Chapter 4 Extensions

Actor-Critic

The TD Target and the TD Residual

Why the Expected TD Residual Vanishes at the True Value Function

In the main chapter, we described the temporal-difference residual as a correction signal and stated that its expected value is zero when the value network represents the true value function. This result follows directly from the Bellman equation.

Recall the definition of the TD residual from Equation 4.4:

$$\delta_t \stackrel{\text{def}}{=} r_{t+1} + \gamma V_{\boldsymbol{\phi}}(o_{t+1}) - V_{\boldsymbol{\phi}}(o_t).$$

Suppose that the network is exact, so that

$$V_{\boldsymbol{\phi}}(o) = V^{\pi_{\boldsymbol{\theta}}}(o)$$

for every observation o. Under this assumption, the residual becomes

$$\delta_t = r_{t+1} + \gamma V^{\pi_{\boldsymbol{\theta}}}(o_{t+1}) - V^{\pi_{\boldsymbol{\theta}}}(o_t).$$

Now condition on the current observation o_t. The action is still sampled from $\pi_{\boldsymbol{\theta}}(\cdot \mid o_t)$, and the environment may still produce different rewards and next observations, so the first two terms remain random. The value $V^{\pi_{\boldsymbol{\theta}}}(o_t)$, however, is fixed once o_t is given. Therefore,

$$\begin{aligned}\mathbb{E}[\delta_t \mid o_t] &= \mathbb{E}[r_{t+1} + \gamma V^{\pi_{\boldsymbol{\theta}}}(o_{t+1}) - V^{\pi_{\boldsymbol{\theta}}}(o_t) \mid o_t] \\ &= \mathbb{E}[r_{t+1} + \gamma V^{\pi_{\boldsymbol{\theta}}}(o_{t+1}) \mid o_t] - V^{\pi_{\boldsymbol{\theta}}}(o_t).\end{aligned}$$

The expectation in the final line is the same expression as the right-hand side of the Bellman equation 4.1:

$$V^{\pi_{\boldsymbol{\theta}}}(o_t) = \mathbb{E}[r_{t+1} + \gamma V^{\pi_{\boldsymbol{\theta}}}(o_{t+1}) \mid o_t].$$

Substituting this equality gives

$$\mathbb{E}[\delta_t \mid o_t] = V^{\pi_{\boldsymbol{\theta}}}(o_t) - V^{\pi_{\boldsymbol{\theta}}}(o_t) = 0.$$

The conclusion concerns the *expected* residual, not every residual observed on an individual transition. Even when the value function is exact, a particular action, reward, or transition may produce a positive or negative δ_t. Across repeated transitions from the same observation under the same policy, those fluctuations average to zero.