Maintenance Planning for Structures Using A Policy Gradient Approach

Zachary Hamida Postdoc
James-A. Goulet Professeur

Polytechnique Montréal, Canada
Département des génies civil, géologique et des mines

December 10, 2021

Funding:
Transportation Ministry of Quebec (MTQ)

Quick Recap

An element with inspections

Quick Recap

An element with inspections

Quick Recap

An element with inspections

An elements with inspections & repairs

Quick Recap

An element with inspections

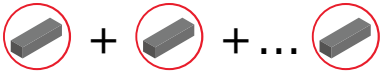
An elements with inspections & repairs

Quick Recap

An element with inspections

An elements with inspections & repairs

Category of elements

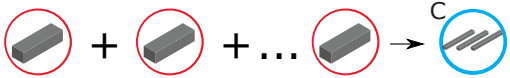


Quick Recap

An element with inspections

An elements with inspections & repairs

Category of elements

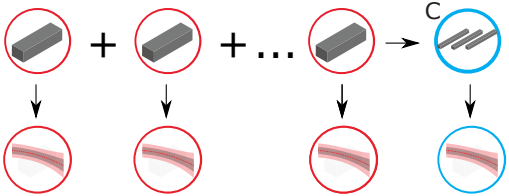


Quick Recap

An element with inspections

An elements with inspections & repairs

Category of elements

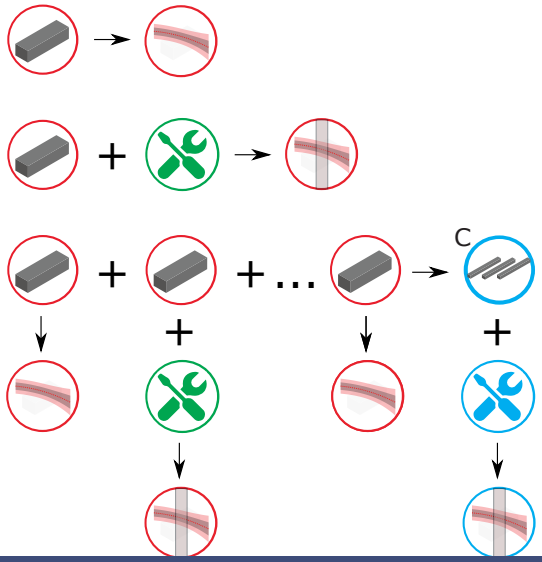


Quick Recap

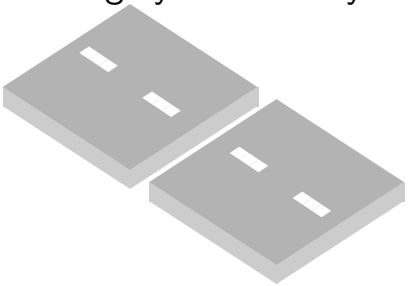
An element with inspections

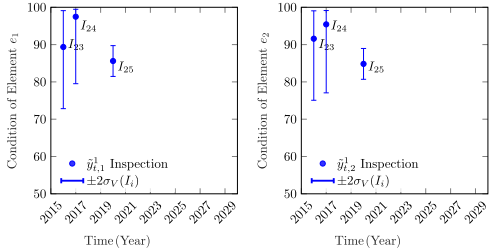
An elements with inspections & repairs

Category of elements

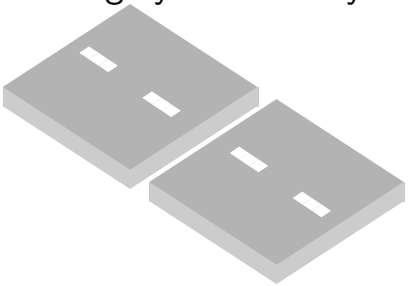


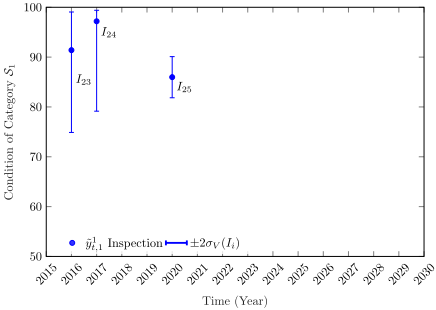
Structural Category-Level Analyses

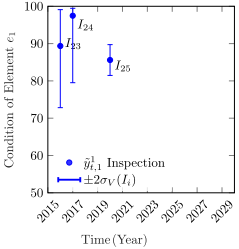


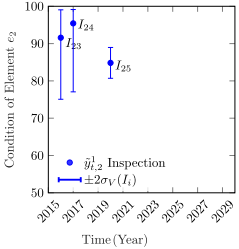


Structural Category-Level Analyses

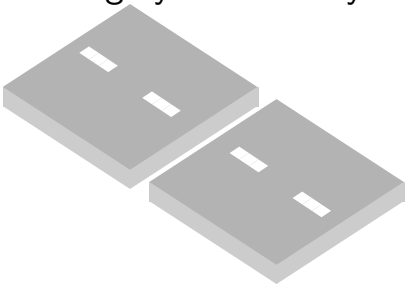


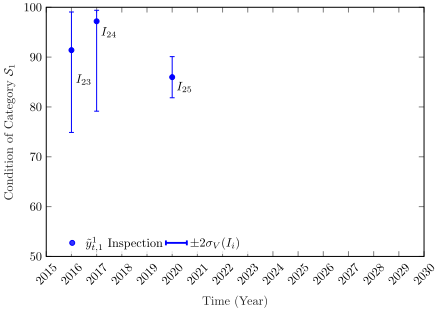


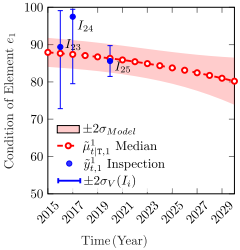


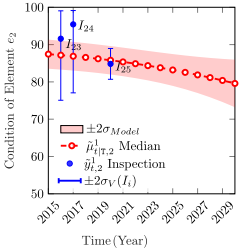


Structural Category-Level Analyses

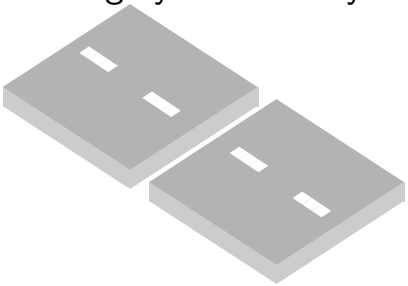


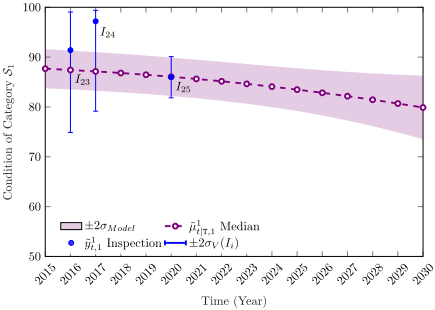


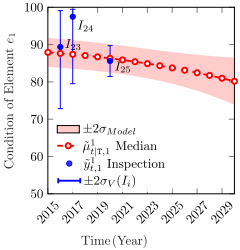


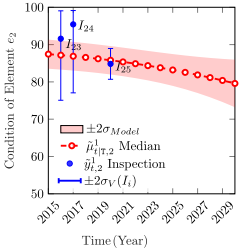


Structural Category-Level Analyses

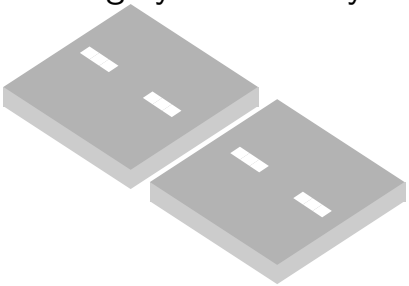


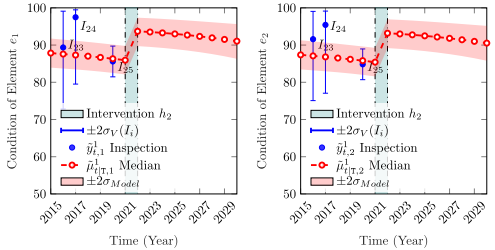




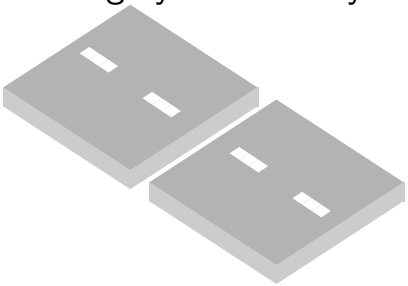


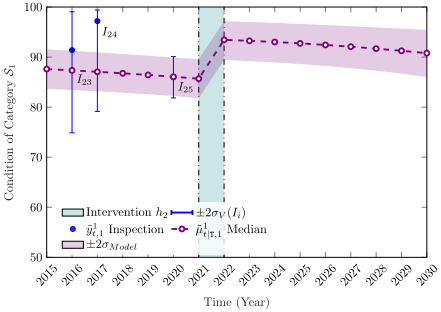
Structural Category-Level Analyses

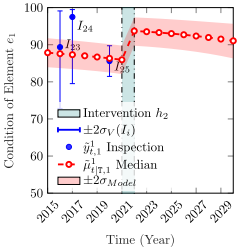


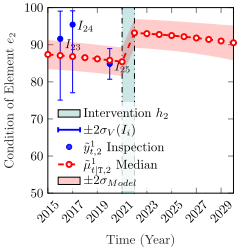


Structural Category-Level Analyses

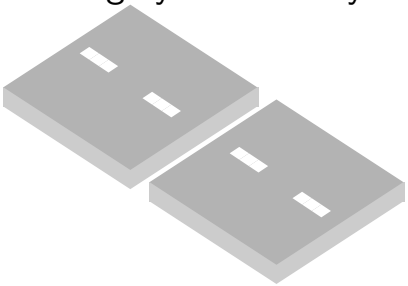


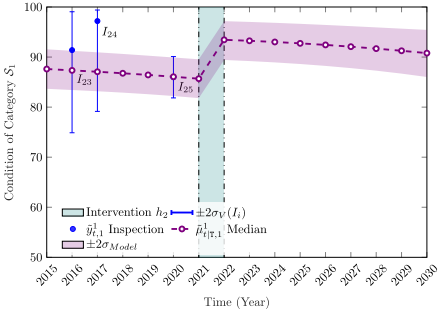


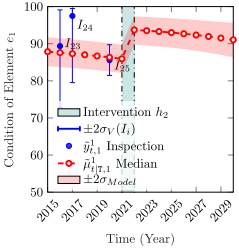


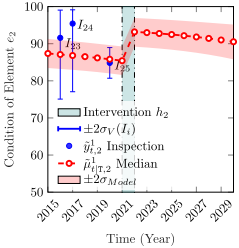


Structural Category-Level Analyses









$$\underbrace{\mu^{\delta}}_{\text{Structure}} = \frac{1}{2}\Big(\underbrace{\mu_1^{\delta}}_{\text{Element 1}} + \underbrace{\mu_2^{\delta}}_{\text{Element 2}}\Big)$$

Research Objective

Research Objective

Find a maintenance policy, which consists in,

Research Objective

Find a maintenance policy, which consists in,

- ▷ High-level policy: $\pi\left(\text{  }, \$\right) \rightarrow$ 

Research Objective

Find a maintenance policy, which consists in,

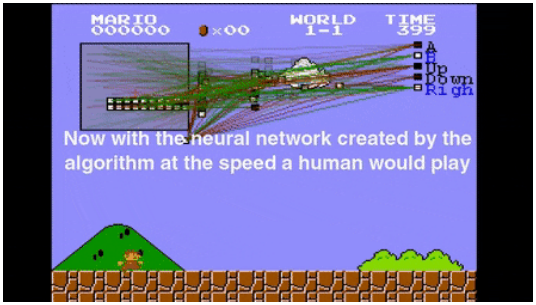
- ▷ High-level policy: $\pi \left(\text{  }, \$ \right) \rightarrow$ 
- ▷ Prioritization: $\rho \left(\text{  }^{1}, \text{  }^{2} \right) \rightarrow$  ^{2},  ^{1}

Research Objective

Find a maintenance policy, which consists in,

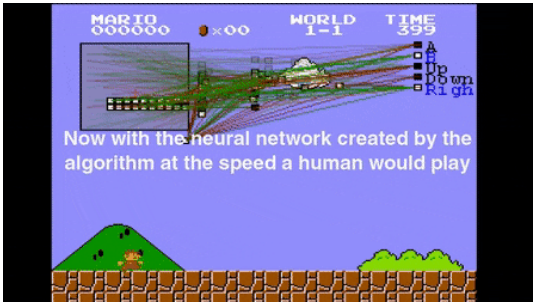
- ▷ High-level policy: $\pi \left(\text{  }, \$ \right) \rightarrow$ 
- ▷ Prioritization: $\rho \left(\text{  }^{1}, \text{  }^{2} \right) \rightarrow$  ^{2},  ^{1}
- ▷ Low-level policy: $\pi^{e} \left(\text{  } \right) \rightarrow$  , 

Reinforcement Learning



Source: Vansh Sethi

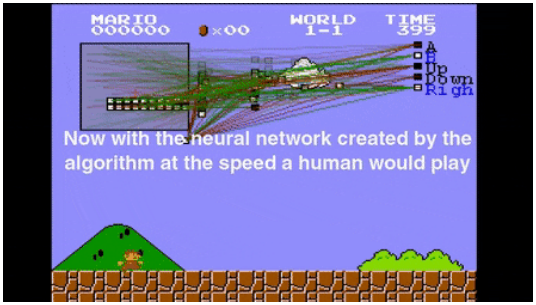
Reinforcement Learning



- ▷ Agent: decision maker.

Source: Vansh Sethi

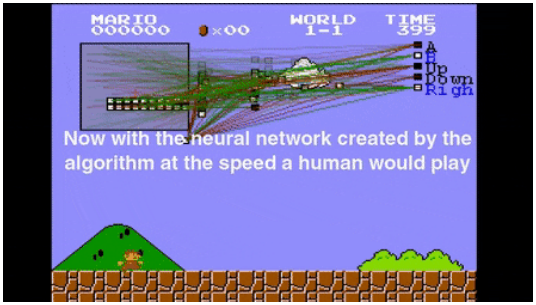
Reinforcement Learning



- ▷ Agent: decision maker.
- ▷ Environment: e.g., video game.

Source: Vansh Sethi

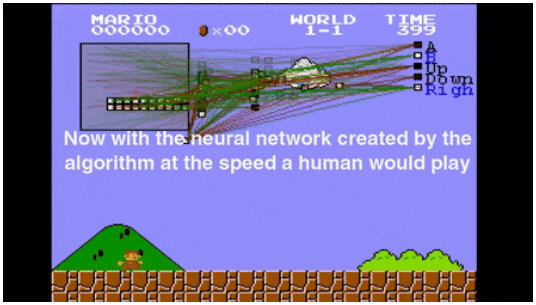
Reinforcement Learning



- ▷ Agent: decision maker.
- ▷ Environment: e.g., video game.
- ▷ Actions: e.g., key press or movement.

Source: Vansh Sethi

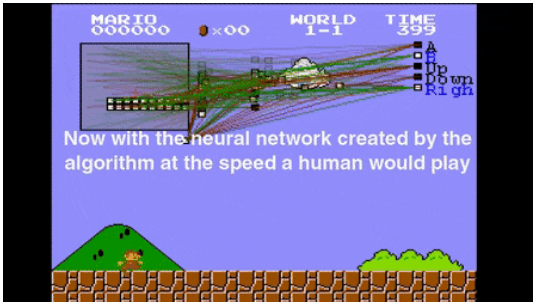
Reinforcement Learning



- ▷ Agent: decision maker.
- ▷ Environment: e.g., video game.
- ▷ Actions: e.g., key press or movement.
- ▷ Rewards: e.g., achieving the goal.

Source: Vansh Sethi

Reinforcement Learning



- ▷ Agent: decision maker.
- ▷ Environment: e.g., video game.
- ▷ Actions: e.g., key press or movement.
- ▷ Rewards: e.g., achieving the goal.

Learn a policy that maximizes the total expected discounted rewards.

Source: Vansh Sethi

The Return and Action-Value Function:

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$G_t = r_{t+1} + r_{t+2} + r_{t+2} \cdots + r_{\mathrm{T}},$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$G_t = \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{T-1} r_T, \text{ where, } \gamma \in [0, 1)$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{\mathrm{T}-1} r_{\mathrm{T}}, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{\mathrm{T}-2} r_{\mathrm{T}}) \end{aligned}$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{\mathrm{T}-1} r_{\mathrm{T}}, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{\mathrm{T}-2} r_{\mathrm{T}}) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{\mathrm{T}-1} r_{\mathrm{T}}, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{\mathrm{T}-2} r_{\mathrm{T}}) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The action-value function:

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{T-1} r_T, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{T-2} r_T) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The action-value function:

$$Q(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{\mathrm{T}-1} r_{\mathrm{T}}, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{\mathrm{T}-2} r_{\mathrm{T}}) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The action-value function:

$$Q(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

The optimal policy & Bellman optimality equation:

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{\mathrm{T}-1} r_{\mathrm{T}}, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{\mathrm{T}-2} r_{\mathrm{T}}) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The action-value function:

$$Q(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

The optimal policy & Bellman optimality equation:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\ Q^*(s, a),$$

The Return and Action-Value Function:

For a sequence of states s_t and rewards r_t, the return at time t:

$$\begin{aligned} G_t &= \gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+2} \cdots + \gamma^{T-1} r_T, \text{ where, } \gamma \in [0, 1) \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots + \gamma^{T-2} r_T) \\ &= r_{t+1} + \gamma G_{t+1} \end{aligned}$$

The action-value function:

$$Q(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

The optimal policy & Bellman optimality equation:

$$\pi^*(s) = \underset{a}{\text{argmax}}\ Q^*(s, a),$$
$$Q^*(s_t, a) = r(s_t, a) + \gamma \max_{a'} Q^*(s_{t+1}, a')$$

Incremental Averaging:

Incremental Averaging:

For a sequence of n samples, $y_1, y_2, \ldots, y_n$, the mean is computed as,

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

Incremental Averaging:

For a sequence of n samples, $y_1, y_2, \ldots, y_n$, the mean is computed as,

$$\begin{aligned}\mu_n &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\ &= \frac{1}{n}\left(y_n + \sum_{i=1}^{n-1} y_i\right),\end{aligned}$$

Incremental Averaging:

For a sequence of n samples, $y_1, y_2, \ldots, y_n$, the mean is computed as,

$$\begin{aligned}\mu_n &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\ &= \frac{1}{n}\left(y_n + \sum_{i=1}^{n-1} y_i\right), \text{by considering: } \left(\sum_{i=1}^{n-1} y_i = (n-1)\mu_{n-1}\right)\end{aligned}$$

Incremental Averaging:

For a sequence of n samples, $y_1, y_2, \ldots, y_n$, the mean is computed as,

$$\begin{aligned}
\mu_n &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\
&= \frac{1}{n}\left(y_n + \sum_{i=1}^{n-1} y_i\right), \text{by considering: } \left(\sum_{i=1}^{n-1} y_i = (n-1)\mu_{n-1}\right) \\
&= \frac{1}{n}\left(y_n + n\mu_{n-1} - \mu_{n-1}\right)
\end{aligned}$$

Incremental Averaging:

For a sequence of n samples, $y_1, y_2, \ldots, y_n$, the mean is computed as,

$$\begin{aligned}
\mu_n &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\
&= \frac{1}{n}\left(y_n + \sum_{i=1}^{n-1} y_i\right), \text{by considering: } \left(\sum_{i=1}^{n-1} y_i = (n-1)\mu_{n-1}\right) \\
&= \frac{1}{n}\left(y_n + n\mu_{n-1} - \mu_{n-1}\right) \\
\mu_n &= \mu_{n-1} + \frac{1}{n}\left(y_n - \mu_{n-1}\right)
\end{aligned}$$

Q-Value Update:

From the incremental averaging equation,

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$
$$\mu_n = \mu_{n-1} + \alpha(y_n - \mu_{n-1})$$

Q-Value Update:

From the incremental averaging equation,

$$\begin{aligned}\mu_n &= \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1}) \\ &= Q(s_t, a) + \alpha(y_n - Q(s_t, a))\end{aligned}$$

Q-Value Update:

From the incremental averaging equation,

$$\begin{aligned}\mu_n &= \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1}) \\ &= Q(s_t, a) + \alpha\left(Q^*(s_t, a) - Q(s_t, a)\right)\end{aligned}$$

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

$$Q(s_t, a) = Q(s_t, a) + \alpha\left(Q^*(s_t, a) - Q(s_t, a)\right)$$

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

$$Q(s_t, a) = Q(s_t, a) + \alpha\left(Q^*(s_t, a) - Q(s_t, a)\right)$$

$$Q(s_t, a) = Q(s_t, a) + \alpha(\underbrace{r(s_t, a) + \gamma \max_{a'} Q(s_{t+1}, a')}_{\text{Target}} - Q(s_t, a))$$

Q-Value Update:

From the incremental averaging equation,

$$\mu_n = \mu_{n-1} + \frac{1}{n}(y_n - \mu_{n-1})$$

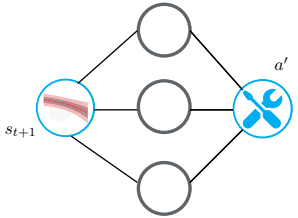
$$Q(s_t, a) = Q(s_t, a) + \alpha\left(Q^*(s_t, a) - Q(s_t, a)\right)$$

$$Q(s_t, a) = Q(s_t, a) + \alpha(\underbrace{r(s_t, a) + \gamma \max_{a'} Q(s_{t+1}, a')}_{\text{Target}} - Q(s_t, a))$$

If a is continuous, then estimating $\max_{a'} Q(s_{t+1}, a')$ is challenging.

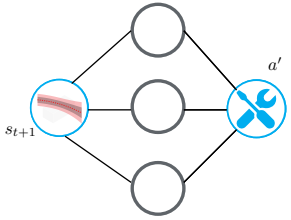
Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

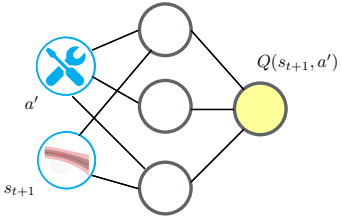


Actor Target

Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

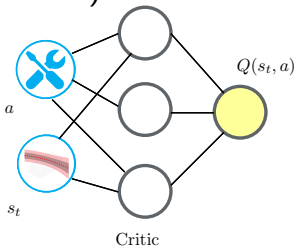


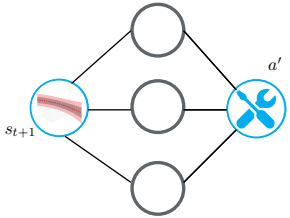
Actor Target



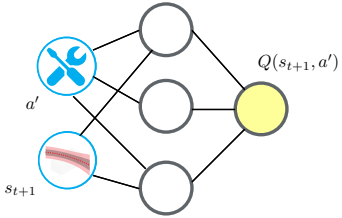
Critic Target

Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:



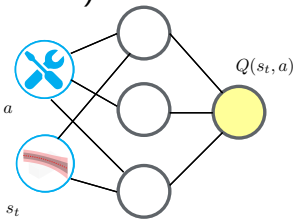


Actor Target



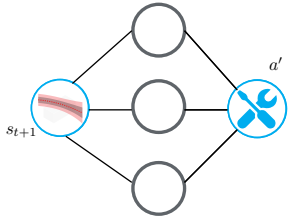
Critic Target

Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

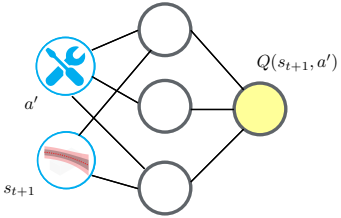


Critic

$$\text{Critic Loss} = \text{MSE}(r + \gamma Q(s_{t+1}, a') - Q(s_t, a))$$

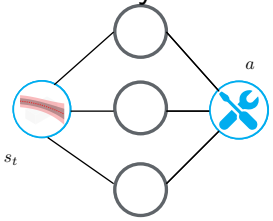


Actor Target

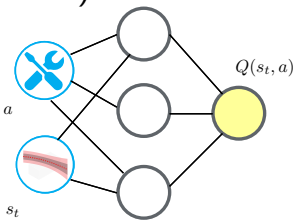


Critic Target

Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

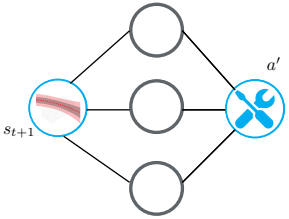


Actor

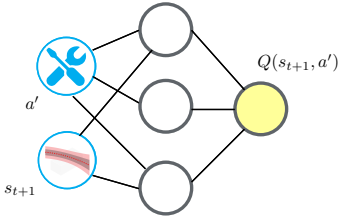


Critic

$$\text{Critic Loss} = \text{MSE}(r + \gamma Q(s_{t+1}, a') - Q(s_t, a))$$

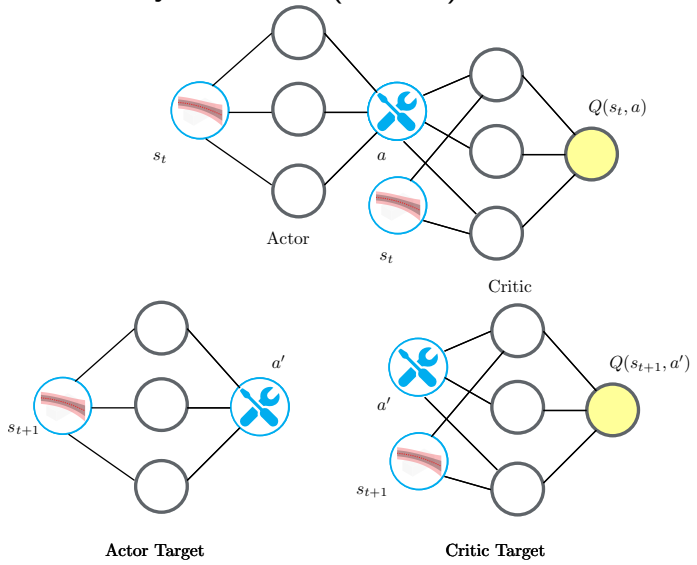


Actor Target

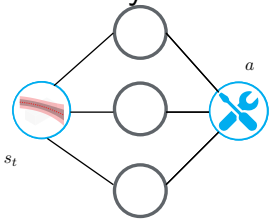


Critic Target

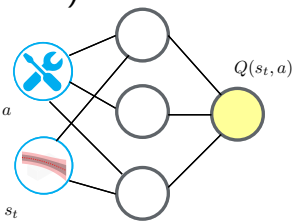
Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:



Deep Deterministic Policy Gradient (DDPG) and Actor-Critic Method:

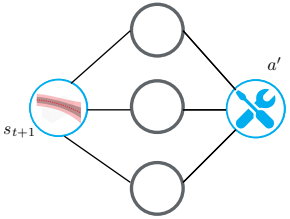


Actor

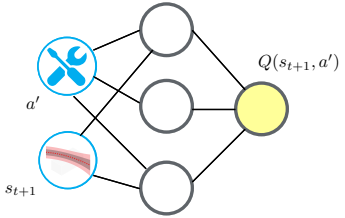


Critic

Delayed soft update for the target model parameters



Actor Target



Critic Target

Next Steps & Current Issues

- Design the reward functions.

Next Steps & Current Issues

- ▷ Design the reward functions.
- ▷ Improve the sampling approach and/or the experience replay itself.

Next Steps & Current Issues

- Design the reward functions.
- Improve the sampling approach and/or the experience replay itself.
- Account for the budget in the framework.