On considère les notations suivantes

$$P(s', r \mid s, a) = P[S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a]$$

$$R(s, a) = E[R_{t+1} \mid S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} r P(r \mid S_t = s, A_t = a)$$

$$V_\pi(s) = E_\pi \left[\sum_{t \geq 0} R(S_t) \mid S_0 = s \right]$$

On souhaite montrer que $V_\pi(s) = E_\pi \left[\sum_{t \geq 0} R(S_t) \mid S_0 = s \right)$

$$= E_\pi [R(s, \pi(s))] + \sum_{s' \in S} P(s' \mid s, \pi(s)) V_\pi(s')$$

On note G_t la somme des futur rewards :

$$G_t = \sum_{k=t+1}^{T} R_k \quad \text{avec } R(S_t) = R_{t+1}$$

On peut alors réécrire $V_\pi(s)$ selon t tel que :

$$V_t(s) = E_\pi [G_t \mid S_t]$$

$$V_{t\pi}(s) = E_\pi [R_{t+1} + G_{t+1} \mid S_t = s]$$

$$= E_\pi [R_{t+1} \mid S_t = s] + E_\pi [G_{t+1} \mid S_t = s] \quad (a)$$

En développant le terme $E_\pi [R_{t+1} \mid S_t = s]$:

$$E_\pi [R_{t+1} \mid S_t = s] = \sum_{r \in \mathcal{R}} r P(r \mid S_t = s)$$

$$= \sum_{r \in \mathcal{R}} r \sum_{s' \in S} P(r, s' \mid s) = \sum_{r \in \mathcal{R}} r \sum_{s' \in S} \frac{P(R_{t+1} = r, S_{t+1} = s', S_t = s)}{P(S_t = s)}$$

$$= \sum_{r \in \mathcal{R}} r \sum_{s' \in S} \sum_{a \in A} \frac{P(R_{t+1} = r, S_{t+1} = s', S_t = s, A_t = a)}{P(S_t = s)}$$

$$= \sum_{r \in \mathcal{R}} r \sum_{s' \in S} \sum_{a \in A} \frac{P(R_{t+1} = r, S_{t+1} = s', S_t = s, A_t = a)}{P(S_t = s)} \times \frac{P(S_t = s, A_t = a)}{P(S_t = s, A_t = a)}$$

$$= \sum_{r \in \mathcal{R}} r \sum_{s' \in S} \sum_{a \in A} P(s', r \mid s, a) \times P(A_t = a \mid S_t = s)$$

On note $\pi(a \mid s) = P(A_t = a \mid S_t = s)$, Ainsi on peut écrire

$$E_\pi [R_{t+1} \mid S_t = s] = \sum_{r \in \mathcal{R}} r \sum_{s' \in S} \sum_{a \in A} P(s', r \mid s, a) \pi(a \mid s)$$

En regardant désormais le second terme de (1) : $E_\pi[G_{t+1} | S_t = s]$

(On aimerait bien le réécrire selon $V_{(t+1)\pi}(s') = E_\pi[G_{t+1} | S_{t+1} = s']$)

Pour avoir une expression de récurrence entre chaque $V_{t\pi}(\cdot)$

$$E_\pi[G_{t+1} | S_t = s] = \sum_{g \in G} g\, P[g | S_t = s] = \sum_{g \in G} g \sum_{s' \in S} P(g, S_{t+1} = s' | S_t = s)$$

$$= \sum_{g \in G} g \sum_{s' \in S} \sum_{a \in A} P(g, S_{t+1} = s', A_t = a | S_t = s)$$

$$= \sum_{g \in G} g \sum_{s' \in S} \sum_{a \in A} \sum_{r \in R} P(g, S_{t+1} = s', A_t = a, R_{t+1} = r | S_t = s)$$

$$= \sum_{g \in G} g \sum_{s'} \sum_{a} \sum_{r} \frac{P(G_{t+1} = g, S_{t+1} = s', A_t = a, R_{t+1} = r, S_t = s)}{P(S_t = s)}$$

$$= \sum_{g \in G} g \sum_{s'} \sum_{a} \sum_{r} \frac{P(g, s', a, r, s)}{P(s)} \times \frac{P(s', a, r, s)}{P(s', a, r, s)} = \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} \frac{P(g, s', a, r, s)}{P(s', a, r, s)} \times \frac{P(s', a, r, s)}{P(s)}$$

$$= \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} P(g | s', a, r, s) \times P(s', a, r | s)$$

$P(g | s', a, r, s) = P(G_{t+1} = g | S_{t+1} = s', A_t = a, R_{t+1} = r, S_t = s)$

On rappel que $G_{t+1} = R_{t+2} + \dots + R_T$

De manière explicite, on sait que R_{t+2} ne dépend que de A_{t+1}, S_{t+1}

De manière plus formel on utilise les propriétés d'un processus Markovien pour réécrire :

$$P(G_{t+1} = g | S_{t+1} = s', A_t = a, R_{t+1} = r, S_t = s) = P(G_{t+1} = g | S_{t+1} = s')$$

Ainsi on a : $E_\pi[G_{t+1} | S_t = s] = \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} P(g | s')\, P(s', a, r | s)$

$$= \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} P(g | s') \times \frac{P(s', a, r, s)}{P(s)} = \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} P(g | s') \times \frac{P(s', a, r, s) P(a, s)}{P(s)} \times \frac{1}{P(a, s)}$$

(e) $$= \sum_{g} g \sum_{s'} \sum_{a} \sum_{r} P(g | s') \times \frac{P(s', a, r, s)}{P(a, s)} \times \frac{P(a, s)}{P(s)}$$

$$= \sum_g g \sum_{s'} \sum_a \sum_r P(g|s') \times P(s', r|a, s) \times \pi(a|s)$$

$$= \sum_r \sum_{s'} \sum_a \left(\sum_{g \in G} g P(g|s') \right) \times P(s', r|a, s) \times \pi(a|s)$$

$$= \sum_r \sum_{s'} \sum_a \mathbb{E}_\pi [G_{t+1} | S_{t+1} = s'] \times P(s', r|a, s) \times \pi(a|s)$$

$$= \sum_r \sum_{s'} \sum_a V_{(t+1)\pi}(s') \times P(s', r|a, s) \times \pi(a|s)$$

Ainsi on a donc l'expression de $V_{t+\pi}(s)$ suivant:

$$V_\pi(s) = \sum_{r \in R} \sum_{s' \in S} \sum_{a \in A} \left[r P(r, s'|s, a) \pi(a|s) + V_{(t+1)\pi}(s') P(r, s'|s, a) \pi(a|s) \right]$$

$$V_\pi(s) = \sum_{r \in R} \sum_{s' \in S} \sum_{a \in A} \pi(a|s) \left[r P(r, s'|s, a) + V_{(t+1)\pi}(s') P(r, s'|s, a) \right]$$

$$V_{\pi(t)}(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} \sum_{r \in R} P(r, s'|s, a) \left[r + V_{(t+1)\pi}(s') \right]$$

Pour retrouver la forme voulue: $\mathbb{E}_\pi [R(A_\pi(s))] + \sum_{s' \in S} P(s'|s, \pi(s)) V_{(t+1)}^{\pi}(s')$

$$V_{\pi,(t)}(s) = \sum_{r \in R} \sum_{s' \in S} \sum_{a \in A} \pi(a|s) r \times P(r, s'|s, a) + \sum_{r \in R} \sum_{s' \in S} \sum_{a \in A} P(r, s'|s, a) \pi(a|s) V_{\pi(t+1)}(s')$$

$$= \sum_{r \in R} \sum_{s' \in S} \sum_{a \in A} r \times P(r, s'|s, a) + \sum_{s' \in S} \sum_{a \in A} \pi(a|s) V_{\pi(t+1)}(s') \sum_{r \in R} P(r, s'|s, a) \quad \text{(PT)}$$

$$= \sum_{r \in R} r \sum_{a \in A} \sum_{s' \in S} P(r, s'|s, a) + \sum_{s' \in S} \sum_{a \in A} \pi(a|s) V_{\pi(t+1)}(s') P(s'|s, a) \quad \text{(PT)}$$

$$= \sum_r r \sum_a P(r|s, a) + \sum_{s'} \sum_a V_{\pi(t+1)}(s') P(s'|s, a) \pi(a|s)$$

$$= \sum_a \mathbb{E}_\pi [R_{t+1} | S_t = s, A_t = a] + \sum_{s'} V_{\pi(t+1)}(s') \sum_a P(s'|s, a) \pi(a|s)$$

PT: proba totale

En considérant que $(A_t)_t$ est pris par la policy $\pi(s)$, on a donc

$$\sum_a \mathbb{E}[R_{t+1} \mid S_t = s, A_t = \pi(s)) = \sum_a R(s, \pi(s)) = \mathbb{E}_\pi[R(s, \pi(s))]$$

De plus :

$$\sum_{s'} V_{\pi(t+1)}(s') \sum_a P(s' \mid s, a)\, \pi(a \mid s)$$

$$= \sum_{s'} V_{\pi(t+1)}(s') P(s' \mid s, \pi(s))$$

Ainsi on retrouve bien la "formule"

$$V_{\pi(t)}(s) = \mathbb{E}_\pi[R(s, \pi(s))] + \sum_{s' \in S} V_{\pi(t+1)}(s')\, P(s' \mid s, \pi(s))$$