DISTRIBUTIONAL REINFORCEMENT LEARNING

SPRING 2024

Hyunin Lee
Ph.D. student
UC Berkeley
hyunin@berkeley.edu

Contents

1 Chapter 2

1.1 Random Variables and Their Probability Distributions

1.2 Markov Decision Processes

Definition 1.1 (Transition dynamics)**.** We define transition dynamics $\boldsymbol{P} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R} \times \mathcal{X})$ that provides the joint probabiltiy distirbuiotn of R_t and X_{t+1} in ertns of state X_t and action A_t.

$$R_t, X_{t+1} \sim \boldsymbol{P}(\cdot, \cdot | X_t, A_t)$$

Definition 1.2 (Reward distribution)**.** $R_t \sim \boldsymbol{P}_{\mathcal{R}}(\cdot \mid X_t, A_t)$

Definition 1.3 (Transition kernel)**.** $X_{t+1} \sim \boldsymbol{P}_{\mathcal{X}}(\cdot \mid X_t, A_t)$

Definition 1.4 (Markov Decision Process (MDP))**.** MDP is a tuple $(\mathcal{X}, \mathcal{A}, \xi_0, \boldsymbol{P}_{\mathcal{X}}, \boldsymbol{P}_{\mathcal{R}})$

Definition 1.5 (Policy)**.** A policy is a maaping $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ rom state to probabilty distributions over actions.

$$A_t \sim \pi(\cdot | X_t)$$

1.3 The Pinball Model

1.4 The Return

Definition 1.6 (Return G)**.** $G = \sum_{t=0}^{\infty} \gamma^t R_t$

The return is a sum of scaled, real-valued random variables and is therefore itself a random variable.

Assumption 1.7. For each state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, the reward distribution $\boldsymbol{P}_{\mathcal{R}}(\cdot \mid x, a)$ has finite first moment. This is if $R \sim \boldsymbol{P}_{\mathcal{R}}(\cdot \mid x, a)$, then

$$\mathbb{E}\left[|R|\right] < \infty.$$

Proposition 1.8. Under Assumption 1.7, the random return G exists and is finite with proabbility 1, in the sense that

$$\mathbb{P}_\pi\left(G \in (-\infty, \infty)\right) = 1.$$

1.5 Properties of the Random Trajectory

Definition 1.9 (Probablity distribution of random variable Z)**.** We denote $\mathcal{D}(Z)$ as the probability distribution of random variable Z. When Z is real-valued, then for $S \in \mathbb{R}$, we have

$$\mathcal{D}(Z)(S) = \mathbb{P}(Z \in S)$$

Also, we denote $\mathcal{D}_\pi(Z)$ as

$$\mathcal{D}_\pi(Z)(S) = \mathbb{P}_\pi(Z \in S)$$

1.6 The Random-Variable Bellman Equation

Definition 1.10 (Return-variable function)**.** $G^\pi = \sum_{t=0}^{\infty} \gamma^t R_t,\ X_0 = x.$

Formally, G^π is a collection of random variables indexed by an initial state x, each generated by a random trajectory $(X_t, A_t, R_t)_{t\geq 0}$ under the distribution $\boldsymbol{P}(\cdot|X_0 = x)$.

Proposition 1.11 (The random-variable Bellman equation)**.** Let G^π be the return-variable function of policy π. For a sample transition $(X = x, A, R, X')$, it holds that for any state $x \in \mathcal{X}$,

$$G^\pi(x) \overset{\mathcal{D}}{=} R + \gamma G^\pi(X')$$

1.7 From Random Variables to Probability Distributions

Recall the notation that for a real-valued cariable Z with probablity distribution $\nu \in \mathscr{P}(\mathbb{R})$, we define

$$\nu(S) = \mathbb{P}(Z \in S),\ S \subseteq \mathbb{R}.$$

In a same way, for each state $x \in \mathcal{X}$, let us denote the distribution of the random variable $G^\pi(x)$ by $\eta^\pi(x)$. Using this notation ,we have

$$\eta^\pi(x)(S) = \mathbb{P}(G^\pi(x) \in S),\ S \subseteq \mathbb{R}.$$

We call the collection of these per-state distribution the return-distirbuion function. Note that $\eta^\pi(x) \in \mathscr{P}(\mathbb{R})^{\mathcal{X}}$.

1.7.1 Mixing

Recall that for return-variable G^π and return-distribution function η^π, we have defined

$$\mathcal{D}_\pi(G^\pi(X')|X = x)(S) \overset{\text{def}}{=} \mathbb{P}_\pi(G^\pi(X') \in S|X = x).$$

Now, let's take a look at $\mathbb{P}_\pi$ term.

$$\begin{aligned}
\mathcal{D}_\pi(G^\pi(X')|X=x)(S) &\stackrel{\text{def}}{=} \mathbb{P}_\pi(G^\pi(X') \in S|X=x) \\
&= \sum_{x' \in \mathcal{X}} \mathbb{P}_\pi(X'=x'|X=x)\mathbb{P}_\pi(G^\pi(X') \in S|X'=x', X=x) \\
&= \sum_{x' \in \mathcal{X}} \mathbb{P}_\pi(X'=x'|X=x)\mathbb{P}_\pi(G^\pi(x') \in S) \\
&= \left(\sum_{x' \in \mathcal{X}} \mathbb{P}_\pi(X'=x'|X=x)\eta^\pi(x')\right)(S)
\end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned}
\mathcal{D}_\pi(G^\pi(X')|X=x)(S) &= \sum_{x' \in \mathcal{X}} \mathbb{P}_\pi(X'=x'|X=x)\eta^\pi(x') \\
&= \mathbb{E}_\pi\left[\eta^\pi(X') \mid X=x\right]
\end{aligned}$$

hte indexing step also has a implse expression interms of cumulative distribution functinos