CHAPTER 1

DEEP REINFORCEMENT LEARNING

1 Neural Networks

...

2 Distributional Reinforcement Learning

Remember that we defined $\mathbb{P}^{\pi}_{s,a} := \mathbb{P}^{\pi} \otimes \delta_{S_0}(s) \otimes \delta_{A_0}(a)$ as the probability measure of the Markov reward process (S, A, R) started in (s, a). We define the random variable of the return under policy π as

$$Z^{\pi} := \sum_{t=0}^{\infty} \gamma^t R_t, \quad \gamma \in (0, 1).$$

Unlike the methods before, where we were interested in the expected reward $Q^{\pi}(s, a) = \mathbb{E}^{\pi}_{s,a}[Z^{\pi}]$, we are now interested in the distribution of these cumulative rewards. For that define

$$\eta^{\pi}_{s,a}(B) := \mathbb{P}^{\pi}_{s,a}(Z^{\pi} \in B), \quad B \in \mathcal{B}(\mathbb{R}).$$

In analogy to weak solutions for PDEs, i.e. in sense of distribution, the return distribution $\eta^{\pi}_{s,a}$ satisfies the Bellman equation in sense of distribution:

$$\forall \phi \in C_b(\mathbb{R}) : \quad \int_{\mathbb{R}} \phi(z) d\eta^{\pi}_{s,a}(z) = \mathbb{E}^{\pi}_{s,a}[\int_{\mathbb{R}} \phi(R + \gamma z) d\eta_{S',A'}(z)].$$

We define $f_{r,\gamma}(z) := r + \gamma z$ and the push forward

$$((\eta^{\pi}_{s,a})_{f_{r,\gamma}})(B) := \eta^{\pi}_{s,a}(f^{-1}_{r,\gamma}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

then, the above can be written as

$$\begin{aligned} &\forall \phi \in C_b(\mathbb{R}) : \quad \int_{\mathbb{R}} \phi(z) d\eta^{\pi}_{s,a}(z) = \mathbb{E}^{\pi}_{s,a}[\int_{\mathbb{R}} \phi(R + \gamma z) d\eta_{S',A'}(z)] = \mathbb{E}^{\pi}_{s,a}[\int_{\mathbb{R}} \phi(z) d(\eta_{S',A'})_{f_{R,\gamma}}(z)] \\ \iff &\forall \phi \in C_b(\mathbb{R}) : \quad \langle \phi, \eta^{\pi}_{s,a} \rangle = \mathbb{E}^{\pi}_{s,a}[\langle \phi, (\eta^{\pi}_{S',A'})_{f_{R,\gamma}} \rangle]. \end{aligned}$$

For any $\phi \in C_b(\mathbb{R})$ and any measure ν we have

$$\begin{aligned} (\nu * \delta_r)(\phi) &= \int_{\mathbb{R}} \phi(z) d(\nu * \delta_r)(z) = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x + y) d\nu(x) d\delta_r(y) \\ &= \int_{\mathbb{R}} \phi(x + r) d\nu(x) = (\nu(\cdot - r))(\phi). \end{aligned}$$

Next, define $D_\gamma(x) := \gamma x$, then if $Z \sim \eta^\pi(x,a)$ then $\gamma Z \sim (\eta^\pi(s,a))_{D_\gamma}$. In total

$$Z\gamma + r \sim (\eta^\pi(s,a))_{D_\gamma}(\cdot - r) = ((\eta^\pi(s,a))_{D_\gamma} * \delta_r)$$

holds in sense of distribution. Now define the distributional Bellman operator in sense of distribution as

$$\mathcal{T}^\pi Z^\pi(s,a) = R(s,a) + \gamma Z^\pi(S', A'),$$

i.e. the distribution

$$\mathcal{T}^\pi : (\mathcal{P}(\mathbb{R}))^{\mathcal{S}\times\mathcal{A}} \to (\mathcal{P}(\mathbb{R}))^{\mathcal{S}\times\mathcal{A}}, (\nu_{s,a})_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mapsto ((\mathcal{T}^\pi \nu)_{s,a})_{(s,a)\in\mathcal{S}\times\mathcal{A}}$$

is point wise defined for all $\phi \in C_b(\mathbb{R})$ as

$$(\mathcal{T}^\pi \nu)_{s,a}(\phi) := \mathbb{E}^\pi_{s,a}[((\nu(S', A'))_{D_\gamma} * \delta_R)(\phi)] = \sum_{s',a',r} \pi(a'; s')p(s', r; s, a)((\nu_{s',a'})_{D_\gamma} * \delta_r)(\phi)$$