The suboptimality of μ_T^* with respect to μ^* is bounded

Damien Ernst, François Rozet and Valentin Vermeylen

February 11, 2021

University of Liège – School of Engineering

Environment

Let consider a deterministic environment described by the dynamics

$$f : X \times U \mapsto X$$

and the reward function

$$r : X \times U \mapsto \mathbb{R}^+$$

where X is the state space and U the action space.

Return J^μ

The return $J^\mu : X \mapsto \mathbb{R}^+$ of a stationary policy $\mu : X \mapsto U$ is defined as

$$J^\mu(x) = \lim_{T \to \infty} \sum_{t=0}^{T} \gamma^t r(x_t, u_t) \tag{1}$$

with $u_t = \mu(x_t)$, $x_{t+1} = f(x_t, u_t)$ and $x_0 = x$. An interesting property is that

$$\|J^\mu\|_\infty \leq \frac{B_r}{1-\gamma} \tag{2}$$

where $B_r = \|r\|_\infty$ and iff $\gamma \in [0; 1)$. Indeed,

$$\lim_{T \to \infty} \sum_{t=0}^{T} \gamma^t = \lim_{T \to \infty} \frac{1 - \gamma^{T+1}}{1 - \gamma} = \frac{1}{1-\gamma}$$

$\|F\|_\infty = \sup_{y \in Y} |F(y)|, \quad \forall F : Y \mapsto \mathbb{R}$

Truncated return J_T^μ

We define the truncated return $J_T^\mu : X \mapsto \mathbb{R}^+$ as the recurrence

$$J_T^\mu(x) = r(x, \mu(x)) + \gamma J_{T-1}^\pi(f(x, \mu(x))), \quad \forall T \geq 1 \qquad (3)$$

with $J_0^\mu(x) \equiv 0$.

By definition, we have

$$J^\mu(x) = \lim_{T \to \infty} J^\mu_T(x) \tag{4}$$

for all $x \in X$. Similarly, we can also write

$$J^\mu(x) = J^\mu_T(x) + \gamma^T J^\mu(x_T) \tag{5}$$

which leads, using (2),

$$\|J^\mu - J^\mu_T\|_\infty \leq \gamma^T \|J^\mu\|_\infty \leq \frac{\gamma^T B_r}{1-\gamma} \tag{6}$$

State-action value function Q_T

We define the state-action value $Q_T : X \times U \mapsto \mathbb{R}^+$ by the recurrence

$$Q_T(x, u) = r(x, u) + \gamma \max_{u' \in U} Q_{T-1}(f(x, u), u'), \quad \forall T \geq 1 \quad (7)$$

with $Q_0(x) \equiv 0$. Similarly to J^μ, we have

$$Q(x, u) = \lim_{T \to \infty} Q_T(x, u) \quad (8)$$

which is the state-action value over an infinite number of steps.

Optimal policy μ^*

A stationary policy μ^* is optimal iff it selects an optimal action when there remains an infinite number of steps.

$$\mu^*(x) \in \arg\max_{u \in U} Q(x, u) \tag{9}$$

or, equivalently,

$$Q(x, \mu^*(x)) = \max_{u \in U} Q(x, u)$$

Thus,

$$\begin{aligned} Q(x, \mu^*(x)) &= r(x, \mu^*(x)) + \gamma \max_{u \in U} Q(f(x, \mu^*(x)), u) \\ &= r(x, \mu^*(x)) + \gamma Q(x', \mu^*(x')) \\ &= r(x, \mu^*(x)) + \gamma r(x', \mu^*(x')) + \gamma^2 Q(x'', \mu^*(x'')) \\ &= J^{\mu^*}(x) \end{aligned}$$

with $x' = f(x, \mu^*(x))$, $x'' = f(x', \mu^*(x'))$, ...

T-optimal policy μ_T^*

In contrary, a stationary policy is T-optimal if it selects an optimal action when there remains exactly T steps.

$$\mu_T^*(x) \in \arg\max_{u \in U} Q_T(x, u) \tag{10}$$

or, equivalently,

$$Q_T(x, \mu_T^*(x)) = \max_{u \in U} Q_T(x, u)$$

Necessarily, μ_T^* is suboptimal with respect to μ^*, *i.e.*

$$J^{\mu^*}(x) \geq J^{\mu_T^*}(x) \tag{11}$$

Optimal truncated return $J_T^{\pi^*}$

However, choosing $u_t = \mu^*_{T-t}(x_t)$ for all $t \leq T$ is optimal. Then, we define

$$J_T^{\pi^*}(x) = \sum_{t=0}^{T} \gamma^t r(x_t, \mu^*_{T-t}(x_t)) \tag{12}$$

or, recurrently,

$$J_T^{\pi^*}(x) = r(x, \mu^*_T(x)) + \gamma J_{T-1}^{\pi^*}(f(x, \mu^*_T(x))) \tag{13}$$

with $J_0^{\pi^*}(x) \equiv 0$. Interestingly,

$$J_T^{\pi^*}(x) = \max_{u \in U} Q_T(x, u) \geq J_T^{\mu^*}(x)$$

The notation π indicates a non-stationary policy.

Summary

- Optimal policy μ^*
- T-optimal policy μ_T^*
- Return of the optimal policy J^{μ^*}
- Return of the T-optimal policy $J^{\mu_T^*}$
- Truncated return of the optimal policy $J_T^{\mu^*}$
- Optimal truncated return $J_T^{\pi^*}$

Theorem

The suboptimality of μ_T^* with respect to μ^* is bounded.

$$\left\| J^{\mu^*} - J^{\mu_T^*} \right\|_\infty \leq \frac{2\gamma^T B_r}{(1-\gamma)^2} \tag{14}$$

By definition and given (13), we have

$$J^{\mu}(x) = r(x, \mu(x)) + \gamma J^{\mu}(f(x, \mu(x)))$$

$$\begin{aligned} J^{\pi^*}_{T}(x) &= r(x, \mu^*_T(x)) + \gamma J^{\pi^*}_{T-1}(f(x, \mu^*_T(x))) \\ &\geq r(x, \mu^*(x)) + \gamma J^{\pi^*}_{T-1}(f(x, \mu^*(x))) \end{aligned}$$

Therefore,

$$\begin{aligned} J^{\mu^*}(x) - J^{\mu^*_T}(x) \leq\ & J^{\mu^*}(x) - \left[r(x, \mu^*(x)) + \gamma J^{\pi^*}_{T-1}(f(x, \mu^*(x))) \right] \\ & + \left[r(x, \mu^*_T(x)) + \gamma J^{\pi^*}_{T-1}(f(x, \mu^*_T(x))) \right] - J^{\mu^*_T}(x) \\ \leq\ & \gamma \left[J^{\mu^*}(f(x, \mu^*(x))) - J^{\pi^*}_{T-1}(f(x, \mu^*(x))) \right] \\ & + \gamma \left[J^{\pi^*}_{T-1}(f(x, \mu^*_T(x))) - J^{\mu^*_T}(f(x, \mu^*_T(x))) \right] \end{aligned}$$

Proof ii

Thus, in norm,

$$\begin{aligned}
\left\| J^{\mu^*} - J^{\mu^*_T} \right\|_\infty &\leq \gamma \left\| J^{\mu^*} - J^{\pi^*}_{T-1} \right\|_\infty + \gamma \left\| J^{\pi^*}_{T-1} - J^{\mu^*_T} \right\|_\infty \\
&\leq \gamma \left\| J^{\mu^*} - J^{\pi^*}_{T-1} \right\|_\infty \\
&+ \gamma \left\| J^{\pi^*}_{T-1} - J^{\mu^*} + J^{\mu^*} - J^{\mu^*_T} \right\|_\infty \\
&\leq 2\gamma \left\| J^{\mu^*} - J^{\pi^*}_{T-1} \right\|_\infty + \gamma \left\| J^{\mu^*} - J^{\mu^*_T} \right\|_\infty \\
&\leq \frac{2\gamma}{1-\gamma} \left\| J^{\mu^*} - J^{\pi^*}_{T-1} \right\|_\infty
\end{aligned}$$

But since

$$\begin{aligned}
J^{\mu^*}(x) - J^{\pi^*}_T(x) &= J^{\mu^*}_T(x) + \gamma^T J^{\mu^*}(x_t) - J^{\pi^*}_T(x) \\
&\leq \gamma^T J^{\mu^*}(x_t)
\end{aligned}$$

We have

$$\left\| J^{\mu^*} - J^{\mu^*_T} \right\|_\infty \leq \frac{2\gamma}{1-\gamma} \left\| J^{\mu^*} - J^{\pi^*}_{T-1} \right\|_\infty$$
$$\leq \frac{2\gamma}{1-\gamma} \gamma^{T-1} \left\| J^{\mu^*} \right\|_\infty$$
$$\leq \frac{2\gamma^T}{1-\gamma} \frac{B_r}{1-\gamma}$$

∎

References i

Damien Ernst, Mevludin Glavic, Florin Capitanescu, and Louis Wehenkel. “Reinforcement learning versus model predictive control: a comparison on a power system problem”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 517–529.