A PI algorithm for deterministic MDPs

Bruno Scherrer

March 22, 2022

1 variable step

For any policy π such that $\mathbb{1}_x (P_\pi)^c = \mathbb{1}_x$, then

$$\begin{aligned} v_\pi(x) - v(x) &= \mathbb{1}_x (I - (\gamma P_\pi)^c)^{-1}((T_\pi)^c v - v) \\ &\leq \frac{\mathbb{1}_x}{1-\gamma^c}(T^c v - v). \end{aligned}$$

Therefore, if we find a set of policies such that

$$\begin{aligned} T_{\pi_1} \dots T_{\pi_c} v &= T^c v, \\ \mathbb{1}_x P_{\pi_1} \dots P_{\pi_c} &= \mathbb{1}_x, \end{aligned}$$

then with $\vec{\pi} = \pi_1 \dots \pi_n$,

$$v_{\vec{\pi}}(x) \geq v_\pi(x).$$

So in n^2 iterations, we can find the optimal policy of a deterministic MDP.

2 n-step

μ_* against $\vec{\nu}$ enters a cycle in some state y after p steps. Let $\vec{\nu}_c$ be the subpart of $\vec{\nu}$ involved. Therefore:

$$\begin{aligned} v_{\mu_*}(y) - v(y) &= v_{\mu_*, \vec{\nu}_c}(y) - v(y) \\ &= \frac{\mathbb{1}_y}{1-\gamma^c}(T_{\mu_*, \vec{\nu}_c} v - v) \\ &\leq \frac{\mathbb{1}_y}{1-\gamma^c}(T^c v - v) \\ &\leq \frac{\mathbb{1}_y}{1-\gamma^c}(T^n v - v). \end{aligned}$$

If $\mathbb{1}_y(T^n v - v) = 0$, and y appears on cycle of the play $(\mu_*, \vec{\nu})$, then $v_{\mu_*}(y) = T^n v(y)$.

If v is not optimal on cycles, then there exits y that appears on a cycle in the play $(\mu_*, \vec{\nu}_c)$ and such that $\mathbb{1}_y(T^n v - v) > 0$.

Suppose I try to improve some policy with value v.

$$\begin{aligned} v_{\vec{\mu}'} - v &= \min_{\vec{\nu}'} v_{\vec{\mu}', \vec{\nu}'} - v \\ &= \min_{\vec{\nu}'} (I - \gamma^n P_{\vec{\mu}', \vec{\nu}'})^{-1} (T_{\vec{\mu}', \vec{\nu}'} v - v) \\ &\geq \min_{\vec{\nu}'} (I - \gamma^n P_{\vec{\mu}', \vec{\nu}'})^{-1} (T^n v - v) \end{aligned}$$