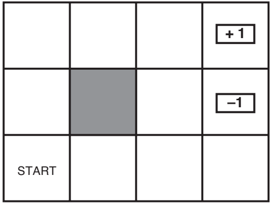
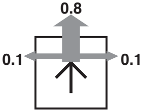
Value/Policy iteration

J. Kostlivá, Z. Straka, P. Švarný, F. Gama





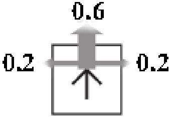
-10	B	30
-10	A	20

We have:

- States: S, Actions: A
- Transition model: $T(s,a,s') \equiv P(s'|s,a)$, we are in state s, take action a and get to state s'
- Reward: $r(s)$, immediate reward
- State value: $V(s)$, Expected sum of rewards when performing optimal actions
- Policy: π, robot/agent behaviour strategy

Example

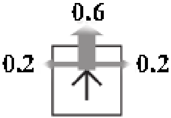
-40	30	-50
-40	*A*	-50
-40	*B*	-50



- Square environment, numbers are rewards
- Red states are terminal
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- Immediate reward $r(A) = r(B) = -1$
- Transition model: see picture
- Forgetting/discount factor: $\gamma = 0.9$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

Task: find optimal policies and values of states A, B

A: $\pi(A) = \uparrow, \pi(B) = \leftarrow$

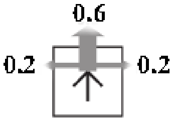
B: $\pi(A) = \rightarrow, \pi(B) = \leftarrow$

C: $\pi(A) = \leftarrow, \pi(B) = \uparrow$

D: $\pi(A) = \uparrow, \pi(B) = \uparrow$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

Task: find optimal policies and values of states A, B

A: $\pi(A) = \uparrow, \pi(B) = \leftarrow$

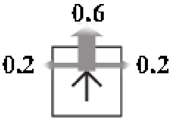
B: $\pi(A) = \rightarrow, \pi(B) = \leftarrow$

C: $\pi(A) = \leftarrow, \pi(B) = \uparrow$

D: $\pi(A) = \uparrow, \pi(B) = \uparrow$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

Task: find optimal policies and values of states A, B

A: $\pi(A) = \uparrow, \pi(B) = \leftarrow$

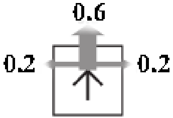
B: $\pi(A) = \rightarrow, \pi(B) = \leftarrow$

C: $\pi(A) = \leftarrow, \pi(B) = \uparrow$

D: $\pi(A) = \uparrow, \pi(B) = \uparrow$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value V(s):

   A: $V(s) = \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

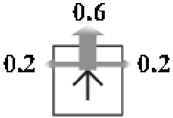
   B: $V(s) = \sum_{s'} \gamma V(s')$

   C: $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

   D: $V(s) = \sum_{s'} [r(s,a,s') + \gamma r(s')]$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value V(s):

   A: $V(s) = \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

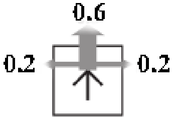
   B: $V(s) = \sum_{s'} \gamma V(s')$

   C: $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

   D: $V(s) = \sum_{s'} [r(s,a,s') + \gamma r(s')]$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value V(s):

   A: $V(s) = \sum_{s'} p(s'|s, a)[r(s, a, s') + \gamma V(s')]$

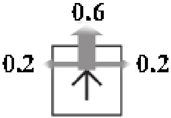
   B: $V(s) = \sum_{s'} \gamma V(s')$

   C: $V(s) = \max_a \sum_{s'} p(s'|s, a)[r(s, a, s') + \gamma V(s')]$

   D: $V(s) = \sum_{s'} [r(s, a, s') + \gamma r(s')]$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

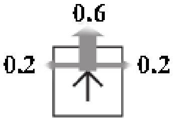
How to do?

1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal strategy $\pi(s)$:

   A: $\pi(s) = \arg\max_a \sum_{s'} [r(s,a,s') + \gamma r(s')]$
   B: $\pi(s) = \sum_{s'} \gamma V(s')$
   C: $\pi(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
   D: $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal strategy $\pi(s)$:

A: $\pi(s) = \arg\max_a \sum_{s'} [r(s,a,s') + \gamma r(s')]$

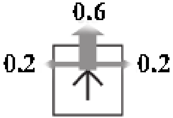
B: $\pi(s) = \sum_{s'} \gamma V(s')$

C: $\pi(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

D: $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

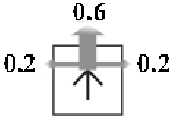
1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal strategy $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] = \arg\max_a V(s)$

We have two methods:

- Value iteration
- Policy iteration

Example

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

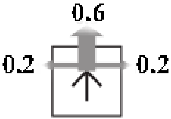
1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal strategy
   $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] = \arg\max_a V(s)$

We have two methods:

- Value iteration
- Policy iteration

Different variants of rewards

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal policy
   $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] = \arg\max_a V(s)$

What if we had another variant of MDP? Let's $r(s)$ be instead $r(s,a,s')$. Then, $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$ changes to:

A: $V(s) = \sum_{s'}[V(s') + \gamma r(s')]$

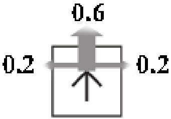
B: $V(s) = r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V(s')$

C: $V(s) = \sum_{s'} p(s'|s,a)[r(s) + \gamma V(s')]$

D: $V(s) = \max\limits_{a \in A(s)} \sum_{s'} \gamma(V(s') + r(s'))$

Different variants of rewards

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

How to do?

1. Calculate the state value $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$
2. Determine the optimal policy $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] = \arg\max_a V(s)$

What about another variant of MDP? Let's $r(s)$ be instead $r(s,a,s')$. Then, $V(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$ changes to:

A: $V(s) = \sum_{s'}[V(s') + \gamma r(s')]$

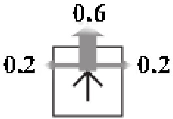
B: $V(s) = r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V(s')$

C: $V(s) = \sum_{s'} p(s'|s,a)[r(s) + \gamma V(s')]$

D: $V(s) = \max\limits_{a \in A(s)} \sum_{s'} \gamma(V(s') + r(s'))$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



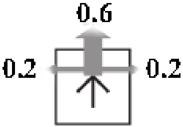
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$

Use Bellman update $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Iterate until the change of state value $V(s)$ between two iterations is lower than ϵ

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

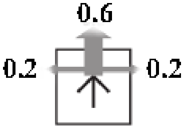
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

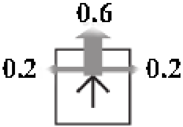
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

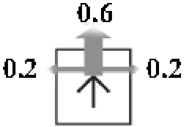
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cc} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cc} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

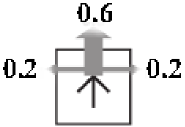
$t = 0: \quad V(A) = 0, V(B) = 0$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

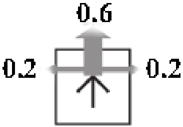
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

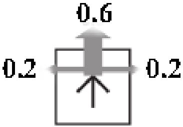
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

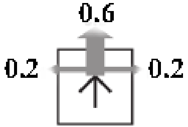
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

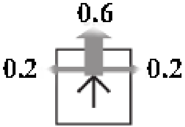
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

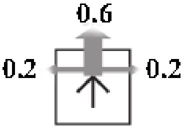
$$t = 0: \quad V(A) = 0, V(B) = 0$$

$$t = 1: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot 0 = -24 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot 0 = -18 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -1 \quad (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 0 + 0.2 \cdot 0 = -30 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 0 + 0.2 \cdot 0 = -24 \\ (\uparrow) & 0.6 \cdot 0 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18 \\ (\downarrow) & 0.6 \cdot 0 + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -18 \end{array} \right\} = -17.2 \quad (\uparrow)/(\downarrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 2: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -27.44 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -21.44 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -1 \ \ (\uparrow)$$

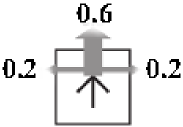
$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -33.64 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -27.64 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -17.74 \ (\uparrow)$$

$$t = 3: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -27.548 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -21.548 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -1 \ \ (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{cl} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -33.748 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -27.748 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -17.74 \ (\uparrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 2: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -27.44 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -21.44 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -1 \ (\uparrow)$$

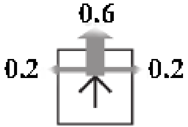
$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -33.64 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -27.64 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -17.74 \ (\uparrow)$$

$$t = 3: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -27.548 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -21.548 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -1 \ (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -33.748 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -27.748 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -17.74 \ (\uparrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 2 : \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -27.44 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -21.44 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -1 \ (\uparrow)$$

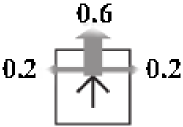
$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -33.64 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -27.64 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -17.74 \ (\uparrow)$$

$$t = 3 : \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -27.548 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -21.548 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -1 \ (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -33.748 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -27.748 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -17.74 \ (\uparrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 2: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -27.44 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -21.44 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -1 \ \ (\uparrow)$$

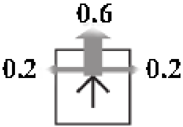
$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -33.64 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -27.64 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{array} \right\} = -17.74 \ (\uparrow)$$

$$t = 3: \quad V(A) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -27.548 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -21.548 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -1 \ \ (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \left\{ \begin{array}{ll} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -33.748 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -27.748 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{array} \right\} = -17.74 \ (\uparrow)$$

Example - Value Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 2: \quad V(A) = -1 + 0.9 \cdot \max \begin{Bmatrix} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -27.44 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.2) = -21.44 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{Bmatrix} = -1 \ (\uparrow)$$

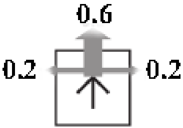
$$V(B) = -1 + 0.9 \cdot \max \begin{Bmatrix} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -33.64 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.2) = -27.64 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.2) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.32 \end{Bmatrix} = -17.74 \ (\uparrow)$$

$$t = 3: \quad V(A) = -1 + 0.9 \cdot \max \begin{Bmatrix} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -27.548 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot 30 + 0.2 \cdot (-17.74) = -21.548 \\ (\uparrow) & 0.6 \cdot 30 + 0.2 \cdot (-40) + 0.2 \cdot (-50) = 0 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{Bmatrix} = -1 \ (\uparrow)$$

$$V(B) = -1 + 0.9 \cdot \max \begin{Bmatrix} (\rightarrow) & 0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -33.748 \\ (\leftarrow) & 0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74) = -27.748 \\ (\uparrow) & 0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50) = -18.6 \\ (\downarrow) & 0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40) = -28.64 \end{Bmatrix} = -17.74 \ (\uparrow$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration:

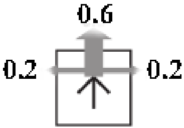
A: 1 step

B: 2 steps

C: 3 steps

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration:

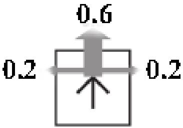
A: 1 step

B: 2 steps

C: 3 steps

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration:

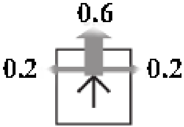
A: 1 step

B: 2 steps

C: 3 steps

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



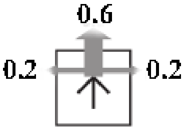
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration: 2 steps

1. Policy Evaluation:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



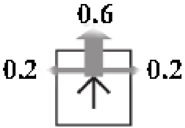
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration: 2 steps

1. Policy Evaluation:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



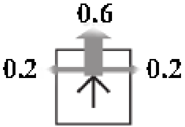
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 1 iteration: 2 steps

1. Policy Evaluation:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



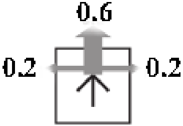
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 2 steps:

1. Policy evaluation:
   - $V^{\pi_i}_{k+1}(s) \leftarrow \sum_{s'} p(s'|s,\pi_i(s))[r(s,\pi_i(s),s') + \gamma V^{\pi_i}_k(s')]$
   - Iteratively or analytically
2. Policy refinement:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



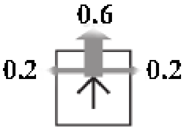
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 2 steps:

1. Policy evaluation:
   - $V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} p(s'|s, \pi_i(s))[r(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s')]$
   - Iteratively or analytically
2. Policy refinement:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



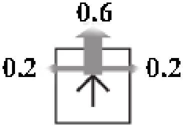
- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Policy iteration - 2 steps:

1. Policy evaluation:
   - $V^{\pi_i}_{k+1}(s) \leftarrow \sum_{s'} p(s'|s, \pi_i(s))[r(s, \pi_i(s), s') + \gamma V^{\pi_i}_k(s')]$
   - Iteratively or analytically
2. Policy refinement:
   - A: Calculate policies
   - B: Calculate state values
   - C: Calculate both

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

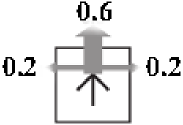
Policy iteration - 2 steps:

1. Policy evaluation:
   - $V^{\pi_i}_{k+1}(s) \leftarrow \sum_{s'} p(s'|s, \pi_i(s))[r(s, \pi_i(s), s') + \gamma V^{\pi_i}_k(s')]$
   - Iteratively or analytically
2. Policy refinement:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V^{\pi_i}_k(s')]$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$t = 0 : \quad \pi(A) = \rightarrow, \pi(B) = \leftarrow$

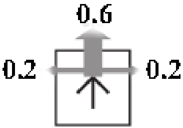
$$
\begin{array}{llrcl}
t = 1 : & PE : & V(A) & = & -1 + 0.9 \cdot \{0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot V(B)\} \\
 & & V(B) & = & -1 + 0.9 \cdot \{0.6 \cdot (-40) + 0.2 \cdot V(A) + 0.2 \cdot V(B)\} \\
\hline
 & & V(A) & = & -22.6 + 0.18 \cdot V(B) \\
 & & V(B) & = & -1 + 0.9 \cdot \{-24 + 0.2 \cdot (-22.6 + 0.18 \cdot V(B)) + 0.2 \cdot V(B)\} = \\
 & & & = & -26.668 + 0.2124 \cdot V(B) \\
 & & V(B) & = & \frac{-26.668}{0.7876} = -33.86 \\
 & & V(A) & = & -28.69
\end{array}
$$

$$
PR : \quad \pi(A) = \arg\max_a \left\{ \begin{array}{cl}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -28.69 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -23.29 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (30) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -1 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48
\end{array} \right\} = (\uparrow)
$$

$$
\pi(B) = \arg\max_a \left\{ \begin{array}{cl}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -39.26 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -33.86 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (-28.69) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -32.69 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48
\end{array} \right\} = (\uparrow)
$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$t = 0: \quad \pi(A) = \rightarrow, \pi(B) = \leftarrow$

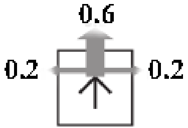
$$
\begin{array}{lllcl}
t = 1: & PE: & V(A) & = & -1 + 0.9 \cdot \{0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot V(B)\} \\
 & & V(B) & = & -1 + 0.9 \cdot \{0.6 \cdot (-40) + 0.2 \cdot V(A) + 0.2 \cdot V(B)\} \\
\hline
 & & V(A) & = & -22.6 + 0.18 \cdot V(B) \\
 & & V(B) & = & -1 + 0.9 \cdot \{-24 + 0.2 \cdot (-22.6 + 0.18 \cdot V(B)) + 0.2 \cdot V(B)\} = \\
 & & & = & -26.668 + 0.2124 \cdot V(B) \\
 & & V(B) & = & \frac{-26.668}{0.7876} = -33.86 \\
 & & V(A) & = & -28.69
\end{array}
$$

$$
PR: \quad \pi(A) = \arg\max_a \left\{ \begin{array}{ll}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -28.69 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -23.29 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (30) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -1 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48
\end{array} \right\} = (\uparrow)
$$

$$
\pi(B) = \arg\max_a \left\{ \begin{array}{ll}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -39.26 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -33.86 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (-28.69) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -32.69 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48
\end{array} \right\} = (\uparrow)
$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

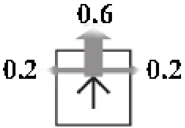
$$
\begin{array}{lllll}
t = 0: & \pi(A) = \rightarrow, \pi(B) = \leftarrow & & & \\
t = 1: & PE: & V(A) & = & -1 + 0.9 \cdot \{0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot V(B)\} \\
 & & V(B) & = & -1 + 0.9 \cdot \{0.6 \cdot (-40) + 0.2 \cdot V(A) + 0.2 \cdot V(B)\} \\
\hline
 & & V(A) & = & -22.6 + 0.18 \cdot V(B) \\
 & & V(B) & = & -1 + 0.9 \cdot \{-24 + 0.2 \cdot (-22.6 + 0.18 \cdot V(B)) + 0.2 \cdot V(B)\} = \\
 & & & = & -26.668 + 0.2124 \cdot V(B) \\
 & & V(B) & = & \frac{-26.668}{0.7876} = -33.86 \\
 & & V(A) & = & -28.69
\end{array}
$$

$$
PR: \quad \pi(A) = \arg\max_a \left\{ \begin{array}{ll} (\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -28.69 \\ (\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -23.29 \\ (\uparrow) & -1 + 0.9\{0.6 \cdot (30) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -1 \\ (\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48 \end{array} \right\} = (\uparrow)
$$

$$
\pi(B) = \arg\max_a \left\{ \begin{array}{ll} (\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -39.26 \\ (\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -33.86 \\ (\uparrow) & -1 + 0.9\{0.6 \cdot (-28.69) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -32.69 \\ (\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48 \end{array} \right\} = (\uparrow)
$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

$$t = 0: \quad \pi(A) = \rightarrow, \pi(B) = \leftarrow$$

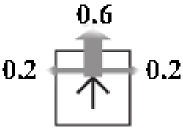
$$t = 1: \quad PE: \quad \begin{array}{lcl} V(A) & = & -1 + 0.9 \cdot \{0.6 \cdot (-50) + 0.2 \cdot 30 + 0.2 \cdot V(B)\} \\ V(B) & = & -1 + 0.9 \cdot \{0.6 \cdot (-40) + 0.2 \cdot V(A) + 0.2 \cdot V(B)\} \\ \hline V(A) & = & -22.6 + 0.18 \cdot V(B) \\ V(B) & = & -1 + 0.9 \cdot \{-24 + 0.2 \cdot (-22.6 + 0.18 \cdot V(B)) + 0.2 \cdot V(B)\} = \\ & = & -26.668 + 0.2124 \cdot V(B) \\ V(B) & = & \frac{-26.668}{0.7876} = -33.86 \\ V(A) & = & -28.69 \end{array}$$

$$PR: \quad \pi(A) = \arg\max_a \left\{ \begin{array}{ll} (\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -28.69 \\ (\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (30) + 0.2 \cdot (-33.86)\} = -23.29 \\ (\uparrow) & -1 + 0.9\{0.6 \cdot (30) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -1 \\ (\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48 \end{array} \right\} = (\uparrow)$$

$$\pi(B) = \arg\max_a \left\{ \begin{array}{ll} (\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -39.26 \\ (\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (-28.69) + 0.2 \cdot (-33.86)\} = -33.86 \\ (\uparrow) & -1 + 0.9\{0.6 \cdot (-28.69) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -32.69 \\ (\downarrow) & -1 + 0.9\{0.6 \cdot (-33.86) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -35.48 \end{array} \right\} = (\uparrow)$$

Example - Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

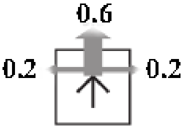
$$
t = 2: \quad PE: \quad
\begin{array}{lcl}
V(A) & = & -1 + 0.9 \cdot \{0.6 \cdot (30) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} \\
V(B) & = & -1 + 0.9 \cdot \{0.6 \cdot V(A) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} \\
\hline
V(A) & = & -1 \\
V(B) & = & -1 + 0.9 \cdot \{-0.6 - 18\} = -17.74
\end{array}
$$

$$
PR: \quad \pi(A) = \arg\max_a \left\{
\begin{array}{cl}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (30) + 0.2 \cdot (-17.74)\} = -25.79 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (30) + 0.2 \cdot (-17.74)\} = -20.39 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (30) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -1 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -26.78
\end{array}
\right\} = (\uparrow)
$$

$$
\pi(B) = \arg\max_a \left\{
\begin{array}{cl}
(\rightarrow) & -1 + 0.9\{0.6 \cdot (-50) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74)\} = -31.37 \\
(\leftarrow) & -1 + 0.9\{0.6 \cdot (-40) + 0.2 \cdot (-1) + 0.2 \cdot (-17.74)\} = -25.97 \\
(\uparrow) & -1 + 0.9\{0.6 \cdot (-1) + 0.2 \cdot (-40) + 0.2 \cdot (-50)\} = -17.74 \\
(\downarrow) & -1 + 0.9\{0.6 \cdot (-17.74) + 0.2 \cdot (-50) + 0.2 \cdot (-40)\} = -26.78
\end{array}
\right\} = (\uparrow)
$$

Example - Comparison Value/Policy Iteration

-40	30	-50
-40	$r(A) = -1$	-50
-40	$r(B) = -1$	-50



- Actions: $\{\leftarrow, \rightarrow, \uparrow, \downarrow\}$
- $\gamma = 0.9$
- $V_{k+1}(s) \leftarrow r(s) + \gamma \max\limits_{a \in A(s)} \sum_{s'} p(s'|s,a) V_k(s')$

Final policy:

A: It may be different

B: It must be the same

C: It shouldn't be different