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

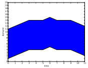
(An introduction to Reinforcement Learning / Optimal control)

Bruno Scherrer

INRIA (Institut National de Recherche en Informatique et ses Applications)
IECL (Institut Elie Cartan de Lorraine)

Example: The Retail Store Management Problem

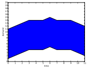
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

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

- $t = 0, 1, \ldots, 11$, $H = 12$
- 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

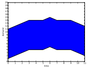
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

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

- $t = 0, 1, \ldots, 11$, $H = 12$
- 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

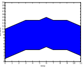
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

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

- $t = 0, 1, \dots, 11$, $H = 12$
- State space: $x \in X = \{0, 1, \dots, M\}$
- Action space: At state x, $a \in A(x) = \{0, 1, \dots, M - x\}$
- 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

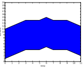
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

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

- $t = 0, 1, \ldots, 11$, $H = 12$
- 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

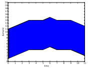
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

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

- $t = 0, 1, \ldots, 11$, $H = 12$
- 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

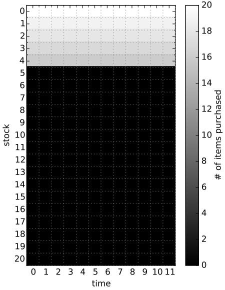
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the beginning of each month t, the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items 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 $g(x)$.

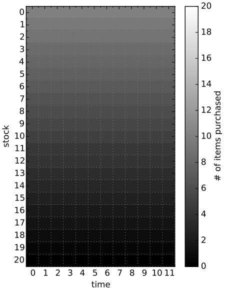
$M = 20$, $f(x) = x$, $g(x) = 0.25x$, $h(x) = 0.25x$, $C(a) = (1 + 0.5a)\mathbb{1}_{a>0}$, $w_t \sim$ 

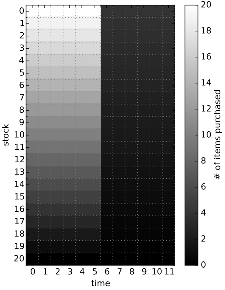
- $t = 0, 1, \ldots, 11$, $H = 12$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state x, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(\ x_t + a_t - w_t\ ,\ 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$ and $R(x) = g(x)$.

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

$$
\begin{aligned}
v_{\pi,s}(x) &= \mathbb{E}_\pi\left[\sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}_\pi[r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] \\
&+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ v_{\pi,s+1}(y).
\end{aligned}
$$

Policy evaluation

$$
\begin{aligned}
v_{\pi,s}(x) &= \mathbb{E}_\pi\left[\sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}_\pi[r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] \\
&+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x, x_{s+1} = y\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ v_{\pi,s+1}(y).
\end{aligned}
$$

Policy evaluation

$$
\begin{aligned}
v_{\pi,s}(x) &= \mathbb{E}_\pi\left[\sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}_\pi[r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] \\
&+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x,\ x_{s+1} = y\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ v_{\pi,s+1}(y).
\end{aligned}
$$

Policy evaluation

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

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

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

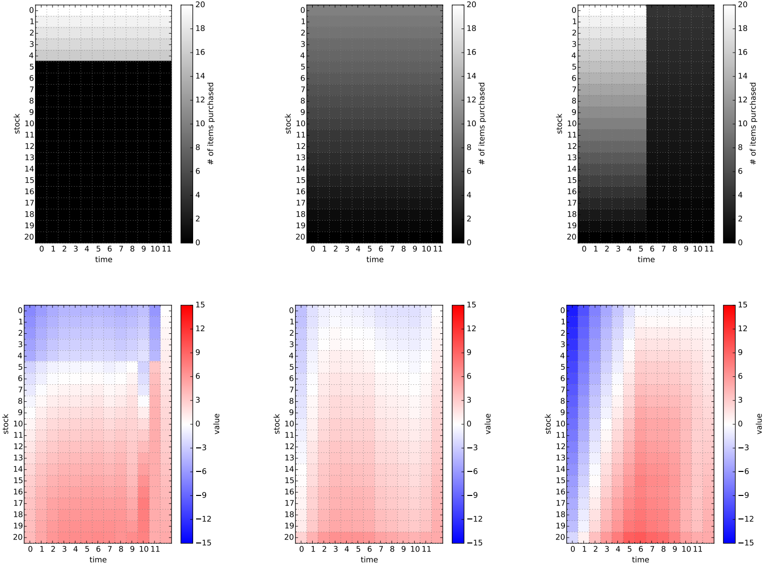
$$+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t,a_t,w_t) + R(x_H) \mid x_s = x,\ x_{s+1} = y\right]$$

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

Policy evaluation

$$
\begin{aligned}
v_{\pi,s}(x) &= \mathbb{E}_\pi\left[\sum_{t=s}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}_\pi[r_s(x_s, a_s, w_s) \mid x_s = x] + \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] \\
&+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ \mathbb{E}_\pi\left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \mid x_s = x,\ x_{s+1} = y\right] \\
&= \mathbb{E}[r_s(x, \pi(x), w_s)] + \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = \pi(x_s))\ v_{\pi,s+1}(y).
\end{aligned}
$$

Example: the Retail Store Management Problem



Policy optimization

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

$$= \max_{\pi_s,\pi_{s+1},\dots} \mathbb{E}_{\pi_s,\pi_{s+1},\dots} \Bigg\{ r_s(x_s, a_s, w_s)$$

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

$$= \max_a \Bigg\{ \mathbb{E}\big[r_s(x, a, w_s) \big]$$

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

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

Policy optimization

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

$$= \max_{\pi_s,\pi_{s+1},\dots} \mathbb{E}_{\pi_s,\pi_{s+1},\dots} \Bigg\{ r_s(x_s, a_s, w_s)$$

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

$$= \max_a \Bigg\{ \mathbb{E}\big[r_s(x, a, w_s)\big]$$

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

$$= \max_a \Big\{ \mathbb{E}\big[r_s(x, a, w_s)\big] + \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = a)\ v_{*,s+1}(y) \Big\}.$$

Policy optimization

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

$$= \max_{\pi_s,\pi_{s+1},\dots} \mathbb{E}_{\pi_s,\pi_{s+1},\dots} \Bigg\{ r_s(x_s, a_s, w_s)$$

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

$$= \max_a \Bigg\{ \mathbb{E}\big[r_s(x, a, w_s) \big]$$

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

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

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

$$= \max_{\pi_s,\pi_{s+1},\dots} \mathbb{E}_{\pi_s,\pi_{s+1},\dots} \Bigg\{ r_s(x_s, a_s, w_s)$$

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

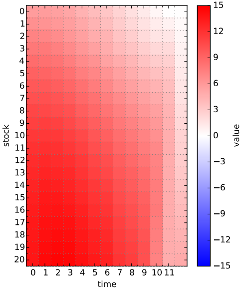
$$= \max_a \Bigg\{ \mathbb{E}\big[r_s(x, a, w_s) \big]$$

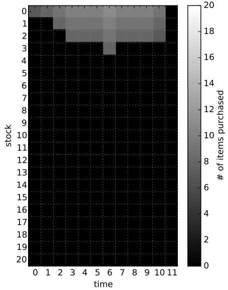
$$+ \sum_y \mathbb{P}(x_{s+1} = y | x_s = x, a_s = a) \max_{\pi_{s+1},\dots} \mathbb{E}_{\pi_{s+1},\dots} \left[\sum_{t=s+1}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) \,\middle|\, x_{s+1} = y \right] \Bigg\}$$

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

Example: the Retail Store Management Problem

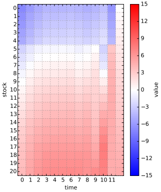
Optimal value and policy

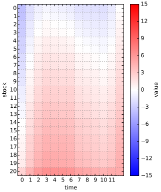


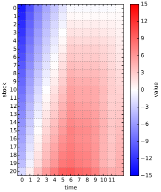


vs

values of policies $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}$







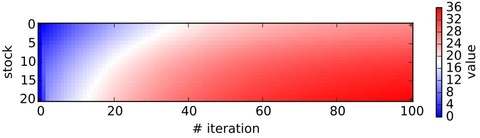
Example: the Retail Store Management Problem

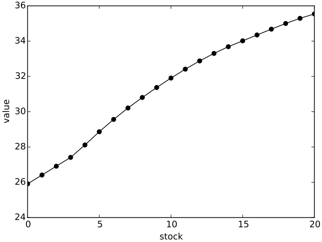
Each month t, a store contains x_t items (maximum capacity M) of a specific goods and the demand for that goods is w_t. At the end of each month the manager of the store can order a_t more items from his supplier. The cost of maintaining an inventory of x is $h(x)$. The cost to order a items is $C(a)$. 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 $\alpha = 3\% = 0.03$.

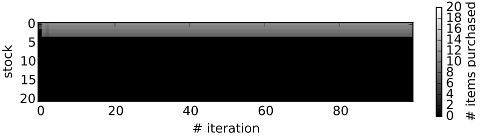
$M = 20$, $f(x) = x$, $g(x) = 0.25x$, $h(x) = 0.25x$, $C(a) = (1 + 0.5a)\mathbb{1}_{a>0}$, $w_t \sim \mathrm{U}(\{5, 6, \ldots, 15\})$, $\gamma = \frac{1}{1+\alpha}$

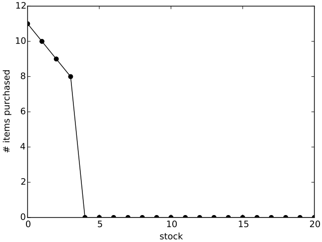
- $t = 0, 1, \ldots$
- State space: $x \in X = \{0, 1, \ldots, M\}$
- Action space: At state x, $a \in A(x) = \{0, 1, \ldots, M - x\}$
- Dynamics: $x_{t+1} = \max(\; x_t + a_t - w_t \;,\; 0)$
- Reward: $r(x_t, a_t, w_t) = -C(a_t) - h(x_t + a_t) + f(\min(w_t, x_t + a_t))$.

Example: the Retail Store Management Problem









Example: the Optimal Replacement Problem

State: level of wear (x) of an object (e.g., a car).
Action: $\{(\mathrm{R})\text{eplace}, (\mathrm{K})\text{eep}\}$.
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: $\{(\text{R})\text{eplace}, (\text{K})\text{eep}\}$.
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: $\{(\text{R})\text{eplace}, (\text{K})\text{eep}\}$.
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: $\{(\text{R})\text{eplace}, (\text{K})\text{eep}\}$.
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: $\{(\text{R})\text{eplace}, (\text{K})\text{eep}\}$.
Cost:

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

Dynamics:

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

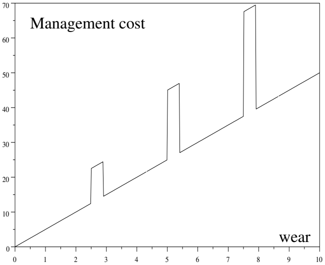
Problem: Minimize the discounted expected cost over an infinite horizon.

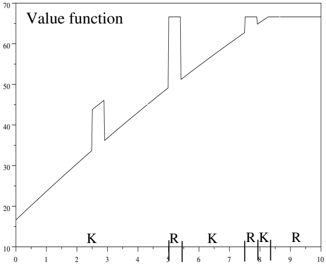
Example: the Optimal Replacement Problem

The optimal value function satisfies

$$v_*(x) = \min\Big\{ \underbrace{c(x) + \gamma \int_0^\infty d(y-x) v_*(y) dy}_{(K)eep},\ \underbrace{C + \gamma \int_0^\infty d(y) v_*(y) dy}_{(R)eplace} \Big\}$$

Optimal policy: action that attains the minimum





Example: the Optimal Replacement Problem

Linear approximation space

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

Collect N samples on a uniform grid:

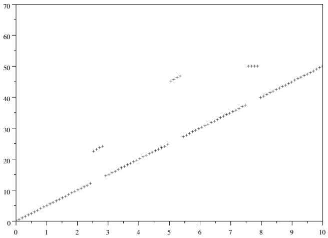


Figure: Left: the *target* values computed as $\{Tv_0(x_n)\}_{1\leq n\leq N}$. Right: the approximation $v_1 \in \mathcal{F}$ of the target function Tv_0.

Example: the Optimal Replacement Problem

Linear approximation space

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

Collect N samples on a uniform grid:

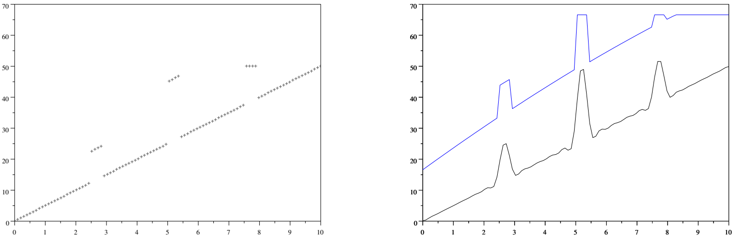
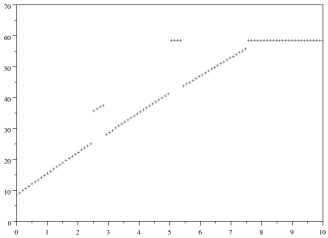


Figure: Left: the *target* values computed as $\{Tv_0(x_n)\}_{1\leq n\leq N}$. Right: the approximation $v_1 \in \mathcal{F}$ of the target function Tv_0.

Example: the Optimal Replacement Problem

One more step:



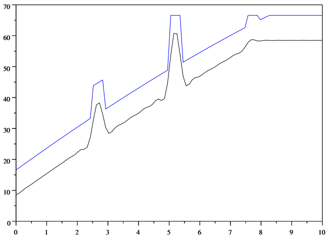


Figure: Left: the *target* values computed as $\{Tv_1(x_n)\}_{1\leq n\leq N}$. Right: the approximation $v_2 \in \mathcal{F}$ of Tv_1.

Example: the Optimal Replacement Problem

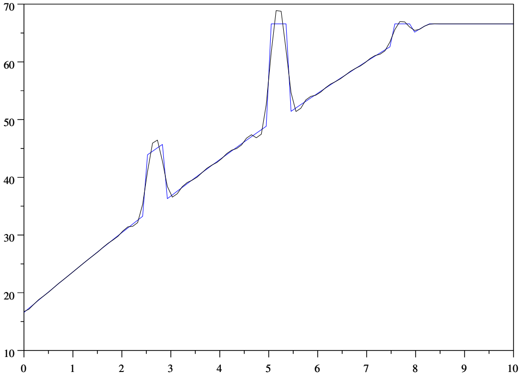


Figure: The approximation $v_{20} \in \mathcal{F}$.

Error propagation for AVI

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

$$\begin{aligned}\|v_* - v_k\|_\infty &= \|v_* - Tv_{k-1} - \epsilon_k\|_\infty \\ &\le \|Tv_* - Tv_{k-1}\|_\infty + \epsilon \\ &\le \gamma\|v_* - v_{k-1}\|_\infty + \epsilon \\ &\le \frac{\epsilon}{1-\gamma}.\end{aligned}$$

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

$$\begin{aligned}\|v_* - v_{\pi_{k+1}}\|_\infty &\le \|Tv_* - T_{\pi_{k+1}}v_k\|_\infty + \|T_{\pi_{k+1}}v_k - T_{\pi_{k+1}}v_{\pi_{k+1}}\|_\infty \\ &\le \|Tv_* - Tv_k\|_\infty + \gamma\|v_k - v_{\pi_{k+1}}\|_\infty \\ &\le \gamma\|v_* - v_k\|_\infty + \gamma\left(\|v_k - v_*\|_\infty + \|v_* - v_{\pi_{k+1}}\|_\infty\right) \\ &\le \frac{2\gamma}{1-\gamma}\|v_* - v_k\|_\infty.\end{aligned}$$

Error propagation for AVI

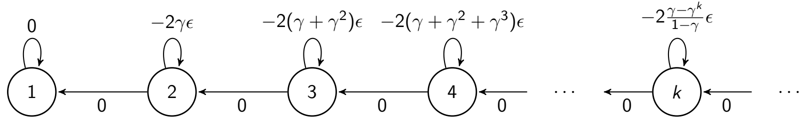
1. Bounding: $\|v_* - v_k\|_\infty$:

$$\begin{aligned}\|v_* - v_k\|_\infty &= \|v_* - T v_{k-1} - \epsilon_k\|_\infty \\ &\le \|T v_* - T v_{k-1}\|_\infty + \epsilon \\ &\le \gamma \|v_* - v_{k-1}\|_\infty + \epsilon \\ &\le \frac{\epsilon}{1-\gamma}.\end{aligned}$$

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

$$\begin{aligned}\|v_* - v_{\pi_{k+1}}\|_\infty &\le \|T v_* - T_{\pi_{k+1}} v_k\|_\infty + \|T_{\pi_{k+1}} v_k - T_{\pi_{k+1}} v_{\pi_{k+1}}\|_\infty \\ &\le \|T v_* - T v_k\|_\infty + \gamma \|v_k - v_{\pi_{k+1}}\|_\infty \\ &\le \gamma \|v_* - v_k\|_\infty + \gamma \left(\|v_k - v_*\|_\infty + \|v_* - v_{\pi_{k+1}}\|_\infty\right) \\ &\le \frac{2\gamma}{1-\gamma} \|v_* - v_k\|_\infty.\end{aligned}$$

Tightness of the bound for AVI

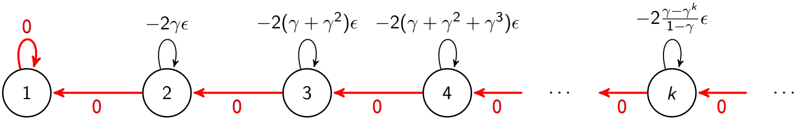


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

Tightness of the bound for AVI

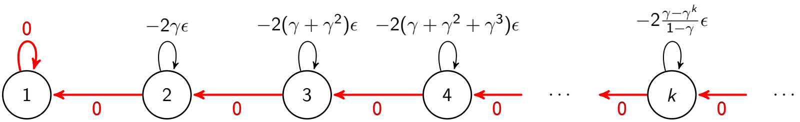


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

Tightness of the bound for AVI

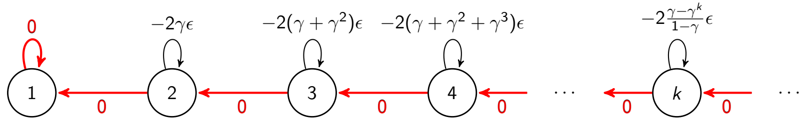


	1	2	3	4	…
v_0	0	0	0	0	…
v_1	$-\epsilon$	ϵ	0	0	…
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	…
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	…
…	…	…	…	…	…

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

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

Tightness of the bound for AVI

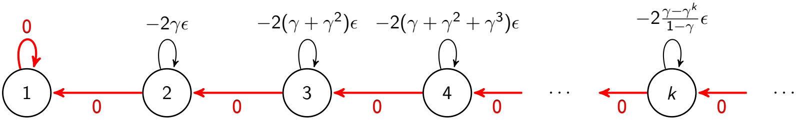


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

Tightness of the bound for AVI



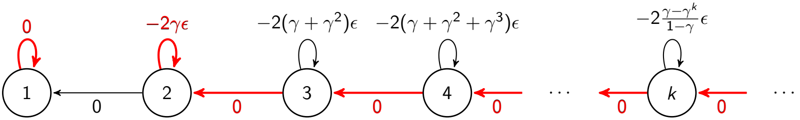
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

$$v_{\pi_k}(k) = \sum_{t=0}^{\infty}\gamma^t\left(-2\frac{\gamma-\gamma^k}{1-\gamma}\epsilon\right) = -2\frac{\gamma-\gamma^k}{(1-\gamma)^2}\epsilon \overset{k\to\infty}{\longrightarrow} -\frac{2\gamma}{(1-\gamma)^2}\epsilon$$

Tightness of the bound for AVI



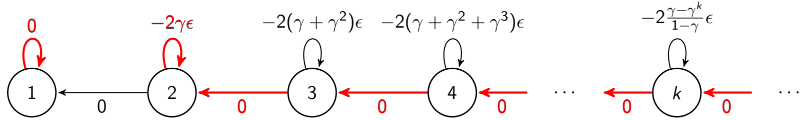
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

$$v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma-\gamma^k}{1-\gamma}\epsilon\right) = -2\frac{\gamma-\gamma^k}{(1-\gamma)^2}\epsilon \overset{k\to\infty}{\longrightarrow} -\frac{2\gamma}{(1-\gamma)^2}\epsilon$$

Tightness of the bound for AVI



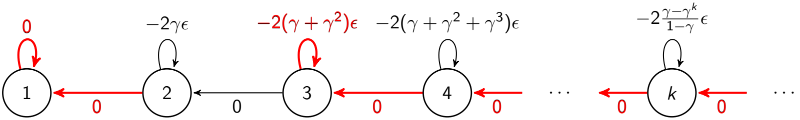
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon - \gamma\epsilon$	$\epsilon + \gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon - \gamma\epsilon - \gamma^2\epsilon$	$\epsilon + \gamma\epsilon + \gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

$$v_{\pi_k}(k) = \sum_{t=0}^{\infty} \gamma^t \left(-2\frac{\gamma - \gamma^k}{1-\gamma}\epsilon\right) = -2\frac{\gamma - \gamma^k}{(1-\gamma)^2}\epsilon \overset{k\to\infty}{\longrightarrow} -\frac{2\gamma}{(1-\gamma)^2}\epsilon$$

Tightness of the bound for AVI



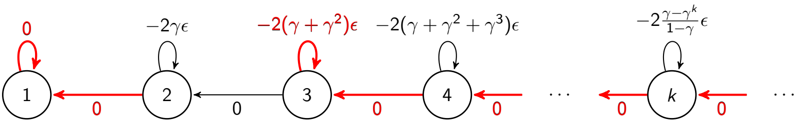
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

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

Tightness of the bound for AVI

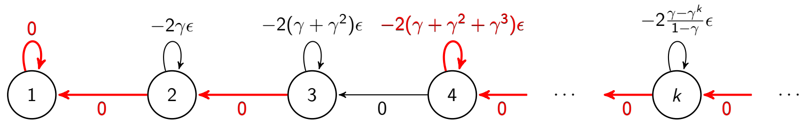


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

Tightness of the bound for AVI

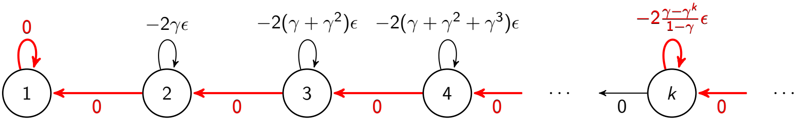


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

$$v_{\pi_k}(k) = \sum_{t=0}^{\infty}\gamma^t\left(-2\frac{\gamma-\gamma^k}{1-\gamma}\epsilon\right) = -2\frac{\gamma-\gamma^k}{(1-\gamma)^2}\epsilon \overset{k\to\infty}{\longrightarrow} -\frac{2\gamma}{(1-\gamma)^2}\epsilon$$

Tightness of the bound for AVI

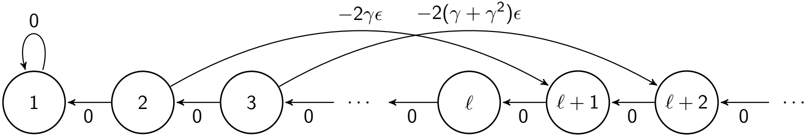


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

Tightness of the bound (Lesner and Scherrer, 2014)



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

$$k, \quad k+\ell-1, \quad k+\ell-2, \quad k+1, \quad k, \quad \dots$$

and therefore

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