On the use of non-stationary policies for stationary optimal control problem

(An introduction to Reinforcement Learning / Optimal control)

Bruno Scherrer

INRIA (Institut National de Recherche en Informatique et ses Applications)
IECL (Institut Elie Cartan de Lorraine)
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

$$M = 20, \quad f(x) = x, \quad g(x) = 0.25x, \quad h(x) = 0.25x, \quad C(a) = (1 + 0.5a)1_{a>0}, \quad w_t \sim \text{Normal}(0, \sigma^2).$$

- $t = 0, 1, \ldots, 11, \ H = 12$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(x_t + a_t - w_t, 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$ and $R(x) = g(x)$. 
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

$M = 20, \ f(x) = x, \ g(x) = 0.25x, \ h(x) = 0.25x, \ C(a) = (1 + 0.5a)\mathbb{1}_{a > 0}, \ w_t \sim \ldots$

- $t = 0, 1, \ldots, 11, \ H = 12$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max( x_t + a_t - w_t, 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$
  and $R(x) = g(x)$.
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

$M = 20$, $f(x) = x$, $g(x) = 0.25x$, $h(x) = 0.25x$, $C(a) = (1 + 0.5a)\mathbb{1}_{a > 0}$, $w_t \sim \mathcal{I}$

- $t = 0, 1, \ldots, 11$, $H = 12$
- **State space:** $x \in X = \{0, 1, \ldots, M\}$
- **Action space:** At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- **Dynamics:** $x_{t+1} = \max(x_t + a_t - w_t, 0)$
- **Reward:** $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$ and $R(x) = g(x)$. 
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

$$M = 20, \quad f(x) = x, \quad g(x) = 0.25x, \quad h(x) = 0.25x, \quad C(a) = (1 + 0.5a)I_{a>0}, \quad w_t \sim \ldots$$

- $t = 0, 1, \ldots, 11, \quad H = 12$
- State space: $x \in \mathcal{X} = \{0, 1, \ldots, M\}$
- Action space: At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(x_t + a_t - w_t, 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$
  and $R(x) = g(x)$. 
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

\[ M = 20, \quad f(x) = x, \quad g(x) = 0.25x, \quad h(x) = 0.25x, \quad C(a) = (1 + 0.5a) 1_{a > 0}, \quad w_t \sim \text{uniform} \]

- $t = 0, 1, \ldots, 11, \quad H = 12$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(x_t + a_t - w_t, 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$ and $R(x) = g(x)$. 
Example: The Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the beginning of each month $t$, the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$.

$M = 20, f(x) = x, g(x) = 0.25x, h(x) = 0.25x, C(a) = (1 + 0.5a)_{a > 0}, w_t \sim$ 

- $t = 0, 1, \ldots, 11, H = 12$
- **State space:** $x \in X = \{0, 1, \ldots, M\}$
- **Action space:** At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- **Dynamics:** $x_{t+1} = \max( x_t + a_t - w_t, 0)$
- **Reward:** $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$
  and $R(x) = g(x)$. 
Example: The Retail Store Management Problem

2 stationary policies and 1 non-stationary policy:

\[\pi^{(2)}(x) = \max\{(M-x)/2-x; 0\}\]

\[\pi^{(1)}(x) = \begin{cases} M - x & \text{if } x < M/4 \\ 0 & \text{otherwise} \end{cases}\]

\[\pi_t^{(3)}(x) = \begin{cases} M - x & \text{if } t < 6 \\ \lfloor(M - x)/5\rfloor & \text{otherwise} \end{cases}\]
Policy evaluation

\[ v_{\pi,s}(x) = \mathbb{E}_{\pi} \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}_{\pi} [r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_{\pi} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E} [r_s(x, \pi(x), w_s)] \]

\[ + \sum_y P(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) \mathbb{E}_{\pi} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y \right] \]

\[ = \mathbb{E} [r_s(x, \pi(x), w_s)] + \sum_y P(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) v_{\pi,s+1}(y). \]
Policy evaluation

\[ v_{\pi,s}(x) = \mathbb{E}_\pi \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}_\pi [r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) v_{\pi,s+1}(y). \]
Policy evaluation

\[ v_{\pi, s}(x) = \mathbb{E}_\pi \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}_\pi [r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) v_{\pi, s+1}(y). \]
Policy evaluation

\[ v_{\pi,s}(x) = \mathbb{E}_\pi \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}_\pi [r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) \mathbb{E}_\pi \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) v_{\pi,s+1}(y). \]
Policy evaluation

\[ v_{\pi,s}(x) = \mathbb{E}_{\pi} \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}_{\pi}[r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_{\pi} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) \mathbb{E}_{\pi} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y \right] \]

\[ = \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi(x_s)) v_{\pi,s+1}(y). \]
Example: the Retail Store Management Problem
Policy optimization

\[ v_{\pi,s}(x) = \max_{\pi_s,\ldots} \mathbb{E}_{\pi_s,\ldots} \left\{ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \right\} = \max_{\pi_s,\ldots} \mathbb{E}_{\pi_s,\ldots} \left\{ r_s(x_s, a_s, w_s) \right. \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = \pi_s(x_s)) \left( \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \right) \bigg| x_s = x, x_{s+1} = y \bigg\} \]

\[ = \max_{\pi_s,\ldots} \mathbb{E}_{\pi_s,\ldots} \left\{ r_s(x, a, w_s) \right\} \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = a) \max_{\pi_{s+1},\ldots} \mathbb{E}_{\pi_{s+1},\ldots} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \bigg| x_{s+1} = y \right] \}

\[ = \max_{a} \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = a) v_{\pi,s+1}(y) \right\} . \]
Policy optimization

\[ v_{\pi_s}(x) = \max_{s, \pi_s, \ldots} \left\{ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \middle| x_s = x \right\} \]

\[ = \max_{\pi_s, \pi_{s+1}, \ldots} \left\{ r_s(x_s, a_s, w_s) + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = \pi_s(x_s)) \left( \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \right) \middle| x_s = x, x_{s+1} = y \right\} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = a) \max_{\pi_{s+1}, \ldots} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \middle| x_{s+1} = y \right] \right\} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y|x_s = x, a_s = a) v_{\pi_s}(y) \right\}. \]
Policy optimization

\[ v_{*,s}(x) = \max_{\pi_s, \ldots} \mathbb{E}_{\pi_s, \ldots} \left\{ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right\} \]

\[ = \max_{\pi_s, \pi_{s+1}, \ldots} \mathbb{E}_{\pi_s, \pi_{s+1}, \ldots} \left\{ r_s(x_s, a_s, w_s) \right. \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi_s(x_s)) \left( \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \right) \bigg| x_s = x, x_{s+1} = y \left. \right\} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] \right. \]

\[ + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = a) \max_{\pi_{s+1}, \ldots} \mathbb{E}_{\pi_{s+1}, \ldots} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_{s+1} = y \right] \left. \right\} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = a) v_{*,s+1}(y) \right\}. \]
Policy optimization

\[ v_{*,s}(x) = \max_{\pi_s,\ldots} \{ \mathbb{E}_{\pi_s,\ldots} \left[ \sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x \right] \} \]

\[ = \max_{\pi_s,\pi_{s+1},\ldots} \{ \mathbb{E}_{\pi_s,\pi_{s+1},\ldots} \left[ r_s(x_s, a_s, w_s) + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = \pi_s(x_s)) \left( \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \right) \right] \mid x_s = x, x_{s+1} = y \} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = a) \max_{\pi_{s+1},\ldots} \left[ \mathbb{E}_{\pi_{s+1},\ldots} \left[ \sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_{s+1} = y \right] \right] \right\} \]

\[ = \max_a \left\{ \mathbb{E}[r_s(x, a, w_s)] + \sum_y \mathbb{P}(x_{s+1} = y \mid x_s = x, a_s = a) v_{*,s+1}(y) \right\}. \]
Example: the Retail Store Management Problem

Optimal value and policy vs values of policies $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}$
Example: the Retail Store Management Problem

Each month $t$, a store contains $x_t$ items (maximum capacity $M$) of a specific goods and the demand for that goods is $w_t$. At the end of each month the manager of the store can order $a_t$ more items from his supplier. The cost of maintaining an inventory of $x$ is $h(x)$. The cost to order $a$ items is $C(a)$. The income for selling $q$ items is $f(q)$. If the demand $w$ is bigger than the available inventory $x$, customers that cannot be served leave. The value of the remaining inventory at the end of the year is $g(x)$. The rate of inflation is $\alpha = 3\% = 0.03$.

$M = 20$, $f(x) = x$, $g(x) = 0.25x$, $h(x) = 0.25x$, $C(a) = (1 + 0.5a)1_{a>0}$, $w_t \sim U(\{5, 6, \ldots, 15\})$, $\gamma = \frac{1}{1+\alpha}$

- $t = 0, 1, \ldots$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state $x$, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(x_t + a_t - w_t, 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$. 
Example: the Retail Store Management Problem
Example: the Optimal Replacement Problem

**State:** level of wear \((x)\) of an object (e.g., a car).

**Action:** \(\{(R)eplace, (K)eep\}\).

**Cost:**
- \(c(x, R) = C\)
- \(c(x, K) = c(x)\) maintenance plus extra costs.

**Dynamics:**
- \(p(y|x, R) \sim d(y) = \beta \exp^{-\beta y} \mathbb{1}\{y \geq 0\}\),
- \(p(y|x, K) \sim d(y - x) = \beta \exp^{-\beta(y-x)} \mathbb{1}\{y \geq x\}\).

**Problem:** Minimize the discounted expected cost over an infinite horizon.
Example: the Optimal Replacement Problem

**State:** level of wear \((x)\) of an object (e.g., a car).

**Action:** \{**(R)eplace, (K)eep**\}.

**Cost:**

- \(c(x, R) = C\)
- \(c(x, K) = c(x)\) maintenance plus extra costs.

**Dynamics:**

- \(p(y|x, R) \sim d(y) = \beta \exp^{-\beta y} 1\{y \geq 0\}\),
- \(p(y|x, K) \sim d(y - x) = \beta \exp^{-\beta(y-x)} 1\{y \geq x\}\).

**Problem:** Minimize the discounted expected cost over an infinite horizon.
Example: the Optimal Replacement Problem

**State:** level of wear \((x)\) of an object (e.g., a car).

**Action:** \(\{(R)eplace, (K)eep\}\).

**Cost:**

- \(c(x, R) = C\)
- \(c(x, K) = c(x)\) maintenance plus extra costs.

**Dynamics:**

- \(p(y|x, R) \sim d(y) = \beta \exp^{-\beta y} 1\{y \geq 0\}\),
- \(p(y|x, K) \sim d(y - x) = \beta \exp^{-\beta(y-x)} 1\{y \geq x\}\).

**Problem:** Minimize the discounted expected cost over an infinite horizon.
Example: the Optimal Replacement Problem

**State:** level of wear \((x)\) of an object (e.g., a car).

**Action:** \(\{(R)eplace, (K)eep\}\).

**Cost:**

- \(c(x, R) = C\)
- \(c(x, K) = c(x)\) maintenance plus extra costs.

**Dynamics:**

- \(p(y|x, R) \sim d(y) = \beta \exp^{-\beta y} 1\{y \geq 0\}\),
- \(p(y|x, K) \sim d(y - x) = \beta \exp^{-\beta(y-x)} 1\{y \geq x\}\).

**Problem:** Minimize the discounted expected cost over an infinite horizon.
Example: the Optimal Replacement Problem

**State:** level of wear \((x)\) of an object (e.g., a car).

**Action:** \{\text{(R)eplace}, \text{(K)eep}\}.

**Cost:**
- \(c(x, R) = C\)
- \(c(x, K) = c(x)\) maintenance plus extra costs.

**Dynamics:**
- \(p(y|x, R) \sim d(y) = \beta \exp^{-\beta y} \mathbb{1}\{y \geq 0\}\),
- \(p(y|x, K) \sim d(y - x) = \beta \exp^{-\beta(y-x)} \mathbb{1}\{y \geq x\}\).

**Problem:** Minimize the discounted expected cost over an infinite horizon.
Example: the Optimal Replacement Problem

The optimal value function satisfies

\[ v^*(x) = \min \left\{ c(x) + \gamma \int_0^\infty d(y - x)v^*(y)dy, \ C + \gamma \int_0^\infty d(y)v^*(y)dy \right\} \]

(K)eeep \quad (R)eplace

Optimal policy: action that attains the minimum
Example: the Optimal Replacement Problem

Linear approximation space

\[ \mathcal{F} := \left\{ v_n(x) = \sum_{k=0}^{19} \alpha_k \cos(k\pi \frac{x}{x_{\text{max}}}) \right\}. \]

Collect \( N \) samples on a uniform grid:

**Figure:** Left: the target values computed as \( \{ T v_0(x_n) \}_{1 \leq n \leq N} \). Right: the approximation \( v_1 \in \mathcal{F} \) of the target function \( T v_0 \).
Example: the Optimal Replacement Problem

Linear approximation space

\[ \mathcal{F} := \left\{ v_n(x) = \sum_{k=0}^{19} \alpha_k \cos \left( k \pi \frac{x}{x_{\text{max}}} \right) \right\}. \]

Collect \( N \) samples on a uniform grid:

**Figure:** Left: the target values computed as \( \{ T v_0(x_n) \}_{1 \leq n \leq N} \). Right: the approximation \( v_1 \in \mathcal{F} \) of the target function \( T v_0 \).
Example: the Optimal Replacement Problem

One more step:

Figure: Left: the target values computed as \( \{ T v_1(x_n) \}_{1 \leq n \leq N} \). Right: the approximation \( v_2 \in F \) of \( T v_1 \).
Example: the Optimal Replacement Problem

**Figure:** The approximation $v_{20} \in \mathcal{F}$.
Error propagation for AVI

1 Bounding: $\|v_* - v_k\|_{\infty}$:

\[
\|v_* - v_k\|_{\infty} = \|v_* - Tv_k - \varepsilon_k\|_{\infty} \\
\leq \|Tv_* - Tv_{k-1}\|_{\infty} + \varepsilon \\
\leq \gamma \|v_* - v_{k-1}\|_{\infty} + \varepsilon \\
\leq \frac{\varepsilon}{1 - \gamma}.
\]

2 From $\|v_* - v_k\|_{\infty}$ to $\|v_* - v_{\pi k+1}\|_{\infty}$ ($\pi_{k+1} = G v_k$):

\[
\|v_* - v_{\pi k+1}\|_{\infty} \leq \|Tv_* - T_{\pi_{k+1}} v_k\|_{\infty} + \|T_{\pi_{k+1}} v_k - T_{\pi_{k+1}} v_{\pi_{k+1}}\|_{\infty} \\
\leq \|Tv_* - Tv_k\|_{\infty} + \gamma \|v_k - v_{\pi_{k+1}}\|_{\infty} \\
\leq \gamma \|v_* - v_k\|_{\infty} + \gamma \left(\|v_k - v_*\|_{\infty} + \|v_* - v_{\pi_{k+1}}\|_{\infty}\right) \\
\leq \frac{2\gamma}{1 - \gamma} \|v_* - v_k\|_{\infty}.
\]
Error propagation for AVI

1 Bounding: \( \|v_* - v_k\|_\infty \):

\[
\|v_* - v_k\|_\infty = \|v_* - Tv_{k-1} - \epsilon_k\|_\infty \\
\leq \|Tv_* - Tv_{k-1}\|_\infty + \epsilon \\
\leq \gamma \|v_* - v_{k-1}\|_\infty + \epsilon \\
\leq \frac{\epsilon}{1 - \gamma}.
\]

2 From \( \|v_* - v_k\|_\infty \) to \( \|v_* - v_{\pi_{k+1}}\|_\infty \) \((\pi_{k+1} = Gv_k)\):

\[
\|v_* - v_{\pi_{k+1}}\|_\infty \leq \|Tv_* - Tv_{\pi_{k+1}}v_k\|_\infty + \|Tv_{\pi_{k+1}}v_k - Tv_{\pi_{k+1}}v_{\pi_{k+1}}\|_\infty \\
\leq \|Tv_* - Tv_k\|_\infty + \gamma \|v_k - v_{\pi_{k+1}}\|_\infty \\
\leq \gamma \|v_* - v_k\|_\infty + \gamma \left( \|v_k - v_*\|_\infty + \|v_* - v_{\pi_{k+1}}\|_\infty \right) \\
\leq \frac{2\gamma}{1 - \gamma} \|v_* - v_k\|_\infty.
\]
Tightness of the bound for AVI

State 2: $0 + \gamma (-\epsilon) = -2\gamma \epsilon + \gamma \epsilon$
State 3: $0 + \gamma (-\epsilon - \gamma \epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma (\epsilon + \gamma \epsilon)$

$$v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1-\gamma} \epsilon\right) = -2 \frac{\gamma - \gamma^k}{(1-\gamma)^2} \epsilon \xrightarrow{k\to\infty} -\frac{2\gamma}{(1-\gamma)^2} \epsilon$$
Tightness of the bound for AVI

\[
\begin{align*}
\pi_k(k) &= \sum_{t=0}^{\infty} \gamma^t \left( -2 \frac{\gamma - \gamma^k}{1 - \gamma} \epsilon \right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \\
&\quad \xrightarrow{k \to \infty} -2 \frac{2\gamma}{(1 - \gamma)^2} \epsilon
\end{align*}
\]

<table>
<thead>
<tr>
<th>State</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>(v_0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>(v_1)</td>
<td>-(\epsilon)</td>
<td>(\epsilon)</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>(v_2)</td>
<td>-(\gamma\epsilon)</td>
<td>-(\epsilon - \gamma\epsilon)</td>
<td>(\epsilon + \gamma\epsilon)</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>(v_3)</td>
<td>-(\gamma^2\epsilon)</td>
<td>-(\gamma^2\epsilon)</td>
<td>-(\epsilon - \gamma\epsilon - \gamma^2\epsilon)</td>
<td>(\epsilon + \gamma\epsilon + \gamma^2\epsilon)</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

State 2: \(0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon\)
State 3: \(0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)\)
Tightness of the bound for AVI

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & \ldots & k \\
\downarrow & \downarrow & \downarrow & \downarrow & & \downarrow \\
0 & -2\gamma\epsilon & -2(\gamma + \gamma^2)\epsilon & -2(\gamma + \gamma^2 + \gamma^3)\epsilon & & -2\frac{\gamma - \gamma^k}{1 - \gamma} \epsilon \\
\end{array}
\]

\[
\begin{array}{cccccc}
\begin{array}{cccc}
\vdots & \vdots & \vdots & \vdots \\
v_0 & v_1 & v_2 & v_3 \\
0 & -\epsilon & -\gamma\epsilon & -\gamma^2\epsilon \\
0 & \epsilon & -\epsilon - \gamma\epsilon & -\gamma^2\epsilon \\
0 & 0 & \epsilon + \gamma\epsilon & -\epsilon - \gamma\epsilon - \gamma^2\epsilon \\
0 & 0 & 0 & \epsilon + \gamma\epsilon + \gamma^2\epsilon \\
\end{array}
\end{array}
\]

State 2: \(0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon\)

State 3: \(0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)\)

\[
v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1 - \gamma} \epsilon\right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \quad k \to \infty \quad -\frac{2\gamma}{(1 - \gamma)^2} \epsilon
\]
Tightness of the bound for AVI

\[ \nu_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left( -2 \frac{\gamma - \gamma^k}{1 - \gamma} \epsilon \right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \quad k \to \infty \to -2 \frac{2\gamma}{(1 - \gamma)^2} \epsilon \]
Tightness of the bound for AVI

\[
\begin{array}{cccccc}
0 & -2\gamma\epsilon & -2(\gamma + \gamma^2)\epsilon & -2(\gamma + \gamma^2 + \gamma^3)\epsilon & \ldots & -2\frac{\gamma - \gamma^k}{1 - \gamma}\epsilon \\
1 & 2 & 3 & 4 & \ldots & k
\end{array}
\]

<table>
<thead>
<tr>
<th>(v_0)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
</tr>
<tr>
<td>(v_1)</td>
<td>(-\epsilon)</td>
<td>(\epsilon)</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
</tr>
<tr>
<td>(v_2)</td>
<td>(-\gamma\epsilon)</td>
<td>(-\epsilon - \gamma\epsilon)</td>
<td>(\epsilon + \gamma\epsilon)</td>
<td>0</td>
<td>\ldots</td>
</tr>
<tr>
<td>(v_3)</td>
<td>(-\gamma^2\epsilon)</td>
<td>(-\gamma^2\epsilon)</td>
<td>(-\epsilon - \gamma\epsilon - \gamma^2\epsilon)</td>
<td>(\epsilon + \gamma\epsilon + \gamma^2\epsilon)</td>
<td>\ldots</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>

State 2: \(0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon\)

State 3: \(0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)\)

\[
v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1 - \gamma}\epsilon\right) = -2\frac{\gamma - \gamma^k}{1 - \gamma}^2 \epsilon \quad k \to \infty \quad -\frac{2\gamma}{(1 - \gamma)^2} \epsilon
\]
Tightness of the bound for AVI

\[
\begin{array}{cccccc}
\text{State 2: } 0 + \gamma(-\epsilon) &=& -2\gamma\epsilon + \gamma\epsilon \\
\text{State 3: } 0 + \gamma(-\epsilon - \gamma\epsilon) &=& -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)
\end{array}
\]

\[
\nu_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2 \frac{\gamma - \gamma^k}{1-\gamma}\epsilon\right) = -2 \frac{\gamma - \gamma^k}{(1-\gamma)^2}\epsilon \quad k \to \infty \to -2 \frac{2\gamma}{(1-\gamma)^2}\epsilon
\]
### Tightness of the bound for AVI

#### State Transition Diagram

![State Transition Diagram](image)

#### Transition Probabilities

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu_0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>$\nu_1$</td>
<td>$-\epsilon$</td>
<td>$\epsilon$</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>$\nu_2$</td>
<td>$-\gamma \epsilon$</td>
<td>$-\epsilon - \gamma \epsilon$</td>
<td>$\epsilon + \gamma \epsilon$</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>$\nu_3$</td>
<td>$-\gamma^2 \epsilon$</td>
<td>$-\gamma^2 \epsilon$</td>
<td>$-\epsilon - \gamma \epsilon - \gamma^2 \epsilon$</td>
<td>$\epsilon + \gamma \epsilon + \gamma^2 \epsilon$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**State 2:** $0 + \gamma(-\epsilon) = -2\gamma \epsilon + \gamma \epsilon$

**State 3:** $0 + \gamma(-\epsilon - \gamma \epsilon) = -2(\gamma + \gamma^2) \epsilon + \gamma(\epsilon + \gamma \epsilon)$

#### Value Function

$$
\nu_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1 - \gamma} \epsilon\right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \quad \overset{k \to \infty}{\longrightarrow} \quad -\frac{2\gamma}{(1 - \gamma)^2} \epsilon
$$
Tightness of the bound for AVI

\[ \pi_k(k) = \sum_{t=0}^{\infty} \gamma^t \left( -2 \frac{\gamma - \gamma^k}{1 - \gamma} \epsilon \right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \rightarrow_{k \to \infty} -2 \frac{2\gamma}{(1 - \gamma)^2} \epsilon \]

State 2: \( 0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon \)
State 3: \( 0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon) \)
Tightness of the bound for AVI

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & \cdots & k \\
0 & -2\gamma\epsilon & -2(\gamma + \gamma^2)\epsilon & -2(\gamma + \gamma^2 + \gamma^3)\epsilon & \cdots & -2\frac{\gamma - \gamma^k}{1-\gamma}\epsilon
\end{array}
\]

<table>
<thead>
<tr>
<th>(v_0)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>\cdots</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>\cdots</td>
</tr>
</tbody>
</table>

| \(v_1\) | \(-\epsilon\) | \(\epsilon\) | 0 | 0 | \cdots |
| \(v_2\) | \(-\gamma\epsilon\) | \(-\epsilon - \gamma\epsilon\) | \(\epsilon + \gamma\epsilon\) | 0 | \cdots |
| \(v_3\) | \(-\gamma^2\epsilon\) | \(-\gamma^2\epsilon\) | \(-\epsilon - \gamma\epsilon - \gamma^2\epsilon\) | \(\epsilon + \gamma\epsilon + \gamma^2\epsilon\) | \cdots |

State 2: \(0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon\)
State 3: \(0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)\)

\[
v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1-\gamma}\epsilon\right) = -2\frac{\gamma - \gamma^k}{(1-\gamma)^2} \epsilon \quad k \to \infty \quad -\frac{2\gamma}{(1-\gamma)^2} \epsilon
\]
Tightness of the bound for AVI

\[ v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left( -2 \frac{\gamma - \gamma^k}{1 - \gamma} \epsilon \right) = -2 \frac{\gamma - \gamma^k}{(1 - \gamma)^2} \epsilon \rightarrow_{k \to \infty} \frac{2\gamma}{(1 - \gamma)^2} \epsilon \]

State 2: \( 0 + \gamma(-\epsilon) = -2\gamma\epsilon + \gamma\epsilon \)
State 3: \( 0 + \gamma(-\epsilon - \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon) \)
Tightness of the bound for AVI

\[
0 + \gamma(\epsilon) = -2\gamma\epsilon + \gamma\epsilon
\]

State 2: \[0 + \gamma(\epsilon + \gamma\epsilon) = -2(\gamma + \gamma^2)\epsilon + \gamma(\epsilon + \gamma\epsilon)\]

\[
v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1 - \gamma}\epsilon\right) = -2\frac{\gamma - \gamma^k}{(1 - \gamma)^2}\epsilon \quad k \to \infty \quad \frac{2\gamma}{(1 - \gamma)^2}\epsilon
\]
Tightness of the bound (Lesner and Scherrer, 2014)

For any $m$ and $\ell$, NSMPI generates a sequence of policies $(\pi_k)_{k \geq 1}$ such that $\pi_k$ acts optimally except in state $k$. Thus, $\pi_{k,\ell} = \pi_k \pi_{k-1} \cdots \pi_{k-\ell+1}$ gets stuck in the loop

$$k, \ k + \ell - 1, \ k + \ell - 2, \ k + 1, \ k, \ \ldots$$

and therefore

$$\nu_{\pi_{k,\ell}}(k) = -\frac{2\gamma - \gamma^k}{(1 - \gamma)(1 - \gamma^\ell)} \epsilon.$$