3.3 Returns and Episodes

G_t = Expected Reward Return T = final time step
$= R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$

S = set of all states
S^+ = Terminal state

Discounting is agent tries to select actions such that sum of discounted rewards over time is maximized.

$$G_t = R_{t+1} + \lambda R_{t+2} + \lambda^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \lambda^k R_{t+k+1}$$

λ = discount rate $0 \leq \lambda \leq 1$

$$\lambda = \begin{cases} < 1 & \text{sum has finite value as long as } R_t \text{ is bounded} \\ = 0 & \text{myopic = only cares about } R_{t+1} \end{cases}$$

As $\lambda \to 1$, takes future rewards more strongly.

$$G_t = R_{t+1} + \lambda G_{t+1}$$

Ex 3.8
$G_0 = R_1 + \lambda G_1 = 2$
$G_1 = R_2 + \lambda G_2 = 6$
$G_2 = R_3 + \lambda G_3 = 8$
$G_3 = R_4 + \lambda G_4 = 4$
$G_4 = R_5 + \lambda G_5 = 2$
$G_5 = R_6 + R_7 + \dots = 0$

3.5 Policies and Value Functions

Value Function: How good agent to be in current state

Policy: Mapping of states to probabilities 1. can only depend on current state

$V_\pi(s)$ = expected return when starting in s and following π

* state-value function

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \lambda^k R_{t+k+1} \mid S_t = s\right] \quad \forall s \in S$$

$q_\pi(s,a)$ = expected return when starting in s, taking action a, then following policy π

* action-value policy

$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \lambda^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

This must hold between value of s and possible successor states.

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

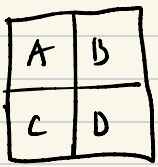
$$= E_\pi[R_{t+1} + \lambda G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a)\left[r + \lambda E_\pi[G_{t+1} \mid S_{t+1} = s']\right]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \lambda V_\pi(s')\right]$$

Bellman Equation look ahead

Ex:



$$V_\pi(A) = \frac{1}{4}(0 + 0.7 V_\pi(C)) + \frac{1}{4}(5 + 0.7 V_\pi(B)) + \frac{1}{2}(0 + 0.7 V_\pi(A))$$

Ex 3.14

$$V_\pi(0.7) = \frac{1}{4}(0.9 \cdot 0.9) + \frac{1}{4}(0.9 \cdot -0.9) + \frac{1}{4}(0.9 \cdot 0.7) + \frac{1}{4}(0.9 \cdot 2.3)$$
$$\approx 0.7$$

3.6 Optimal Policies and Optimum Value Function

V_* = optimal policy — $\pi_1 \geq \pi_2$ if $V_{\pi_1}(s) \geq V_{\pi_2}(s)$ $\forall s \in S$

$= \max_\pi V_\pi(s)$ $\forall s \in S$ - Always an optimal policy

q_* = optimum action-value function

$= \max_\pi q_\pi(s,a)$

$= E\left[R_{t+1} + \lambda V_*(S_{t+1}) \mid S_t = s, A_t = a\right]$

Bellman Optimum Equality

Value of a state under policy must be expected return for best action from state

$$V_*(s) = \max_{a \in A(s)} q_{\pi_*}(a,s)$$

$$= \max_a \sum_{s',r} p(s',r|s,a)\left[r + \lambda V_*(s')\right]$$

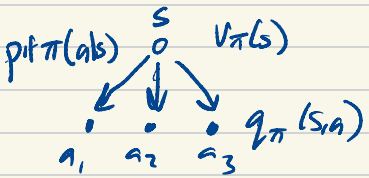
$$Q_*(s) = \sum_{s'} P(s'|s,a)\left[R(s,a) + \lambda \max_{a'} Q^*(s',a')\right]$$

Ex 3.11

$$E_\pi[R_{t+1}|S_t] = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)[r]$$

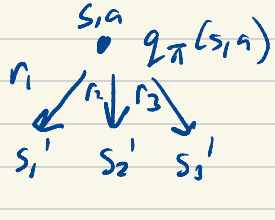
Ex 3.12

$$V_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$



Ex 3.13

$$q_\pi(s) = \sum_{s' \in S} \sum_{r \in R} p(s',r|s,a) \cdot [r + \lambda V_\pi(s_{t+1})]$$



$$\pi_*(s) = \arg\max_a \sum_{s'} \sum_r p(s',r|s,a)[r + \lambda V_*(s')]$$

w/ state

$$\pi_*(s) = \arg\max_a q_*(s,a)$$

w/ action-value