On Periodic Markov Decision Processes

Bruno Scherrer

INRIA, Institut Elie Cartan, Nancy, FRANCE

EWRL, December 3rd, 2016

Outline

1. Markov Decision Processes
2. Periodic Markov Decision Processes
3. Approximate Dynamic Programming

Markov Decision Process (MDP)

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

Controlled and rewarded dynamical system:

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

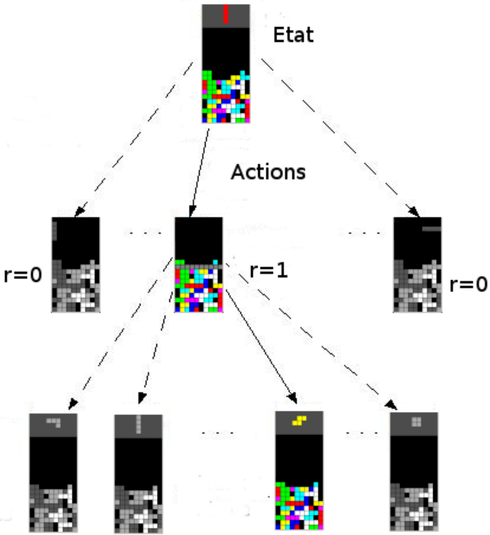
Markov Decision Process (MDP):

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

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**:

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

- $T_\pi : \mathbb{R}^X \to \mathbb{R}^X$ 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

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

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

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)

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

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)

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

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)

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

Asynchronous Dynamic Programming

Subsets: $X_1, X_2, \ldots, X_k, \ldots \qquad X_i \subset X$

At each iteration k,

- Either update the value on X_k

$$\begin{aligned} \pi_{k+1} &= \pi_k \\ v_{k+1}(x) &= \begin{cases} [T_{\pi_k} v_k](x) & \text{if } x \in X_k \\ v_k(x) & \text{otherwise} \end{cases} \end{aligned}$$

- or update the policy on X_k

$$\begin{aligned} \pi_{k+1}(x) &= \begin{cases} [\mathcal{G} v_k](x) & \text{if } x \in X_k \\ \pi_k(x) & \text{otherwise} \end{cases} \\ v_{k+1} &= v_k \end{aligned}$$

Convergence if all states are updated infinitely often (Bertsekas & Tsitsiklis, 1996).

Outline

ℓ-periodic MDP (Riis, 1965)

Let $\ell \geq 1$. Write $[t] = t \bmod \ell$.
At time t, the reward and the transition used are $r_{[t]}$ and $p_{[t]}$:

- $r_0, r_1, \ldots, r_{\ell-1}$, with $r_i : X \times A \to \mathbb{R}$
- $p_0, p_1, \ldots, p_{\ell-1}$, with $p_i : X \times A \to \Delta_X$

An ℓ-periodic MDP is an MDP on

$$X \times \{0, 1, \ldots, \ell - 1\} \quad = \quad X_0 \cup X_1 \cup \cdots \cup X_{\ell-1}$$

$\Rightarrow$ There exists a deterministic ℓ-periodic optimal policy

$$\pi^* = (\pi_0^*, \pi_1^*, \ldots, \pi_{\ell-1}^*)$$

and the optimal decisions are $a_t = \pi_{[t]}^*(x_t)$.

ℓ-periodic MDP (Riis, 1965)

Let $\ell \geq 1$. Write $[t] = t \bmod \ell$.
At time t, the reward and the transition used are $r_{[t]}$ and $p_{[t]}$:

- $r_0, r_1, \ldots, r_{\ell-1}$, with $r_i : X \times A \to \mathbb{R}$
- $p_0, p_1, \ldots, p_{\ell-1}$, with $p_i : X \times A \to \Delta_X$

An ℓ-periodic MDP is an MDP on

$$X \times \{0, 1, \ldots, \ell - 1\} \quad = \quad X_0 \cup X_1 \cup \cdots \cup X_{\ell-1}$$

$\Rightarrow$ There exists a deterministic ℓ-periodic optimal policy

$$\pi^* = (\pi_0^*, \pi_1^*, \ldots, \pi_{\ell-1}^*)$$

and the optimal decisions are $a_t = \pi_{[t]}^*(x_t)$.

ℓ-periodic MDP (Riis, 1965)

Bellman operators: for all $i \in \{0, 1, \ldots, \ell - 1\}$,

- $T_{i,\pi} v = r_i + \gamma P_{i,\pi} v$
- $T_i v = \max_\pi T_{i,\pi} v$
- $\pi \in \mathcal{G}_i v \quad \Leftrightarrow \quad T_{i,\pi} v = T_i v$

For all t, for all periodic policy $\pi = (\pi_0, \pi_1, \ldots \pi_{\ell-1})$, the value when starting at time t satisfies

$$v_{[t],\pi} = T_{[t],\pi_{[t]}} T_{[t+1],\pi_{[t+1]}} \ldots T_{[t+\ell-1],\pi_{[t+\ell-1]}} v_{[t],\pi}$$

The optimal value function $v^* = (v^*_0, v^*_1, \ldots, v^*_{\ell-1})$ satisfies

$$\forall t, \quad v^*_{[t]} = T_{[t]} v^*_{[t+1]}$$

and an optimal policy $v^* = (\pi^*_0, \pi^*_1, \ldots, \pi^*_{\ell-1})$ is such that $\pi^*_{[t]} \in \mathcal{G}_{[t]} v^*_{[t+1]}$.

Dynamic Programming Algorithms

Store in memory:
$$\begin{aligned} \pi &= (\pi_0, \quad \pi_1, \quad \dots, \quad \pi_{\ell-2}, \quad \pi_{\ell-1}) \\ v &= (v_0, \quad v_1, \quad \dots, \quad v_{\ell-2}, \quad v_{\ell-1}) \end{aligned}$$

Value Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow T_{[-k]} v_{[-k+1]} = T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Policy Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \cdots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{\infty} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Modified Policy Iteration ($m \geq 0$)

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \cdots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{m} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Dynamic Programming Algorithms

Store in memory:
$$\pi = (\pi_0, \quad \pi_1, \quad \ldots, \quad \pi_{\ell-2}, \quad \pi_{\ell-1})$$
$$v = (v_0, \quad v_1, \quad \ldots, \quad v_{\ell-2}, \quad v_{\ell-1})$$

Value Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow T_{[-k]} v_{[-k+1]} = T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Policy Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \cdots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{\infty} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Modified Policy Iteration ($m \geq 0$)

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \cdots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{m} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Dynamic Programming Algorithms

Store in memory:
$$\pi = (\pi_0, \quad \pi_1, \quad \ldots, \quad \pi_{\ell-2}, \quad \pi_{\ell-1})$$
$$v = (v_0, \quad v_1, \quad \ldots, \quad v_{\ell-2}, \quad v_{\ell-1})$$

Value Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow T_{[-k]} v_{[-k+1]} = T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Policy Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \ldots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{\infty} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Modified Policy Iteration ($m \geq 0$)

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \ldots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{m} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Dynamic Programming Algorithms

Store in memory:
$$\begin{aligned} \pi &= (\pi_0, \quad \pi_1, \quad \dots, \quad \pi_{\ell-2}, \quad \pi_{\ell-1}) \\ v &= (v_0, \quad v_1, \quad \dots, \quad v_{\ell-2}, \quad v_{\ell-1}) \end{aligned}$$

Value Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow T_{[-k]} v_{[-k+1]} = T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Policy Iteration

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \dots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{\infty} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Modified Policy Iteration ($m \geq 0$)

$$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$$
$$v_{[-k]} \leftarrow v_{[-k],\pi} = \left(T_{[-k],\pi_{[-k]}} \dots T_{[-k+l-1],\pi_{[-k+l-1]}}\right)^{m} T_{[-k],\pi_{[-k]}} v_{[-k+1]}$$

Outline

1. Markov Decision Processes
2. Periodic Markov Decision Processes
3. **Approximate Dynamic Programming**

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:

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

- "$\pi(\cdot) \leftarrow [\mathcal{G} f](\cdot)$" approximated by (cost-sensitive) classification

$$\min_{\pi \in \Pi \subset A^X} \sum_x \mu(x) \left(\max_a [T_a f](x) - [T_\pi f](x) \right)$$

Approximate Algorithms (stationary)

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

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

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 (stationary)

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

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

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 (stationary)

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

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

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 (periodic)

App. Value Iteration

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow T_{[-k],\pi_{[-k]}} v_{[-k+1]} + \epsilon_k$

App. Policy Iteration

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow v_{[-k],\pi} + \epsilon_k$

App. Modified Policy Iteration ($0 \leq m$)

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow \left(T_{[-k],\pi_{[-k]}} \dots T_{[-k+l-1],\pi_{[-k+l-1]}} \right)^m T_{[-k],\pi_{[-k]}} v_{[-k+1]} + \epsilon_k$

Theorem

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. Asymptotically, the loss due to running the policy π produced instead of the optimal policy π_* satisfies:

$$\forall 0 \leq i < \ell, \quad \|v_{i,\pi^*} - v_{i,\pi}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

Approximate Algorithms (periodic)

App. Value Iteration

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow T_{[-k],\pi_{[-k]}} v_{[-k+1]} + \epsilon_k$

App. Policy Iteration

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow v_{[-k],\pi} + \epsilon_k$

App. Modified Policy Iteration ($0 \leq m$)

$\pi_{[-k]} \leftarrow \mathcal{G}_{[-k]} v_{[-k+1]}$
$v_{[-k]} \leftarrow (T_{[-k],\pi_{[-k]}} \dots T_{[-k+l-1],\pi_{[-k+l-1]}})^m T_{[-k],\pi_{[-k]}} v_{[-k+1]} + \epsilon_k$

Theorem

Assume $\|\epsilon_k\|_\infty \leq \epsilon$. Asymptotically, the loss due to running the policy π produced instead of the optimal policy π_* satisfies:

$$\forall 0 \leq i < \ell, \quad \|v_{i,\pi^*} - v_{i,\pi}\|_\infty \leq \frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon.$$

The non-stationary trick

- The bigger the period ℓ, the better the bound $\frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon$
- Any stationary MDP is a ℓ-periodic MDP for any $\ell \geq 1$
  (Scherrer & Lesner, 2012; Lesner & Scherrer, 2015; Perolat *et al.* , 2016)
- Any ℓ-periodic MDP is a $\ell\ell'$-periodic MDP for any $\ell' \geq 1$

The non-stationary trick

- The bigger the period ℓ, the better the bound $\frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon$
- Any stationary MDP is a ℓ-periodic MDP for any $\ell \geq 1$
  (Scherrer & Lesner, 2012; Lesner & Scherrer, 2015; Perolat *et al.* , 2016)
- Any ℓ-periodic MDP is a $\ell\ell'$-periodic MDP for any $\ell' \geq 1$

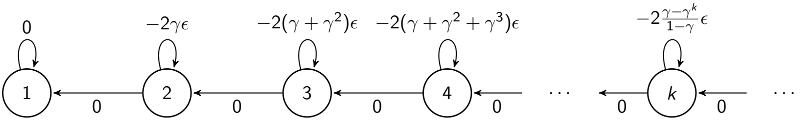
The non-stationary trick

- The bigger the period ℓ, the better the bound $\frac{2\gamma}{(1-\gamma^\ell)(1-\gamma)}\epsilon$

- Any stationary MDP is a ℓ-periodic MDP for any $\ell \geq 1$
(Scherrer & Lesner, 2012; Lesner & Scherrer, 2015; Perolat *et al.* , 2016)

- Any ℓ-periodic MDP is a $\ell\ell'$-periodic MDP for any $\ell' \geq 1$

Tightness of the bound for $m = 0, \ell = 1$

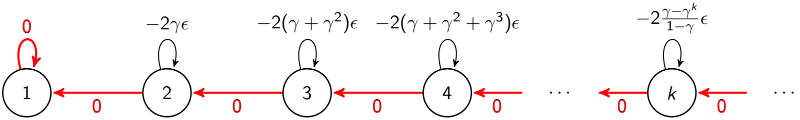


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$

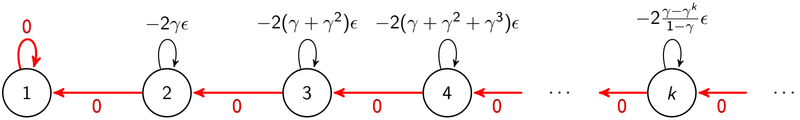


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$

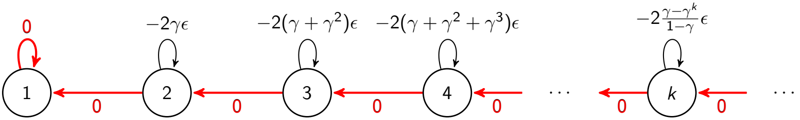


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$



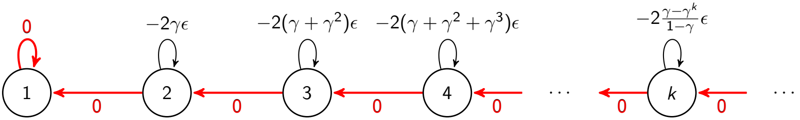
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$



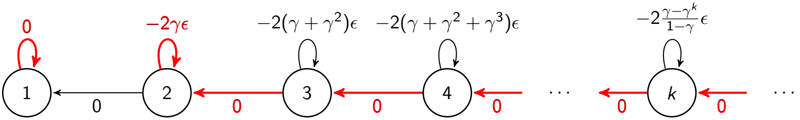
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$



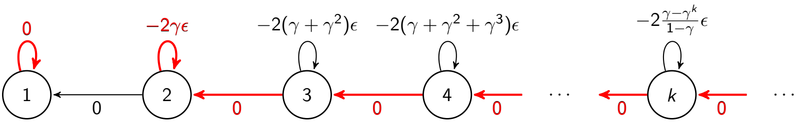
	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$

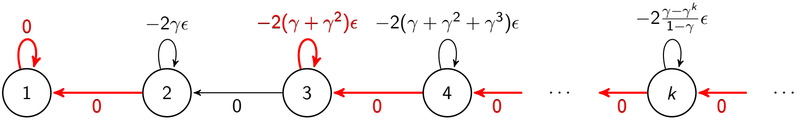


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon - \gamma\epsilon$	$\epsilon + \gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon - \gamma\epsilon - \gamma^2\epsilon$	$\epsilon + \gamma\epsilon + \gamma^2\epsilon$	...
...	...	...	...	...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

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 $m = 0, \ell = 1$

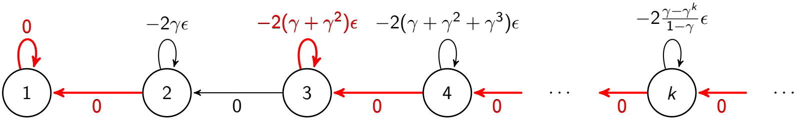


	1	2	3	4	. . .
v_0	0	0	0	0	. . .
v_1	$-\epsilon$	ϵ	0	0	. . .
v_2	$-\gamma\epsilon$	$-\epsilon - \gamma\epsilon$	$\epsilon + \gamma\epsilon$	0	. . .
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon - \gamma\epsilon - \gamma^2\epsilon$	$\epsilon + \gamma\epsilon + \gamma^2\epsilon$	. . .
. . .	. .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

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 $m = 0, \ell = 1$

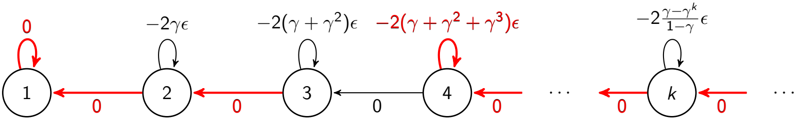


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$

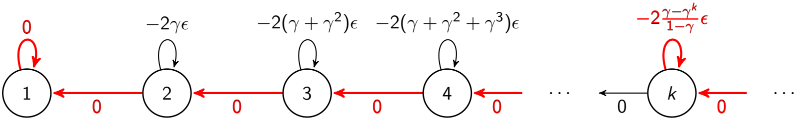


	1	2	3	4	...
v_0	0	0	0	0	...
v_1	$-\epsilon$	ϵ	0	0	...
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	...
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	...
...	...	...	...	...	...

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 $m = 0, \ell = 1$

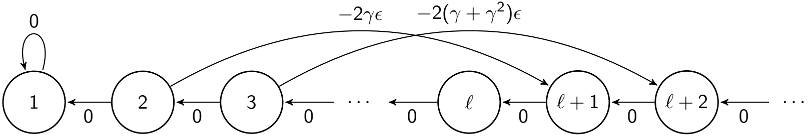


	1	2	3	4	…
v_0	0	0	0	0	…
v_1	$-\epsilon$	ϵ	0	0	…
v_2	$-\gamma\epsilon$	$-\epsilon-\gamma\epsilon$	$\epsilon+\gamma\epsilon$	0	…
v_3	$-\gamma^2\epsilon$	$-\gamma^2\epsilon$	$-\epsilon-\gamma\epsilon-\gamma^2\epsilon$	$\epsilon+\gamma\epsilon+\gamma^2\epsilon$	…
…	…	…	…	…	…

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, ADP generates a sequence of policies $(\pi_{[-k]})_{k\geq 1}$ such that $\pi_{[-k]}$ acts optimally except in state k.
Thus, the resulting policy $\pi = (\pi_0, \ldots, \pi_{\ell-1})$ gets stuck in the loop

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

and therefore

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

Simulations

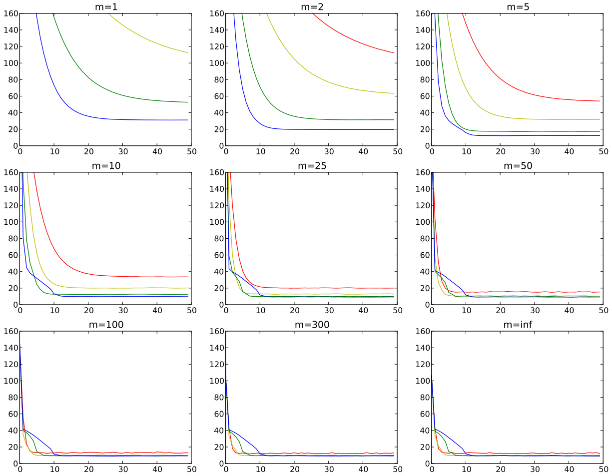


Figure: Average error of policy π per iteration k ADP. $\ell = 1$, $\ell = 2$, $\ell = 5$, $\ell = 10$.

Conclusion

This talk

- Periodic Markov Decision Processes
- The bigger the period, the better the (tight) performance guarantee
- (Stationary) MDPs are also periodic MDPs

Beyond deterministic stationary policies

- periodic deterministic policies
- probabilistic policies (Conservative Policy Iteration) (Kakade & Langford, 2002)

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.

Kakade, S.M., & Langford, J. 2002.
Approximately Optimal Approximate Reinforcement Learning.
Pages 267–274 of: International Conference on Machine Learning.

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

Perolat, Julien, Piot, Bilal, Scherrer, Bruno, & Pietquin, Olivier. 2016 (May).
On the Use of Non-Stationary Strategies for Solving Two-Player Zero-Sum Markov Games.
In: 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016).

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).

References II

Riis, Jens Ove. 1965.
Discounted markov programming in a periodic process.
Operations research, **13**(6), 920–929.

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.

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)

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

2. Sampling Scheme: play games