VALUE ITERATION, POLICY ITERATION: PRELIMINARY RESULTS

Recall

$$V^*(x) = \min_{\mu} \left[\sum_{t=0}^{+\infty} \gamma^t \mathbb{E}\left[g(x_t, \mu(x_t), w_t) \right] \right] \qquad x_0 = x$$

↳ optimal policy is stationary

$$V^{\mu}(x) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}\left[g(x_t, \mu(x_t), w_t) \right]$$

$$\min_u \mathbb{E}\left[g(x,u,w) + \gamma V^*(f(x,u,w)) \right]$$

we know that (Bellman's equation)

$$V^*(x) = \min_u \left[\mathbb{E}\left[g(x,u,w_0) \right] + \gamma \sum_{i=1}^{N} p(i \mid x,u) \, V^*(i) \right] \qquad \forall x \in \mathcal{X}$$

$$\Rightarrow T^*[V] : \mathbb{R}^N \to \mathbb{R}^N$$

$$\min_u \mathbb{E}\left[g(x,u,w) + \gamma V(f(x,u,w)) \right]$$

$$T^*[V]_x = \min_u \left[\mathbb{E}\left[g(x,u,w_0) \right] + \gamma \sum_{i=1}^{N} p(i \mid x,u) \, V(i) \right] \qquad \forall x \in \mathcal{X}$$

V^* = solution to $V = T^*[V]$ (fixed point)

similarly, $\forall x \in \mathcal{X}$

$$V^{\mu}(x) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}\left[g(x_t, \mu(x_t), w_t) \right] = \mathbb{E}\left[g(x, \mu(x), w_0) \right] + \gamma \sum_{t=1}^{+\infty} \gamma^{t-1} \mathbb{E}\left[g(x_t, \mu(x_t), w_t) \right] \qquad (\tau = t-1)$$

$$= \mathbb{E}\left[g(x, \mu(x), w_0) + \gamma V^{\mu}(x_1) \right] = \mathbb{E}\left[g(x, \mu(x), w) \right] + \gamma \sum_{i=1}^{N} p(i \mid x, \mu(x)) \, V^{\mu}(i) \qquad (x_1 = f(x, \mu(x), w_1))$$

$$\Rightarrow T^{\mu}[V] : \mathbb{R}^N \to \mathbb{R}^N \qquad T^{\mu}[V]_x = \mathbb{E}\left[g(x, \mu(x), w_0) \right] + \gamma \sum_{i=1}^{N} p(i \mid x, \mu(x)) \, V(i) \qquad x \in \mathcal{X}$$

$$\mathbb{E}\left[g(x, \mu(x), w) + \gamma V(f(x, \mu(x), w)) \right]$$

V^{μ} = solution to $V = T^{\mu}[V]$ (fixed point)

result $V = T^*[V]$ and $V = T^{\mu}[V]$ ... solution exist and is unique ⇐ Banach-Caccioppoli theorem (the tool)

(S, d) = complete metric space $\quad F : S \to S$ contraction mapping

i.e. $\exists \rho < 1 : \; d(F(s_1), F(s_2)) \leq \rho \cdot d(s_1, s_2) \quad \forall s_1, s_2 \in S$ ↳ clearly this implies continuity

$s = F(s)$ always admits one and only one solution

Proof: uniqueness

By contradiction, suppose that $s_1 = F(s_1)$ and $s_2 = F(s_2)$. Then

$$d(s_1, s_2) = d(F(s_1), F(s_2)) \leq \rho\, d(s_1, s_2) < d(s_1, s_2) \quad \text{... contradiction!}$$
$$\rho < 1$$

existence

let s_0 be an arbitrary point in S. and define

$$s_{k+1} = F(s_k) = F^{k+1}(s_0) \qquad k = 0, 1, 2, \dots$$

Then $d(s_{k+1}, s_k) = d(F(s_k), F(s_{k-1})) \leq \rho\, d(s_k, s_{k-1}) \leq \dots \leq \rho^k d(s_1, s_0)$

For any $n > m$ (triangle ineq.)

$$d(s_m, s_n) \leq d(s_n, s_{n-1}) + d(s_{n-1}, s_n) \leq d(s_n, s_{n-1}) + d(s_{n-1}, s_{n-2}) + d(s_{n-2}, s_n) \leq \dots$$
$$\dots \leq \sum_{k=m}^{n-1} d(s_{k+1}, s_k) \leq \sum_{k=m}^{n-1} \rho^k d(s_1, s_0) \leq \sum_{k=m}^{\infty} \rho^k d(s_1, s_0) = \frac{\rho^m}{1-\rho} d(s_1, s_0)$$

so for any $n > m$, $d(s_m, s_n) \xrightarrow[m \to \infty]{} 0$, i.e. $\{s_m\}$ is Cauchy convergent

completeness $\Rightarrow \exists\, \bar{s} : s_m \xrightarrow[m \to \infty]{} \bar{s}$

however, $\bar{s} = \lim_{k \to \infty} s_{k+1} = \lim_{k \to \infty} F(s_k) = F(\bar{s})$ i.e. $\bar{s}$ is a fixed point.

QED

s_0 = arbitrary
$s_{k+1} = F(s_k)$ $\Rightarrow$ fixed-point iteration (subsequent approximations) $\rightsquigarrow$ algorithm to find fixed point $\bar{s} = F(\bar{s})$

$$d(s_{k+1}, \bar{s}) = d(F(s_k), F(\bar{s})) \leq \rho\, d(s_k, \bar{s}) \leq \dots \leq \rho^{k+1} d(s_0, \bar{s})$$ (exponential convergence)

Fundamental property

Both T^* and T^μ are contractions from $\mathbb{R}^n$ to $\mathbb{R}^n$ with respect to the max-norm distance

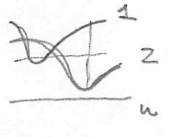
T^*: let V_1 and V_2 two value functions (i.e. two vectors in $\mathbb{R}^n$)

$$\max_{x \in X} \left| T^*[V_1] - T^*[V_2] \right| = \max_{x \in X} \left| \min_u \left[E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V_1(i) \right] + \right.$$
$$\left. - \min_u \left[E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V_2(i) \right] \right| \leq$$

$\leq$ assume $\min_u [\dots V_2(i)] \leq \min_u [\dots V_1(i)]$ and attained for u_2^*

(if $\min_u [\dots V_1(i)] \leq \min_u [\dots V_2(i)]$ evaluate everything in u_1^* where $\min_u [\dots V_1(i)]$ is attained)

replace u by u



not min → distance increases ← min

$$\leq \max_{x \in X} \left| \mathbb{E}[g(x,u_2^*,w)] + \gamma \sum_{i=1}^{m} p(i|x,u_2^*) V_1(i) - \mathbb{E}[g(x,u_2^*,w)] - \gamma \sum_{i=1}^{m} p(i|x,u_2^*) V_2(i) \right|$$

$$= \max_{x \in X} \left| \gamma \sum_{i=1}^{m} p(i|x,u_2^*) \left(V_1(i) - V_2(i) \right) \right| \leq \max_{x \in X} \gamma \sum_{i=1}^{m} p(i|x,u_2^*) \left| V_1(i) - V_2(i) \right|$$

$$\leq \max_{i \in X} \left| V_1(i) - V_2(i) \right| \cdot \max_{x \in X} \gamma \sum_{i=1}^{m} p(i|x,u_2^*) = \quad \left(\sum_{i=1}^{m} p(i|x,u_2^*) = 1\right)$$

$$= \gamma \cdot \max_{x \in X} \left| V_1(x) - V_2(x) \right|$$

T^μ: same, simpler V_1 and $V_2 \in \mathbb{R}^m$

$$\max_{x \in X} \left| T^\mu[V_1] - T^\mu[V_2] \right| =$$

$$= \max_{x \in X} \left| \mathbb{E}[g(x,\mu(x),w)] + \gamma \sum_{i=1}^{m} p(i|x,\mu(x)) V_1(i) - \mathbb{E}[g(x,\mu(x),w)] - \gamma \sum_{i=1}^{m} p(i|x,\mu(x)) V_2(i) \right|$$

$$= \max_{x \in X} \left| \gamma \sum_{i=1}^{m} p(i|x,\mu(x)) \left(V_1(i) - V_2(i) \right) \right| \leq$$

$$\leq \gamma \max_{x \in X} \sum_{i=1}^{m} p(i|x,\mu(x)) \left| V_1(i) - V_2(i) \right| \leq \gamma \max_{i \in X} \left| V_1(i) - V_2(i) \right| \cdot \max_{x \in X} \sum_{i=1}^{m} p(i|x,\mu(x)) \quad (=1)$$

$$= \gamma \max_{x \in X} \left| V_1(x) - V_2(x) \right|$$

Thus, $V = T^*[V]$ and $V = T^\mu[V]$ admits one and only one solution $\rightsquigarrow V^*(x)$ and $V^\mu(x)$

optimal policy $\Longleftrightarrow$ optimal value function $V^*(x)$

policy $^{\mu}$ iteration $\Longleftrightarrow$ value function for μ $V^\mu(x)$

} Key problem: compute $V^*(x)$, $V^\mu(x)$

VALUE ITERATION AND POLICY ITERATION ALGORITHMS

- $V^*(x) = \min_u \Big[\mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V^*(i) \Big]$ $\quad x = 1,2,\dots,n$

  piecewise linear in V^* $\Rightarrow$ system of <u>non-linear eqs</u> ↯ tough (impossible, cause of dimensionality)

- $V^\mu(x) = \mathbb{E}[g(x,\mu(x),w)] + \gamma \sum_{i=1}^{n} p(i|x,\mu(x))\, V^\mu(i)$

  $\Downarrow$ linear in V^μ $\Rightarrow$ system of <u>linear eqs</u>

$(I - \gamma \cdot \mathbb{P})\, V^\mu = \Big[\mathbb{E}[g(x,\mu(x),w)] \Big]_{x=1}^{n}$ <u>better</u> but still tough if n is large

$\Rightarrow$ use fixed-point iterations! $\Rightarrow$ <u>VALUE ITERATION</u> algorithm (VI)

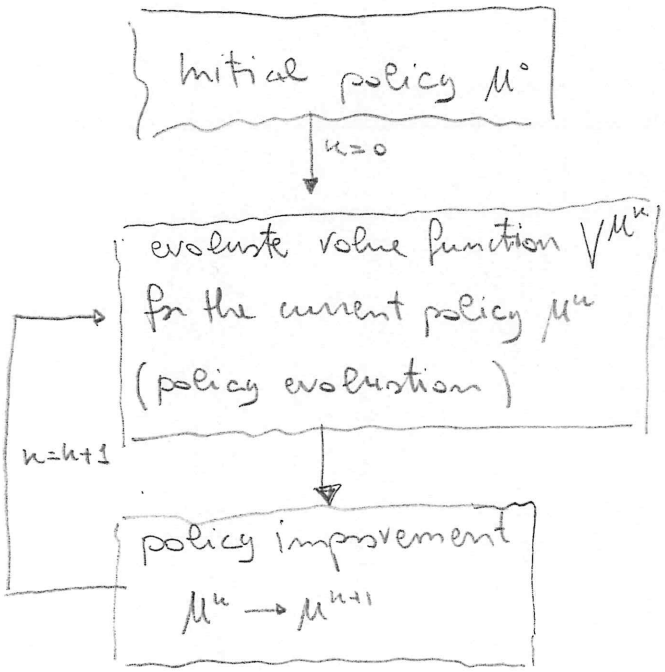
$V^0(x)$ $x \in X$ <u>arbitrarily</u> chosen (need not be a value function associated to some policy)

- $V^{k+1}(x) = T^*[V^k(x)]$ converges to $V^*(x)$
- $V^{k+1}(x) = T^\mu[V^k(x)]$ converges to $V^\mu(x)$

exponentially fast

OBS: $V^k(x) = [T^*]^k [V^0(x)]$ or $V^k(x) = [T^\mu]^k [V^0(x)]$

Alternative to VI to compute V^* $\Rightarrow$ POLICY ITERATION (PI)



- sequence of policies, improving at each iteration ($V^{\mu^{k+1}} \leq V^{\mu^k}$ hint)
- convergence in a <u>finite</u> no. of iterations ↳ can be very large
- faster convergence than VI for large state space ↳ big improvement in the first few iterations

approximate

property : MONOTONICITY

Suppose $V_1(x) \leq V_2(x) \quad \forall x \in X$. Then, for every μ we have

$T^\mu[V_1]_x \leq T^\mu[V_2]_x \quad \forall x \in X$ and $T^*[V_1]_x \leq T^*[V_2]_x \quad \forall x \in X$

Proof :

$$T^\mu[V_1]_x = \mathbb{E}[g(x,\mu(x),w)] + \gamma \sum_{i=1}^{M} p(i|x,\mu(x))\, V_1(i)$$

$$\leq \mathbb{E}[g(x,\mu(x),w)] + \gamma \sum_{i=1}^{M} p(i|x,\mu(x))\, V_2(i) = T^\mu[V_2]_x$$

$$\mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V_1(i) \leq \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V_2(i)$$

($p(i|x,u) \geq 0$) $\forall x \in X \quad \forall u \in U$. Hence,

$$T^*[V_1]_x = \min_{u \in U} \left[\mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V_1(i) \right]$$

$$\leq \min_{u \in U} \left[\mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V_2(i) \right] = T^*[V_2]_x$$

Start with an initial policy $\mu^0: X \to U$ arbitrary

Policy evaluation: solve $V^{\mu^n}(x) = \mathbb{E}\left[g(x,\mu^n(x),w)\right] + \alpha \sum_{i=1}^{M} p(i|x,\mu^n(x)) \cdot V^{\mu^n}(i)$ $\quad \forall x \in X$

↳ either solve system of linear eqs. or use VI for V^{μ^n} (exact till convergence) → solve $V = T^{\mu^n}[V]$

Policy improvement: $\mu^{n+1}(x) = \underset{u \in U}{\operatorname{argmin}}\ \mathbb{E}\left[g(x,u,w)\right] + \alpha \sum_{i=1}^{M} p(i|x,u)\, V^{\mu^n}(i)$

Properties:

1. $V^{\mu^{n+1}}(x) \leq V^{\mu^n}(x) \quad \forall x \in X$ $\qquad V^0 = T^{\mu^{n+1}}[V^{\mu^n}]$

Define $\forall x \in X$, $V^0(x) = \mathbb{E}\left[g(x,\mu^{n+1}(x),w)\right] + \alpha \sum_{i=1}^{M} p(i|x,\mu^{n+1}(x)) \cdot V^{\mu^n}(i)$

$= \min_{u \in U} \mathbb{E}\left[g(x,u,w)\right] + \alpha \sum_{i=1}^{M} p(i|x,u)\, V^{\mu^n}(i)$

$\leq \mathbb{E}\left[g(x,\mu^n(x),w)\right] + \alpha \sum_{i=1}^{M} p(i|x,\mu^n(x)) V^{\mu^n}(i) = V^{\mu^n}(x)$

so $V^0(x) \leq V^{\mu^n}(x) \quad \forall x \in X$

Then $V^{k+1} = T^{\mu^{n+1}}[V^k] \quad k = 0,1,\ldots \overset{VI}{\Rightarrow} V^k \xrightarrow[k\to\infty]{} V^{\mu^{n+1}}$

$V^1 = T^{\mu^{n+1}}[V^0] \leq$ (monotonicity, $V^0 \leq V^{\mu^n}$) $\leq T^{\mu^{n+1}}[V^{\mu^n}] = V^0$

$V^2 = T^{\mu^{n+1}}[V^1] \leq T^{\mu^{n+1}}[V^0] = V^1 \leq V^0$ (monotonicity, $V^1 \leq V^0$)

$\vdots$

$V^k \leq V^0 \leq V^{\mu^n} \Rightarrow V^{\mu^{n+1}} = \lim_{k\to\infty} V^k \leq V^{\mu^n}$

2. no. of (stationary) policies is finite ($\leq m^M$) $\Rightarrow V^{\mu^{n+1}}(x) < V^{\mu^n}(x)$ for some x

occurs for a finite no. of iterations, then for some k

$V^{\mu^{n+1}}(x) = V^{\mu^n}(x) \quad \forall x \in X \to V^{\mu^{n+1}} = V^k$ $\forall k$ in previous iterations

$\Rightarrow V^{\mu^n}(x) = V^{\mu^{n+1}}(x) = V^0(x) = \min_u \left[\mathbb{E}\left[g(x,u,w)\right] + \alpha \sum_{i=1}^{M} p(i|x,u)\, V^{\mu^n}(x)\right]$

i.e $V^{\mu^u}(x)$ satisfies the Bellman eq. for optimal value function $\Rightarrow$ $\mu^u = \mu^*$ optimal policy, $V^{\mu^u} = V^*$ optimal value function

PI converges in finite time to the optimal policy !

Q-FACTORS (action-value functions)

$$Q^* : \mathcal{X} \times \mathcal{U} \to \mathbb{R} \qquad Q^\mu : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$$

$$Q^*(x,u) = \mathbb{E}\left[g(x,u,w_0)\right] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V^*(i) \qquad x \in \mathcal{X} \quad u \in \mathcal{U}$$

$$Q^\mu(x,u) = \mathbb{E}\left[g(x,u,w_0)\right] + \gamma \sum_{i=1}^{M} p(i|x,u)\, V^\mu(i) \qquad x \in \mathcal{X} \quad u \in \mathcal{U}$$

<u>interpretation</u>: total discounted cost achieved when at $t=0$ (initial time) $x_0 = x$ and $u_t = u$ (action u is used) and then:

$*$: optimal policy μ^* is applied for $t \geq 1$

μ: policy μ is applied for $t \geq 1$

Clearly, thanks to Bellman equations

- $V^*(x) = \min_{u \in \mathcal{U}} Q^*(x,u)$ $\qquad \mu^*(x) = \operatorname{argmin}_{u \in \mathcal{U}} Q^*(x,u)$
- $V^\mu(x) = Q^\mu(x, \mu(x))$

$\rightsquigarrow$ Q^* summarizes all the info we need to determine the optimal policy

Bellman equations for Q-factors (from definitions and relations above)

$$Q^*(x,u) = \mathbb{E}\left[g(x,u,w)\right] + \gamma \sum_{i=1}^{M} p(i|x,u) \cdot \min_{v \in \mathcal{U}} Q^*(i,v)$$

solution to $Q = T^*[Q]$ where $T^*[Q] = \mathbb{E}\left[g(x,u,w)\right] + \gamma \sum_{i=1}^{M} p(i|x,u) \min_{v} Q(i,v)$

$$Q^\mu(x,u) = \mathbb{E}\left[g(x,u,w)\right] + \gamma \sum_{i=1}^{M} p(i|x,u)\, Q^\mu(i,\mu(i))$$

solution to $Q = T^\mu[Q]$ where $T^\mu[Q] = \mathbb{E}\left[g(x,u,w)\right] + \gamma \sum_{i=1}^{M} p(i|x,u)\, Q(i,\mu(i))$

property: T^* and T^μ (for Q-factors) are contractions with respect to the $\max_{x,u}$ distance

$Q_1(x,u)$, $Q_2(x,u)$ given.

$$\max_{x,u} \left| T^*[Q_1] - T^*[Q_2] \right| = \max_{x,u} \left| \cancel{E[g(x,u,w)]} + \gamma \sum_{i=1}^{n} p(i|x,u) \min_v Q_1(i,v) - \cancel{E[g(x,u,w)]} - \gamma \sum_{i=1}^{n} p(i|x,u) \min_v Q_2(i,v) \right|$$

$$\leq \max_{x,u} \gamma \sum_{i=1}^{n} p(i|x,u) \left| \min_v Q_1(i,v) - \min_v Q_2(i,v) \right|$$

$$\leq \max_{x,u} \gamma \sum_{i=1}^{n} p(i|x,u) \left| Q_1(i,v(i)) - Q_2(i,v(i)) \right| \qquad v(i) = \begin{cases} \arg\min_v Q_1(i,v) & \text{if } \min_u Q_1(i,u) \leq \min_u Q_2(i,u) \\ \arg\min_v Q_2(i,v) & \text{else} \end{cases}$$

$$\leq \max_{i,v} \left| Q_1(i,v) - Q_2(i,v) \right| \cdot \max_{x,u} \gamma \cancel{\sum_{i=1}^{n} p(i|x,u)}^{=1}$$

$$= \gamma \max_{x,u} \left| Q_1(x,u) - Q_2(x,u) \right|$$

$$\max_{x,u} \left| T^\mu[Q_1] - T^\mu[Q_2] \right| = \max_{x,u} \left| \cancel{E[\cdot]} + \gamma \sum_{i=1}^{n} p(i|x,u) Q_1(i,\mu(i)) - \cancel{E[\cdot]} - \gamma \sum_{i=1}^{n} p(i|x,u) Q_2(i,\mu(i)) \right|$$

$$\leq \max_{x,u} \gamma \cdot \sum_{i=1}^{n} p(i|x,u) \left| Q_1(i,\mu(i)) - Q_2(i,\mu(i)) \right| \overset{\text{as before}}{\leq} \cdots \leq \gamma \max_{x,u} \left| Q_1(x,u) - Q_2(x,u) \right|$$

property: T^* and T^μ (for Q-factors) are monotonic

$Q_1(x,u) \leq Q_2(x,u) \quad \forall x \in X, u \in U$ Then, $\forall x \in X, u \in U$

$$T^*[Q_1](x,u) = E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u) \min_v Q_1(i,v)$$

$$\leq E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u) \min_v Q_2(i,v) = T^*[Q_2](x,u)$$

$$T^\mu[Q_1](x,u) = E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u) Q_1(i,\mu(i))$$

$$\leq E[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u) Q_2(i,\mu(i)) = T^\mu[Q_2](x,u)$$

FACT: since T^* and T^μ are contractions, the Bellman eqs. for Q-factors univocally determine $Q^*(x,u)$ and $Q^\mu(x,u)$ always (can be proved otherwise) and shows the validity of the following Value Iteration method for Q-factors.

in a model-based environment $\Rightarrow$ V^* and Q^* are equivalent

in a model-free (RL):

$$\mu^*(x) = \operatorname*{argmin}_u Q^*(x,u) = \operatorname*{argmin}_u \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V^*(i)$$

($Q^*(x,u)$: that's all we need to compute the optimal policy; $\mathbb{E}[g(x,u,w)]$: model needed besides V^*; $p(i|x,u)$: model needed besides V^*)

Moreover,

$$Q^*(x,u) = T^*[Q^*](x,u) = \mathbb{E}\left[g(x,u,w) + \gamma \min_v Q^*(f(x,u,w), v) \right]$$

($\mathbb{E}$: expected value outside; $\min_v$: min inside $\Rightarrow$ suitable for incremental computation)

$$V^*(x) = T^*[V^*](x) = \min_u \left[\mathbb{E}[g(x,u,w) + \gamma V^*(f(x,u,w))] \right]$$

($\min_u$: outside; $\mathbb{E}$: inside $\rightsquigarrow$ not suitable for incremental computation)

VALUE ITERATION for Q-factors

$Q^0(x,u) : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ arbitrary

$$Q^{k+1}(x,u) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u) \cdot \min_v Q^k(i,v) \xrightarrow[k\to\infty]{} Q^*(x,u)$$

$$= T^*[Q^k](x,u)$$

$$Q^{k+1}(x,u) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, Q^k(i,\mu(i)) \xrightarrow[k\to\infty]{} Q^\mu(x,u)$$

$$= T^\mu[Q^k](x,u)$$

POLICY ITERATION for Q-factors

$\mu^0(x) : \mathcal{X} \to \mathcal{U}$ initial policy (arbitrary)

(either solve sys. of linear equations or use VI)

policy evaluation: solve $Q^{\mu^k}(x,u) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, Q^{\mu^k}(i,\mu^k(i))$

$$Q^{\mu^k}(x,u) = T^{\mu^k}[Q^{\mu^k}](x,u) \quad \forall x \in \mathcal{X}\ u \in \mathcal{U}$$

policy improvement: $\mu^{k+1}(x) = \operatorname*{argmin}_u Q^{\mu^k}(x,u)$

OBS: the fact that

$$Q^{\mu^k}(x,u) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, Q^{\mu^k}(i,\mu^k(i)) \qquad (\star)$$

always admit one and only one solution can be also seen by noting that

$$Q^{\mu^k}(x,\mu^k(x)) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, Q^{\mu^k}(i,\mu^k(i)) \quad \text{is the}$$

Bellman equation for value functions, which admits one and only one solution. Hence, $(\star)$ univocally determines $Q^{\mu^k}(x,\mu^k(x))$ and $Q^{\mu^k}(x,\mu^k(x)) = V^{\mu^k}(x)$ (solution to). However, once $Q^{\mu^k}(x,\mu^k(x))$ are determined, $(\star)$ determines $Q^{\mu^k}(x,u)$ for the other values of u since right-hand side of $(\star)$ becomes given.

At every policy evaluation: $Q^{\mu^k}(x,u) : Q^{\mu^k}(x,\mu^k(x)) = V^{\mu^k}(x)$

and therefore $Q^{\mu^k}(x,u) = \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V^{\mu^k}(i)$

Hence, at every policy improvement

$$\mu^{k+1}(x) = \arg\min_{u} \mathbb{E}[g(x,u,w)] + \gamma \sum_{i=1}^{n} p(i|x,u)\, V^{\mu^k}(i)$$

same as PI for value iteration

PI for Q-factor and PI for value function give the same policy at every iteration

convergence and other properties inherited