Artificial Intelligence, Reinforcement Learning & Sequential Decision Problems

(CIV6540 - Probabilistic Machine Learning for Civil Engineers)

Professor: James-A. Goulet

Département des génies civil, géologique et des mines
Polytechnique Montréal

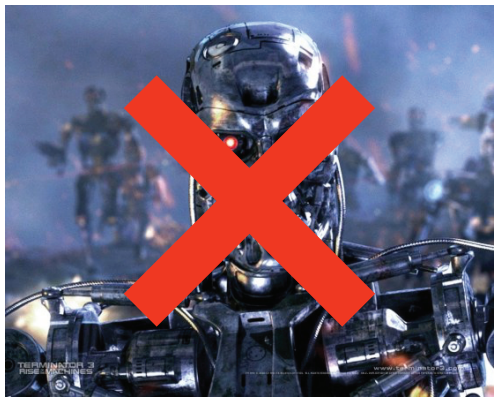
Chapter 14 – Goulet (2020)
Probabilistic Machine Learning for Civil Engineers
MIT Press

Chapter 17 – Russell, S. and Norvig, P. (1995)
Artificial Intelligence, A modern approach
Prentice-Hall

Sutton, R. S. and Barto, A. G. (2018)
Reinforcement learning: An introduction
MIT Press, 2nd edition

Artificial intelligence – AI

AI : reproduction of an intelligent behavior by a machine.
$\rightarrow$ perception and dynamic interaction with its environment through the process of **taking actions** for achieving goals.



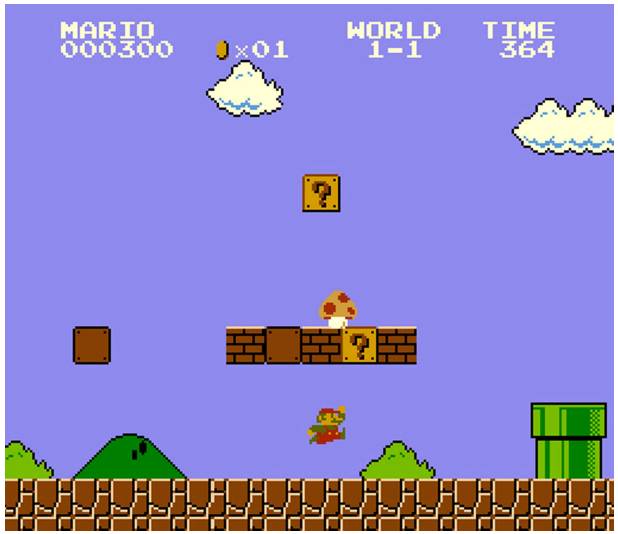
Level 1 - Limited AI
Beyond human capacity in a specific tasks

Behind limited AI:
Machine learning

Behind machine learning:
Probability theory / Decision Theory / Linear Algebra / Optimization/...

[Google images]

AI $\rightarrow$ Reinforcement Learning $\rightarrow$ Sequential Decisions



$$\mathcal{S} = \{ \text{  } \}$$

$$\mathcal{A} = \{ \text{  } \}$$

$$s_t = \{ \text{  } \}$$

$$r(s_t) = \{ \text{fct}(\text{  }) \}$$

$$p(s_{t+1}|s_t, a) = \text{Unknown...}$$

Reinforcement learning $\rightarrow$ **Long term** expected utility to take an **action** *a* if we are in a **state** *s*

[https://youtu.be/QVyu9oVyh9Q, Google images]

Anomaly detection & Structural Health Monitoring

$$\mathcal{S} = \{\text{SHM system}\}$$

$$\mathcal{A} = \{\text{alarm}, \neg \text{ alarm}\}$$

$$s_t = \{\boldsymbol{x}_t\} \text{ (e.g. SSM, M7)}$$

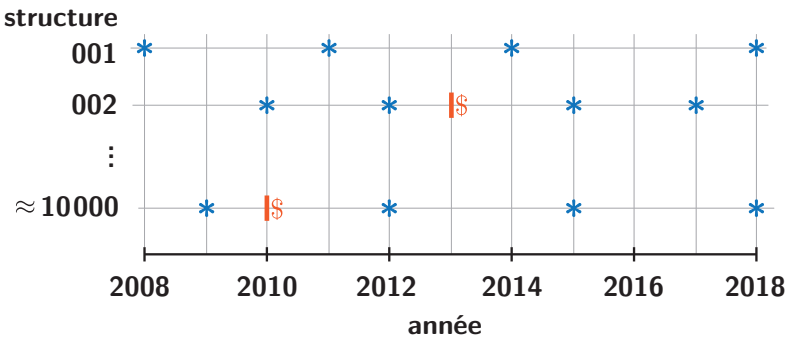
$$r(s_t, a_t) = \{\text{fct}(s_t, a_t)\}$$

$$p(s_{t+1}|s_t, a) = \text{unknown...}$$

Reinforcement learning $\rightarrow$ **Long term** expected utility to take an **action** *a* if we are in a **state** *s*

[Hydro-Quebec]

Network-Scale Infrastructure Maintenance Planning



$$\mathcal{S} = \{\text{Infrastructure network}\}$$

$$\mathcal{A} = \{\text{wait, repair, replace}\}$$

$$s_t = \{\%A, \%B, \%C, \%D\}$$

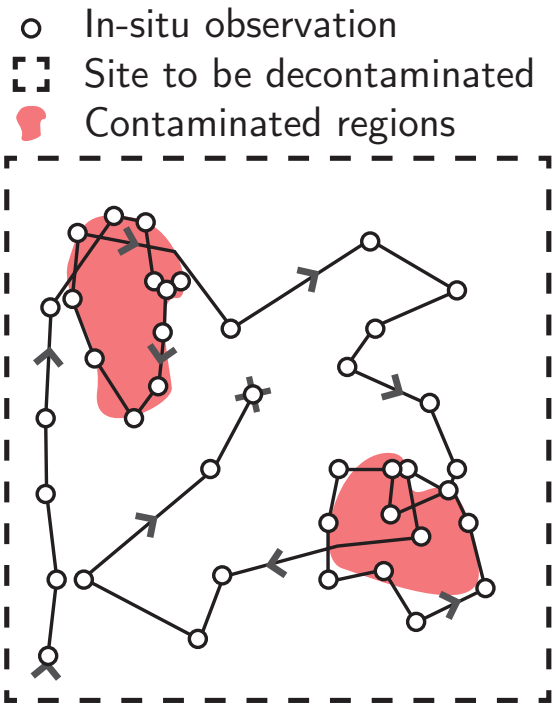
$$r(s_t) = \{\text{fct}(s_t)\}$$

$$p(s_{t+1}|s_t, a) = \text{unknown...}$$

Reinforcement learning $\rightarrow$ **Long term** expected utility to take an **action** *a* if we are in a **state** *s*

[Google images]

Contaminated site characterization



$$\mathcal{S} = \{\text{Contaminated site}\}$$

$$\mathcal{A} = \{\text{Sample } (\Delta x, \Delta y), \text{ stop}\}$$

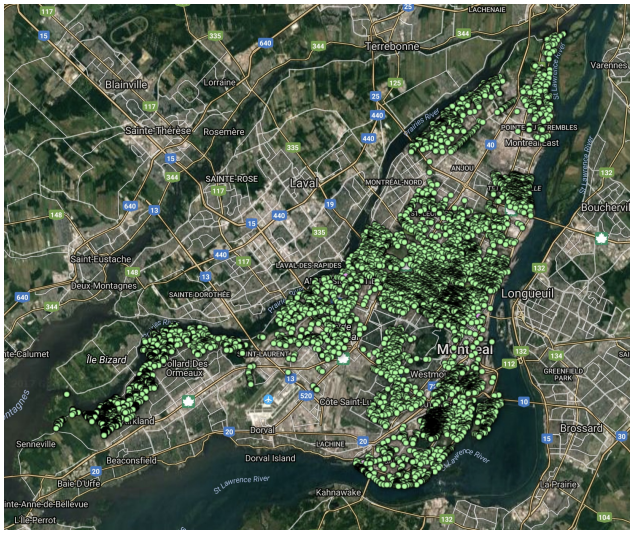
$$s_t = \{[\text{Hg}]_{\mathcal{S}}\}$$

$$r(s_t, a_t) = \{\text{fct}(s_t, a_t)\}$$

$$p(s_{t+1}|s_t, a) = \text{unknown...}$$

Reinforcement learning $\rightarrow$ **Long term** expected utility to take an **action** *a* if we are in a **state** *s*

Emerald ash borer – Tree cutting policy



$$\mathcal{S} = \{\text{Tree population}\}$$

$$\mathcal{A} = \{\text{cut, wait}\}$$

$$s_t = \{[\text{insects}]_{\mathcal{S}}\}$$

$$r(s_t) = \{\text{fct}(s_t)\}$$

$$p(s_{t+1}|s_t, a) = \text{unknown...}$$

Reinforcement learning $\rightarrow$ **Long term** expected utility to take an **action** *a* if we are in a **state** *s*

[Google images]

Reinforcement learning

When we can explicitly represent $p(s_{t+1}|s_t, a)$
$\rightarrow$ **Model-based RL: MDP/POMDP, etc...**

When we cannot represent $p(s_{t+1}|s_t, a)$ explicitly
$\rightarrow$ **Model-free RL: Q-learning, SARSA, etc...**

Module #10 Outline

Intro

Nomenclature

Markov Decision Process, MDP

Q-Learning

Topics organization

Background
- 1 Revision probability & linear algebra
- 2 Probability distributions

Machine Learning Basics
- 0 Introduction
- 3 Bayesian Estimation $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- 4 MCMC sampling & Newton

Supervised learning
- 5 Regression
- 6 Classification
- 7 LSTM networks for time series

Unsupervised learning
- 7 State-space model for time-series

Decision Making & RL
- 8 Decision Theory
- 9 AI & Sequential decision problems

Section Outline

Nomenclature

2.1 Variables

2.2 Example #1 - Infrastructure Maintenance Planning

RL sequential decision problem – Nomenclature

States: $\underbrace{s \in \mathbb{R}}_{\text{continuous}}$ or $\underbrace{s_j \in \mathcal{S} = \{s_1, s_2, \cdots, s_{\mathtt{S}}\}}_{\text{discrete}}$

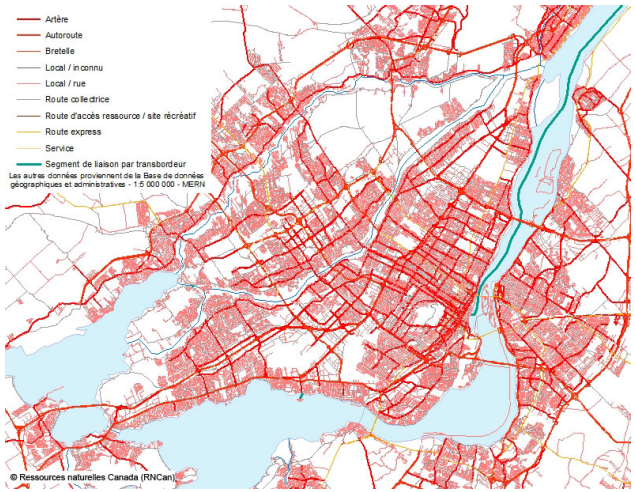
Actions: $a_i \in \mathcal{A} = \{a_1, a_2, \cdots, a_{\mathtt{A}}\}$

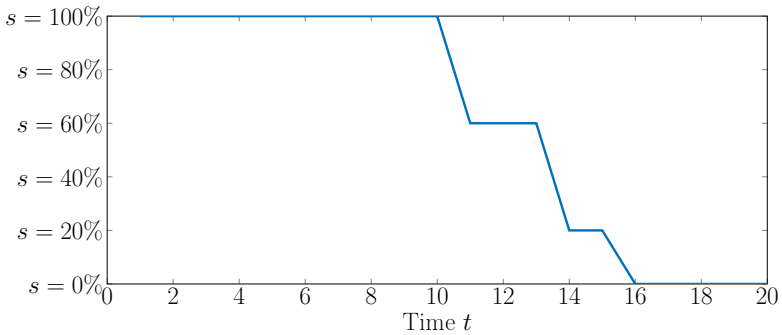
Transition model: $p(s'|s_j, a_i) = \Pr(s' = s_k|s_j, a_i), \forall k = 1 : \mathtt{S}$

Reward: $r(s, a, s') \in \mathbb{R}, \quad (\equiv r(s, a), \equiv r(s))$

Policy: $\pi = \{a(s_1), a(s_2), \cdots, a(s_{\mathtt{S}})\}$

Infrastructure Maintenance - Definitions





Bridges are inspected every two years:

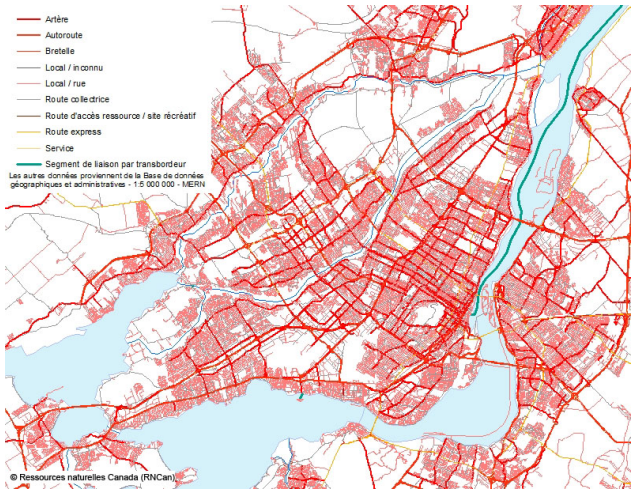
$$s \in \mathcal{S} = \left\{ \begin{array}{l} 100\% \\ 80\% \\ 60\% \\ 40\% \\ 20\% \\ 0\% \end{array} \right\}$$

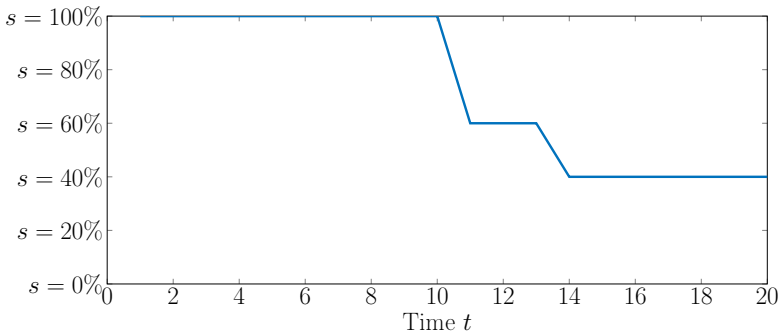
Data

$$\mathcal{D} = \left\{ \begin{array}{c} \{s_1, s_2, \cdots, s_T\}_1 \\ \{s_1, s_2, \cdots, s_T\}_2 \\ \vdots \\ \{s_1, s_2, \cdots, s_T\}_N \end{array} \right\}_{N \times T}$$

[Google images]

Infrastructure Maintenance - Definitions





Bridges are inspected every two years:

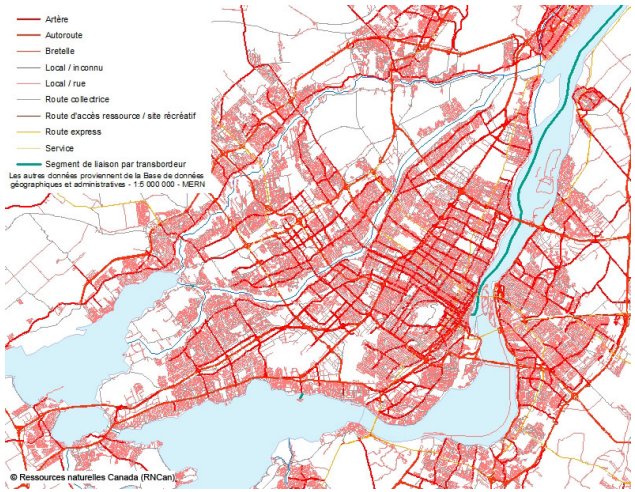
$$s \in \mathcal{S} = \left\{ \begin{array}{l} 100\% \\ 80\% \\ 60\% \\ 40\% \\ 20\% \\ 0\% \end{array} \right\}$$

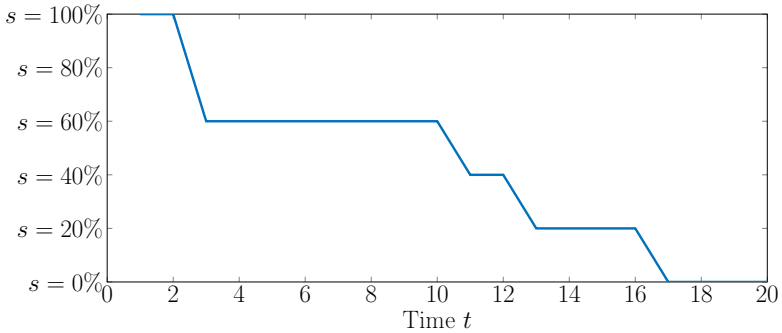
Data

$$\mathcal{D} = \left\{ \begin{array}{c} \{s_1, s_2, \cdots, s_T\}_1 \\ \{s_1, s_2, \cdots, s_T\}_2 \\ \vdots \\ \{s_1, s_2, \cdots, s_T\}_N \end{array} \right\}_{N \times T}$$

[Google images]

Infrastructure Maintenance - Definitions





Bridges are inspected every two years:

$$s \in \mathcal{S} = \left\{ \begin{array}{l} 100\% \\ 80\% \\ 60\% \\ 40\% \\ 20\% \\ 0\% \end{array} \right\}$$

Data

$$\mathcal{D} = \left\{ \begin{array}{c} \{s_1, s_2, \cdots, s_T\}_1 \\ \{s_1, s_2, \cdots, s_T\}_2 \\ \vdots \\ \{s_1, s_2, \cdots, s_T\}_N \end{array} \right\}_{N \times T}$$

[Google images]

Infrastructure Maintenance - Actions

$$a \in \mathcal{A} = \{\text{do nothing}, \text{maintain}, \text{replace}\}$$

Infrastructure Maintenance - Model

Employ the dataset $\mathcal{D}$ to build a **Markovian model**

$$p(s_{t+1}|s_t, a) \equiv p(s_{t+1}|s_{1:t}, a)$$

$$\hat{\Pr}(s_{t+1} = k|s_t = i, a = \text{do nothing}) = \frac{\#\{s_t = i, s_{t+1} = k\}}{\#\{s_t = i\}}$$

$$p(s_{t+1}|s_t = 1, a = \text{do nothing}) = \begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix} \text{S}$$

Infrastructure Maintenance - Model

Employ the dataset $\mathcal{D}$ to build a **Markovian model**

$$p(s_{t+1}|s_t, a) \equiv p(s_{t+1}|s_{1:t}, a)$$

$$\hat{\Pr}(s_{t+1} = k|s_t = i, a = \text{do nothing}) = \frac{\#\{s_t = i, s_{t+1} = k\}}{\#\{s_t = i\}}$$

$$p(s_{t+1}|s_t = 2, a = \text{do nothing}) = \begin{bmatrix} 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ & & & & & \\ & & & & & \end{bmatrix}_{S}$$

Infrastructure Maintenance - Model

Employ the dataset $\mathcal{D}$ to build a **Markovian model**

$$p(s_{t+1}|s_t, a) \equiv p(s_{t+1}|s_{1:t}, a)$$

$$\hat{\Pr}(s_{t+1} = k|s_t = i, a = \text{do nothing}) = \frac{\#\{s_t = i, s_{t+1} = k\}}{\#\{s_t = i\}}$$

$$p(s_{t+1}|s_t, a = \text{do nothing}) = \begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{\mathrm{S}\times\mathrm{S}}$$

Infrastructure Maintenance - Model (cont.)

$$p(s_{t+1}|s_t, a = \text{maintain}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}_{\mathrm{S}\times\mathrm{S}}$$

$$p(s_{t+1}|s_t, a = \text{replace}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{\mathrm{S}\times\mathrm{S}}$$

Infrastructure Maintenance - Rewards

Rewards for being in a state s or for taking an action a

$$\begin{aligned} r(\mathcal{S}) &= \overbrace{\text{AADTF}10^5 \text{ users} \cdot 365 \text{ days} \cdot 3\$/\text{user} \cdot \text{capacity}(\mathcal{S})}^{\text{AATF}} \\ &= \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$ \end{aligned}$$

$$\text{capacity}(\mathcal{S}) = \{ \overbrace{1}^{s=100\%}, \overbrace{1}^{80\%}, \overbrace{1}^{60\%}, \overbrace{0.90}^{40\%}, \overbrace{0.75}^{20\%}, \overbrace{0}^{0\%} \}$$

$$r(\mathcal{A}) = \{ \overbrace{0}^{\text{do nothing}}, \overbrace{-5}^{\text{repair}}, \overbrace{-20}^{\text{replace}} \} M\$$$

$$r(s, a, s') = r(s, a) = r(s) + r(a)$$

Illustrative example - Optimal actions

Given that the transition model is **Markovian**, the goal of the **Markovian Decision Process** is to identify the **optimal actions** a to be taken for each state s so that a *policy*

$$\pi^*(\mathcal{S}) = \{a(s_1), a(s_2), \cdots, a(s_S)\}$$

Section Outline

Markov Decision Process, MDP

Planning horizon

Finite planning horizon:
The rewards and utilities are considered over a fixed period of time.
The optimal policy is **non-stationary**; $\pi_t^*(s)$ depends on t.

Infinite planning horizon:
The rewards and utilities are considered over an infinite period.
The optimal policy is **stationary**; $\pi^*(s)$ does not depend on t.

We will only study problems with an **infinite planning horizon**

Utility for an infinite planning horizon

For a discount factor $\gamma \in (0, 1[$, the discounted sum of rewards over an infinite planning horizon is finite (note: interest rate $= \frac{1}{\gamma} - 1$)

$$\begin{aligned}
\mathbb{U}(s_{t=0}, s_{t=1}, \cdots, s_{t=\infty}) &= r(s_0) + \gamma r(s_1) + \gamma^2 r(s_2) + \cdots \\
&= \sum_{t=0}^{\infty} \gamma^t r(s_t) \\
&\leq \frac{\max_{s \in \mathcal{S}} r(s)}{1-\gamma}
\end{aligned}$$

⚠ **The issue is that we do not know $s_1, s_2, \cdots$**
we only know $\underbrace{p(s_{t+1}|s_t, a) \equiv p(s'|s, a)}_{\text{transition model}}$

Expected utility for an infinite planning horizon

Given $p(s'|s,a)$ and $r(s)$, we can compute the expected utility

Expected utility

$$\overline{\mathbb{U}}(s) \equiv \overline{\mathbb{U}}(s,\pi) \equiv \mathbb{E}[\mathbb{U}(s,\pi)] = r(s) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(S_t)\right]$$

Optimal policy

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot \left(r(s,a,s') + \gamma \overline{\mathbb{U}}(s',\pi^*)\right)$$

⚠ π^* appears on both sides of the equality $\rightarrow$ **iterative resolution**

Utility for an infinite planning horizon – Bellman

Bellman equation: "*The utility of a state is the immediate reward for that state plus the expected discounted utility of the next state given that the optimal action is taken*"

$$\begin{aligned}\overline{\mathbb{U}}(s) &= \max_{a\in\mathcal{A}} \mathbb{E}\left[r(s,a,s') + \gamma\overline{\mathbb{U}}(S')\right] \\ &= \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} p(s'|s,a)\cdot\left(r(s,a,s') + \gamma\overline{\mathbb{U}}(s')\right)\end{aligned}$$

Bellman update & Value iteration

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} p(s'|s,a)\cdot\left(r(s,a,s') + \gamma\overline{\mathbb{U}}^{(i)}(s')\right)$$

Algorithm 1: Value iteration

define: states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$;
define: transition model $p(s'|s,a)$;
define: rewards $r(s,a,s')$, discount γ, convergence tol. η;
initialize: $\overline{\mathbb{U}}'(s) = \overline{\mathbb{U}}(s) = 0$;
while $\overline{\mathbb{U}}(s) = 0$ *or* $|\overline{\mathbb{U}}'(s) - \overline{\mathbb{U}}(s)| \geq \eta$ **do**
    $\overline{\mathbb{U}}(s) \leftarrow \overline{\mathbb{U}}'(s)$;
    for $s \in \mathcal{S}$ **do**
        for $a \in \mathcal{A}$ **do**
            $\overline{\mathbb{U}}(s,a) = \sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot \big(r(s,a,s') + \gamma \overline{\mathbb{U}}(s')\big)$;
        $\pi^*(s) \leftarrow a^* = \arg\max_{a \in \mathcal{A}} \overline{\mathbb{U}}(s,a)$;
        $\overline{\mathbb{U}}'(s) \leftarrow \overline{\mathbb{U}}(s,a^*)$;
return: $\overline{\mathbb{U}}(s) = \overline{\mathbb{U}}'(s),\ \pi^*(s)$;

Infrastructure Maintenance - Summary

$$s \in \mathcal{S} = \left\{ \begin{array}{c} 100\% \\ 80\% \\ 60\% \\ 40\% \\ 20\% \\ 0\% \end{array} \right\}$$

$$a \in \mathcal{A} = \{\text{Do nothing, Maintain, Replace}\}$$

$$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$$

$$r(\mathcal{A}) = \{0, -5, -20\} M\$$$

$$r(s, a, s') = r(s) + r(a)$$

$$p(s_{t+1}|s_t, a = \text{Do nothing}) = \begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{S \times S}$$

$$p(s_{t+1}|s_t, a = \text{Maintain}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}_{S \times S}$$

$$p(s_{t+1}|s_t, a = \text{Replace}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{S \times S}$$

Infrastructure Maintenance - Value iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing, Maintain, Replace}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 1 , $\forall s \in \mathcal{S}$

$\overline{\mathbb{U}}_0(\mathcal{S}) = \{0, 0, 0, 0, 0, 0\}\ M\$$

$$\overline{\mathbb{U}}_1(s) = \max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s, a_1) \cdot (r(s, a_1) + \gamma \overline{\mathbb{U}}_0(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_2) \cdot (r(s, a_2) + \gamma \overline{\mathbb{U}}_0(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_3) \cdot (r(s, a_3) + \gamma \overline{\mathbb{U}}_0(s')) \end{cases}$$

$$\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$$

Infrastructure Maintenance - Value iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing, Maintain, Replace}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 2 , $\forall s \in \mathcal{S}$

$\overline{\mathbb{U}}_1(\mathcal{S}) = \{110, 211, 309, 393, 459, 440\}\ M\$$

$$\overline{\mathbb{U}}_2(s) = \max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s,a_1) \cdot (r(s,a_1) + \gamma \overline{\mathbb{U}}_1(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_2) \cdot (r(s,a_2) + \gamma \overline{\mathbb{U}}_1(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_3) \cdot (r(s,a_3) + \gamma \overline{\mathbb{U}}_1(s')) \end{cases}$$

$$\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$$

Infrastructure Maintenance - Value iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing, Maintain, Replace}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 3 , $\forall s \in \mathcal{S}$

$\overline{\mathbb{U}}_2(\mathcal{S}) = \{223, 329, 430, 511, 572, 550\}\ M\$$

$$\overline{\mathbb{U}}_3(s) = \max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s,a_1) \cdot (r(s,a_1) + \gamma \overline{\mathbb{U}}_2(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_2) \cdot (r(s,a_2) + \gamma \overline{\mathbb{U}}_2(s')) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_3) \cdot (r(s,a_3) + \gamma \overline{\mathbb{U}}_2(s')) \end{cases}$$

$$\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$$

Value iteration - $\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$

$$
\begin{aligned}
& & s{=}100\% & & 80\% & & 60\% & & 40\% & & 20\% & & 0\% & \\
\overline{\mathbb{U}}^{(0)}(\mathcal{S}) &= \{ \overbrace{0}, & \overbrace{0}, & & \overbrace{0}, & & \overbrace{0}, & & \overbrace{0}, & & \overbrace{0} & \} M\$ \\
\overline{\mathbb{U}}^{(1)}(\mathcal{S}) &= \{ 110, & 211, & & 309, & & 393, & & 459, & & 440 & \} M\$ \\
\overline{\mathbb{U}}^{(2)}(\mathcal{S}) &= \{ 223, & 329, & & 430, & & 511, & & 572, & & 550 & \} M\$ \\
& \vdots & & & & & & & & & & & & \\
\overline{\mathbb{U}}^{(536)}(\mathcal{S}) &= \{ 3640, & 3635, & & 3630, & & 3615, & & 3592, & & 3510 & \} M\$ \\
\overline{\mathbb{U}}^{(537)}(\mathcal{S}) &= \{ 3640, & 3635, & & 3630, & & 3615, & & 3592, & & 3510 & \} M\$
\end{aligned}
$$

$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$

$r(\mathcal{A}) = \{0, -5, -20\} M\$$

$p(s'|s, \text{do nothing}) =$

$$
\begin{bmatrix}
0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\
0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\
0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\
0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\
0 & 0 & 0 & 0 & 0.6 & 0.4 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

$$
\begin{aligned}
\overline{\mathbb{U}}^{(2)}(1) &\leftarrow r(1) + \max_{a \in \mathcal{A}} \left(\sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot (r(a) + \gamma \overline{\mathbb{U}}^{(1)}(s')) \right) \\
&\leftarrow \overbrace{109.5 M\$}^{r(s=1)} + \max_{a \in \mathcal{A}} \begin{cases} 0.95 \cdot (0.97 \cdot 110 M\$) + 0.03 \cdot (0.97 \cdot 211 M\$) + 0.02 \cdot (0.97 \cdot 309 M\$) = \boxed{113.5\ M\$} \\ 1 \cdot (-5 + 0.97 \cdot 110 M\$) = 101.7\ M\$ \\ 1 \cdot (-20 + 0.97 \cdot 110 M\$) = 86.7\ M\$ \end{cases} \\
&= 109.5\$ + 113.5\ M\$ \\
&= 223\ M\$ \quad (\pi^*(s=1) = a^* = \text{do nothing})
\end{aligned}
$$

Value iteration - $\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$

$$
\begin{array}{rcl}
 & & \overbrace{}^{s=100\%}\ \overbrace{}^{80\%}\ \overbrace{}^{60\%}\ \overbrace{}^{40\%}\ \overbrace{}^{20\%}\ \overbrace{}^{0\%} \\
\overline{\mathbb{U}}^{(0)}(\mathcal{S}) & = & \{0,\ 0,\ 0,\ 0,\ 0,\ 0\}\ M\$ \\
\overline{\mathbb{U}}^{(1)}(\mathcal{S}) & = & \{110,\ 211,\ 309,\ 393,\ 459,\ 440\}\ M\$ \\
\overline{\mathbb{U}}^{(2)}(\mathcal{S}) & = & \{223,\ 329,\ 430,\ 511,\ 572,\ 550\}\ M\$ \\
 & \vdots & \\
\overline{\mathbb{U}}^{(536)}(\mathcal{S}) & = & \{3640,\ 3635,\ 3630,\ 3615,\ 3592,\ 3510\}\ M\$ \\
\overline{\mathbb{U}}^{(537)}(\mathcal{S}) & = & \{3640,\ 3635,\ 3630,\ 3615,\ 3592,\ 3510\}\ M\$
\end{array}
$$

$$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\}M\$$$

$$r(\mathcal{A}) = \{0, -5, -20\}M\$$$

$$p(s'|s, \text{repair}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$
\begin{array}{rcll}
\overline{\mathbb{U}}^{(2)}(1) & \leftarrow & r(1) & + \max\limits_{a \in \mathcal{A}} \left(\sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot (r(a) + \gamma \overline{\mathbb{U}}^{(1)}(s')) \right) \\
 & \leftarrow & \overbrace{109.5 M\$}^{r(s=1)} & + \max\limits_{a \in \mathcal{A}} \begin{cases} 0.95 \cdot (0.97 \cdot 110 M\$) + 0.03 \cdot (0.97 \cdot 211 M\$) + 0.02 \cdot (0.97 \cdot 309 M\$) = \boxed{113.5\ M\$} \\ 1 \cdot (-5 + 0.97 \cdot 110 M\$) = 101.7\ M\$ \\ 1 \cdot (-20 + 0.97 \cdot 110 M\$) = 86.7\ M\$ \end{cases} \\
 & = & 109.5\$ & + 113.5\ M\$ \\
 & = & 223\ M\$ & (\pi^*(s=1) = a^* = \text{do nothing})
\end{array}
$$

Value iteration - $\overline{\mathbb{U}}^{(1)}(1) = 109.5\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 223\ M\$$

$$
\begin{array}{rcl}
 & & \overbrace{}^{s=100\%}\ \overbrace{}^{80\%}\ \overbrace{}^{60\%}\ \overbrace{}^{40\%}\ \overbrace{}^{20\%}\ \overbrace{}^{0\%} \\
\overline{\mathbb{U}}^{(0)}(\mathcal{S}) & = & \{0,\ 0,\ 0,\ 0,\ 0,\ 0\}\ M\$ \\
\overline{\mathbb{U}}^{(1)}(\mathcal{S}) & = & \{110,\ 211,\ 309,\ 393,\ 459,\ 440\}\ M\$ \\
\overline{\mathbb{U}}^{(2)}(\mathcal{S}) & = & \{223,\ 329,\ 430,\ 511,\ 572,\ 550\}\ M\$ \\
 & \vdots & \\
\overline{\mathbb{U}}^{(536)}(\mathcal{S}) & = & \{3640,\ 3635,\ 3630,\ 3615,\ 3592,\ 3510\}\ M\$ \\
\overline{\mathbb{U}}^{(537)}(\mathcal{S}) & = & \{3640,\ 3635,\ 3630,\ 3615,\ 3592,\ 3510\}\ M\$
\end{array}
$$

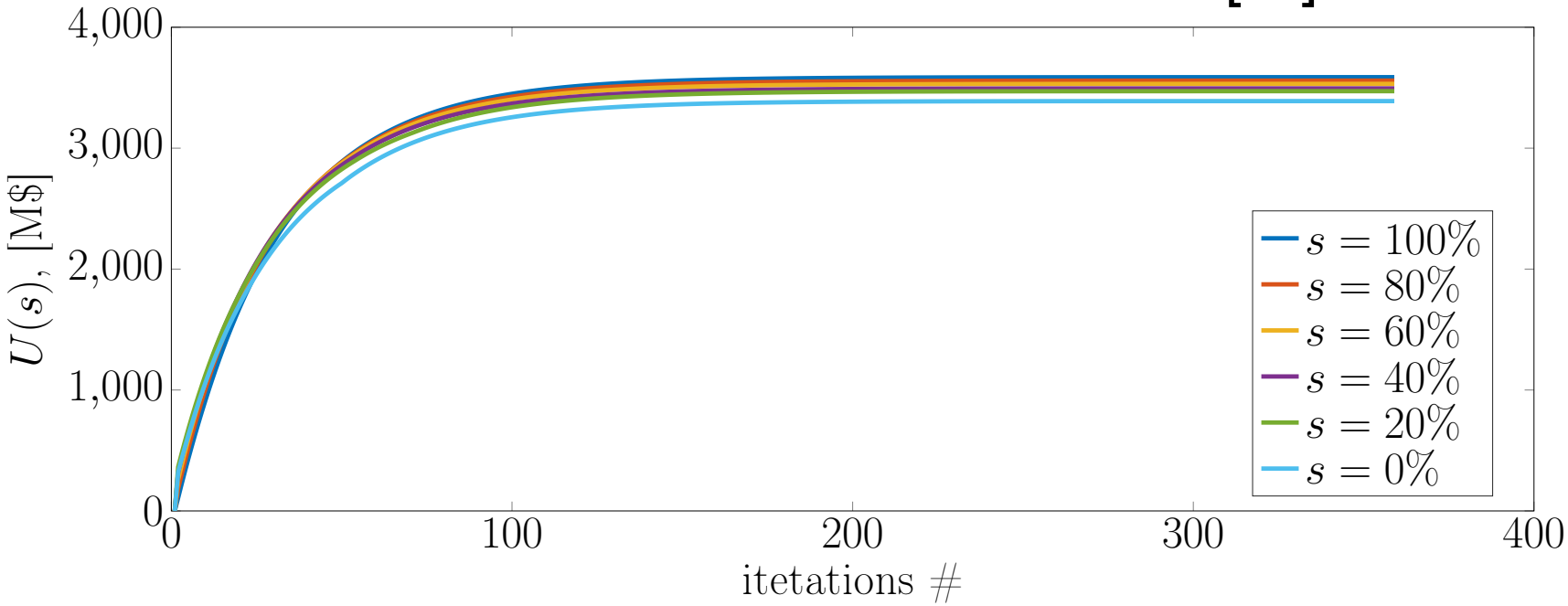
$$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\}M\$$$

$$r(\mathcal{A}) = \{0, -5, -20\}M\$$$

$$p(s'|s, \text{replace}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$
\begin{array}{rcll}
\overline{\mathbb{U}}^{(2)}(1) & \leftarrow & r(1) & + \max\limits_{a \in \mathcal{A}} \left(\sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot (r(a) + \gamma \overline{\mathbb{U}}^{(1)}(s')) \right) \\
 & \leftarrow & \overbrace{109.5 M\$}^{r(s=1)} & + \max\limits_{a \in \mathcal{A}} \begin{cases} 0.95 \cdot (0.97 \cdot 110 M\$) + 0.03 \cdot (0.97 \cdot 211 M\$) + 0.02 \cdot (0.97 \cdot 309 M\$) = \boxed{113.5\ M\$} \\ 1 \cdot (-5 + 0.97 \cdot 110 M\$) = 101.7\ M\$ \\ 1 \cdot (-20 + 0.97 \cdot 110 M\$) = 86.7\ M\$ \end{cases} \\
 & = & 109.5\$ & +113.5\ M\$ \\
 & = & 223\ M\$ & (\pi^*(s=1) = a^* = \text{do nothing})
\end{array}
$$

Infrastructure Maintenance - Value iteration [.m]



$$\overline{\mathbb{U}}(\mathcal{S}) = \{3640, 3635, 3630, 3615, 3592, 3510\}\ M\$$$

$$\pi^*(\mathcal{S}) = \{\underbrace{\text{Do nothing}}_{s=100\%}, \underbrace{\text{Maintain}}_{s=80\%}, \underbrace{\text{Maintain}}_{s=60\%}, \underbrace{\text{Maintain}}_{s=40\%}, \underbrace{\text{Replace}}_{s=20\%}, \underbrace{\text{Replace}}_{s=0\%}\}$$

[CIV_ML/Value_iteration_bridge.m]

Bellman update & Value iteration

Value iteration

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot \left(r(s,a,s') + \gamma \overline{\mathbb{U}}^{(i)}(s')\right)$$

The max operation is computationally expensive if $\mathtt{A}$ is large

solution: policy iteration

A policy $\pi = \{a(s_1), a(s_2), \cdots, a(s_S)\}$ defines an action to be taken for each state $s \in \mathcal{S}$

Policy iteration

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \sum_{s' \in \mathcal{S}} p(s'|s, \underbrace{\pi_i(s)}_{a^*}) \cdot \left(r(s, \underbrace{\pi_i(s)}_{a^*}, s') + \gamma \overline{\mathbb{U}}^{(i)}(s')\right)$$

Algorithm 2: Policy iteration

define: states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$;
define: transition model $p(s'|s,a)$;
define: rewards $r(s,a,s')$, discount γ;
initialize: $\overline{\mathbb{U}}(s) = 0$, $\pi'(s) = \pi(s) \in \mathcal{A}$;
while $\pi'(s) \neq \pi(s)$ **do**
    $\pi(s) = \pi'(s)$;
    $\overline{\mathbb{U}}(s) = \text{Policy eval.}(\overline{\mathbb{U}}(s), \pi(s), p(s'|s,a), r(s,a,s'), \gamma)$;
    for $s \in \mathcal{S}$ **do**
        $\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s'|s,a) \cdot \left(r(s,a,s') + \gamma \overline{\mathbb{U}}(s')\right)$;
return: $\overline{\mathbb{U}}(s)$, $\pi^*(s) = \pi'(s)$;

Infrastructure Maintenance - Policy iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing (1), Maintain (2), Replace (3)}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 1 , $\forall s \in \mathbb{S}$

$\pi_0(\mathcal{S}) = \{1, 1, 1, 1, 1, 1\}$

$\overline{\mathbb{U}}_1(\mathcal{S}) = \{2063, 1290, 768, 455, 197, 0\}\ M\$$

$$\pi_1(s) = \arg\max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s, a_1)\left(r(s, a_1) + \gamma \overline{\mathbb{U}}_1(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_2)\left(r(s, a_2) + \gamma \overline{\mathbb{U}}_1(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_3)\left(r(s, a_3) + \gamma \overline{\mathbb{U}}_1(s')\right) \end{cases}$$

$\pi_5(\mathcal{S}) = \pi_6(\mathcal{S}) = \pi^*(\mathcal{S}) = \{1, 2, 2, 2, 3, 3\}$

Infrastructure Maintenance - Policy iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing (1), Maintain (2), Replace (3)}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 2 , $\forall s \in \mathcal{S}$

$\pi_1(\mathcal{S}) = \{2, 2, 3, 3, 3, 3\}$

$\overline{\mathbb{U}}_2(\mathcal{S}) = \{3483, 3483, 3468, 3457, 3441, 3359\} M\$$

$$\pi_2(s) = \arg\max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s, a_1)\left(r(s, a_1) + \gamma \overline{\mathbb{U}}_2(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_2)\left(r(s, a_2) + \gamma \overline{\mathbb{U}}_2(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_3)\left(r(s, a_3) + \gamma \overline{\mathbb{U}}_2(s')\right) \end{cases}$$

$\pi_5(\mathcal{S}) = \pi_6(\mathcal{S}) = \pi^*(\mathcal{S}) = \{1, 2, 2, 2, 3, 3\}$

Infrastructure Maintenance - Policy iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing (1), Maintain (2), Replace (3)}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 3 , $\forall s \in \mathbb{S}$

$\pi_2(\mathcal{S}) = \{1, 1, 2, 2, 3, 3\}$

$\overline{\mathbb{U}}_3(\mathcal{S}) = \{3622, 3606, 3603, 3588, 3576, 3494\}\ M\$$

$$\pi_3(s) = \arg\max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s,a_1)\left(r(s,a_1) + \gamma \overline{\mathbb{U}}_3(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_2)\left(r(s,a_2) + \gamma \overline{\mathbb{U}}_3(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s,a_3)\left(r(s,a_3) + \gamma \overline{\mathbb{U}}_3(s')\right) \end{cases}$$

$\pi_5(\mathcal{S}) = \pi_6(\mathcal{S}) = \pi^*(\mathcal{S}) = \{1, 2, 2, 2, 3, 3\}$

Infrastructure Maintenance - Policy iteration

$s \in \mathcal{S} = \{100\%, 80\%, 60\%, 40\%, 20\%, 0\%\}$
$a \in \mathcal{A} = \{\text{Do nothing (1), Maintain (2), Replace (3)}\}$
$r(\mathcal{S}) = \{109.5, 109.5, 109.5, 98.6, 82.1, 0\} M\$$
$r(\mathcal{A}) = \{0, -5, -20\} M\$$
$\gamma = 0.97$

$p(s'|s,a) =$

$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Iteration: 6 , $\forall s \in \mathbb{S}$

$\pi_5(\mathcal{S}) = \{1, 2, 2, 2, 3, 3\}$

$\overline{\mathbb{U}}_4(\mathcal{S}) = \{3640, 3635, 3630, 3615, 3592, 3510\}\ M\$$

$$\pi_6(s) = \arg\max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s, a_1)\big(r(s, a_1) + \gamma \overline{\mathbb{U}}_5(s')\big) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_2)\big(r(s, a_2) + \gamma \overline{\mathbb{U}}_5(s')\big) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_3)\big(r(s, a_3) + \gamma \overline{\mathbb{U}}_5(s')\big) \end{cases}$$

$\pi_5(\mathcal{S}) = \pi_6(\mathcal{S}) = \pi^*(\mathcal{S}) = \{1, 2, 2, 2, 3, 3\}$

Policy eval. - $\overline{\mathbb{U}}^{(1)}(1) = 2063\ M\$ \rightarrow \overline{\mathbb{U}}^{(2)}(1) = 3483\ M\$$

$$\overline{\mathbb{U}}^{(1)}(\mathcal{S}) = \{2063, 1290, 768, 455, 197, 0\}\ M\$, \quad \pi^{(1)}(\mathcal{S}) = \{2, 2, 3, 3, 3, 3\}$$

Policy Iteration loop
Policy eval. loop
State
Policy Iteration loop-1
$\pi^{(1)}(1)$=2, maintain

$$\begin{aligned}
\overline{\mathbb{U}}^{(2)(1)}(1) &= 1 \cdot (109.5M\$ - 5 + 0.97 \cdot 2063M\$) &&= 2106\ M\$ \\
\overline{\mathbb{U}}^{(2)(2)}(1) &= 1 \cdot (109.5M\$ - 5 + 0.97 \cdot 2106M\$) &&= 2147\ M\$ \\
&\vdots \\
\overline{\mathbb{U}}^{(2)(350)}(1) &= 1 \cdot (109.5M\$ - 5 + 0.97 \cdot 3483M\$) &&= 3483\ M\$. \\
\overline{\mathbb{U}}^{(2)}(1) = \overline{\mathbb{U}}^{(2)(351)}(1) &= 1 \cdot (109.5M\$ - 5 + 0.97 \cdot 3483M\$) &&= 3483\ M\$.
\end{aligned}$$

Policy iteration - $\pi^{(1)}(1) = 2 \rightarrow \pi^{(2)}(1) = 1$

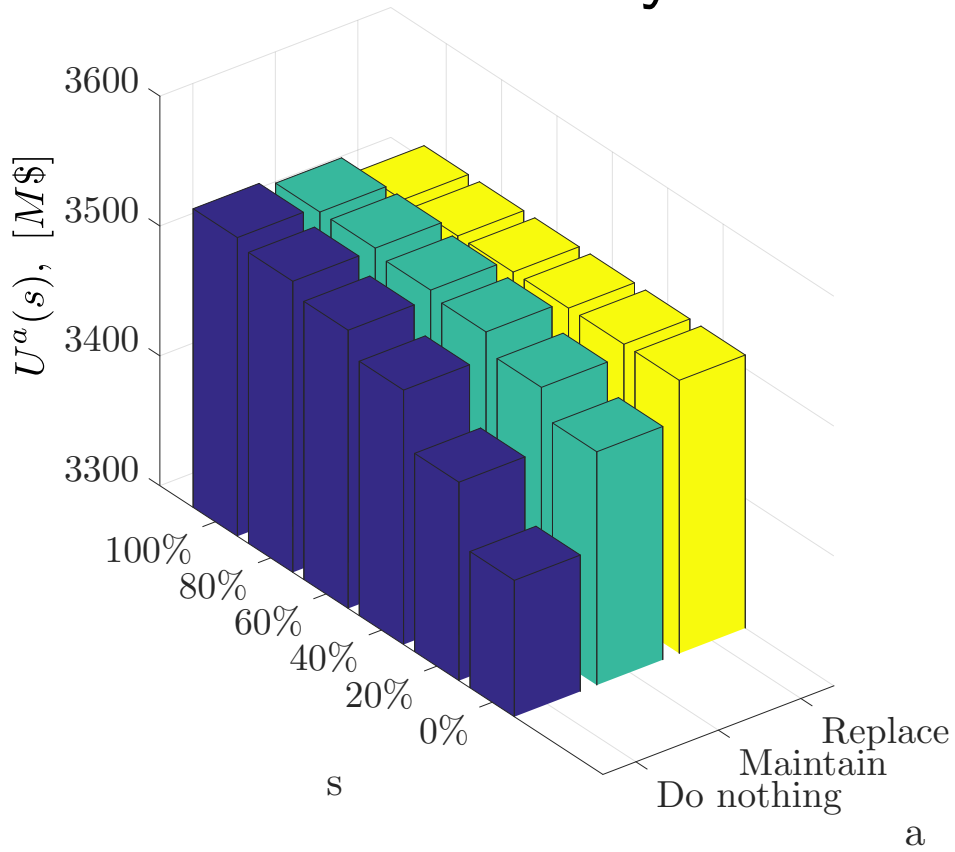
$$\overline{\mathbb{U}}^{(2)}(\mathcal{S}) = \{3483, 3483, 3468, 3457, 3441, 3359\}\ M\$$$

$$\pi^{(2)}(1) = \arg\max_{a \in \mathcal{A}} \begin{cases} \sum_{s' \in \mathcal{S}} p(s'|s, a_1) \cdot \left(r(s, a_1) + \gamma \overline{\mathbb{U}}_2(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_2) \cdot \left(r(s, a_2) + \gamma \overline{\mathbb{U}}_2(s')\right) \\ \sum_{s' \in \mathcal{S}} p(s'|s, a_3) \cdot \left(r(s, a_3) + \gamma \overline{\mathbb{U}}_2(s')\right) \end{cases}$$

$$= \arg\max_{a \in \mathcal{A}} \begin{cases} 109.5 + 0\ M\$ + 0.97(0.95 \cdot 3483 M\$ + 0.03 \cdot 3483 M\$ + 0.02 \cdot 3468 M\$) = \boxed{3488\ M\$} \\ 109.5 - 5\ M\$ + 0.97 \cdot (1 \cdot 3483 M\$) = 3483\ M\$ \\ 109.5 - 20\ M\$ + 0.97 \cdot (1 \cdot 3483 M\$) = 3468\ M\$ \end{cases}$$

$$= 1.$$

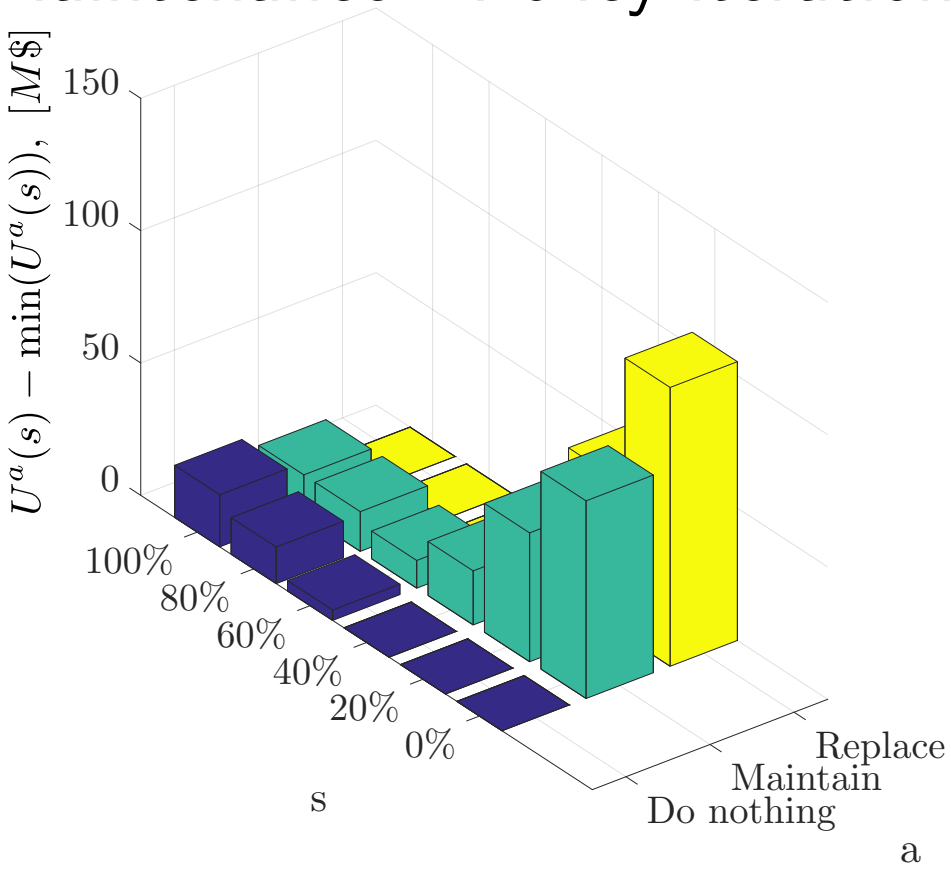
Infrastructure Maintenance - Policy iteration [.m]



$$\pi^*(\mathcal{S}) = \{\underbrace{\text{Do nothing}}_{s=100\%}, \underbrace{\text{Maintain}}_{s=80\%}, \underbrace{\text{Maintain}}_{s=60\%}, \underbrace{\text{Maintain}}_{s=40\%}, \underbrace{\text{Replace}}_{s=20\%}, \underbrace{\text{Replace}}_{s=0\%}\}$$

[CIV_ML/Policy_iteration_bridge.m]

Infrastructure Maintenance - Policy iteration [.m]



$$\pi^*(\mathcal{S}) = \{\underbrace{\text{Do nothing}}_{s=100\%}, \underbrace{\text{Maintain}}_{s=80\%}, \underbrace{\text{Maintain}}_{s=60\%}, \underbrace{\text{Maintain}}_{s=40\%}, \underbrace{\text{Replace}}_{s=20\%}, \underbrace{\text{Replace}}_{s=0\%}\}$$

[CIV_ML/Policy_iteration_bridge.m]

Policy v.s. Value iteration

π: Policy iteration $\equiv$ π: Value iteration

Time: Policy iteration $\ll$ Time: Value iteration

Partially Observable MDP (POMDP)

One limitation of MDP is that at a time t, you must know exactly in what state (s) you are.

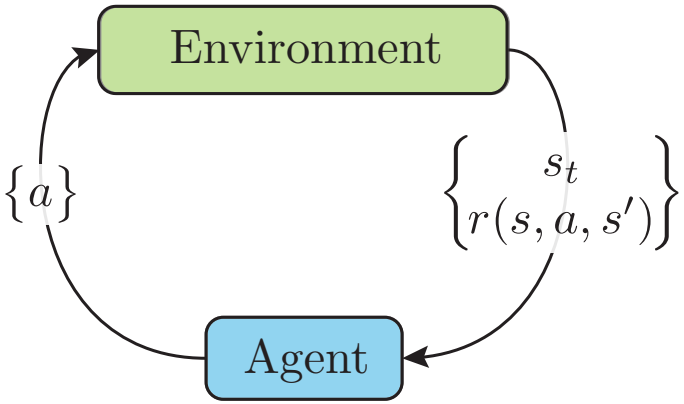
When this is not the case $\rightarrow$ **POMDP**
(Note: computationally more demanding than MDP)

Chapter 17 – Russell, S. and Norvig, P. (1995).
Artificial Intelligence, A modern approach. Prentice-Hall.

Model-based v.s. Model-free

When we can explicitly represent $p(s_{t+1}|s_t, a)$
$\rightarrow$ **Model-based RL: MDP/POMDP, etc...**

When we cannot represent $p(s_{t+1}|s_t, a)$ explicitly
$\rightarrow$ **Model-free RL: Q-learning**, SARSA, etc...



Interaction between the **agent** and the **environment** over multiple **episodes**

[Google images]

Section Outline

Expected Utility

Utility $\leftarrow$ sum of discounted reward

$$\mathbb{U}(s,\pi) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}), \quad a_t = \pi(s_t)$$

Expected Utility

$$\mathbb{E}[\mathbb{U}(s,\pi)] \equiv \overline{\mathbb{U}}(s,\pi) \equiv \overline{\mathbb{U}}(s)$$

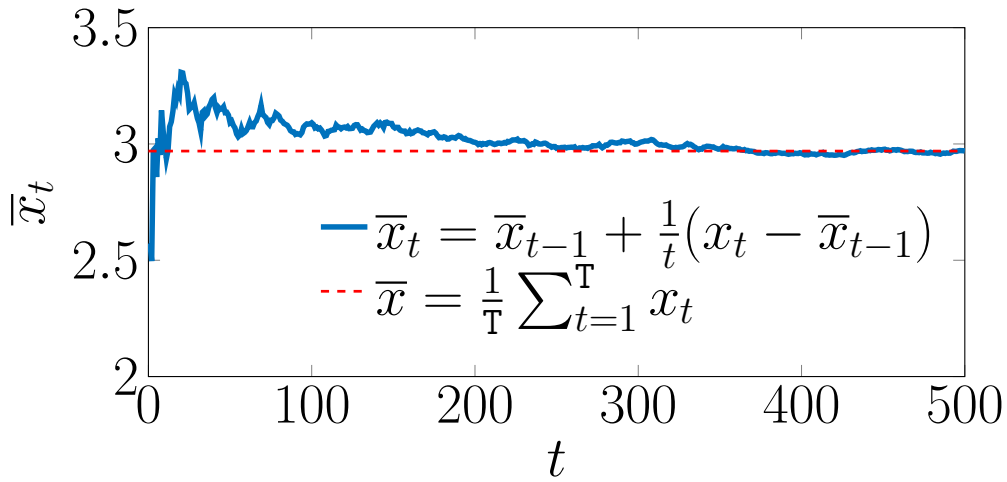
$\overline{\mathbb{U}}(s,\pi) \leftarrow$ **average over realizations of** $\mathbb{U}(s,\pi)$

Running average

The average $\overline{x}_{\mathrm{T}}$ of a set $\{x_1, x_2, \cdots, x_{\mathrm{T}}\}$ can be calculated incrementally by following

$$\overline{x}_t = \overline{x}_{t-1} + \frac{1}{t}(x_t - \overline{x}_{t-1})$$

e.g.: a set of $\mathrm{T} = 500$ realizations of $x_t : X \sim \mathcal{N}(x; 3, 1)$



Temporal difference (TD) learning

Running average

$$\overline{x}_t = \overline{x}_{t-1} + \frac{1}{t}(x_t - \overline{x}_{t-1})$$

Expected Utility

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \overline{\mathbb{U}}^{(i)}(s) + \frac{1}{\mathtt{N}(s)}\left(\mathbb{U}^{(i)}(s) - \overline{\mathbb{U}}^{(i)}(s)\right)$$

TD learning

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \overline{\mathbb{U}}^{(i)}(s) + \alpha\Big(\underbrace{r(s,a,s') + \gamma\overline{\mathbb{U}}^{(i)}(s')}_{\text{TD-target}} - \overline{\mathbb{U}}^{(i)}(s)\Big)$$

Learning rate – α

$$\left.\begin{aligned} \sum_{\mathtt{N}(s,a)=1}^{\infty} \alpha(\mathtt{N}(s,a)) &= \infty \\ \sum_{\mathtt{N}(s,a)=1}^{\infty} \alpha^2(\mathtt{N}(s,a)) &< \infty \end{aligned}\right\} \text{e.g. } \alpha(\mathtt{N}(s,a)) = \frac{c}{c + \mathtt{N}(s,a)},\ c \in \mathbb{R}^+$$

Action-utility function

$\mathbb{Q}(s,a)$: Action-utility function such that

$$\overline{\mathbb{U}}(s) \equiv \overline{\mathbb{U}}(s,\pi^*) = \max_{a\in\mathcal{A}} \mathbb{Q}(s,a)$$

$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} \mathbb{Q}(s,a)$$

Temporal difference Q-learning (off-policy)

$$\mathbb{Q}^{(i+1)}(s,a) \leftarrow \mathbb{Q}^{(i)}(s,a) + \alpha\left(r(s,a,s') + \gamma \max_{a'\in\mathcal{A}} \mathbb{Q}^{(i)}(s',a') - \mathbb{Q}^{(i)}(s,a)\right)$$

Exploration-exploitation tradeoff

When transiting from state s to state s':

- you cannot always rely on the current $\pi^*(s) \rightarrow$ **local maximum**
- you cannot always act randomly $\rightarrow$ **poor performance**

An ϵ-**greedy agent** selects an action following

$$a = \begin{cases} \mathrm{a}_i : p(a) = \frac{1}{\mathrm{A}}, \forall a, & \text{with } \Pr(\epsilon(\mathrm{N}(s))) & \text{Random action} \\ \arg\max_{a \in \mathcal{A}} \mathbb{Q}(s,a), & \text{with } 1 - \Pr(\epsilon(\mathrm{N}(s))) & \text{Greedy action,} \end{cases}$$

$$\text{where } \epsilon(\mathrm{N}(s)) \rightarrow 0 \text{ as } \mathrm{N}(s) \rightarrow \infty$$

Algorithm 3: Temporal difference Q-learning

define: $s \in \mathcal{S}$, $a \in \mathcal{A}$, $r(s,a,s')$, γ, ϵ, α

initialize: $\mathbb{Q}^{(0)}(s,a)$, $\mathtt{N}(s,a) = 0$, $\forall\{s \in \mathcal{S}, a \in \mathcal{A}\}$, $i = 0$

for *episodes* $e \in \{1 : \mathtt{E}\}$ **do**

  initialize: s

  for *time* $t \in \{1 : \mathtt{T}\}$ **do**

    $u : U \sim \mathcal{U}(0,1)$

$$a = \begin{cases} \mathrm{a}_i : p(a) = \frac{1}{\mathtt{A}}, \forall a, & \text{if } u < \epsilon\big(\mathtt{N}(s)\big) \quad \text{Random} \\ \arg\max_{a \in \mathcal{A}} \mathbb{Q}^{(i)}(s,a), & \text{if } u \geq \epsilon\big(\mathtt{N}(s)\big) \quad \text{Greedy} \end{cases}$$

    observe: $r(s,a,s')$, s'

$$\mathbb{Q}^{(i+1)}(s,a) \leftarrow \mathbb{Q}^{(i)}(s,a) + \alpha\big(\mathtt{N}(s,a)\big)\left(r(s,a,s') + \gamma \max_{a' \in \mathcal{A}} \mathbb{Q}^{(i)}(s',a') - \mathbb{Q}^{(i)}(s,a)\right)$$

    $s = s'$, $\mathtt{N}(s,a) = \mathtt{N}(s,a) + 1$, $i = i + 1$

return: $\overline{\mathbb{U}}(s) = \max_{a \in \mathcal{A}} \mathbb{Q}^{(i)}(s,a)$

$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \mathbb{Q}^{(i)}(s,a)$

Infrastructure Maintenance – Episodes

Learning rate: $\alpha\big(\mathbb{N}(s,a)\big) = \frac{70}{70+\mathbb{N}(s,a)}$

ϵ-Greedy: $\epsilon\big(\mathbb{N}(s)\big) = \frac{70}{70+\mathbb{N}(s)}$

$\mathbb{Q}^{(0)}(s,a) = \$0\, M,\ \forall\{s \in \mathcal{S}, a \in \mathcal{A}\}$

Environement:  $\equiv p(s'|s,a)$

500 episodes, each comprising 100 steps

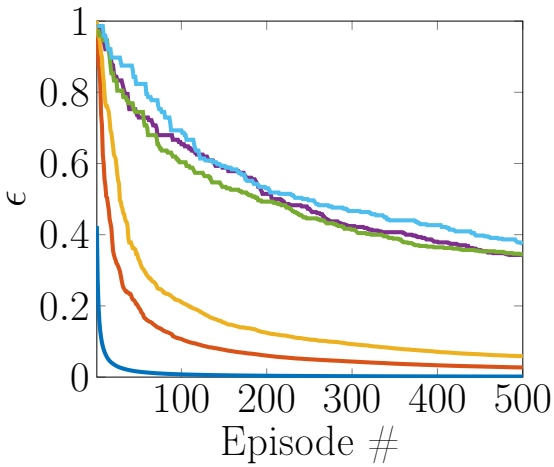
$p(s'|s,a) =$

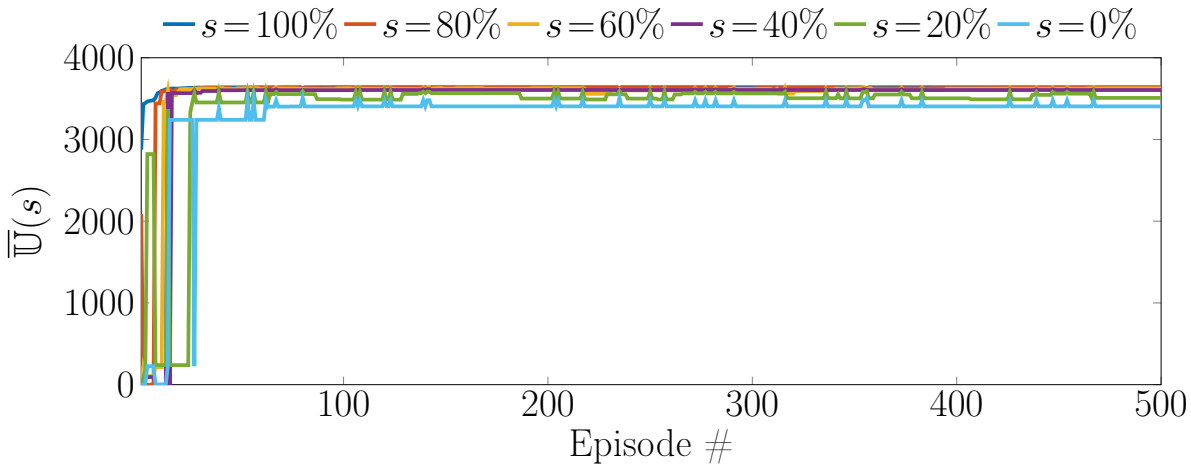
$$\begin{bmatrix} 0.95 & 0.03 & 0.02 & 0 & 0 & 0 \\ 0 & 0.9 & 0.05 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0.8 & 0.12 & 0.05 & 0.03 \\ 0 & 0 & 0 & 0.7 & 0.25 & 0.05 \\ 0 & 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Infrastructure Maintenance – Q-Learning

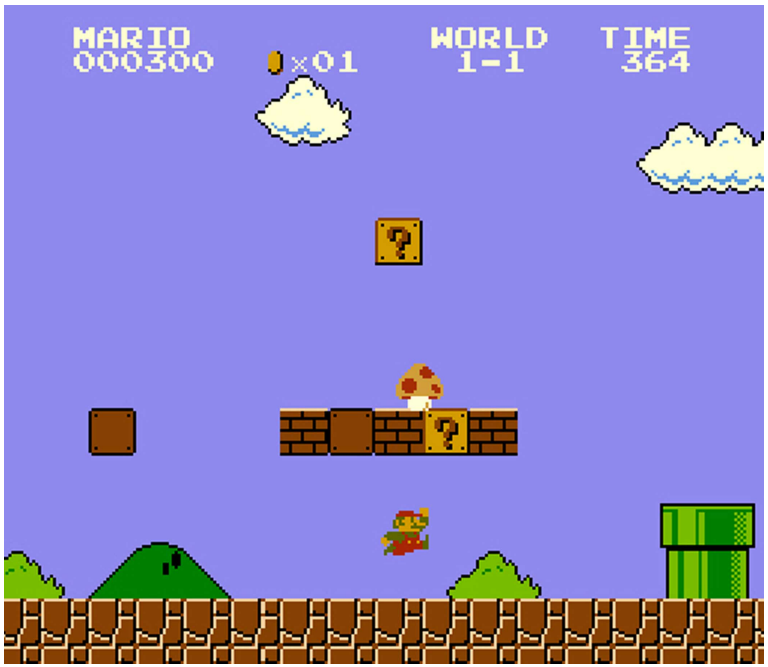




$$\overline{\mathbb{U}}^{(50K)}(\mathcal{S}) = \{3640, 3635, 3630, 3615, 3592, 3510\}\,\$M.$$

$$\pi^*(\mathcal{S}) = \{\underbrace{\text{Do nothing}}_{s=100\%}, \underbrace{\text{Maintain}}_{s=80\%}, \underbrace{\text{Maintain}}_{s=60\%}, \underbrace{\text{Maintain}}_{s=40\%}, \underbrace{\text{Replace}}_{s=20\%}, \underbrace{\text{Replace}}_{s=0\%}\}$$

Q-learning with function approximation



$$\mathcal{A} = \{\text{  }\}$$

$$\mathcal{S} = \{\text{  }\}$$

$$r(\mathcal{S}) = \{\text{fct}(\text{  })\}$$

$$p(s_{t+1}|s_t, a) = \text{unknown...}$$

$$\overline{\mathbb{U}}(s) \rightarrow \mathbb{Q}(s, a)$$
(Neural Network)

Chapter 21 – Russell, S. and Norvig, P. (1995)
Artificial Intelligence, A modern approach. Prentice-Hall.

Sutton, R. S. and Barto, A. G. (2018)
Reinforcement learning: An introduction, MIT Press, 2nd edition

[Google images]

Summary

AI : reproduction of an intelligent behavior by a machine.
$\rightarrow$ perception and dynamic interaction with its environment through the process of **making rational decisions**.

States: $\underbrace{s \in \mathbb{R}}_{\text{continuous}}$ or $\underbrace{s_j \in \mathcal{S} = \{s_1, s_2, \cdots, s_{\mathtt{S}}\}}_{\text{discrete}}$

Actions: $a_i \in \mathcal{A} = \{a_1, a_2, \cdots, a_{\mathtt{A}}\}$

Model: $p(s'|s_j, a_i) = \Pr(s' = s_k|s_j, a_i),\ \forall k = 1 : \mathtt{S}$

Reward: $r(s, a, s') \in \mathbb{R}$

Policy: $\pi = \{a(s_1), a(s_2), \cdots, a(s_{\mathtt{S}})\}$

When we can explicitly represent $p(s_{t+1}|s_t, a)$
$\rightarrow$ **Model-based RL: MDP/POMDP, etc...**

When we cannot represent $p(s_{t+1}|s_t, a)$ explicitly
$\rightarrow$ **Model-free RL: Q-learning, SARSA, etc...**

Infinite planning horizon:
The rewards and utilities are considered over an infinite period of time.
The optimal policy is stationary; $\pi^*(s)$ does not depend on t.

Bellman equation: "*The utility of a state is the immediate reward for that state plus the expected discounted utility of the next state given that the optimal action is taken*"

$$\overline{\mathbb{U}}(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s'|s, a) \cdot \left(r(s, a, s') + \gamma \overline{\mathbb{U}}(s')\right)$$

Bellman update & Value iteration

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s'|s, a) \cdot \left(r(s, a, s') + \gamma \overline{\mathbb{U}}^{(i)}(s')\right)$$

Policy iteration

$$\overline{\mathbb{U}}^{(i+1)}(s) \leftarrow \sum_{s' \in \mathcal{S}} p(s'|s, \pi_i(s)) \cdot \left(r(s, \pi_i(s), s') + \gamma \overline{\mathbb{U}}^{(i)}(s')\right)$$

π: Policy iteration $\equiv$ π: Value iteration
Time: Policy iteration $\ll$ Time: Value iteration

TD Q-learning:

$$\mathbb{Q}^{(i+1)}(s,a) \leftarrow \mathbb{Q}^{(i)}(s,a) + \alpha\left(\mathtt{N}(s,a)\right)\left(r(s,a,s') + \gamma \max_{a' \in \mathcal{A}} \mathbb{Q}^{(i)}(s',a') - \mathbb{Q}^{(i)}(s,a)\right)$$