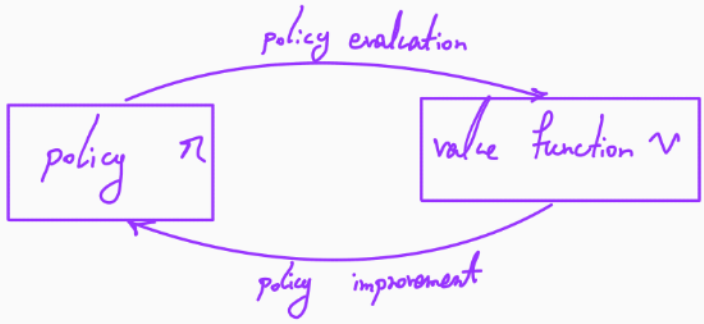
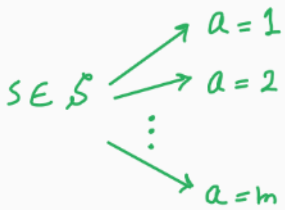
Policy Iteration

PI aims to find an (approximately) optimal policy π^* through iterative policy evaluation and policy improvement.



Policy Improvement: Suppose there exists a deterministic, stationary policy π with value function v^π.

*** Question:** Would changing the policy at a state improve the policy?



$\pi(s) = a$

Would setting $\pi(s) = a' \neq a$ be helpful?

Recall

$$Q^\pi(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s,a)} \left[v^\pi(s') \right]$$

If $\exists a \in \mathcal{A}$ s.t. $Q^\pi(s,a) > v^\pi(s)$, then changing the policy at state s to a improves the policy.

Note: If for all $s \in \mathcal{S}$, $\max_{a \in \mathcal{A}} Q^\pi(s,a) = v^\pi(s)$, then

$$v^\pi(s) = v^*(s) \quad \forall s \in \mathcal{S}$$

π an optimal policy

Policy Improvement Theorem: Let π and π' be a pair of deterministic, stationary policies such that

$$Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \qquad \forall s \in S.$$

Then, π' must be as good as or better than π, i.e.,

$$V^{\pi'} \geq V^{\pi}.$$

Policy Iteration Algorithm

- Initialize $\pi_0 : S \longrightarrow A$
- For $t = 0, 1, 2, \ldots, T-1$:

policy evaluation
- Evaluate π_t by computing V^{π_t}

policy improvement
- Improve the policy by

$$\pi_{t+1}(s) = \underset{a \in A}{\operatorname{argmax}} \left[R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^{\pi_t}(s') \right] \right] \qquad \forall s \in S$$

- Return π_T

Lemma: Let V^{π_t} represent the value function of policy π_t at iteration t of PI. It holds that

$$V^{\pi_{t+1}} \geq V^{\pi_t} \qquad \boxed{\text{Monotonic Improvement}}$$

element-wise

Proof:

$$\forall s \in S: \quad V^{\pi_{t+1}}(s) - V^{\pi_t}(s) = R(s, \pi_{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi_{t+1}(s))}\left[V^{\pi_{t+1}}(s')\right]$$

$$- \left(R(s, \pi_t(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi_t(s))}\left[V^{\pi_t}(s')\right]\right)$$

$$\geqslant R(s, \pi_{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi_{t+1}(s))}\left[V^{\pi_{t+1}}(s')\right]$$

$$- \left(R(s, \pi_{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi_{t+1}(s))}\left[V^{\pi_t}(s')\right]\right)$$

$$= \gamma \mathbb{E}_{s' \sim P(\cdot|s, \pi_{t+1})}\left[V^{\pi_{t+1}}(s') - V^{\pi_t}(s')\right]$$

$$\longrightarrow V^{\pi_{t+1}} - V^{\pi_t} \geqslant \gamma P^{\pi_{t+1}} (V^{\pi_{t+1}} - V^{\pi_t})$$

$$\geqslant \gamma P^{\pi_{t+1}} \left(\gamma P^{\pi_{t+1}} (V^{\pi_{t+1}} - V^{\pi_t})\right) = \gamma^2 (P^{\pi_{t+1}})^2 (V^{\pi_{t+1}} - V^{\pi_t})$$

$$\vdots$$

$$\geqslant \gamma^k (P^{\pi_{t+1}})^k (V^{\pi_{t+1}} - V^{\pi_t})$$

Let $k \longrightarrow \infty \implies V^{\pi_{t+1}} - V^{\pi_t} \geqslant 0 \implies V^{\pi_{t+1}} \geqslant V^{\pi_t}$ ■

Theorem: The value of the final policy V^{π_T} returned by the PI algorithm satisfies $\|V^{\pi_T} - V^*\|_\infty \leq \gamma^T \|V^{\pi_0} - V^*\|_\infty$.

Proof :

$$v^*(s) - v^{\pi_t}(s) = \max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s')\right]\right]$$

$$- \left(R(s,\pi_t(s)) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,\pi_t(s))}\left[v^{\pi_t}(s')\right]\right)$$

$$\leq \max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s')\right]\right]$$

$$- \left(R(s,\pi_t(s)) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,\pi_t(s))}\left[v^{\pi_{t-1}}(s')\right]\right)$$

$$= \max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s')\right]\right]$$

$$- \max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^{\pi_{t-1}}(s')\right]\right]$$

$$\Rightarrow \forall s\in S: \left|v^*(s) - v^{\pi_t}(s)\right| \leq \Big|\max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s')\right]\right]$$

$$- \max_{a\in A}\left[R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^{\pi_{t-1}}(s')\right]\right]\Big|$$

$$\leq \max_{a\in A}\Big|\left(R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s')\right]\right)$$

$$- \left(R(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^{\pi_{t-1}}(s')\right]\right)\Big|$$

$$= \gamma \max_{a\in A}\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[v^*(s') - v^{\pi_{t-1}}(s')\right]\right|$$

$$\leq \gamma \max_{a\in A} \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\left|v^*(s') - v^{\pi_{t-1}}(s')\right|\right]$$

$$\leq \gamma \max_{s'\in S}\left|v^*(s') - v^{\pi_{t-1}}(s')\right|$$

$$\Rightarrow \max_s \left| v^*(s) - v^{\pi_t}(s) \right| \leq \gamma \max_{s \in S} \left| v^*(s) - v^{\pi_{t-1}}(s) \right|$$

$$\begin{aligned} \Rightarrow \| v^{\pi_t} - v^* \|_\infty &\leq \gamma \| v^{\pi_{t-1}} - v^* \|_\infty \\ &\leq \gamma^2 \| v^{\pi_{t-2}} - v^* \|_\infty \\ &\vdots \\ &\leq \gamma^t \| v^{\pi_0} - v^* \|_\infty \end{aligned}$$

■

Value Iteration vs. Policy Iteration

Value Iteration	Policy Iteration
• Initialize v_0 • Each iteration: Apply Bellman optimality operator • Return v_T & π_T	• Initialize π_0 • Each iteration: Apply policy evaluation Apply policy improvement • Return π_T

search over the space of

value functions, $v \in \mathbb{R}^{\|S\|}$	deterministic stationary policies, $\|A\|^{\|S\|}$ policies

convergence rate of $\| v^{\pi_T} - v^* \|_\infty$

geometric in T	geometric in T

Computational complexity

polynomial, $O\left(\frac{\|S\|^2 \|A\|}{1-\gamma} \log\left(\frac{1}{(1-\gamma)\varepsilon}\right)\right)$	strongly polynomial, $O\left(\frac{\|S\|^4 \|A\|^2}{1-\gamma} \log\left(\frac{1}{1-\gamma}\right)\right)$

Finite-Horizon MDP

Goal: The learning agent aims to find a policy π that maximizes the expected cumulative reward over a finite horizon

$$\pi^* \in \arg\max_{\pi \in \Pi} \; \mathbb{E}_{\substack{S_0 \sim \mu_0 \\ A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \Big[\overbrace{\sum_{t=0}^{T-1} R(S_t, A_t)}^{\text{return}} \Big]$$

γ^t

Value functions

- (state) value function of a policy π at time $t \in \{0, 1, \dots, T-1\}$,

  $V_t^\pi : \mathcal{S} \longrightarrow \mathbb{R}$, is defined as

$$V_t^\pi(s) = \mathbb{E}_{\substack{A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \Big[\sum_{t'=t}^{T-1} R(S_{t'}, A_{t'}) \;\Big|\; S_t = s \Big] \qquad \forall s \in \mathcal{S}$$

$\gamma^{t'-t}$

- (state-) action value function of a policy π at time $t \in \{0, 1, \dots, T-1\}$,

  $Q_t^\pi : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$, is defined as

$$Q_t^\pi(s, a) = \mathbb{E}_{\substack{A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \Big[\sum_{t'=t}^{T-1} R(S_{t'}, A_{t'}) \;\Big|\; S_t = s, A_t = a \Big] \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

$\gamma^{t'-t}$

Theorem: There exists an optimal, deterministic, history-independent, time-varying policy π^*, i.e.,

$$\exists \pi^* = (\pi_0^*, \pi_1^*, \dots, \pi_{T-1}^*), \quad \pi_t^* : S \longrightarrow A$$

$$\text{s.t.} \quad V_t^{\pi^*}(s) \geqslant V_t^{\pi}(s) \quad \forall s \in S, \forall \pi \in \Pi$$

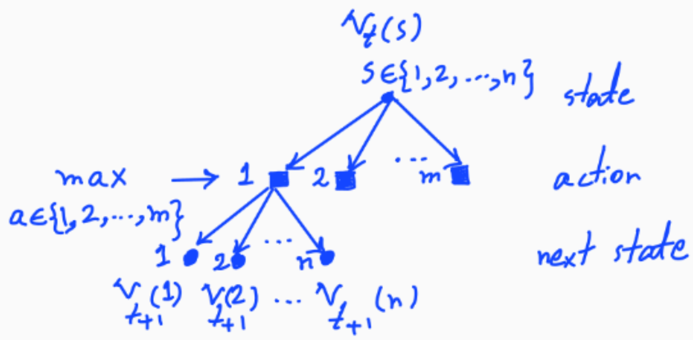
$\longrightarrow$ we look for finding an optimal policy $\pi^* = (\pi_0^*, \pi_1^*, \dots, \pi_{T-1}^*)$, which is a sequence of deterministic policies dependent on state and time.

* Bellman (consistency) Equations:

$$V_t^{\pi}(s) = \mathbb{E}_{a \sim \pi_t(s)}\left[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V_{t+1}^{\pi}(s') \right]\right]$$

$$Q_t^{\pi}(s,a) = R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V_{t+1}^{\pi}(s') \right]$$

* Bellman Optimality Equations:



$$V_t^*(s) = \max_{a \in A} \left[R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V_{t+1}^*(s') \right]\right]$$

$$Q_t^*(s,a) = R(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a' \in A} Q_{t+1}^*(s',a') \right]$$

Note: The optimal value function is also a sequence, i.e.,

$$v^* = (\underbrace{v_0^*}_{\text{objective}}, v_1^*, \dots, \underbrace{v_T^*}_{0})$$

Value Iteration (for finite-Horizon MDP)

VI aims to find the (exact) optimal value function and an optimal policy through dynamic programming (backward induction).

Iterative method:

- Initialize $v_T^* = \mathbf{0}_{|S| \times 1}$
- For $t \in \{T-1, T-2, \dots, 0\}$:

$$v_t^*(s) = \max_{a \in A} \left[R_t(s,a) + \overset{\gamma}{\mathbb{E}}_{s' \sim P_t(\cdot|s,a)} \left[v_{t+1}^*(s') \right] \right] \quad \forall s \in S$$

$$\pi_t^*(s) = \operatorname*{argmax}_{a \in A} \left[R_t(s,a) + \underset{\gamma}{\mathbb{E}}_{s' \sim P_t(\cdot|s,a)} \left[v_{t+1}^*(s') \right] \right] \quad \forall s \in S$$

if time-varying

- Return $v^* = (\underbrace{v_0^*}_{\text{objective}}, v_1^*, \dots, v_T^*)$

and $\pi^* = (\pi_0^*, \pi_1^*, \dots, \pi_{T-1}^*)$

Note: These expressions and the VI solution for finite-horizon MDP can be extended to the case of time-varying MDPs where the transition P_t and reward R_t are time-dependent.

Effective Horizon

Question: Can we approximate an infinite-horizon discounted return with a finite-horizon (truncated) return?

$$G = \underbrace{\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)}_{\text{true return } G} = \underbrace{\sum_{t=0}^{T-1} \gamma^t R(S_t, A_t)}_{\text{truncated return } \hat{G}} + \underbrace{\sum_{t=T}^{\infty} \gamma^t R(S_t, A_t)}_{\text{error } G - \hat{G}}$$

If $|R(s,a)| \leq R_{max} \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$

$$\leq \gamma^T \sum_{t=T}^{\infty} \gamma^{t-T} R_{max}$$

$$\longrightarrow \quad |\text{error}| = |G - \hat{G}| \leq \frac{\gamma^T R_{max}}{1-\gamma} \leq \varepsilon$$

$$\longrightarrow \quad T \geq \frac{\log \frac{R_{max}}{\varepsilon(1-\gamma)}}{\log \frac{1}{\gamma}} \qquad \log \tfrac{1}{\gamma} \geq 1-\gamma$$

effective horizon

$$\xrightarrow[\text{simplicity}]{\text{for}} \quad T_{\gamma,\varepsilon} := \frac{\log \frac{R_{max}}{\varepsilon(1-\gamma)}}{1-\gamma}$$

Note: The notion of effective horizon can be used to create approximations between infinite-horizon discounted MDPs and finite-horizon MDPs.