UN EXEMPLE DE PROCESSUS DE DÉCISION MARKOVIEN

On souhaite maximiser le revenu provenant de la production d'une machine. Or, le niveau de production dépend de l'état de la machine, et l'état de la machine dépend de son entretien. Les données du problème sont les suivantes:

- Les états sont: 0 (neuf), 1 (bon état), 2 (mauvais état) et 3 (en panne).
- Les rendements respectifs (par période) sont de 30, 15, 5 et 0.
- Les actions possibles sont: entretenir (1), ne rien faire (2), rénover (3).
- Le revenu par objet produit est de 100$.
- Le taux d'actualisation est $\alpha = 0,8$.

Le tableau suivant fournit les informations sur les revenus associés aux différentes combinaisons d'états et de décision, ainsi quet les probabilités de transition.

état	action	coût	probabilité de transition	nouvel état	revenu (gain − coût)
0	entretenir	500$	3/4	0	2500$
			1/4	1	2500$
	ne rien faire	0$	4/5	1	3000$
			1/5	3	3000$
1	entretenir	1000$	4/7	1	500$
			2/7	2	500$
			1/7	3	500$
	ne rien faire	0$	4/5	2	1500$
			1/5	3	1500$
	rénover	3000$	1	0	−500$
2	entretenir	1000$	3/4	2	−1000$
			1/4	3	−1000$
	ne rien faire	0$	1/2	2	500$
			1/2	3	500$
	rénover	3000$	1	0	−2500$
3	rénover	3000$	1	0	−3000$

On considère la politique δ:

- entretenir une machine neuve ($\delta(0) = 1$);
- ne rien faire si la machine est en bon état ($\delta(1) = 2$);
- entretenir une machine en mauvais état ($\delta(2) = 1$);
- réparer une machine en panne ($\delta(3) = 3$).

On obtient alors:

$$P^\delta = \begin{pmatrix} 3/4 & 1/4 & 0 & 0 \\ 0 & 0 & 4/5 & 1/5 \\ 0 & 0 & 3/4 & 1/4 \\ 1 & 0 & 0 & 0 \end{pmatrix} \qquad R^\delta = \begin{pmatrix} 2500\$ \\ 1500\$ \\ -1000\$ \\ -3000\$ \end{pmatrix}.$$

On évalue le revenu moyen v^δ associé à la politique δ en résolvant le système linéaire

$$(I - \alpha P^\delta) v^\delta = R^\delta,$$

c'est-à-dire

$$\begin{pmatrix} 8/20 & -4/20 & 0 & 0 \\ 0 & 1 & -16/25 & -4/25 \\ 0 & 0 & 8/20 & -4/20 \\ -4/5 & 0 & 0 & 1 \end{pmatrix} v^\delta = \begin{pmatrix} 2500 \\ 1500 \\ -1000 \\ -3000 \end{pmatrix},$$

dont la solution approximative est

$$v^\delta = \begin{pmatrix} 6782 \\ 1064 \\ -1287 \\ 2426 \end{pmatrix}.$$

pour l'état $i = 2$, tentons maintenant d'améliorer la politique en améliorant δ sur *une seule* étape:

$$\max_k \Big\{ R_i^k + \alpha \sum_{j=0}^{3} P_{ij}^k v_j^k \Big\}.$$

$$\begin{aligned} \max \Big\{ \quad & 500\$ + 0,8 \left[\frac{1}{2} \times (-1287\$) + \frac{1}{2} \times (2426\$) \right] \quad \text{(ne rien faire)}, \\ & -1000\$ + 0,8 \left[\frac{3}{4} \times (-1287\$) + \frac{1}{4} \times (2426\$) \right] \quad \text{(entretien)}, \\ & -2500\$ + 0,8 \left[1 \times (2426\$) \right] \qquad\qquad\qquad\qquad \text{(réparer)} \quad \Big\} \\ = \quad \max \quad & \Big\{ 955\$,\ -1287\$,\ -560\$ \Big\} \quad = \quad 955\$. \end{aligned}$$

Le maximum étant atteint pour l'action $k = 1$ (entretenir), on modifie la politique en conséquence. On effectue ensuite les mêmes minimisations pour les états 1 et 3 pour obtenir une politique améliorée. On retourne alors à l'étape d'évaluation de la politique améliorée.

Dans l'algorithme "value iteration", on se contente de mettre à jour le vecteur de revenus *sans évaluer la politique améliorée.* Les calculs sont grandement simplifiés car on n'a pas à résoudre de système linéaire. Par contre, les vecteurs de revenus ne correspondent plus aux politiques. Néanmoins, on a vu qu'ils convergeaient vers le vecteur de revenus optimal.