Two Simple Tricks for Improving the Solution to Large RL Problems

Bruno Scherrer

INRIA, Institut Elie Cartan, Nancy, FRANCE

AWRL'17, November 15th, 2017

Outline

1. Markov Decision Processes and Approximate Dynamic Programming
2. Trick 1: Use a lower discount factor
3. Trick 2: Use a periodic non-stationary policy

Markov Decision Process (MDP)

(Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998)

Controlled and rewarded dynamical system:

$$x_0,\ a_0,\ r_0,\ x_1,\ a_1,\ r_1,\ x_2,\ a_2,\ r_2,\ x_3,\ \ldots$$

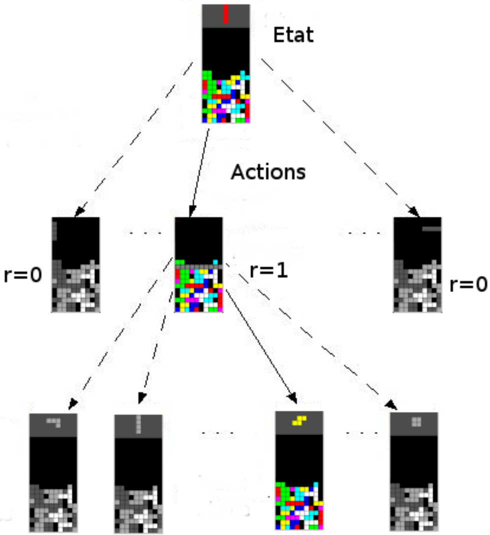
Markov Decision Process (MDP):

- X is the state space,
- A is the action space,
- $r : X \times A \to \mathbb{R}$ is the reward function, $\qquad (r_t = r(x_t, a_t))$
- $p : X \times A \to \Delta_X$ is the transition kernel. $\qquad (x_{t+1} \sim p(\cdot|x_t, a_t))$

Goal: Find a stationary deterministic policy $\pi : X \to A$ that maximizes the value $v_\pi(x)$ *for all* x:

$$v_\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| x_0 = x,\ \{\forall t,\ a_t = \pi(x_t)\}\right]. \qquad (\gamma \in (0,1))$$

Illustration: Tetris



Bellman Equations/Operators

- For any policy π, v_π is the unique solution of the **Bellman equation**:

$$\forall x,\; v_\pi(x) = r(x,\pi(x)) + \gamma \sum_{y\in X} p(y|x,\pi(x)) v_\pi(y) \quad \Leftrightarrow \quad v_\pi = T_\pi v_\pi.$$

- The **optimal value** v_* is the unique solution of the **Bellman optimality equation**:

$$\forall x,\; v_*(x) = \max_{a\in A} \Big(r(x,a) + \gamma \sum_{y\in X} p(y|x,a) v_*(y) \Big) \quad \Leftrightarrow \quad v_* = T v_*.$$

- $T_\pi : \mathbb{R}^X \to \mathbb{R}^X$ and $T : \mathbb{R}^X \to \mathbb{R}^X$ are γ-contraction mappings w.r.t. the max norm $\|v\|_\infty = \max_s |v(s)|$.
- For any v, π is a **greedy policy** w.r.t. v, written $\pi = \mathcal{G}v$, iff

$$\forall x,\; \pi(x) \in \arg\max_{a\in A} \Big(r(x,a) + \gamma \sum_{y\in X} p(y|x,a) v(y) \Big) \quad \Leftrightarrow \quad T_\pi v = T v.$$

- $\pi_* = \mathcal{G} v_*$

Bellman Equations/Operators

- For any policy π, v_π is the unique solution of the **Bellman equation**:

$$\forall x,\ v_\pi(x) = r(x,\pi(x)) + \gamma \sum_{y \in X} p(y|x,\pi(x)) v_\pi(y) \quad \Leftrightarrow \quad v_\pi = T_\pi v_\pi.$$

- The **optimal value** v_* is the unique solution of the **Bellman optimality equation**:

$$\forall x,\ v_*(x) = \max_{a \in A} \left(r(x,a) + \gamma \sum_{y \in X} p(y|x,a) v_*(y) \right) \quad \Leftrightarrow \quad v_* = T v_*.$$

- $T_\pi : \mathbb{R}^X \to \mathbb{R}^X$ and $T : \mathbb{R}^X \to \mathbb{R}^X$ are γ-contraction mappings w.r.t. the max norm $\|v\|_\infty = \max_s |v(s)|$.
- For any v, π is a **greedy policy** w.r.t. v , written $\pi = \mathcal{G} v$, iff

$$\forall x,\ \pi(x) \in \arg\max_{a \in A} \left(r(x,a) + \gamma \sum_{y \in X} p(y|x,a) v(y) \right) \quad \Leftrightarrow \quad T_\pi v = T v.$$

- $\pi_* = \mathcal{G} v_*$

Dynamic Programming Algorithms

Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k = T_{\pi_{k+1}} v_k$

Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^{\infty} v_k$

Modified Policy Iteration (Puterman & Shin, 1978)

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^{m} v_k \qquad (1 \leq m \leq \infty)$

Dynamic Programming Algorithms

Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k = T_{\pi_{k+1}} v_k$

Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^{\infty} v_k$

Modified Policy Iteration (Puterman & Shin, 1978)

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^{m} v_k \qquad (1 \leq m \leq \infty)$

Dynamic Programming Algorithms

Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k = T_{\pi_{k+1}} v_k$

Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^\infty v_k$

Modified Policy Iteration (Puterman & Shin, 1978)

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^m v_k \qquad (1 \leq m \leq \infty)$

Approximate Dynamic Programming

- $[(T_\pi)^m v](x)$ approximated by Monte-Carlo:

$$[(T_\pi)^m v](x) = \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t r(x_t, a_t) + \gamma^m v(x_m) \middle| x_0 = x, \{\forall t, a_t = \pi(x_t)\} \right]$$

- "$v(\cdot) \leftarrow [Au](\cdot)$" approximated by regression:

$$\min_{v \in \mathcal{F} \subset \mathbb{R}^X} \sum_x \mu(x) |v(x) - \widehat{[Au](x)}|^p$$

where $\widehat{[Au](x)}$ is an unbiased sample of $[Au](x)$.

Approximate MPI

(q_k) are represented in $\mathcal{F} \subseteq \mathbb{R}^{X \times A}$

- $\pi_{k+1} \quad \leftarrow \mathcal{G} q_k$
- $q_{k+1} \quad \leftarrow (T_{\pi_{k+1}})^m q_k$

Policy update

In any state x, the greedy action is: $\pi_{k+1}(x) = \arg\max_{a \in A} q_k(x, a)$

Value function update

1. **Point-wise estimation** through rollouts of length m:
For $1 \leq i \leq N$, sample state-action pairs $(x^{(i)}, a^{(i)}) \sim \mu$ and trajectory $\left(x^{(i)}, a^{(i)}, x_1^{(i)}, \ldots, a_{m-1}^{(i)}, x_m^{(i)}\right)$ with $a_t^{(i)} = \pi_{k+1}(x_t^{(i)})$
and deduce an unbiased estimate $\widehat{q}_{k+1}^{(i)}$ of $\left[\left(T_{\pi_{k+1}}\right)^m v_k\right](x^{(i)}, a^{(i)})$:

$$\widehat{q}_{k+1}^{(i)} = r(x^{(i)}, a^{(i)}) + \sum_{t=1}^{m-1} \gamma^t r(x_t^{(i)}, a_t^{(i)}) + \gamma^m q_k(x_m^{(i)}, \pi_{k+1}(x_m^{(i)}))$$

2. **Generalisation** through regression:
q_{k+1} is computed as the best fit of these estimates in $\mathcal{F}$

$$q_{k+1} = \arg\min_{q \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \left(q(x^{(i)}, a^{(i)}) - \widehat{q}_{k+1}^{(i)}\right)^2$$

Approximate MPI

(q_k) are represented in $\mathcal{F} \subseteq \mathbb{R}^{X \times A}$

- $\pi_{k+1} \leftarrow \mathcal{G} q_k$
- $q_{k+1} \leftarrow (T_{\pi_{k+1}})^m q_k$

Policy update

In any state x, the greedy action is: $\pi_{k+1}(x) = \arg\max_{a \in A} q_k(x, a)$

Value function update

1. **Point-wise estimation** through rollouts of length m:
   For $1 \leq i \leq N$, sample state-action pairs $(x^{(i)}, a^{(i)}) \sim \mu$ and trajectory $\left(x^{(i)}, a^{(i)}, x_1^{(i)}, \ldots, a_{m-1}^{(i)}, x_m^{(i)}\right)$ with $a_t^{(i)} = \pi_{k+1}(x_t^{(i)})$
   and deduce an unbiased estimate $\widehat{q}_{k+1}^{(i)}$ of $\left[\left(T_{\pi_{k+1}}\right)^m v_k\right](x^{(i)}, a^{(i)})$:

$$\widehat{q}_{k+1}^{(i)} = r(x^{(i)}, a^{(i)}) + \sum_{t=1}^{m-1} \gamma^t r(x_t^{(i)}, a_t^{(i)}) + \gamma^m q_k(x_m^{(i)}, \pi_{k+1}(x_m^{(i)}))$$

2. **Generalisation** through regression:
   q_{k+1} is computed as the best fit of these estimates in $\mathcal{F}$

$$q_{k+1} = \arg\min_{q \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \left(q(x^{(i)}, a^{(i)}) - \widehat{q}_{k+1}^{(i)}\right)^2$$

Approximate MPI

(q_k) are represented in $\mathcal{F} \subseteq \mathbb{R}^{X \times A}$

- $\pi_{k+1} \leftarrow \mathcal{G} q_k$
- $q_{k+1} \leftarrow (T_{\pi_{k+1}})^m q_k$

Policy update

In any state x, the greedy action is: $\pi_{k+1}(x) = \arg\max_{a \in A} q_k(x, a)$

Value function update

1. **Point-wise estimation** through rollouts of length m:
   For $1 \le i \le N$, sample state-action pairs $(x^{(i)}, a^{(i)}) \sim \mu$ and trajectory $\left(x^{(i)}, a^{(i)}, x_1^{(i)}, \ldots, a_{m-1}^{(i)}, x_m^{(i)}\right)$ with $a_t^{(i)} = \pi_{k+1}(x_t^{(i)})$
   and deduce an unbiased estimate $\widehat{q}_{k+1}^{(i)}$ of $\left[\left(T_{\pi_{k+1}}\right)^m v_k\right](x^{(i)}, a^{(i)})$:

$$\widehat{q}_{k+1}^{(i)} = r(x^{(i)}, a^{(i)}) + \sum_{t=1}^{m-1} \gamma^t r(x_t^{(i)}, a_t^{(i)}) + \gamma^m q_k(x_m^{(i)}, \pi_{k+1}(x_m^{(i)}))$$

2. **Generalisation** through regression:
   q_{k+1} is computed as the best fit of these estimates in $\mathcal{F}$

$$q_{k+1} = \arg\min_{q \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \left(q(x^{(i)}, a^{(i)}) - \widehat{q}_{k+1}^{(i)} \right)^2$$

Approximate Algorithms

App. Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k + \epsilon_k = T_{\pi_{k+1}} v_k + \epsilon_k$

App. Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^\infty v_k + \epsilon_k$

App. Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^m v_k + \epsilon_k \qquad (1 \leq m \leq \infty)$

Theorem (Singh & Yee, 1994; Gordon, 1995; Bertsekas & Tsitsiklis, 1996; Scherrer *et al.* , 2012; Scherrer *et al.* , 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The *loss* due to running policy π_k instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_{\pi_*} - v_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\epsilon.$$

Approximate Algorithms

App. Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k + \epsilon_k = T_{\pi_{k+1}} v_k + \epsilon_k$

App. Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^\infty v_k + \epsilon_k$

App. Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^m v_k + \epsilon_k \qquad (1 \leq m \leq \infty)$

Theorem (Singh & Yee, 1994; Gordon, 1995; Bertsekas & Tsitsiklis, 1996; Scherrer *et al.* , 2012; Scherrer *et al.* , 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The *loss* due to running policy π_k instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_{\pi_*} - v_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\epsilon.$$

Approximate Algorithms

App. Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T v_k + \epsilon_k = T_{\pi_{k+1}} v_k + \epsilon_k$

App. Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{\pi_{k+1}} = (T_{\pi_{k+1}})^\infty v_k + \epsilon_k$

App. Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\pi_{k+1}})^m v_k + \epsilon_k \qquad (1 \leq m \leq \infty)$

Theorem (Singh & Yee, 1994; Gordon, 1995; Bertsekas & Tsitsiklis, 1996; Scherrer *et al.* , 2012; Scherrer *et al.* , 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The *loss* due to running policy π_k instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_{\pi_*} - v_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\epsilon.$$

Outline

1. Markov Decision Processes and Approximate Dynamic Programming
2. **Trick 1: Use a lower discount factor**
3. Trick 2: Use a periodic non-stationary policy

Trick 1

Use a discount factor $\beta < \gamma$ to compute a policy π^{β} and run (and evaluate) it on the original γ-discounted MDP.

Intuition: find a policy that solves a shorter-horizon (simpler) problem.

$v_{\pi}^{\alpha} \overset{def}{=}$ value of policy π on problem with discount α.
$\pi_{*}^{\alpha} \overset{def}{=}$ an optimal on problem with discount α ?
$v_{*}^{\alpha} = v_{\pi_{*}^{\alpha}}^{\alpha}$ is the optimal value function with discount α

Can we have $\|v_{*}^{\gamma} - v_{\pi^{\beta}}^{\gamma}\| \leq \|v_{*}^{\gamma} - v_{\pi^{\gamma}}^{\gamma}\|$?

Trick 1

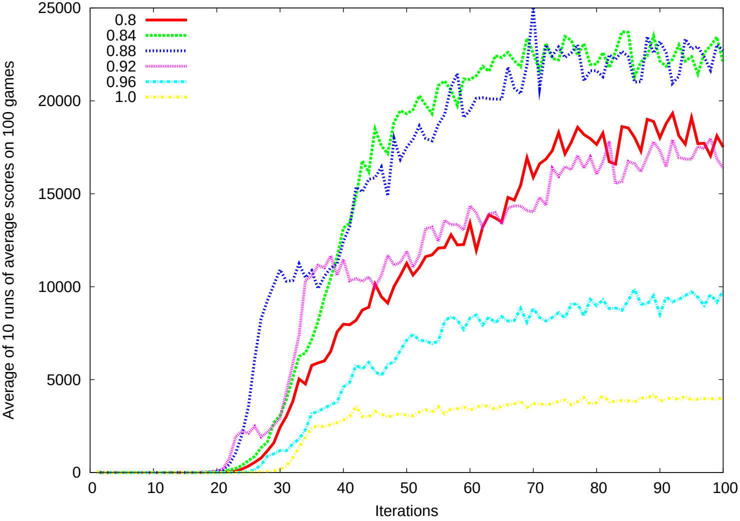
Use a discount factor $\beta < \gamma$ to compute a policy π^β and run (and evaluate) it on the original γ-discounted MDP.

Intuition: find a policy that solves a shorter-horizon (simpler) problem.

$v_\pi^\alpha \overset{def}{=}$ value of policy π on problem with discount α.
$\pi_*^\alpha \overset{def}{=}$ an optimal on problem with discount α ?
$v_*^\alpha = v_{\pi_*^\alpha}^\alpha$ is the optimal value function with discount α

Can we have $\|v_*^\gamma - v_{\pi^\beta}^\gamma\| \leq \|v_*^\gamma - v_{\pi^\gamma}^\gamma\|$?

Experiment on Tetris (AVI)



Assume w.l.o.g. that $\|r\|_\infty = 1$.

$$\begin{aligned} & \|v_*^\gamma - v_{\pi^\beta}^\gamma\| \\ \leq \quad & \|v_*^\gamma - v_*^\beta\| + \|v_*^\beta - v_{\pi^\beta}^\beta\| + \|v_{\pi^\beta}^\beta - v_{\pi^\beta}^\gamma\| \\ \leq \quad & 2 \times \frac{\gamma - \beta}{(1-\gamma)(1-\beta)} + \frac{2\beta}{(1-\beta)^2}\epsilon \\ \leq \quad & \frac{2\gamma}{(1-\gamma)^2}\epsilon \end{aligned}$$

for ϵ sufficiently big (since $\beta < \gamma$)

Theory

Assume w.l.o.g. that $\|r\|_\infty = 1$.

$$
\begin{aligned}
& \|v_*^\gamma - v_{\pi^\beta}^\gamma\| \\
\leq \quad & \|v_*^\gamma - v_*^\beta\| + \|v_*^\beta - v_{\pi^\beta}^\beta\| + \|v_{\pi^\beta}^\beta - v_{\pi^\beta}^\gamma\| \\
\leq \quad & 2 \times \frac{\gamma - \beta}{(1-\gamma)(1-\beta)} + \frac{2\beta}{(1-\beta)^2}\epsilon \\
\leq \quad & \frac{2\gamma}{(1-\gamma)^2}\epsilon
\end{aligned}
$$

for ϵ sufficiently big (since $\beta < \gamma$)

Assume w.l.o.g. that $\|r\|_\infty = 1$.

$$
\begin{aligned}
& \|v_*^\gamma - v_{\pi^\beta}^\gamma\| \\
\leq \quad & \|v_*^\gamma - v_*^\beta\| + \|v_*^\beta - v_{\pi^\beta}^\beta\| + \|v_{\pi^\beta}^\beta - v_{\pi^\beta}^\gamma\| \\
\leq \quad & 2 \times \frac{\gamma - \beta}{(1-\gamma)(1-\beta)} + \frac{2\beta}{(1-\beta)^2}\epsilon \\
\leq \quad & \frac{2\gamma}{(1-\gamma)^2}\epsilon
\end{aligned}
$$

for ϵ sufficiently big (since $\beta < \gamma$)

Outline

1. Markov Decision Processes and Approximate Dynamic Programming

2. Trick 1: Use a lower discount factor

3. **Trick 2: Use a periodic non-stationary policy**

Non-Stationary Value Iteration

AVI generates a sequence of values/policies ($\pi_{i+1} \in \mathcal{G} v_i$)

$$\begin{array}{cccccccc} v_0 & v_1 & v_2 & \dots & v_{k-\ell} & \dots & v_{k-2} & v_{k-1} \\ \pi_1 & \pi_2 & \pi_3 & \dots & \pi_{k-\ell+1} & \dots & \pi_{k-1} & \pi_k \end{array}$$

Return the following periodic non-stationary policy

$$(\sigma_{k,\ell})^\infty = \underbrace{\pi_k \ \pi_{k-1} \ \cdots \ \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \underbrace{\pi_k \ \pi_{k-1} \ \cdots \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \cdots\cdots$$

Theorem (Scherrer & Lesner, 2012)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. For all ℓ, the loss due to running the non-stationary policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies:

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

Non-Stationary Value Iteration

AVI generates a sequence of values/policies ($\pi_{i+1} \in \mathcal{G} v_i$)

$$\begin{array}{cccccccc} v_0 & v_1 & v_2 & \cdots & v_{k-\ell} & \cdots & v_{k-2} & v_{k-1} \\ \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_{k-\ell+1} & \cdots & \pi_{k-1} & \pi_k \end{array}$$

Return the following periodic non-stationary policy

$$(\sigma_{k,\ell})^\infty = \underbrace{\pi_k \; \pi_{k-1} \; \cdots \; \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \; \underbrace{\pi_k \; \pi_{k-1} \; \cdots \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \; \cdots\cdots$$

Theorem (Scherrer & Lesner, 2012)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. For all ℓ, the loss due to running the non-stationary policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies:

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

Non-Stationary Value Iteration

AVI generates a sequence of values/policies ($\pi_{i+1} \in \mathcal{G} v_i$)

$$\begin{array}{cccccccc} v_0 & v_1 & v_2 & \dots & v_{k-\ell} & \dots & v_{k-2} & v_{k-1} \\ \pi_1 & \pi_2 & \pi_3 & \dots & \pi_{k-\ell+1} & \dots & \pi_{k-1} & \pi_k \end{array}$$

Return the following periodic non-stationary policy

$$(\sigma_{k,\ell})^\infty = \underbrace{\pi_k \ \pi_{k-1} \ \cdots \ \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \underbrace{\pi_k \ \pi_{k-1} \ \cdots \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \cdots\cdots$$

Theorem (Scherrer & Lesner, 2012)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. For all ℓ, the loss due to running the non-stationary policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies:

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

Non-Stationary Value Iteration

AVI generates a sequence of values/policies ($\pi_{i+1} \in \mathcal{G} v_i$)

$$\begin{array}{cccccccc} v_0 & v_1 & v_2 & \dots & v_{k-\ell} & \dots & v_{k-2} & v_{k-1} \\ \pi_1 & \pi_2 & \pi_3 & \dots & \pi_{k-\ell+1} & \dots & \pi_{k-1} & \pi_k \end{array}$$

Return the following periodic non-stationary policy

$$(\sigma_{k,\ell})^\infty = \underbrace{\pi_k \ \pi_{k-1} \ \cdots \ \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \underbrace{\pi_k \ \pi_{k-1} \ \cdots \pi_{k-\ell+1}}_{\sigma_{k,\ell}\text{: last } \ell \text{ policies}} \ \cdots\cdots$$

Theorem (Scherrer & Lesner, 2012)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. For all ℓ, the loss due to running the non-stationary policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies:

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

API with a non-stationary policy of period ℓ

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{(\sigma_{k+1,\ell})^\infty} + \epsilon_k \quad$ (by solving $v_{k+1} \simeq T_{\sigma_{k+1,\ell}} v_{k+1}$)

where $\pi_{\ell,\ell} = \pi_\ell \; \pi_{\ell-1} \; \dots \; \pi_1 \; \pi_\ell \; \pi_{\ell-1} \; \dots \; \pi_1 \; \dots$
with arbitrary $\pi_0, \pi_{-1}, \dots \pi_{-\ell+1}$ and

$$\forall v, \quad T_{\sigma_{k,\ell}} v = T_{\pi_k} T_{\pi_{k-1}} \dots T_{\pi_{k-\ell+1}} v.$$

Output as a function of k:

$$(\sigma_0^\ell)^\infty = (\pi_0 \; \pi_{-1} \; \dots \; \pi_{-\ell+2} \; \pi_{-\ell+1})^\infty$$
$$(\sigma_1^\ell)^\infty = (\pi_1 \; \pi_0 \; \dots \; \pi_{-\ell+3} \; \pi_{-\ell+2})^\infty$$
$$(\sigma_2^\ell)^\infty = (\pi_2 \; \pi_1 \; \dots \; \pi_{-\ell+4} \; \pi_{-\ell+3})^\infty$$
$$\vdots \qquad \qquad \vdots$$
$$(\sigma_k^\ell)^\infty = (\pi_k \; \pi_{k-1} \; \dots \; \pi_{k-\ell+2} \; \pi_{k-\ell+1})^\infty$$

Non-Stationary API

API with a non-stationary policy of period ℓ

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow v_{(\sigma_{k+1,\ell})^\infty} + \epsilon_k \quad$ (by solving $v_{k+1} \simeq T_{\sigma_{k+1,\ell}} v_{k+1}$)

where $\pi_{\ell,\ell} = \pi_\ell \ \pi_{\ell-1} \ \ldots \ \pi_1 \ \pi_\ell \ \pi_{\ell-1} \ \ldots \ \pi_1 \ \ldots$
with arbitrary $\pi_0, \pi_{-1}, \ldots \pi_{-\ell+1}$ and

$$\forall v, \quad T_{\sigma_{k,\ell}} v = T_{\pi_k} T_{\pi_{k-1}} \ldots T_{\pi_{k-\ell+1}} v.$$

Theorem (Scherrer & Lesner, 2012)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The loss due to running the non-stationary policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies:

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

Non Stationary Modified Policy Iteration

NS Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T_{\pi_{k+1}} v_k + \epsilon_k$

NS Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^\infty T_{\pi_{k+1}} v_k + \epsilon_k$

NS Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^m T_{\pi_{k+1}} v_k + \epsilon_k \qquad (0 \leq m \leq \infty)$

Theorem (Lesner & Scherrer, 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The loss due to running policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

The algorithms above are algorithms for ℓ-periodic MDPs.
Intuition: more degrees of freedom

Non Stationary Modified Policy Iteration

NS Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T_{\pi_{k+1}} v_k + \epsilon_k$

NS Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^\infty T_{\pi_{k+1}} v_k + \epsilon_k$

NS Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^m T_{\pi_{k+1}} v_k + \epsilon_k$ $\quad (0 \leq m \leq \infty)$

Theorem (Lesner & Scherrer, 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The loss due to running policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

The algorithms above are algorithms for ℓ-periodic MDPs.
Intuition: more degrees of freedom

Non Stationary Modified Policy Iteration

NS Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T_{\pi_{k+1}} v_k + \epsilon_k$

NS Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^\infty T_{\pi_{k+1}} v_k + \epsilon_k$

NS Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^m T_{\pi_{k+1}} v_k + \epsilon_k \qquad (0 \leq m \leq \infty)$

Theorem (Lesner & Scherrer, 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The loss due to running policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

The algorithms above are algorithms for ℓ-periodic MDPs.
Intuition: more degrees of freedom

Non Stationary Modified Policy Iteration

NS Value Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow T_{\pi_{k+1}} v_k + \epsilon_k$

NS Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^{\infty} T_{\pi_{k+1}} v_k + \epsilon_k$

NS Modified Policy Iteration

$\pi_{k+1} \leftarrow \mathcal{G} v_k$
$v_{k+1} \leftarrow (T_{\sigma_{k+1,\ell}})^{m} T_{\pi_{k+1}} v_k + \epsilon_k \qquad (0 \leq m \leq \infty)$

Theorem (Lesner & Scherrer, 2015)

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. The loss due to running policy $(\sigma_{k,\ell})^\infty$ instead of the optimal policy π_* satisfies

$$\limsup_{k\to\infty} \|v_* - v_{(\sigma_{k,\ell})^\infty}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

The algorithms above are algorithms for ℓ-periodic MDPs.
Intuition: more degrees of freedom

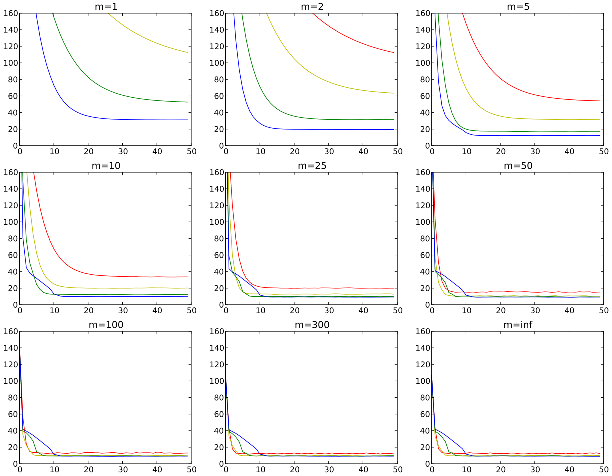


Figure: Average error of policy $(\sigma_{k,\ell})^\infty$ per iteration k of NS-AMPI. $\ell = 1$, $\ell = 2$, $\ell = 5$, $\ell = 10$.

Summary

- Markov Decision Processes
- Approximate Dynamic Programming
- Sometimes, it is easier to solve a problem different from the original problem:
  - Trick 1: a problem with a lower discount factor
  - Trick 2: a periodic variation of the problem
- Bounds matter!

Trick 1 based on a work with Marek Petrik (Petrik & Scherrer, 2008)
Trikh 2 based on a work with Boris Lesner (Scherrer & Lesner, 2012; Lesner & Scherrer, 2015)

References I

Bertsekas, D.P., & Tsitsiklis, J.N. 1996.
Neurodynamic Programming.
Athena Scientific.

Gordon, G.J. 1995.
Stable function approximation in dynamic programming.
Pages 261–268 of: International conference on machine learning.

Lesner, B., & Scherrer, B. 2015 (July).
Non-Stationary Approximate Modified Policy Iteration.
In: ICML 2015.

Petrik, M., & Scherrer, B. 2008.
Biasing Approximate Dynamic Programming with a Lower Discount Factor.
In: Neural Information Processing Systems.

Puterman, M. 1994.
Markov Decision Processes.
Wiley, New York.

Puterman, M., & Shin, M. 1978.
Modified policy iteration algorithms for discounted Markov decision problems.
Management science, **24**(11).

Scherrer, B., & Lesner, B. 2012 (Dec.).
On the use of non-stationary policies for stationary infinite-horizon Markov decision processes.
In: Neural Information Processing Systems.

References II

Scherrer, B., Ghavamzadeh, M., Gabillon, V., & Geist, M. 2012 (June).
Approximate Modified Policy Iteration.
In: 29th International Conference on Machine Learning - ICML 2012.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., & Geist, M. 2015.
Approximate Modified Policy Iteration and its Application to the Game of Tetris.
Journal of Machine Learning Research, 47.
A paraître.

Singh, S., & Yee, R. 1994.
An Upper Bound on the Loss from Approximate Optimal-Value Functions.
Machine learning, **16-3**, 227–233.

Sutton, R.S., & Barto, A.G. 1998.
Reinforcement learning: An introduction.
MIT Press.

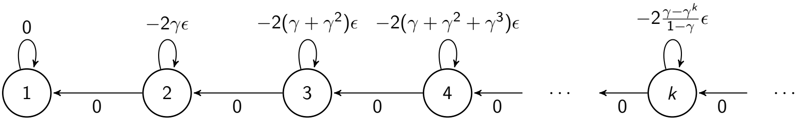
Illustration of approximation on Tetris

1. Approximation architecture for the value and for the score (on which the policy is based)

$$
\begin{aligned}
f_\theta(x) = &\ \theta_0 && \text{Constant} \\
&+ \theta_1 h_1(x) + \theta_2 h_2(x) + \cdots + \theta_{10} h_{10}(x) && \text{column height} \\
&+ \theta_{11} \Delta h_1(x) + \theta_{12} \Delta h_2(x) + \cdots + \theta_{19} \Delta h_9(x) && \text{height variation} \\
&+ \theta_{20} \max_k h_k(x) && \text{max height} \\
&+ \theta_{21} L(x) && \text{\# holes}
\end{aligned}
$$

2. Sampling Scheme: play games

Tightness of the bound for AVI

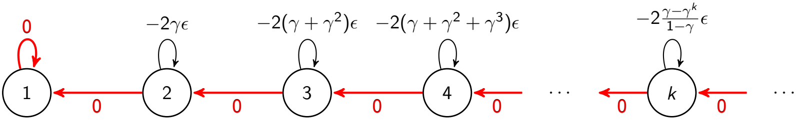


	1	2	3	4	…
v_0	0	0	0	0	…
v_1	$-\epsilon$	ϵ	0	0	…
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	…
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	…
…	…	…	…	…	…

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

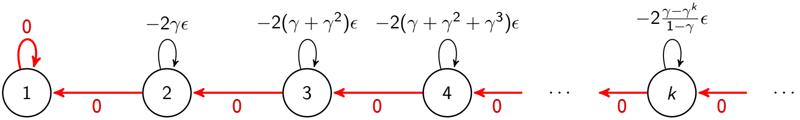


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

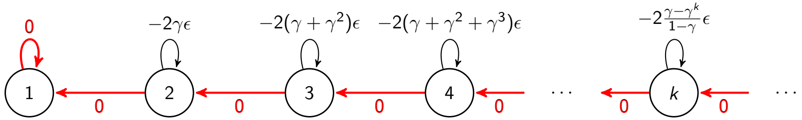


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

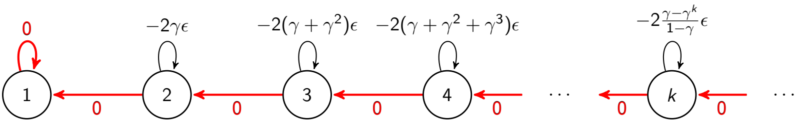


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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



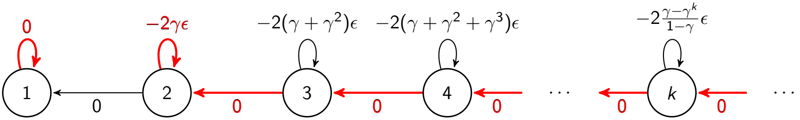
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

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

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

Tightness of the bound for AVI



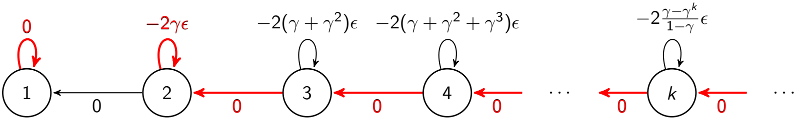
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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



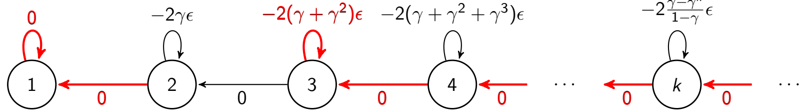
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

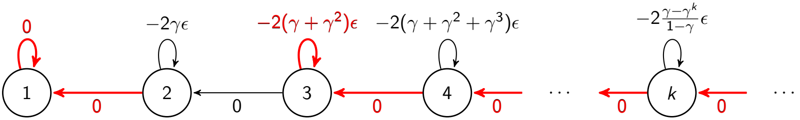


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

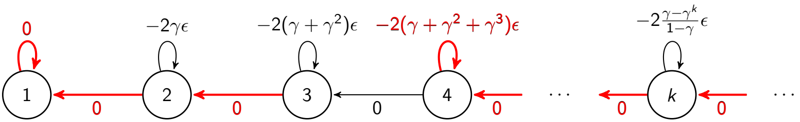


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

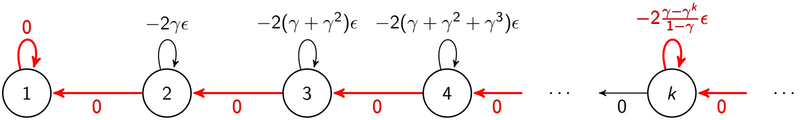


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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

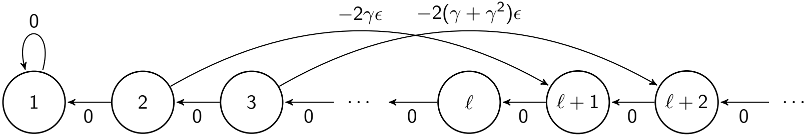


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 (Lesner & Scherrer, 2015)



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, $(\sigma_{k,\ell})^\infty = (\pi_k \pi_{k-1} \dots \pi_{k-\ell+1})^\infty$ gets stuck in the loop

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

and therefore

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