\pi'}\left[G_0\right] - \mathbb{E}_{\pi'}\left[V_\pi(s_0)\right] \\
&= \mathbb{E}_{\pi'}\left[G_0\right] - \mathbb{E}_{p_0(s_0)}\left[V_\pi(s_0)\right] \\
&= V_{\pi'} - V_\pi
\end{aligned}$$

□

2 Policy gradient methods

2.1 Trust region policy optimization

Let's rewrite the equation for a policy-value using states instead of frequencies:

$$V_{\pi'} = V_\pi + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{T} \gamma^t A_\pi(s_t, a_t)\right] \tag{2.1.1}$$

$$= V_\pi + \sum_{t=0}^{T} \gamma^t \left[\mathbb{E}_{\{s_t=s, a_t=a|\pi'\}} A_\pi(s_t, a_t)\right] \tag{2.1.2}$$

$$= V_\pi + \sum_{t=0}^{T} \gamma^t \left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} A_\pi(s, a)\pi'(a|s)\mathbb{P}(s_t = s|\pi')\right] \tag{2.1.3}$$

$$= V_\pi + \sum_{s\in\mathcal{S}} \left[\sum_{t=0}^{T} \gamma^t \mathbb{P}(s_t = s|\pi)p \sum_{a\in\mathcal{A}} A_\pi(s, a)\pi'(a|s)\right] \tag{2.1.4}$$

$$= V_\pi + \sum_{s\in\mathcal{S}} \left[\mathrm{dvf}_{\pi'}(s) \sum_{a\in\mathcal{A}} A_\pi(s, a)\pi'(a|s)\right] \tag{2.1.5}$$

We are considering policy update $\pi \to \pi'$, and want to make sure that $V_{\pi'} \geq V_\pi$. We can see from that if for every state $s \in \mathcal{S}$ we have $\sum_{a\in\mathcal{A}} A_\pi(s, a)\pi'(a|s) > 0$ then policy has indeed improved.

Since eq. (2.1.5) is hard to optimize directly, we introduce a local approximation

$$L_\pi(\pi') = V_\pi + \sum_{s\in\mathcal{S}} \left[\mathrm{dvf}_\pi(s) \sum_{a\in\mathcal{A}} A_\pi(s, a)\pi'(a|s)\right] \tag{2.1.6}$$

TRPO paper claims that $L_\pi(\pi')$ matches $V_{\pi'}$ up to first order, that is, for a parametrized policy π_θ we have:

$$L_{\pi_\theta}(\pi_\theta) = V_{\pi_\theta} \tag{2.1.7}$$

$$\nabla_\theta L_{\pi_{\theta'}}(\pi_\theta)|_{\theta=\theta'} = \nabla_\theta V_{\pi_\theta}|_{\theta=\theta'} \tag{2.1.8}$$

Definition 2.1.1. Variation divergence

For two probability distributions p and q, their *Variation divergence* $D_{TV}(p||q)$ is:

$$D_{TV}(p||q) = \frac{1}{2}\sum_i |p_i - q_i| \tag{2.1.9}$$

For policies π and π' we define *Max variation divergence* to be

$$D_{TV}^{\max}(\pi, \pi') = \max_s D_{TV}\left(\pi(\cdot|s)||\pi'(\cdot|s)\right) \tag{2.1.10}$$

Theorem 2.1.1. *Let* $\alpha = D_{TV}^{\max}(\pi, \pi')$. *Then the following holds:*

$$V_{\pi'} \geq L_\pi(\pi') - \frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{2.1.11}$$

where $\varepsilon = \max_{s,a} |A_\pi(s,a)|$

Theorem 2.1.2. *Following relationship holds between variation divergence and KL-divergence:*

$$D_{KL}(p||q) \geq D_{TV}(p||q)^2 \tag{2.1.12}$$

Theorem 2.1.3. *The following holds:*

$$V_{\pi'} \geq L_\pi(\pi') - CD_{KL}^{\max}(\pi, \pi') \tag{2.1.13}$$

where $C = \frac{4\varepsilon\gamma}{(1-\gamma)^2}$

Notation 2.1.1.

$$\begin{aligned} V_\theta &= V_{\pi_\theta} \\ L_\theta(\theta') &= L_{\pi_\theta}(\pi_{\theta'}) \\ D_{KL}(\theta||\theta') &= D_{KL}(\pi_\theta, \pi_{\theta'}) \end{aligned}$$

By maximizing objective

$$\text{maximize}_{\theta'} \left[L_\pi(\pi') - CD_{KL}^{\max}(\pi, \pi')\right] \tag{2.1.14}$$

we would guarantee monotonically improving policy, but the step size would be very small. Instead authors propoze a *trust-region update*:

$$\text{maximize}_{\theta'} L_\pi(\pi') \tag{2.1.15}$$

$$\text{subject to } D_{KL}^{\max}(\pi, \pi') \leq \delta \tag{2.1.16}$$

For practical purposes we estimate $D_{KL}^{\max}(\pi, \pi')$ by

$$\overline{D}_{KL}^{\pi}(\theta, \theta') = \mathbb{E}_{s\sim\pi}\left[D_{KL}(\pi(\cdot|s), \pi'(\cdot|s))\right]$$