Optimal policy

If

$$v_{\pi_1}(s) \geq v_{\pi_2}(s) \quad \text{for all } s \in \mathcal{S}$$

then π_1 is "better" than π_2.

> Definition:
>
> A policy π^* is optimal if $v_{\pi^*}(s) \geq v_\pi(s)$ for all s and for any other policy π.

Bellman optimality equation (BOE)

Bellman optimality equation (elementwise form):

$$\begin{aligned} v(s) &= \max_\pi \sum_a \pi(a|s)\Big[\sum_r p(r|s,a)r + \gamma \sum_{s'} v(s')p(s'|s,a)\Big] \\ &= \max_\pi \sum_a \pi(a|s)q(s,a) \end{aligned}$$

Bellman optimality equation (matrix-vector form):

$$v = \max_\pi (r_\pi + \gamma P_\pi v)$$

> The form of v consists of elements of the maximum function, represented as, i.e. $[\max, \ldots, \max]$

Maximization on the right-hand side of BOE

Consider:

$$v(s) = \max_\pi \sum_a \pi(a|s)q(s,a)$$

Fix $q(s,a)$ first, because of $\sum_a \pi(a|s) = 1$, we have:

$$\max_\pi \sum_a \pi(a|s)q(s,a) = \max_{a \in \mathcal{A}(s)} q(s,a)$$

where the optimality is achieved when:

$$\pi(a|s) = \begin{cases} 1 & a = a^* \\ 0 & a \neq a^* \end{cases}$$

> IMO: The process is just about finding the best action (i.e. maximize $q(s,a)$, a simple greedy policy) to take.

Solve the Bellman optimality equation

Consider BOE of matrix-vector form. Let:

$$f(v) := \max_{\pi}(r_\pi + \gamma P_\pi v)$$

Then BOE becomes:

$$v = f(v)$$

Preliminaries: Contraction mapping theorem

Consider a function $f(x)$, where $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^d$. A point x^* is called a **fixed point** if

$$f(x^*) = x^*$$

f is called **Contraction mapping** (or contractive function) if there exists $\gamma \in [0, 1)$ such that

$$||f(x_1) - f(x_2)|| \leq \gamma ||x_1 - x_2||$$

$\forall x_1, x_2 \in \mathbb{R}^d$.

Contraction mapping theorem:

Consider a contraction mapping f, then

- There exists a unique vector x^* satisfying $x^* = f(x^*)$.
- x^* can be obtained by the method of successive approximation, starting from any arbitrary initial vector in R^d.

  i.e. Consider the process: $x_{k+1} = f(x_k)$. Then, $x_k \to x^*$ as $k \to \infty$ for any initial x_0.

> Proof:
>
> - Prove that sequence $\{x_k\}_{k=0}^{\infty}$ is convergent.
>
>   we have:
>
> $$\begin{aligned} ||x_{k+1} - x_k|| &\leq \gamma ||x_k - x_{k-1}|| \\ &\vdots \\ &\leq \gamma^k ||x_1 - x_0|| \end{aligned}$$
>
>   Therefore, $\forall m > n$
>
> $$\begin{aligned} ||x_m - x_n|| &= ||x_m - x_{m-1} + \cdots + x_{n+1} - x_n|| \\ &\leq ||x_m - x_{m-1}|| + \cdots + ||x_{n+1} - x_n|| \\ &= \gamma^n(\gamma^{m-1-n} + \cdots + 1)||x_1 - x_0|| \\ &\leq \gamma^n(1 + \cdots + \gamma^{m-1-n} + \gamma^{m-n} + \ldots)||x_1 - x_0|| \\ &= \frac{\gamma^n}{\gamma - 1}||x_1 - x_0|| \end{aligned}$$
>
>   Thus, the sequence is Cauchy. Hence converges to a limit point $x^* = lim_{k\to\infty} x_k$.
>
> - Now show that $x^* = f(x^*)$.
>
>   Since $||f(x_k) - x_k|| \leq \gamma^k ||x_1 - x_0||$, we have $x^* = f(x^*)$ at the limit.
>
> - Show that fixed point is unique.
>
>   Suppose that x^*, y^* are fixed points. Then,

> $$||x^* - y^*|| = ||f(x^*) - f(y^*)|| \leq \rho||x^* - y^*||$$
>
> Since $\rho < 1$, $x^* = y^*$ holds.

Contraction property of the right-hand side of the BOE

Theorem:

$f(v) = \max_\pi(r_\pi + \gamma P_\pi v)$ is a **contraction mapping**. In particular, $\forall v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$, it holds that

$$||f(v_1) - f(v_2)||_\infty \leq \gamma||v_1 - v_2||_\infty$$

$||\cdot||_\infty$ is the maximum norm, which is the maximum absolute value of the elements of a vector.

> Proof:
>
> > The following vector operations are all elementwise. Like $\leq, |\cdot|$.
>
> Consider any two vectors $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$, and $\pi_1^* = \arg\max_\pi(r_\pi + \gamma P_\pi v_1),\ \pi_2^* = \arg\max_\pi(r_\pi + \gamma P_\pi v_2)$.
>
> Then,
>
> $$\begin{aligned} f(v_1) - f(v_2) &= (r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1) - (r_{\pi_2^*} + \gamma P_{\pi_1^*} v_2) \\ &\leq (r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1) - (r_{\pi_2^*} + \gamma P_{\pi_1^*} v_2) \\ &= \gamma P_{\pi_1^*}(v_1 - v_2) \end{aligned}$$
>
> Similarly,
>
> $$f(v_2) - f(v_1) \leq \gamma P_{\pi_2^*}(v_2 - v_1)$$
>
> Therefore,
>
> $$\gamma P_{\pi_2^*}(v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*}(v_1 - v_2)$$
>
> Define
>
> $$z = \max\{\gamma|P_{\pi_2^*}(v_1 - v_2)|, \gamma|P_{\pi_1^*}(v_1 - v_2)|\}$$
>
> implies,
>
> $$|f(v_1) - f(v_2)| \leq z$$
>
> then follows that,
>
> $$||f(v_1) - f(v_2)||_\infty \leq ||z||_\infty$$
>
> And suppose z_i is the ith entry of z, and p_i^T, q_i^T are ith row of $P_{\pi_1^*}, P_{\pi_2^*}$, then
>
> $$z_i = \max\{\gamma|p_i^T(v_1 - v_2)|, \gamma|q_i^T(v_1 - v_2)|\}$$
>
> Since $|p_i^T(v_1 - v_2)| \leq ||v_1 - v_2||_\infty$, also $|q_i^T(v_1 - v_2)| \leq ||v_1 - v_2||_\infty$. Thus,
>
> $$||f(v_1) - f(v_2)||_\infty \leq ||z||_\infty = \max_i |z_i| \leq \gamma||v_1 - v_2||_\infty$$

Policy optimality

Suppose v^* is the solution of Bellman optimality equation. Thus

$$v^* = \max_\pi(r_\pi + \gamma P_\pi v^*)$$

Holds.

Suppose

$$\pi^* = \arg\max_\pi(r_\pi + \gamma P_\pi v^*)$$

Then

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*$$

Theorem (Optimality of v^* and π^*):

$\forall \pi$, it holds that

$$v^* = v_{\pi^*} \geq v_\pi$$

> $\geq$ is elementwise comparison.

> Proof:
>
> We have
>
> $$v^* - v_\pi \geq (r_\pi + \gamma P_{\pi^*} v^*) - (r_\pi + \gamma P_\pi v) = \gamma P_\pi (v^* - v_\pi)$$
>
> And
>
> $$\gamma P_\pi (v^* - v_\pi) \geq \cdots \geq (\gamma P_\pi)^n (v^* - v_\pi)$$
>
> Therefore,
>
> $$v^* - v_\pi \geq \lim_{n\to\infty} (\gamma P_\pi)^n (v^* - v_\pi) = 0$$

Factors that influence optimal policies

According to BOE, optimal state value and optimal policy are determined by:

- immediate reward r
- discount rate: γ
- system model: $p(s'|s,a), p(r|s,a)$

Impact of the discount rate

- Small γ may cause short-sighted, Large γ may cause long-sightedness.
- the states close to the target have greater state values, whereas those far away have lower values.
- ...

Impact of the reward values

- increase the punishment to strictly prohibit the agent from entering any forbidden area
- scale all the rewards or add the same value to all the rewards, the optimal policy remains the same.
- ...

Theorem (Optimal policy invariance):

Consider a Markov decision process with v^* as the optimal state value. If every reward r is changed to $\alpha r + \beta$ (affine transformation, and $\alpha, \beta \in \mathbb{R},\ a > 0$), then corresponding optimal state value v' satisfies:

$$v' = \alpha v^* + \frac{\beta}{1-\gamma}\mathbf{1}$$

> $\mathbf{1} = [1, \ldots, 1]^T$

> Proof:
>
> > I believe the proof in the original book is overly complex, so I have devised an alternative proof. If there are any flaws in my proof, please feel free to point them out.
>
> Consider one element $v(s)$ of vector v^*, since $v(s)$ is the expectation of discounted returns. And we consider the influence of one return
>
> $$G = \sum_{i=0}^{\infty} \gamma^i r_i$$
>
> After reward r is changed to $\alpha r + \beta$,
>
> $$\begin{aligned} G' &= \sum_{i=0}^{\infty} \gamma^i (\alpha r_i + \beta) \\ &= \alpha G + \frac{\beta}{1-\gamma} \end{aligned}$$
>
> According to property of expectation, the corresponding $v'(s)$ is
>
> $$v'(s) = \alpha v(s) + \frac{\beta}{1-\gamma}$$