May 2

①

Textbook Sections 17.1 - 17.3

$$U_h([s_0, s_1, \ldots, s_t, \ldots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t)$$

Note: If for all s $R(s) \le R_{max}$
then $\sum_{t=0}^{\infty} \gamma^t R(s_t) \le \sum_{t=0}^{\infty} \gamma^t R_{max}$
$= \dfrac{R_{max}}{1-\gamma}$

$$= R(s_0) + \sum_{t=1}^{\infty} \gamma^t R(s_t)$$

$$= R(s_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t)$$

$$= R(s_0) + \gamma \sum_{t=0}^{\infty} \gamma^t R(s_{t+1})$$

$$= R(s_0) + \gamma U_h([s_1, s_2, \ldots])$$

Value of a policy:

$$U^{\pi}(s_0) = E(U_h([s_0, s_1, \ldots])) = E\left(\sum_{t=0}^{\infty} \gamma^t R(s_t)\right)$$

where $a_t = \pi(s_t)$
$s_{t+1} = a_t(s_t)$ - stochastic, expectation is over those probabilities

$$= E_{s_1 \in S}(R(s_0) + \gamma U^{\pi}(s_1)) \quad a_0 = \pi(s_0) \quad s_1 = a_0(s_0)$$

$$= R(s_0) + \gamma E_{s_1 \in S}(U^{\pi}(s_1)) = R(s_0) + \gamma \sum_{s_1 \in S} P(s_1 | s_0, a) U(s_1)$$

Optimal Policy

$$\pi^*(s) = \operatorname{argmax}_{\pi} U^{\pi}(s)$$

Optimal Value

$$U^{\pi^*}(s) = U(s)$$

If you have $U(s)$ only, can be used to define π^*

$$\pi^*(s) = \operatorname{argmax}_{a \in A} \sum_{s' \in S} P(s' | s, a) U(s')$$

$U(s) = R(s) + \gamma U(s')$ where $a = \pi^*(s)$ $s' = a(s)$

$= R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) U(s')$ Bellman Equation

Can be used to compute $U(s)$ for all s given known R values and know $P(s'|s,a)$

Start with $U_0 = 0$ for all states

$$U_{i+1}(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) U_i(s')$$

Value-Iteration(S, A, P, R, γ)
  For all $s \in S$ $U'(s) = 0$
  Repeat
    For all $s \in S$ $U(s) \leftarrow U'(s)$
    For all $s \in S$
      $U'(s) \leftarrow R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) U(s')$
  Until stopping condition
  Return U

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s'|s,a) U(s')$$

Can also learn policies directly in a similar way

Start with random π_0

It defines $U_0^{\pi}(s)$

Having π_i simplifies Bellman Equation

$$U_i(s) = R(s) + \gamma \sum_{s' \in S} P(s'|s,a) U_i(s')$$

where $a = \pi_i(s)$