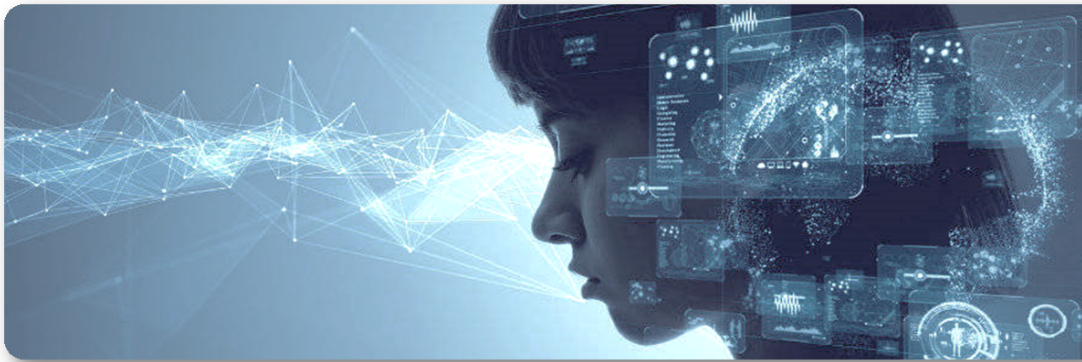


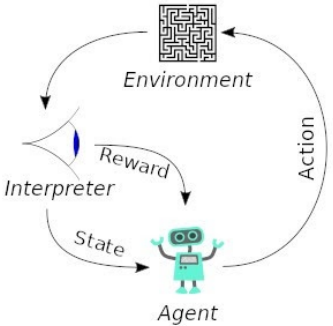
COMP 2211 Exploring Artificial Intelligence
Introduction to Reinforcement Learning

Prof. Song Guo, Dr. Desmond Tsoi & Dr. Huiru Xiao

Department of Computer Science & Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China

Reinforcement Learning

- The goal of Reinforcement Learning (RL) is to design an autonomous/intelligent agent that learns by interacting with an environment.
- In standard RL setting, the agent perceives a state at every time step and chooses an action.
- The action is applied to the environment and the environment returns a reward and a new state. The agent trains a policy to choose actions to maximize the sum of rewards.



Definition of Reinforcement Learning

- RL, a type of machine learning, in which agents take actions in an environment aimed at maximizing their cumulative awards - NVIDIA
- RL is based on rewarding desired behaviors or punishing undesired ones. Instead of one input producing one output, the algorithm produces a variety of outputs and is trained to select the right one based on certain variables - Gartner

The above definitions are technically provided by experts in that field however for someone who is starting with RL, these definitions might feel a little bit difficult.

Definition

Through a series of Trial and Error methods, an agent keeps learning continuously is an interactive environment from its own actions and experiences. The only goal of it is to find a suitable action model which would increase the total cumulative reward of the agent. It learns via interaction and feedback.

Explanation of Reinforcement Learning - Daily Life Example

- Imagine training your dog to complete a task within an environment.
- First, the trainer issues a command, which the dog observes (observation), the dog then responds by taking an action.
- If the action is close to the desired behavior, the trainer will likely provide a reward, such as a food treat or a toy. Otherwise, no reward or a negative reward will be provided.
- At the beginning of training, the dog will likely take more random actions like rolling over when the command given is "sit", as it is trying to associate specific observations with actions and rewards.
- This association, or mapping, between observations and actions is called policy.
- From the dog's perspective, the ideal case would be on in which it would respond correctly to every command, so that it gets as many treats as possible.
- So, the whole meaning of reinforcement learning training is to "tune" the dog's policy so that it learns the desired behaviors that will maximize some reward.
- After training is complete, the dog should be able to observe the owner and take the appropriate action, for example, sitting when commanded to "sit" by using the internal policy it has developed.

Explanation of Reinforcement Learning - Daily Live Example

Question

List the following in reference to the dog training example.

Agent, Environment, Observations, Actions, Rewards, Policy

Answer

- Agent: Your dog
- Environment: Your home, backyard, or any other place where you teach and play with your dog
- Observations: What the dog observes
- Actions: Sit, Roll, Stand, Walk, etc.
- Rewards: Food treat or a toy
- Policy: Generate the correct actions from the observations

Idea of Reinforcement Learning - Another Example

Question

Consider the task of parking a vehicle using automated driving system. List the following in reference to this problem.

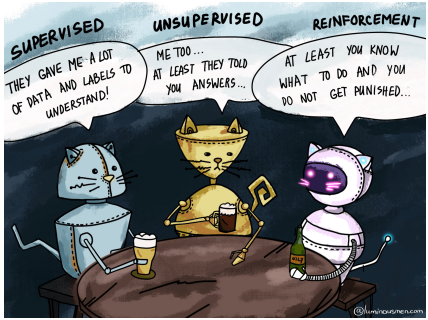
Agent, Environment, Observations, Actions, Rewards, Policy

Answer

- Agent: Vehicle computer
- Environment: Parking area
- Observations: Readings from sensors such as cameras, GPS, and lidar (light detection and ranging)
- Actions: Generate steering, braking, and acceleration commands
- Rewards: Reach the parking point as soon as possible
- Policy: Generate the correct actions from the observations

Part I

Basic Concepts



Example: Academic Life

- An assistant professor gets paid, say, 160K per year.
- How much, in total, will the assistant professor earn in their life?

$$160 + 160 + 160 + 160 + 160 + \ldots = \textit{Infinity}$$

What's wrong with this argument?

Discounted Rewards & Discounted Sum of Future Rewards

- A reward (payment) in the future is not worth quite as much as a reward now, because of inflation. For example, being promised \$10,000 next year is worth only 90% as much as receiving \$10,000 right now.

Question

Assume payment n years in future is worth only $(0.9)^n$ of payment now, what is the assistant professor's future discounted sum of rewards?

$$
\begin{aligned}
& 160 + 160 \times (0.9)^1 + 160 \times (0.9)^2 + 160 \times (0.9)^3 + \ldots \\
=& 160 \times (1 + 0.9 + (0.9)^2 + (0.9)^3 + \ldots) \\
=& 160 \times \left(\frac{1}{1 - 0.9} \right) \\
=& 1600
\end{aligned}
$$

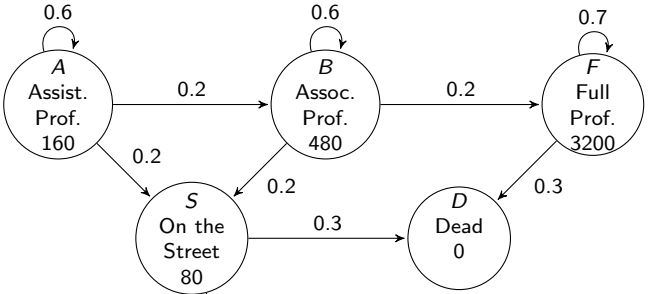
Discounted Rewards & Discounted Sum of Future Rewards

- Discounting is a concept, where a parameter called the discount factor, γ, and $0 \leq \gamma \leq 1$, and a power of it multiplies a reward.
- The "Discounted sum of future rewards" using discount factor γ is

$$\begin{aligned}(\textit{reward now}) &+ \gamma(\textit{reward in } 1 \textit{ time step from now}) \\ &+ \gamma^2(\textit{reward in } 2 \textit{ time step from now}) \\ &+ \gamma^3(\textit{reward in } 3 \textit{ time step from now}) \\ &+ \ldots\end{aligned}$$

- People in economics and probabilistic decision-making do this all the time.

Example: The Academic Life



Define:

- V_A = Expected discounted sum of future rewards starting in state A
- V_B = Expected discounted sum of future rewards starting in state B
- V_F = Expected discounted sum of future rewards starting in state F
- V_S = Expected discounted sum of future rewards starting in state S
- V_D = Expected discounted sum of future rewards starting in state D

Hint: Expected value computation

Suppose we roll a fair 6-sided die. The expected value of our die roll is

$$\left(\frac{1}{6} \times 1\right) + \left(\frac{1}{6} \times 2\right) + \left(\frac{1}{6} \times 3\right) + \left(\frac{1}{6} \times 4\right) + \left(\frac{1}{6} \times 5\right) + \left(\frac{1}{6} \times 6\right) = 3.5$$

Question

Assume discount factor $\gamma = 0.9$. How do we compute V_A, V_B, V_T, V_S, T_D?

Example: The Academic Life - Start from state A

But there are so many different possibilities!!! Each with different probability :(

Sample episodes, all start from A:

- A $\to$ B $\to$ F $\to$ D:

$$160 + (0.2)(0.9)^1(480) + (0.2)(0.2)(0.9)^2(3200) + (0.2)(0.2)(0.3)(0.9)^3(0) = 350.08$$

- A $\to$ A $\to$ S $\to$ D:

$$160 + (0.6)(0.9)^1(160) + (0.6)(0.2)(0.9)^2(80) + (0.6)(0.2)(0.3)(0.9)^3(0) = 254.176$$

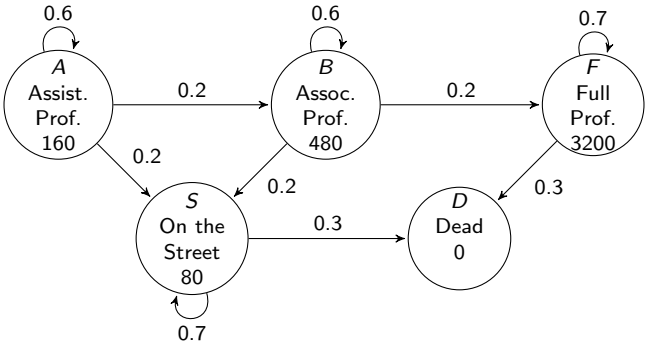
- A $\to$ B $\to$ S $\to$ D:

$$160 + (0.2)(0.9)^1(480) + (0.2)(0.2)(0.9)^2(80) + (0.2)(0.2)(0.3)(0.9)^3(0) = 248.992$$

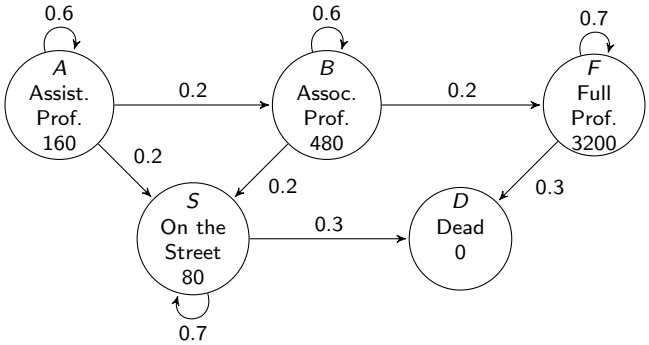
- $\cdots$

It is very difficult to compute V_A. $>.<$

Idea



- Let V_A^1, V_B^1, V_F^1, V_S^1, V_D^1 be the expected discounted sum of rewards over the next 1 time step from now, do you know how to find them?
- Let V_A^2 be the expected discounted sum of rewards over the next 2 time step from now, do you know how to find it if you know V_A^1, V_B^1, V_F^1, V_S^1, and V_D^1?



$$V_A^1 = 160, V_B^1 = 480, V_F^1 = 3200, V_S^1 = 80, V_D^1 = 0$$

$$\begin{aligned} V_A^2 &= 160 + 0.9 \times (P_{AA} V_A^1 + P_{AB} V_B^1 + P_{AF} V_F^1 + P_{AS} V_S^1 + P_{AD} V_D^1) \\ &= 160 + 0.9 \times (0.6(160) + 0.2(480) + 0(3200) + 0.2(80) + 0(0)) = 347.2 \end{aligned}$$

Do you know how to compute V_B^2, V_F^2, V_S^2, V_D^2, ...?

Markov Property

Definition

A state S_t is Markov if and only if

$$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, \ldots, S_t)$$

- The future is independent of the past given the present
- The present state captures all relevant information from the history
- Once the present state is known, the history may be thrown away

Part II

Problem Formulation

A Markov System with Rewards

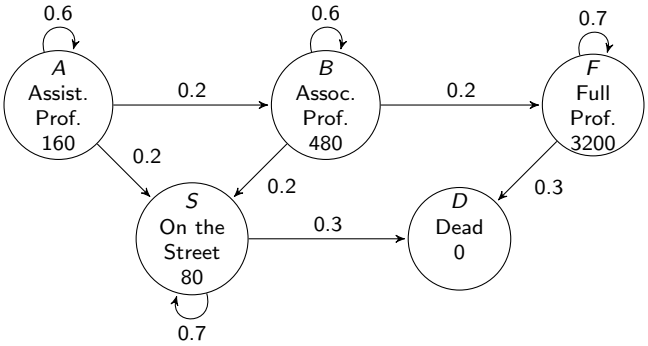
A Markov system consists of the following:

- A set of N states $\{s_1, s_2, \ldots, s_N\}$
- A transition probability matrix (i.e., a 2D array showing the probability of going from one state to another state):

$$T = \begin{array}{c} \\ From \end{array} \begin{array}{c} \begin{array}{cccc} \multicolumn{4}{c}{To} \end{array} \\ \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1N} \\ T_{21} & T_{22} & \cdots & T_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ T_{N1} & T_{N2} & \cdots & T_{NN} \end{bmatrix} \end{array}$$

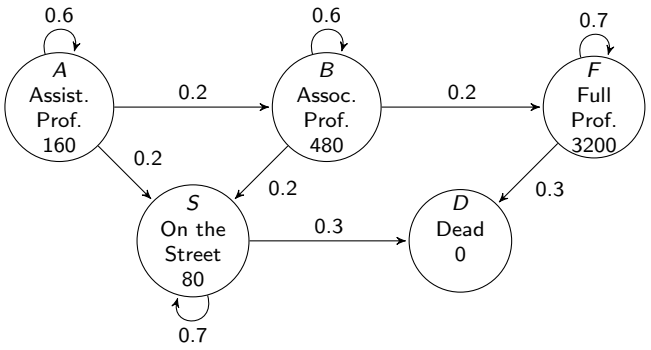
  where $T_{ij} = \text{P}(\text{next state } s_{t+1} = s_j \mid \text{this state } s_t = s_i)$
  Note: Each row of the matrix sums to 1
- Each state has a reward $\{r_1, r_2, \ldots, r_N\}$
- There is a discount factor γ, where $0 < \gamma < 1$
- All future rewards are discounted by γ

Example: The Academic Life



What are the states, transition probability matrix, rewards, discount factor for this problem?

Example: Academic Life



- States: $\{A, B, F, S, D\}$
- Transition probability matrix:

$$\mathrm{T} = \begin{bmatrix} 0.6 & 0.2 & 0 & 0.2 & 0 \\ 0 & 0.6 & 0.2 & 0.2 & 0 \\ 0 & 0 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Rewards: $\{160, 480, 3200, 80, 0\}$
- Discount factor: 0.9

Value Function

Solving a Markov system

- Write $V(s_i)$ = expected discounted sum of future rewards starting in state s_i

$$\begin{aligned} V(s_i) &= r(s_i) + \gamma \cdot \text{Expected future rewards starting from next state} \\ &= r(s_i) + \gamma \cdot (T_{i1}V(s_1) + T_{i2}V(s_2) + \ldots + T_{1N}V(S_N)) \end{aligned}$$

- Using vector notation, we have

$$\begin{bmatrix} V(s_1) \\ V(s_2) \\ \vdots \\ V(s_N) \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} + \gamma \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1N} \\ T_{21} & T_{22} & \cdots & T_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ T_{N1} & T_{N2} & \cdots & T_{NN} \end{bmatrix} \begin{bmatrix} V(s_1) \\ V(s_2) \\ \vdots \\ V(s_N) \end{bmatrix}$$

Solving the System of Linear Equations

- The equation is a linear equation, which can be solved directly.

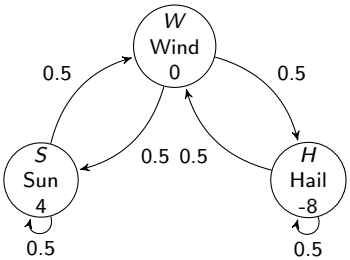
$$\begin{aligned} V &= R + \gamma T V \\ (1 - \gamma T) V &= R \\ V &= (1 - \gamma T)^{-1} R \end{aligned}$$

- The good thing for directly solving the above equation is you get an exact number.
- The bad thing is it is slow if you have a large number of states, i.e., N is big.
- There are many iterative methods for solving the equation, e.g.,
  - Dynamic programming (We will do Value Iteration)
  - Monte-Carlo evaluation
  - Temporal-Difference learning

Value Iteration

- Define
  - $V^1(s_i)$ = Expected discounted sum of rewards over the next 1 time step from now
  - $V^2(s_i)$ = Expected discounted sum of rewards over the next 2 time steps from now
  - $V^3(s_i)$ = Expected discounted sum of rewards over the next 3 time steps from now
  - $\ldots$
  - $V^k(s_i)$ = Expected discounted sum of rewards over the next k time steps from now
- What are the formula to compute all of them?
  - $V^1(s_i) = r(s_i)$
  - $V^2(s_i) = r(s_i) + \gamma(T_{i1}V^1(s_1) + T_{i2}V^1(s_2) + \ldots + T_{iN}V^1(s_N))$
  - $V^3(s_i) = r(s_i) + \gamma(T_{i1}V^2(s_1) + T_{i2}V^2(s_2) + \ldots + T_{iN}V^2(s_N))$
  - $\ldots$
  - $V^k(s_i) = r(s_i) + \gamma(T_{i1}V^{k-1}(s_1) + T_{i2}V^{k-1}(s_2) + \ldots + T_{iN}V^{k-1}(s_N))$

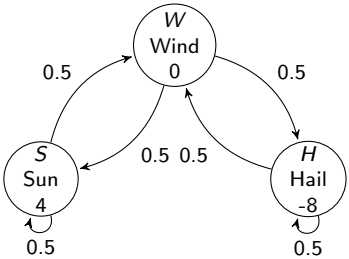
Example: Weather



$$T = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1			
2			
3			
4			
5			

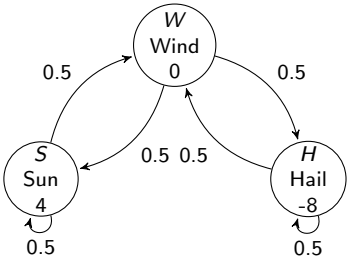
Example: Weather



k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1	4	0	-8
2			
3			
4			
5			

$$V^1(S) = r(S) = 4$$
$$V^1(W) = r(W) = 0$$
$$V^1(H) = r(H) = -8$$
$$V^2(S) = r(S) + \gamma(T_{SS}V^1(S) + T_{WS}V^1(W) + T_{HS}V^1(H))$$
$$V^2(W) = r(W) + \gamma(T_{SW}V^1(S)) + T_{WW}V^1(W) + T_{HW}V^1(H)$$
$$V^2(H) = r(H) + \gamma(T_{SH}V^1(S) + T_{HW}V^1(W) + T_{HH}V^1(H))$$

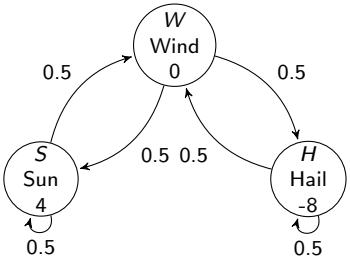
Example: Weather



k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1	4	0	-8
2	5	-1	-10
3			
4			
5			

$$
\begin{aligned}
V^2(S) &= r(S) + \gamma(T_{SS}V^1(S) + T_{WS}V^1(W) + T_{HS}V^1(H)) \\
&= 4 + 0.5(0.5 \times 4 + 0.5 \times 0 + 0 \times (-8)) = 5 \\
V^2(W) &= r(W) + \gamma(T_{SW}V^1(S)) + T_{WW}V^1(W) + T_{HW}V^1(H) \\
&= 0 + 0.5(0.5 \times 4 + 0 \times 0 + 0.5 \times (-8)) = -1 \\
V^2(H) &= r(H) + \gamma(T_{SH}V^1(S) + T_{HW}V^1(W) + T_{HH}V^1(H)) \\
&= -8 + 0.5(0 \times 4 + 0.5 \times 0 + 0.5 \times (-8)) = -10
\end{aligned}
$$

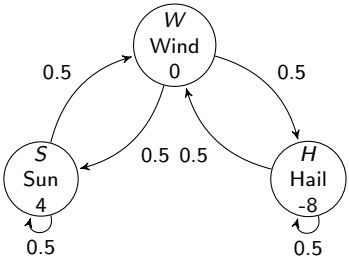
Example: Weather



k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1	4	0	-8
2	5	-1	-10
3	5	-1.25	-10.75
4			
5			

$$
\begin{aligned}
V^3(S) &= r(S) + \gamma(T_{SS}V^2(S) + T_{WS}V^2(W) + T_{HS}V^2(H)) \\
&= 4 + 0.5(0.5 \times 5 + 0.5 \times (-1) + 0 \times (-10)) = 5 \\
V^3(W) &= r(W) + \gamma(T_{SW}V^2(S)) + T_{WW}V^2(W) + T_{HW}V^2(H) \\
&= 0 + 0.5(0.5 \times 5 + 0 \times (-1) + 0.5 \times (-10)) = -1.25 \\
V^3(H) &= r(H) + \gamma(T_{SH}V^2(S) + T_{HW}V^2(W) + T_{HH}V^2(H)) \\
&= -8 + 0.5(0 \times 5 + 0.5 \times (-1) + 0.5 \times (-10)) = -10.75
\end{aligned}
$$

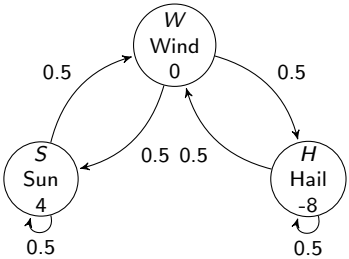
Example: Weather



k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1	4	0	-8
2	5	-1	-10
3	5	-1.25	-10.75
4	4.9375	-1.4375	-11
5			

$$
\begin{aligned}
V^4(S) &= r(S) + \gamma(T_{SS}V^3(S) + T_{WS}V^3(W) + T_{HS}V^3(H)) \\
&= 4 + 0.5(0.5 \times 5 + 0.5 \times (-1.25) + 0 \times (-10.75)) = 4.9375 \\
V^4(W) &= r(W) + \gamma(T_{SW}V^3(S)) + T_{WW}V^3(W) + T_{HW}V^3(H) \\
&= 0 + 0.5(0.5 \times 5 + 0 \times (-1.25) + 0.5 \times (-10.75)) = -1.4375 \\
V^4(H) &= r(H) + \gamma(T_{SH}V^3(S) + T_{HW}V^3(W) + T_{HH}V^3(H)) \\
&= -8 + 0.5(0 \times 5 + 0.5 \times (-1.25) + 0.5 \times (-10.75)) = -11
\end{aligned}
$$

Example: Weather



k	$V^k(S)$	$V^k(W)$	$V^k(H)$
1	4	0	-8
2	5	-1	-10
3	5	-1.25	-10.75
4	4.9375	-1.4375	-11
5	4.875	-1.515625	-11.109375

$$
\begin{aligned}
V^5(S) &= r(S) + \gamma(T_{SS}V^4(S) + T_{WS}V^4(W) + T_{HS}V^4(H)) \\
&= 4 + 0.5(0.5 \times 4.9375 + 0.5 \times (-1.4375) + 0 \times (-11)) = 4.875 \\
V^5(W) &= r(W) + \gamma(T_{SW}V^4(S)) + T_{WW}V^4(W) + T_{HW}V^4(H) \\
&= 0 + 0.5(0.5 \times 4.9375 + 0 \times (-1.4375) + 0.5 \times (-11)) = -1.515625 \\
V^5(H) &= r(H) + \gamma(T_{SH}V^4(S) + T_{HW}V^4(W) + T_{HH}V^4(H)) \\
&= -8 + 0.5(0 \times 4.9375 + 0.5 \times (-1.4375) + 0.5 \times (-11)) = -11.109375
\end{aligned}
$$

Value Iteration for Solving the System of Linear Equations

- Compute $V^1(S_i)$ for each i in range $[1, N]$
- Compute $V^2(S_i)$ for each i in range $[1, N]$
- Compute $V^3(S_i)$ for each i in range $[1, N]$
- $\vdots$
- Compute $V^k(S_i)$ for each i in range $[1, N]$

When to stop?

When the maximum absolute difference between two successive expected discounted sum of rewards (V^k and V^{k-1}) is less than a threshold, ξ, i.e.,

$$\text{Max}_i |V^k(s_i) - V^{k-1}(s_i)| < \xi$$

Markov Decision Process

- A Markov Decision Process (MDP) is a Markov reward process with decisions.
- It is an environment in which all states are Markov.

Definition

A Markov Decision Process is a tuple $\langle S, A, T, R, \gamma \rangle$

- S: A finite set of states $\{s_1, s_2, \ldots, s_N\}$
- A: A finite set of actions $\{a_1, a_2, \ldots, a_M\}$
- T: A transition probability matrix

$$T_{ij}^{a} = P(S_j | S_i, A = a)$$

- R: Each state has a reward $\{r_1, r_2, \ldots, r_N\}$
- γ: A discount factor $0 \leq \gamma \leq 1$

Value Iteration (How to Determine Actions?)

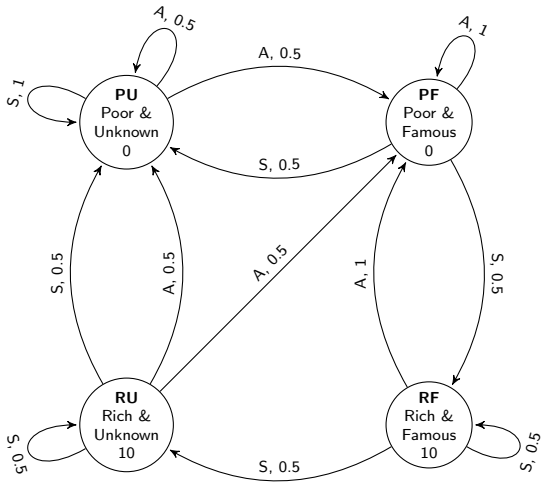
- Define
  - $V^1(s_i)$ = Expected discounted sum of rewards over the next 1 time step from now
  - $V^2(s_i)$ = Expected discounted sum of rewards over the next 2 time steps from now
  - $V^3(s_i)$ = Expected discounted sum of rewards over the next 3 time steps from now
  - $\dots$
  - $V^k(s_i)$ = Expected discounted sum of rewards over the next k time steps from now
- What are the formula to compute all of them?
  - $V^1(s_i) = r(s_i)$
  - $V^2(s_i) = max_a(r(s_i) + \gamma(T^a_{i1}V^1(s_1) + T^a_{i2}V^1(s_2) + \dots + T^a_{iN}V^1(s_N)))$
  - $V^3(s_i) = max_a(r(s_i) + \gamma(T^a_{i1}V^2(s_1) + T^a_{i2}V^2(s_2) + \dots + T^a_{iN}V^2(s_N)))$
  - $\dots$
  - $V^k(s_i) = max_a(r(s_i) + \gamma(T^a_{i1}V^{k-1}(s_1) + T^a_{i2}V^{k-1}(s_2) + \dots + T^a_{iN}V^{k-1}(s_N)))$

Bellman Optimality Equation

$$V^k(s_i) = max_a(r(s_i) + \gamma(T^a_{i1}V^{k-1}(s_1) + T^a_{i2}V^{k-1}(s_2) + \dots + T^a_{iN}V^{k-1}(s_N)))$$

Policies

- A policy is a mapping from states to actions



Policy 1	
State	Action
PU	S
PF	A
RU	S
RF	A

Policy 2	
State	Action
PU	A
PF	A
RU	A
RF	A

Finding the Near Optimal Policy

- Compute $V^k(s_i)$ for all i using value iteration
- Compute (near) optimal policy $\pi(s_i)$ in state s_i as

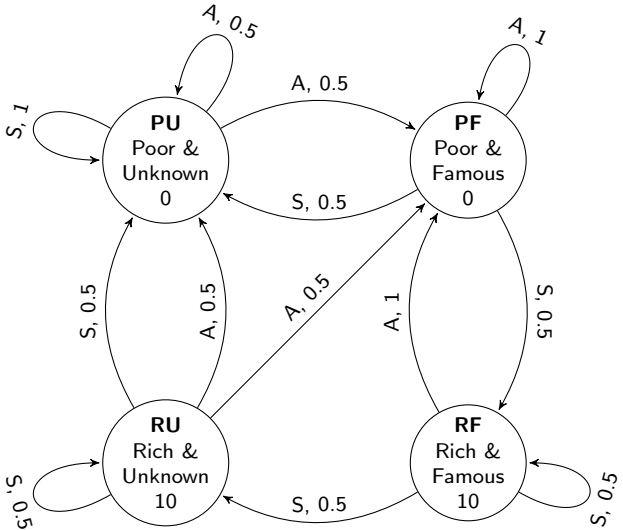
$$\pi(s_i) = argmax_a(r(s_i) + \gamma(T^a_{i1}V^{k-1}(s_1) + T^a_{i2}V^{k-1}(s_2) + \ldots + T^a_{iN}V^{k-1}(s_N)))$$

until

$$\mathrm{Max}_i|V^{k+1}(s_i) - V^k(s_i)| < \xi$$

- Once it is done, the near optimal policy consists of taking the action that leads to the state that has maximum state value.

Example



$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1				
2				
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2				
3				
4				
5				
6				

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2				
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2				
3				
4				
5				
6				

$$V^1(PU) = 0$$
$$V^1(PF) = 0$$
$$V^1(RU) = 10$$
$$V^1(RF) = 10$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0			
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S			
3				
4				
5				
6				

$$V^2(PU) = max(r(PU) + \gamma(T^A_{PU\to PU}V^1(PU) + T^A_{PU\to PF}V^1(PF) + T^A_{PU\to RU}V^1(RU) + T^A_{PU\to RF}V^1(RF)),$$
$$r(PU) + \gamma(T^S_{PU\to PU}V^1(PU) + T^S_{PU\to PF}V^1(PF) + T^S_{PU\to RU}V^1(RU) + T^S_{PU\to RF}V^1(RF)))$$
$$V^2(PU) = max(0 + 0.9 \times (0.5 \times 0 + 0.5 \times 0 + 0 \times 10 + 0 \times 10)),$$
$$0 + 0.9 \times (1 \times 0 + 0 \times 0 + 0 \times 10 + 0 \times 10))$$
$$= max(0, 0) = 0$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5		
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S		
3				
4				
5				
6				

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

$$V^2(PF) = max(r(PF) + \gamma(T^A_{PF\to PU} V^1(PU) + T^A_{PF\to PF} V^1(PF) + T^A_{PF\to RU} V^1(RU) + T^A_{PF\to RF} V^1(RF)),$$
$$r(PF) + \gamma(T^S_{PF\to PU} V^1(PU) + T^S_{PF\to PF} V^1(PF) + T^S_{PF\to RU} V^1(RU) + T^S_{PF\to RF} V^1(RF)))$$
$$V^2(PF) = max(0 + 0.9 \times (0 \times 0 + 1 \times 0 + 0 \times 10 + 0 \times 10)),$$
$$0 + 0.9 \times (0.5 \times 0 + 0 \times 0 + 0.5 \times 10 + 0.5 \times 10))$$
$$= max(0, 4.5) = 4.5$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	
3				
4				
5				
6				

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

$$V^2(RU) = max(r(RU) + \gamma(T^A_{RU\to PU}V^1(PU) + T^A_{RU\to PF}V^1(PF) + T^A_{RU\to RU}V^1(RU) + T^A_{RU\to RF}V^1(RF)), r(RU) + \gamma(T^S_{RU\to PU}V^1(PU) + T^S_{RU\to PF}V^1(PF) + T^S_{RU\to RU}V^1(RU) + T^S_{RU\to RF}V^1(RF)))$$

$$V^2(PF) = max(10 + 0.9 \times (0.5 \times 0 + 0.5 \times 0 + 0 \times 10 + 0 \times 10)), 10 + 0.9 \times (0.5 \times 0 + 0 \times 0 + 0.5 \times 10 + 0 \times 10))$$

$$= max(0, 14.5) = 14.5$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3				
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3				
4				
5				
6				

$$V^2(RF) = max(r(RF) + \gamma(T^A_{RF\to PU}V^1(PU) + T^A_{RF\to PF}V^1(PF) + T^A_{RF\to RU}V^1(RU) + T^A_{RF\to RF}V^1(RF)),$$
$$r(RF) + \gamma(T^S_{RF\to PU}V^1(PU) + T^S_{RF\to PF}V^1(PF) + T^S_{RF\to RU}V^1(RU) + T^S_{RF\to RF}V^1(RF)))$$
$$V^2(PF) = max(10 + 0.9 \times (0 \times 0 + 1 \times 0 + 0 \times 10 + 0 \times 10)),$$
$$10 + 0.9 \times (0 \times 0 + 0 \times 0 + 0.5 \times 10 + 0.5 \times 10))$$
$$= max(0, 19) = 19$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025			
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2				
3				
4				
5				
6				

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

$$V^3(PU) = max(r(PU) + \gamma(T^A_{PU\to PU}V^2(PU) + T^A_{PU\to PF}V^2(PF) + T^A_{PU\to RU}V^2(RU) + T^A_{PU\to RF}V^2(RF)),$$
$$r(PU) + \gamma(T^S_{PU\to PU}V^2(PU) + T^S_{PU\to PF}V^2(PF) + T^S_{PU\to RU}V^2(RU) + T^S_{PU\to RF}V^2(RF)))$$
$$V^3(PU) = max(0 + 0.9 \times (0.5 \times 0 + 0.5 \times 4.5 + 0 \times 14.5 + 0 \times 19)),$$
$$0 + 0.9 \times (1 \times 0 + 0 \times 4.5 + 0 \times 14.5 + 0 \times 19))$$
$$= max(2.205, 0) = 2.025$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55		
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S		
4				
5				
6				

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

$$V^3(PF) = max(r(PF) + \gamma(T^A_{PF\to PU}V^2(PU) + T^A_{PF\to PF}V^2(PF) + T^A_{PF\to RU}V^2(RU) + T^A_{PF\to RF}V^2(RF)),$$
$$r(PF) + \gamma(T^S_{PF\to PU}V^2(PU) + T^S_{PF\to PF}V^2(PF) + T^S_{PF\to RU}V^2(RU) + T^S_{PF\to RF}V^2(RF)))$$
$$V^3(PF) = max(0 + 0.9 \times (0 \times 0 + 1 \times 4.5 + 0 \times 14.5 + 0 \times 19)),$$
$$0 + 0.9 \times (0.5 \times 0 + 0 \times 4.5 + 0 \times 14.5 + 0.5 \times 19))$$
$$= max(4.05, 8.55) = 8.55$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	
4				
5				
6				

$T^A =$

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

$T^S =$

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

$$V^3(RU) = max(r(RU) + \gamma(T^A_{RU\to PU}V^2(PU) + T^A_{RU\to PF}V^2(PF) + T^A_{RU\to RU}V^2(RU) + T^A_{RU\to RF}V^2(RF)),$$
$$r(RU) + \gamma(T^S_{RU\to PU}V^2(PU) + T^S_{RU\to PF}V^2(PF) + T^S_{RU\to RU}V^2(RU) + T^S_{RU\to RF}V^2(RF)))$$
$$V^3(RU) = max(10 + 0.9 \times (0.5 \times 0 + 0.5 \times 4.5 + 0 \times 14.5 + 0 \times 19)),$$
$$10 + 0.9 \times (0.5 \times 0 + 0 \times 4.5 + 0.5 \times 14.5 + 0 \times 19))$$
$$= max(12.025, 16.525) = 16.525$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4				
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4				
5				
6				

$$
\begin{aligned}
V^3(RF) = max(&r(RF) + \gamma(T^A_{RF\to PU}V^2(PU) + \\
&T^A_{RF\to PF}V^2(PF) + T^A_{RF\to RU}V^2(RU) + T^A_{RF\to RF}V^2(RF)), \\
&r(RF) + \gamma(T^S_{RF\to PU}V^2(PU) + \\
&T^S_{RF\to PF}V^2(PF) + T^S_{RF\to RU}V^2(RU) + T^S_{RF\to RF}V^2(RF)))
\end{aligned}
$$

$$
\begin{aligned}
V^3(PF) &= max(10 + 0.9 \times (0 \times 0 + 1 \times 4.5 + 0 \times 14.5 + 0 \times 19)), \\
&\quad\ 10 + 0.9 \times (0 \times 0 + 0 \times 4.5 + 0.5 \times 14.5 + 0.5 \times 19)) \\
&= max(14.05, 25.075) = 25.075
\end{aligned}
$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4	4.75875			
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4	A			
5				
6				

$$V^4(PU) = max(r(PU) + \gamma(T^A_{PU\to PU}V^3(PU)+ T^A_{PU\to PF}V^3(PF) + T^A_{PU\to RU}V^3(RU) + T^A_{PU\to RF}V^3(RF)),$$
$$r(PU) + \gamma(T^S_{PU\to PU}V^3(PU)+ T^S_{PU\to PF}V^3(PF) + T^S_{PU\to RU}V^3(RU) + T^S_{PU\to RF}V^3(RF)))$$
$$V^4(PU) = max(0 + 0.9 \times (0.5 \times 2.025 + 0.5 \times 8.55 + 0 \times 16.525 + 0 \times 25.075)),$$
$$0 + 0.9 \times (1 \times 2.025 + 0 \times 8.55 + 0 \times 16.525 + 0 \times 25.075))$$
$$= max(4.75875, 1.8225) = 4.75875$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4	4.75875	12.195		
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4	A	S		
5				
6				

$$V^4(PF) = max(r(PF) + \gamma(T^A_{PF\to PU}V^3(PU)+ T^A_{PF\to PF}V^3(PF) + T^A_{PF\to RU}V^3(RU) + T^A_{PF\to RF}V^3(RF)),$$
$$r(PF) + \gamma(T^S_{PF\to PU}V^3(PU)+ T^S_{PF\to PF}V^3(PF) + T^S_{PF\to RU}V^3(RU) + T^S_{PF\to RF}V^3(RF)))$$
$$V^4(PF) = max(0 + 0.9 \times (0 \times 2.025 + 1 \times 8.55 + 0 \times 16.525 + 0 \times 25.075)),$$
$$0 + 0.9 \times (0.5 \times 2.025 + 0 \times 8.55 + 0 \times 16.525 + 0.5 \times 25.075))$$
$$= max(7.695, 12.195) = 12.195$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4	4.75875	12.195	18.3475	
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4	A	S	S	
5				
6				

$$V^4(RU) = max(r(RU) + \gamma(T^A_{RU\to PU}V^3(PU)+ T^A_{RU\to PF}V^3(PF) + T^A_{RU\to RU}V^3(RU) + T^A_{RU\to RF}V^3(RF)), r(RU) + \gamma(T^S_{RU\to PU}V^3(PU)+ T^S_{RU\to PF}V^3(PF) + T^S_{RU\to RU}V^3(RU) + T^S_{RU\to RF}V^3(RF)))$$

$$V^4(PF) = max(10 + 0.9 \times (0.5 \times 2.025 + 0.5 \times 8.55 + 0 \times 16.525 + 0 \times 25.075)), 10 + 0.9 \times (0.5 \times 2.025 + 0 \times 8.55 + 0.5 \times 16.525 + 0 \times 25.075)) = max(14.75875, 18.3475) = 18.3475$$

Example

$\gamma = 0.9$

R=

PU	PF	RU	RF
0	0	10	10

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4	4.75875	12.195	18.3475	28.72
5				
6				

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4	A	S	S	S
5				
6				

$$V^4(RF) = max(r(RF) + \gamma(T^A_{RF\to PU}V^3(PU)+ T^A_{RF\to PF}V^3(PF) + T^A_{RF\to RU}V^3(RU) + T^A_{RF\to RF}V^3(RF)),$$
$$r(RF) + \gamma(T^S_{RF\to PU}V^3(PU)+ T^S_{RF\to PF}V^3(PF) + T^S_{RF\to RU}V^3(RU) + T^S_{RF\to RF}V^3(RF)))$$
$$V^4(RF) = max(10 + 0.9 \times (0 \times 2.025 + 1 \times 8.55 + 0 \times 16.525 + 0 \times 25.075)),$$
$$10 + 0.9 \times (0 \times 2.025 + 0 \times 8.55 + 0.5 \times 16.525 + 0.5 \times 25.075))$$
$$= max(17.695, 28.72) = 28.72$$

Example

R=

PU	PF	RU	RF
0	0	10	10

$\gamma = 0.9$

T^A=

	PU	PF	RU	RF
PU	0.5	0.5	0	0
PF	0	1	0	0
RU	0.5	0.5	0	0
RF	0	1	0	0

k	V(PU)	V(PF)	V(RU)	V(RF)
1	0	0	10	10
2	0	4.5	14.5	19
3	2.025	8.55	16.525	25.075
4	4.75875	12.195	18.3475	28.72
5	7.62919	15.0654	20.3978	31.1804
6	10.2126	17.4643	22.6121	33.2102

k	$\pi(PU)$	$\pi(PF)$	$\pi(RU)$	$\pi(RF)$
1				
2	A/S	S	S	S
3	A	S	S	S
4	A	S	S	S
5	A	S	S	S
6	A	S	S	S

T^S=

	PU	PF	RU	RF
PU	1	0	0	0
PF	0.5	0	0	0.5
RU	0.5	0	0.5	0
RF	0	0	0.5	0.5

Practice

Can you calculate the remaining ones? :)

Pros and Cons of Value Iteration

- Pros:
  - Will converge towards optimal values
  - Good for a small set of states
- Cons:
  - Value iteration has to touch every state in every iteration and so if we have a large number of total states, value iteration suffers
  - It is slow because we have to consider all actions at every state, and often, there are many actions

That's all!

Any questions?