RL 2022

Lecture 4. TD & AC

$$\max_{\theta} J^{\theta}(x_0) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(X_k, U_k)\right]$$

$$\text{s.t.} \quad X_{k+2} \sim f(x_{k+2} \mid x_k, u_k)$$

$$U_k \sim \rho^{\theta}(u_k \mid x_k)$$

The HJB for the stated problem reads:

$$\forall x \in X \quad V^*(x) = \max_u \left\{ r(x,u) + \gamma \mathbb{E}\left[V^*(X_+^u)\right] \right\}$$

where $X_+^u \sim f(x_{k+2} \mid x_k, u)$

Verification principle: if a fuction W fulfills the HJB, then it is the optimal value function: $W = V^*$

$$V^*(x_k) = \max_u \left\{ r(x_k, u) + \gamma \mathbb{E}\left[V^*(X_{k+2}^u)\right] \right\}$$

The idea of TD is, roughly, to „equalize the lhs of the HJB with its rhs under a model $\hat{V}^w$ with weights w"!

The goal is $\hat{V}^w \approx V^*$

Let's recall tabular value iteration:

suppose $\mathcal{X}_0 = \{x^j\}_{j=1}^N$

Init $V_0 \equiv 0$, $\rho_0 \equiv 0$

Loop over i

$$\forall x^i \in \mathcal{X}_0 \quad V_{i+1}(x^i) := r(x^i, \rho_i(x^i)) + \gamma \mathbb{E}\left[V_i\left(X^{\rho_i(x^i)}\right) \middle| X = x^i \right]$$

$$\forall x^i \in \mathcal{X}_0 \quad \rho_{i+1}(x^i) := \operatorname{argmax}_u \left\{ r(x,u) + \gamma \mathbb{E}\left[V_{i+1}\left(X_+^u\right) \right] \right\}$$

Instead V_i, let's use a model $\hat{V}^w$ and so the above algorithm will be cast to, while also parametrizing the policy, i.e., ρ^θ:

Init w_0, θ_0

Loop over time steps k

$$w_{k+1} := \arg\min_w \left\{ \frac{1}{2} \left(\hat{V}^w(x_k) - r(x_k, \rho^{\theta_k}(x_k)) - \gamma \mathbb{E}\left[\hat{V}^{w_k}\left(X_{k+1}^{\rho^{\theta_k}(x_k)} \right) \right] \right)^2 \right\}$$

$$\theta_{k+1} := \arg\max_\theta \left\{ r(x_k, \rho^\theta(x_k)) + \gamma \mathbb{E}\left[\hat{V}^{w_{k+1}}\left(X_{k+1}^{\rho^\theta(x_k)} \right) \right] \right\}$$

$\hat{V}^w$ is associated with the so-called critic

ρ^θ is associated with the so-called actor

The difference in the (....) is referred to as the temporal difference

Upgrades and modifications

- Action-value $Q^\rho(x,u)$

$$Q^\rho(x,u) = r(x,u) + \gamma \mathbb{E}\left[J^\rho(X_+^u)\right]$$

$$Q^*(x,u) = r(x,u) + \gamma \mathbb{E}\left[V^*(X_+^u)\right]$$

$$\rho^*(x) = \underset{u}{\operatorname{argmax}}\ Q^*(x,u)$$

Casting this into a model $\hat{Q}^w(x,u)$, e.g., $\langle w, (\varphi_x(x), \varphi_u(u), \varphi_{xu}(x,u)) \rangle$

say, $\varphi_{xu}(x,u) = x \cdot u + x^2 \cdot u^2 + \ldots$

Action-value update

$$w_{k+1} := \arg\min_{w} \frac{1}{2} TD_{k-1}^2$$

$$TD_{k-1} = \hat{Q}^{w}(x_{k-1}, u_{k-1}) - r(x_{k-1}, u_{k-1}) - \hat{Q}^{w_k}(x_k, u_k)$$

$$TD_{k-2}$$

...

over an experience replay

policy update

$$\rho^{\theta_{k+1}}(x_k) := \arg\max_{\theta} \hat{Q}^{w_{k+1}}(x_k, \rho^{\theta}(x_k))$$

- Update target

was $\underbrace{r(x,u) + \gamma \mathbb{E}\left[\hat{V}^{w^-}(x_+^u)\right]}_{Q^{w^-}(x,u)}$

(TD(0))

TD(N):

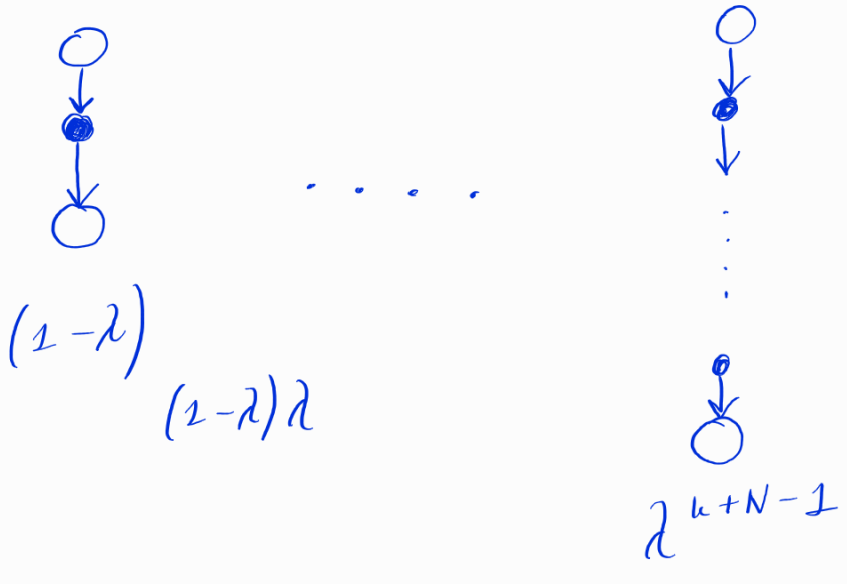
$$Q_N^{\omega_-}(x,u) := r(x,u) + \mathbb{E}\left[\sum_{k=1}^{N-1} \gamma^k r(X_{+k}, U_{+k}) + \gamma^N \hat{V}^{\omega_-}(X_{+N})\right],$$

where X_{+k} is the state k steps ahead of x

There is also

$$Q_\lambda^{\omega_-}(x,u) = (1-\lambda)\sum_{k=1}^{N-1} \lambda^{k-1} Q_k^{\omega_-}(x,u) + \lambda^{N-1} \sum_{j=1}^{N} \gamma^{j-1} \mathbb{E}\left[r(X_{+j}, U_{+j})\right]$$

Visualization: (backup diagrams)



TD(λ)

- Eligibility traces

Introduce z_k (a vector)

$z_0 := 0$

Then, for k

$$z_{k+1} := \gamma \lambda z_k + \nabla_w \hat{V}^w \Big|_{w=w_k}$$