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

1 Optimal control at horizon $H < \infty$

1.1 Problem

$$x_{t+1} = f_t(x_t, a_t, w_t), \qquad t = 0, 1, ..., H-1$$

x_t: state, a_t: action, w_t: random (mutually independent)

$$\mathbb{P}(x_{t+1} = x' | x_t = x, a_t = a) = \mathbb{P}(x_{t+1} = x' | \mathcal{F}_t)$$

Performance measure:

$$v(x) = \mathbb{E}\left[\sum_{t=0}^{H-1} r_t(x_t, a_t, w_t) + R(x_H) | x_0 = x\right]$$

Goal: find a policy (control law) $a_t = \pi_t(x_t)$ that maximizes $v(x)$.

$$\pi = (\pi_0, \pi_1, \ldots, \pi_{H-1})$$

Remarks: closed-loop control, Markov Decision Process
...EXAMPLE...

1.2 Algorithms

...PROOFS+NUMERICAL ILLUSTRATION...

$$v_{\pi,s} = T_{\pi_s,s} v_{\pi,s+1} \quad \text{with: } \forall v, x, [T_{\pi_s,s} v](x) = \underbrace{\mathbb{E}\left[r_s(x), \pi_s(x), w_s\right]}_{r_s(x,\pi_s(x))} + \sum_y p_s(y|x, \pi_s(x)) v(y)$$

$$v_{*,s} = T_s v_{*,s+1} = \max_\pi T_{\pi,s} v_{*,s} \quad \text{with: } \forall v, x, [T_s v](x) = \max_a (r_s(x,a) + \sum_y p_s(y|x,a) v(y))$$

$$\pi_{*,s} = \arg\max_\pi T_{\pi,s} v_{*,s} = \mathcal{G}_s v_{*,s}$$

2 Stationary Optimal control at horizon $H = \infty$

2.1 Problem

$$x_{t+1} = f(x_t, a_t, w_t), \qquad t = 0, 1, ...,$$

w_t: i.i.d.
Performance measure:

$$v(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t, w_t) | x_0 = x\right]$$

$\gamma \in [0, 1[$: discount factor
...EXAMPLE...

2.2 Algorithms

...PROOFS...

$$v_\pi \;=\; T_\pi v_\pi \quad \text{with: } \forall v, x, [T_\pi v](x) \;=\; r(x, \pi(x)) + \sum_y p(y|x, \pi(x)) v(y)$$

$$v_* \;=\; T v_* \;=\; \max_\pi T_\pi v_* \quad \text{with: } \forall v, x, [T v](x) \;=\; \max_a (r(x, a) + \sum_y p(y|x, a) v(y))$$

$$\pi_* = \arg\max_\pi T_\pi v_* \;=\; \mathcal{G} v_*$$

- T_π and T are γ-contraction mappings for $\|.\|_\infty$:

$$\forall v, v', \quad \|Tv - Tv'\| \leq \gamma \|v - v'\|_\infty.$$

- $\pi_0^\infty = (\underbrace{\pi_0, \ldots, \pi_{H-1}}_{\pi_0^{H-1}}, \ldots)$

$$v_{\pi_0^H} = T_{\pi_0} T_{\pi_1} \ldots T_{\pi_{H-1}} 0 \quad \stackrel{H \to \infty}{\longrightarrow} v_{\pi_0^\infty}$$

- $$v_* = \max_{\pi_0^\infty} v_{\pi_0^\infty} = \max_{\pi_0^\infty} \lim_{H \to \infty} v_{\pi_0}^H \stackrel{(*)}{=} \lim_{H \to \infty} \max_{\pi_0^H} v_{\pi_0^H} = \lim_{H \to \infty} T^H 0.$$

...NUMERICAL ILLUSTRATION...
Thm (Bellman, 1957): There exists an optimal policy that is stationary and

$$\pi \text{ optimal} \Leftrightarrow \pi \in \mathcal{G} v_* \Leftrightarrow T_\pi v_* = T v_* = v_*$$

Proof:
$\Leftarrow$: Assume $v_* = T_\pi v_*$. As $v_\pi = T_\pi v_\pi$ and T_π has a unique fixed point, then $v_\pi = v_*$, hence π is optimal.
$\Rightarrow$: Assume $v_\pi = v_*$. We have $T v_* = v_* = v_\pi = T_\pi v_\pi = T_\pi v_*$.

3 Large scale problems

3.1 Approx. Value Iteration

$$v_{k+1} \leftarrow \mathcal{A}Tv_k = Tv_k + \epsilon_k$$

...NUMERICAL ILLUSTRATION...
Thm: Singh & Yee (95) Gordon (95), Bertsekas & Tsitsiklis (95) : If $\|\epsilon_k\| \leq \epsilon$, then

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

3.2 The non-stationary trick

$$\begin{array}{cccccc} v_0 & v_1 & \dots & \pi_{k-\ell} & \dots & v_{k-1} \\ \pi_1 & \pi_1 & \dots & \pi_{k-\ell+1} & \dots & \pi_k \end{array}$$

$$\pi_{k,\ell} = (\pi_k, \pi_{k-1}, \dots, \pi_{k-\ell+1}, \pi_k, \dots)$$

Thm: Scherrer & Lesner (12): Under the same conditions,

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

Interpretation: solving a ℓ periodic problem, less sensitive to perturbations.
Extensions: other algorithms (PI), 2 player-games (min max)