Day 1. Tabular MDPs

SAMSUNG AI

Reinforcement Learning

June 17, 2022

Jaeuk Shin, Mingyu Park

CORE
Control + Optimization Research Lab

Contents

What we are going to do…

- Representation of Markov decision process
  - Basic matrix algebra using **numpy**
- Implementation of **value iteration**
- Implementation of **policy iteration**
- Practice : GridWorld

MDP - Review

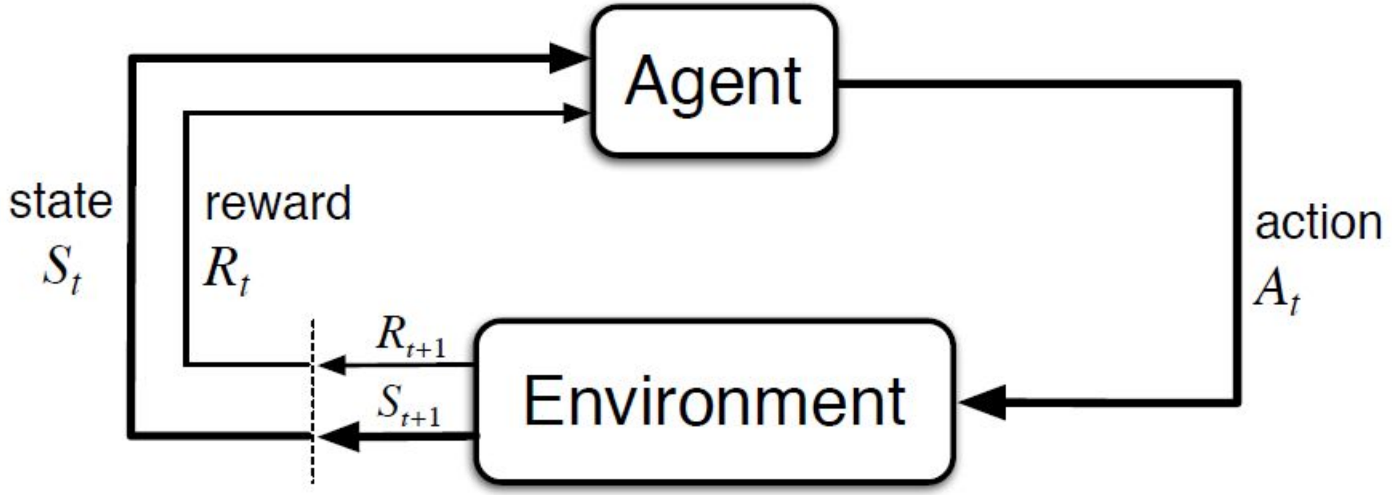
$$\mathcal{S} = \{s_0, \cdots s_{n-1}\}: \text{ state space}$$

$$\mathcal{A} = \{a_0, \cdots a_{m-1}\}: \text{ action space}$$

$$p(s'|s,a): \text{ transition probability}$$

$$r(s,a): \text{ reward function}$$

$$\gamma: \text{ discount rate}$$



MDP - Playing Go

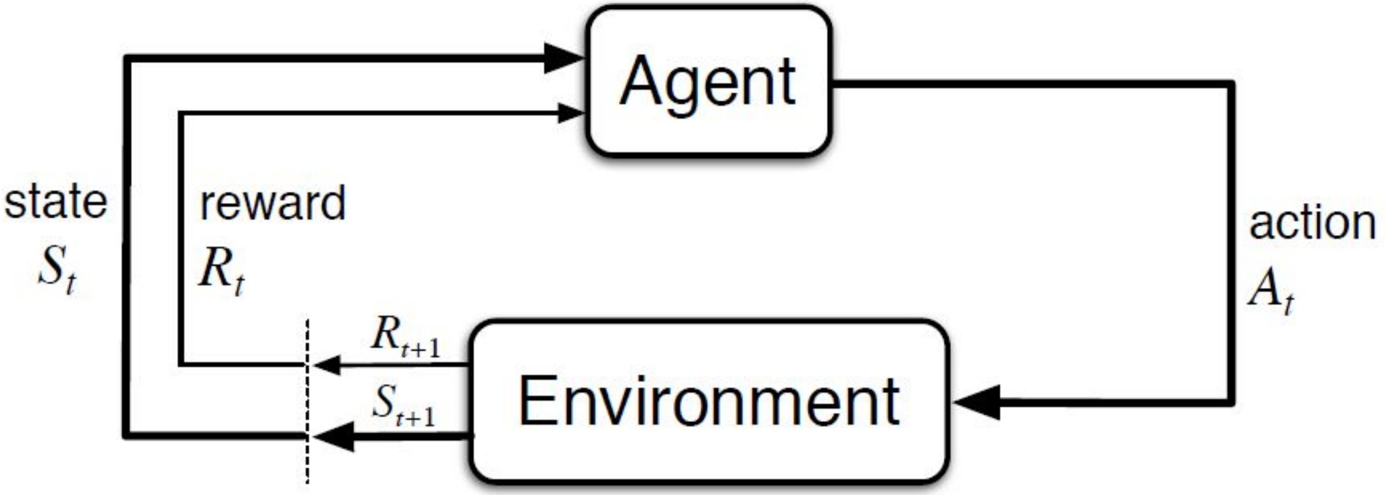
$\mathcal{S} = \{s_0, \cdots s_{n-1}\}$: state space Image

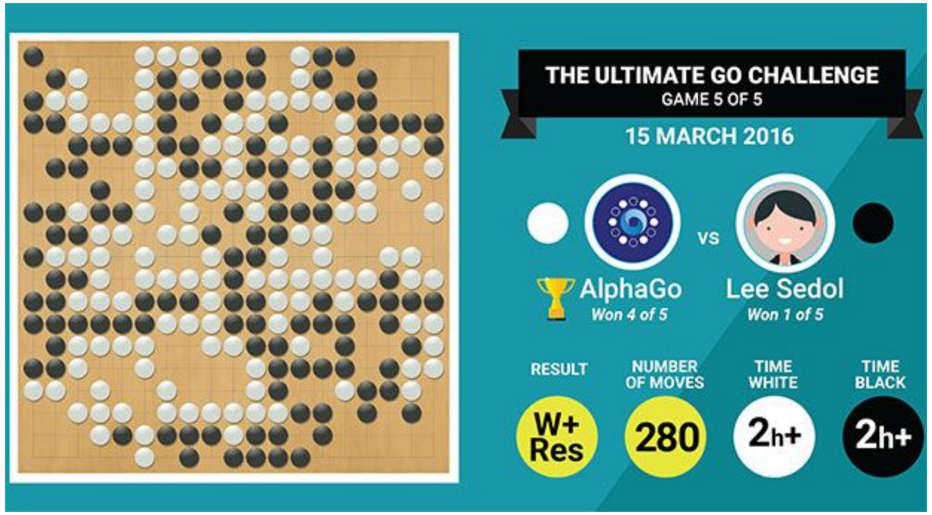
$\mathcal{A} = \{a_0, \cdots a_{m-1}\}$: action space Placing stone

$p(s'|s,a)$: transition probability Opponent action

$r(s,a)$: reward function +1 if win

γ: discount rate





MDP - Breakout

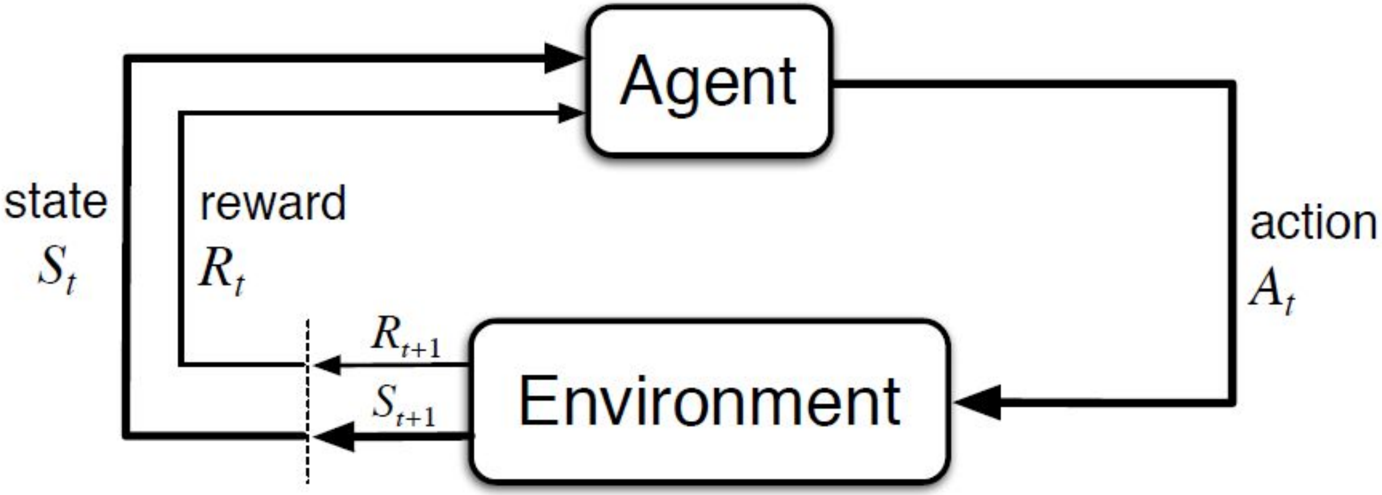
$\mathcal{S} = \{s_0, \cdots s_{n-1}\}$: state space Image

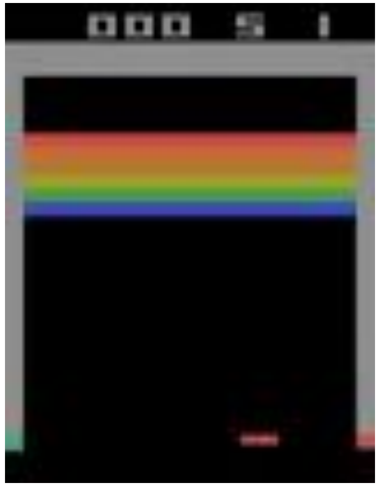
$\mathcal{A} = \{a_0, \cdots a_{m-1}\}$: action space Move bar / Reset stone

$p(s'|s,a)$: transition probability Game engine

$r(s,a)$: reward function Score

γ: discount rate





MDP - Portfolio Optimization

https://arxiv.org/pdf/1706.10059.pdf

$\mathcal{S} = \{s_0, \cdots s_{n-1}\}$: state space — $\boxed{\text{Asset price history}}$ + Portfolio weight history

$\mathcal{A} = \{a_0, \cdots a_{m-1}\}$: action space — Portfolio weight

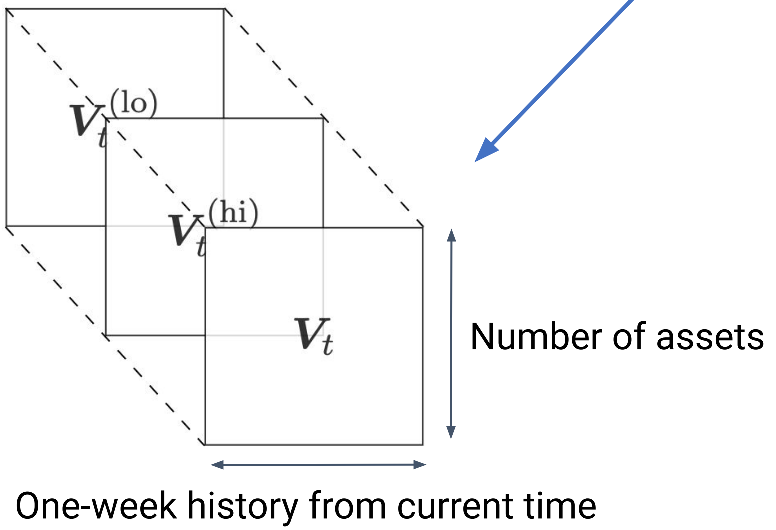
$p(s'|s,a)$: transition probability — Price model

$r(s,a)$: reward function — Log ratio of portfolio value increase

γ: discount rate

$$r_t = \ln \frac{p_t}{p_{t-1}}$$

$$\Rightarrow \sum_{t=1}^{T} r_t = \ln \frac{p_1}{p_0} + \ln \frac{p_2}{p_1} + \cdots + \ln \frac{p_{T-1}}{p_{T-2}} + \ln \frac{p_T}{p_{T-1}} = \ln \frac{p_T}{p_0}.$$



Q1: Is the model stationary?
Q2: Does price history convey all the information for the asset allocation?

MDP - Review

How to represent these data?

1. transition probability $p(s'|s,a)$

2. reward function $r(s,a)$

$\Rightarrow$ We use **NumPy**, a Python package for linear algebra!

MDP - Review

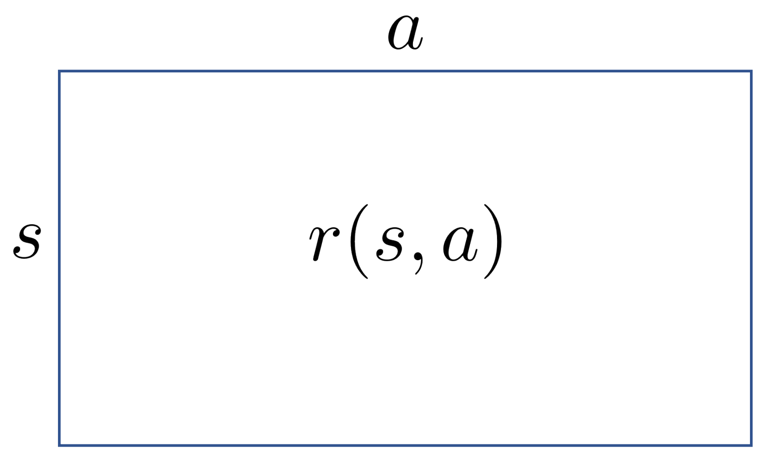
How to represent these data?

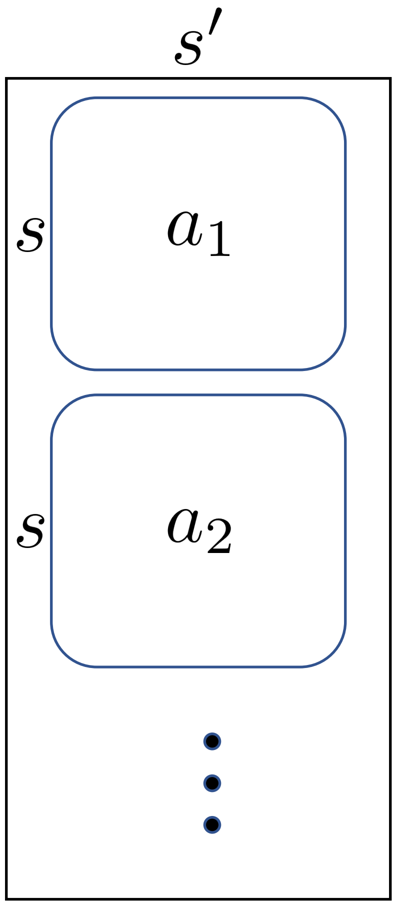
1. transition probability $p(s'|s,a)$

   $\rightarrow$ matrix P of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$

2. reward function $r(s,a)$

   $\rightarrow$ matrix R of size $|\mathcal{S}| \times |\mathcal{A}|$





CORE
Control + Optimization Research Lab

MDP - Review

$$\mathcal{S} = \{s_0, s_1\}, \quad \mathcal{A} = \{a_0, a_1\},$$
$$r(s_0, a_0) = -2, \quad r(s_0, a_0) = -0.5,$$
$$r(s_1, a_0) = -1, \quad r(s_1, a_1) = -3.0,$$
$$p(s_0 | s_0, a_0) = 0.75,$$
$$p(s_0 \,| s_1, a_0) = 0.75,$$
$$p(s_0 \,| s_0, a_1) = 0.25,$$
$$p(s_0 \,| s_1, a_1) = 0.25.$$

```
R = np.array([[-2.0, -0.5],
              [-1.0, -3.0]])

P = np.array([[0.75, 0.25],
              [0.75, 0.25],
              [0.25, 0.75],
              [0.25, 0.75]])
```

Solving Tabular MDPs – Value Iteration

Review : **Bellman operator** $\mathcal{T} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is given by

$$(\mathcal{T}v)(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a)v(s') \right)$$

Given a vector v of size $n \times 1$,

Step 1. compute $r(s,a) + \gamma \sum_{s'} p(s'|s,a)v(s')$,

Step 2. and then take $\max_a$.

Solving Tabular MDPs – Value Iteration

Step 1. compute $r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')$,

```python
def q_ftn(P, R, gamma, v):
    """
    given v, get corresponding q
    """
    return R + gamma * np.reshape(np.matmul(P, v), newshape=R.shape, order='F')
```

Shape of $q(s,a)$?

Step 2. and then take $\max_a$.

```python
def bellman_update(P, R, gamma, v):
    """
    implementation of one-step Bellman update
    return : vector of shape (|S|, 1) which corresponds to Tv, where T is Bellman operator
    """
    q = q_ftn(P, R, gamma, v)
    v_next = np.max(q, axis=1, keepdims=True)    # computation of Bellman operator Tv

    return v_next
```

Solving Tabular MDPs – Value Iteration

$$\pi(s) = \arg\max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$

```python
def greedy(P, R, gamma, v):
    """
    construct greedy policy by pi(s) = argmax_a q(s, a)
    """
    q = q_ftn(P, R, gamma, v)
    pi = np.argmax(q, axis=1)

    return pi
```

Combining all of these, we have...

Solving Tabular MDPs – Value Iteration

```python
def VI(P, R, gamma):
    """
    implementation of value iteration
    """
    EPS = 1e-6
    nS, nA = R.shape
    # initialize v
    v = np.zeros(shape=(nS, 1), dtype=np.float)

    while True:
        v_next = bellman_update(P, R, gamma, v)
        if np.linalg.norm(v_next - v, ord=np.inf) < EPS:
            break
        v = v_next

    pi = greedy(P, R, gamma, v)

    return v, pi
```

(terminal condition)

$$\max_s |v(s) - (\mathcal{T}v)(s)| \leq \epsilon$$

Solving Tabular MDPs – Policy Iteration

Review : any policy π satisfies

$$v^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s))v^{\pi}(s').$$

Step 1. compute v^{π} by solving the above equation (Policy Evaluation)

Step 2. determine π_{next} greedily (Policy Improvement):

$$\pi_{\text{next}}(s) = \arg\max_{a} \left(r(s, a) + \gamma \sum_{s'} p(s'|s, a)v^{\pi}(s') \right)$$

Solving Tabular MDPs – Policy Iteration

Step 1. compute v^π by solving the above equation (Policy Evaluation)

```python
def induced_dynamic(nS, P, R, pi):
    """
    given policy pi, compute induced dynamic P^pi & R^pi
    """
    S = range(nS)
    rows = np.arange(nS) + nS * pi
    P_pi = P[rows]
    R_pi = np.array([[R[s, pi[s]]] for s in range(nS)])

    return P_pi, R_pi
```

```python
def eval_policy(nS, P, R, gamma, pi):
    """
    policy evaluation
    """
    P_pi, R_pi = induced_dynamic(nS, P, R, pi)

    Id = np.identity(nS)

    # discounted reward problem
    v_pi = np.linalg.solve(Id - gamma * P_pi, R_pi)
    return v_pi
```

$$P^\pi = \begin{pmatrix} p(0|0,\pi(0)) & \cdots & p(n-1|0,\pi(0)) \\ \vdots & \vdots & \vdots \\ p(0|n-1,\pi(n-1)) & \cdots & p(n-1|n-1,\pi(n-1)) \end{pmatrix}$$

$$r^\pi = \begin{pmatrix} r(0,\pi(0)) \\ \vdots \\ r(n-1,\pi(n-1)) \end{pmatrix}$$

$$v^\pi = r^\pi + \gamma P^\pi v^\pi$$

$$\downarrow$$

$$(I - \gamma P^\pi) v^\pi = r^\pi$$

Control + Optimization Research Lab

Solving Tabular MDPs – Policy Iteration

Step 2. determine π_{next} greedily (Policy Improvement):

```python
def PI(P, R, gamma):
    """
    implementation of policy iteration
    """
    nS, nA = R.shape

    # initialize policy
    pi = np.random.randint(nA, size=nS)

    while True:
        v = eval_policy(nS, P, R, gamma, pi)
        pi_next = greedy(P, R, gamma, v)
        if (pi_next == pi).all():
            break
        pi = pi_next

    return v, pi
```

terminal condition : $\pi_{k+1} = \pi_k$

Example : GridWorld

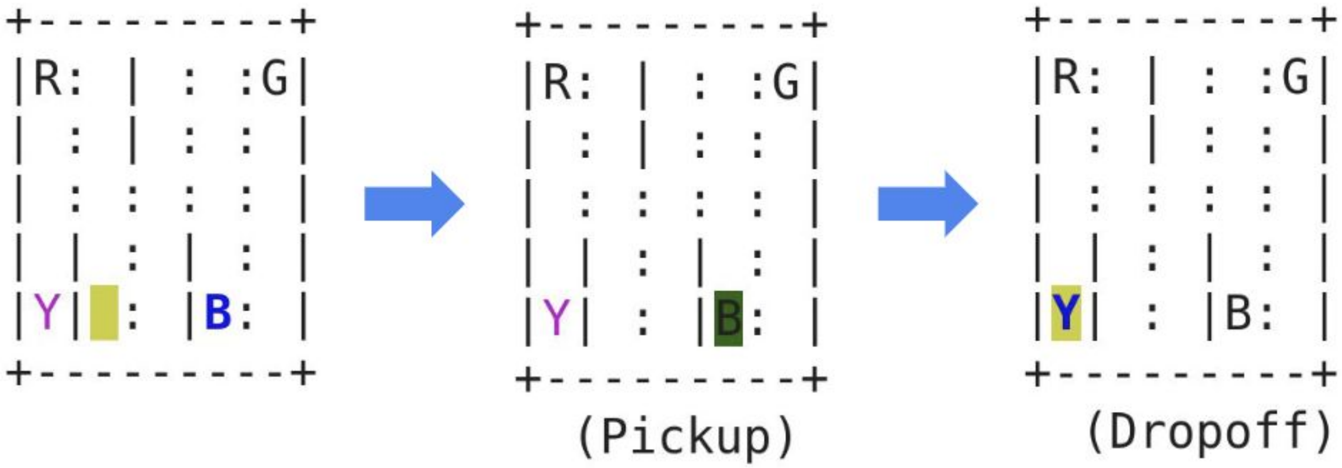
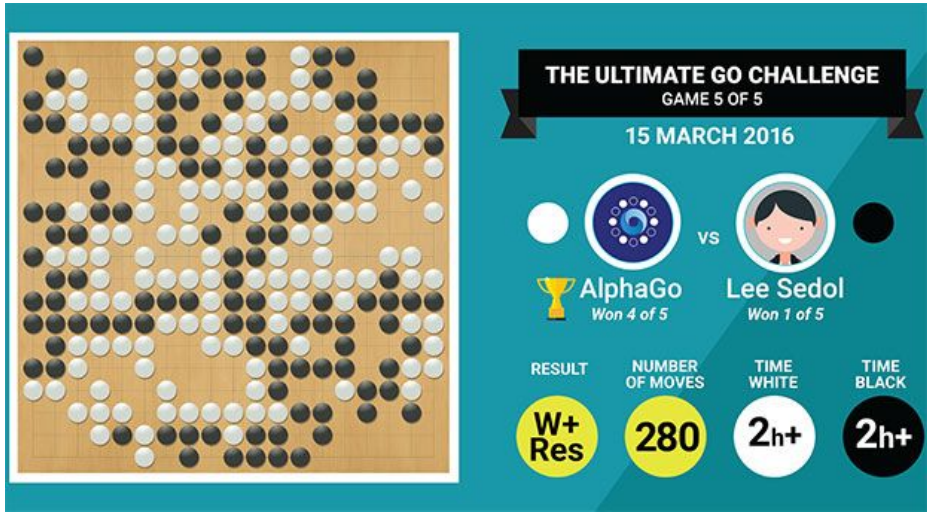


Figure 2: Illustration of the `Taxi-v3`. The box represents taxi, where blue and pink points (B and Y) denotes the passenger and the destination respectively. The problem is **solved** when the robot first moves to B to *pick up* the passenger, and then drops off the passenger at the destination Y with the shortest path.

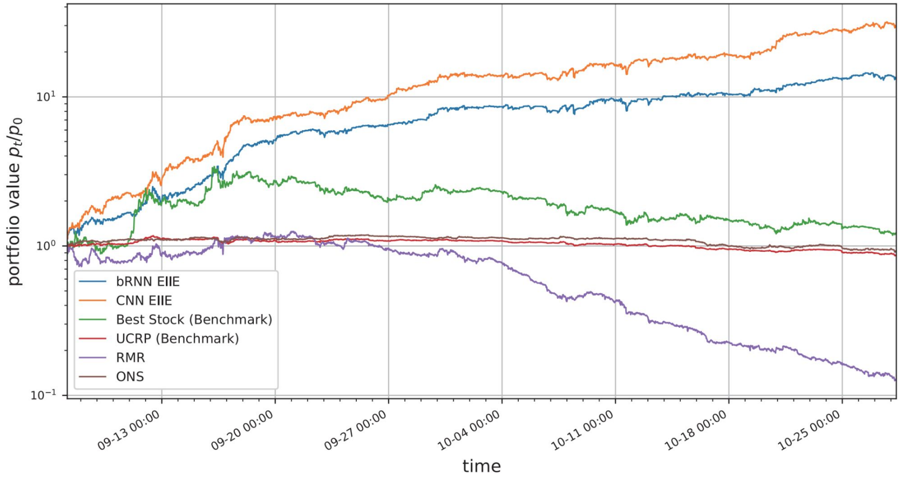
Problems?

Model if VERY difficult to obtain!

Game engine



Human behavior



Price model

- We will use data, instead of model

Thank you!