Bellman Equation

- Bellman equation for optimal $V^*(s)$:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

- Bellman equation for $Q^*(s,a)$:

$$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma(\max_{a'}(Q^*(s',a')))]$$

- Optimal:
  1. Take the first optimal action
  2. Keep being optimal

Policy Extraction

- One step lookahead
- Gets the policy from a value function
- Given an optimal $V^*(s)$ you would use

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

  - $\arg\max$ gives you the index of the optimal value in a list
- Given $Q^*(s,a)$:

$$\pi^* = \arg\max_a Q^*(s,a)$$

  - Actions are easier to select from q-values than values

Fixed Policies

- Utility of a state $s \triangleq V^\pi(s) \mathrel{\widehat{=}}$ expected total discounted rewards starting in s and following π
- Recursive relation:

$$V^\pi(s) = \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V^\pi(s')]$$

- Iterative definition:

$$V_k^\pi = \begin{cases} 0, k = 0 \\ V_{k+1}^\pi(s) = \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V_k^\pi(s')] \end{cases}$$

  - $O(s^2)$ per iteration

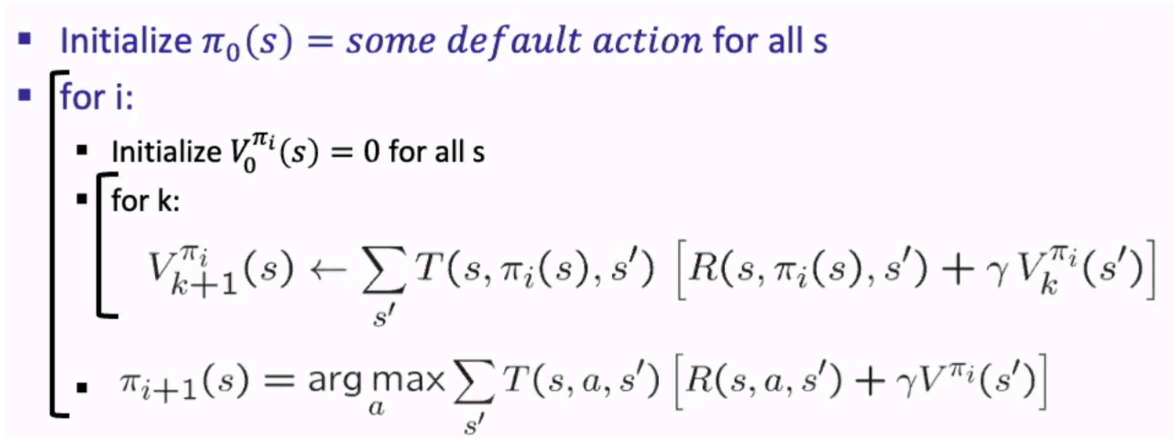
Policy Iteration

- Alternative approach for optimal values
- Alternate between updating policy function and updating value function
- This is what ends up being used
- Steps:
  1. Policy evaluation: calculate utilities for some fixed policy (not optimal utilities) until convergence
  2. Policy improvement: update policy using one-step look-ahead with resulting converged (but not optimal) utilities as future values
  3. Repeat steps until convergence
- Evaluation:

$$V_{k+1}^{\pi_i} = \sum_{s'} T(s,\pi_i(s),s')[R(s,\pi_i(s),s') + \gamma V_k^{\pi_i}(s')]$$

- Improvement:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s, \pi_i(s), s')[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s')]$$



Summary

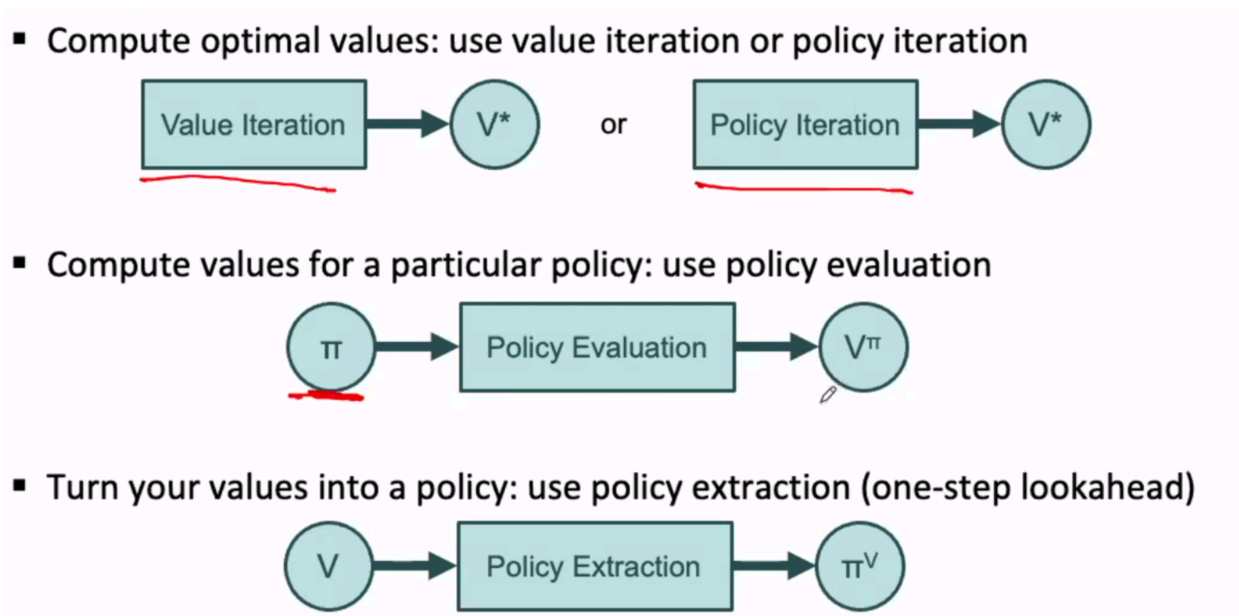


Figure 1: Screenshot_2023-09-21_at_6.21.43_PM.png

-
-

- **Optimal V and Q value functions:**

$$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s')\right] \qquad V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a') \right]$$

- **Value function for fixed policy π:**

$$V^\pi(s) = \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V^\pi(s')]$$

- **Policy π for V and Q value functions:**

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

Figure 2: Screenshot_2023-09-21_at_6.22.54_PM.png