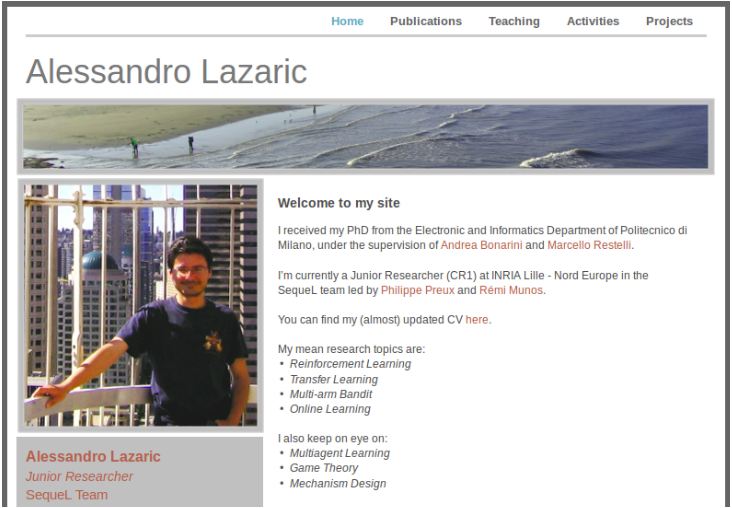
A2DI: Apprentissage de politiques

John Klein

Lille1 Université - CRIStAL UMR CNRS 9189

- d'après un document d'Alessandro Lazaric -

Visitez sa homepage



Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le chapitre précédent, nous avons dérivé le modèle MDP qui repose (entre autres) sur le choix d'une politique.

Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le chapitre précédent, nous avons dérivé le modèle MDP qui repose (entre autres) sur le choix d'une politique. Nous avons introduit la notion de fonction de valeur pour évaluer une politique.

Dans le cadre de l'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r.

Dans le chapitre précédent, nous avons dérivé le modèle MDP qui repose (entre autres) sur le choix d'une politique. Nous avons introduit la notion de fonction de valeur pour évaluer une politique.

$\rightarrow$ Apprenons à présent à maximiser cette fonction !

Plan du chapitre

Avertissement

Dans la suite, on se concentre essentiellement sur des problèmes à horizon inifinie avec affaiblissement.

La plupart des résultats que nous verrons se généralisent aux autres problèmes.

Problème d'optimisation :

$$\max_{\pi} \; V^{\pi}(x_0) =$$

$$\max_{\pi} \; \mathbb{E}\left[r(x_0,\pi(x_0)) + \gamma r(x_1,\pi(x_1)) + \gamma^2 r(x_2,\pi(x_2)) + \ldots\right]$$

$$\Downarrow$$

Recherche exhaustive : infaisable (il faudrait tester $\sharp A^{\sharp S}$ politiques !)

$$\Downarrow$$

il faut s'appuyer sur la structure du MDP et simplifier le problème d'optimisation.

Problème d'optimisation :

$$\max_{\pi} \; V^{\pi}(x_0) =$$

$$\max_{\pi} \; \mathbb{E}\left[r(x_0, \pi(x_0)) + \gamma r(x_1, \pi(x_1)) + \gamma^2 r(x_2, \pi(x_2)) + \dots \right]$$

$$\Downarrow$$

Recherche exhaustive : infaisable (il faudrait tester $\sharp A^{\sharp S}$ politiques !)

$$\Downarrow$$

il faut s'appuyer sur la structure du MDP et simplifier le problème d'optimisation.

Astuce : équation de Bellman

Proposition

Pour toute politique sationnaire $\pi = (\pi, \pi, \dots)$, la fonction de valeur à l'état $x \in X$ satisfait l'équation de Bellman :

$$V^{\pi}(x) = r(x, \pi(x)) + \gamma \sum_{y} p(y|x, \pi(x)) V^{\pi}(y).$$

Astuce : équation de Bellman

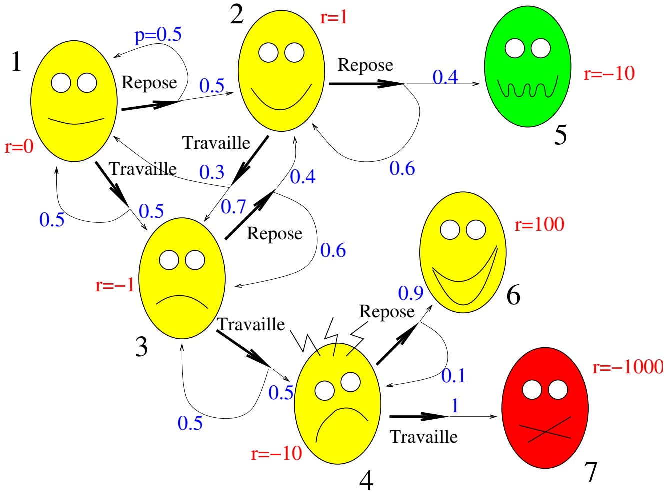
Preuve :

Pour toute politique π, on a

$$
\begin{aligned}
V^\pi(x) &= \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\Big] \\
&= r(x, \pi(x)) + \mathbb{E}\Big[\sum_{t\geq 1} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\Big] \\
&= r(x, \pi(x)) \\
&\qquad + \gamma \sum_{y} \mathbb{P}(x_1 = y \,|\, x_0 = x; \pi(x_0)) \mathbb{E}\Big[\sum_{t\geq 1} \gamma^{t-1} r(x_t, \pi(x_t)) \,|\, x_1 = y; \pi\Big] \\
&= r(x, \pi(x)) + \gamma \sum_{y} p(y|x, \pi(x)) V^\pi(y).
\end{aligned}
$$

$\blacksquare$

Exemple : le dilemme de l'étudiant



Exemple : le dilemme de l'étudiant

- Problème : horizon infinie avec états terminaux.
- Objectif : trouver la politique qui maximise la récompense cumulée attendue avant d'atteindre un état terminal.

Exemple : le dilemme de l'étudiant

- Problème : horizon infinie avec états terminaux.
- Objectif : trouver la politique qui maximise la récompense cumulée attendue avant d'atteindre un état terminal.

Exemple : le dilemme de l'étudiant

- Problème : horizon infinie avec états terminaux.
- Objectif : trouver la politique qui maximise la récompense cumulée attendue avant d'atteindre un état terminal.

Exemple : le dilemme de l'étudiant

- Problème : horizon infinie avec états terminaux.
- Objectif : trouver la politique qui maximise la récompense cumulée attendue avant d'atteindre un état terminal.

Note : L'équation de Bellman reste vraie pour ce type de problème aussi.

Exemple : le dilemme de l'étudiant

- Espace d'état : $X = \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\}$ et les états x_5, x_6, x_7 sont terminaux .

- Espace d'action : $A = \{\text{repos} ; \text{travail}\} \rightarrow \{0; 1\}$.

- Dynamique Markovienne : a_t et x_t sont des stats. suff. pour x_{t+1}

Exemple : le dilemme de l'étudiant

- Espace d'état : $X = \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\}$ et les états x_5, x_6, x_7 sont terminaux .

- Espace d'action : $A = \{\text{repos}\,;\,\text{travail}\} \rightarrow \{0; 1\}$.

- Dynamique Markovienne : a_t et x_t sont des stats. suff. pour x_{t+1}

Exemple : le dilemme de l'étudiant

- Espace d'état : $X = \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\}$ et les états x_5, x_6, x_7 sont terminaux .

- Espace d'action : $A = \{\text{repos}\,;\ \text{travail}\} \rightarrow \{0; 1\}$.

- Dynamique Markovienne : a_t et x_t sont des stats. suff. pour x_{t+1}

$$\mathbb{P}\left(X_{t+1} = y | X_t = x; a = 0\right) = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0.4 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Exemple : le dilemme de l'étudiant

- Espace d'état : $X = \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\}$ et les états x_5, x_6, x_7 sont terminaux .

- Espace d'action : $A = \{\text{repos}\,;\,\text{travail}\} \rightarrow \{0; 1\}$.

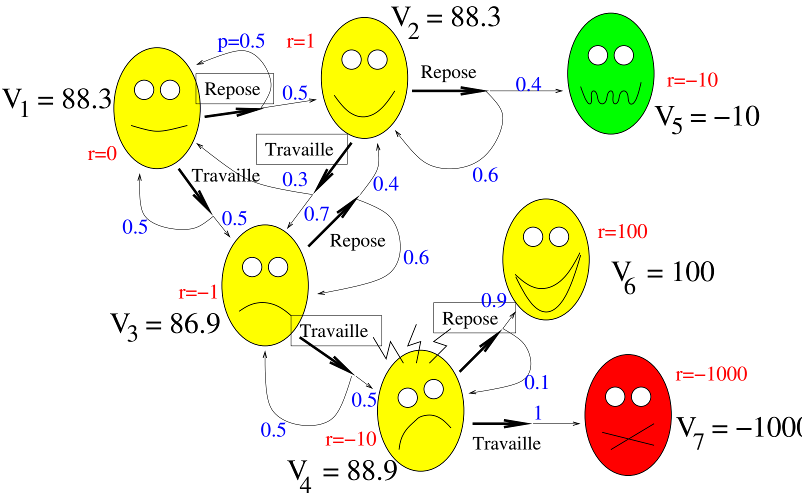
- Dynamique Markovienne : a_t et x_t sont des stats. suff. pour x_{t+1}

$$\mathbb{P}\left(X_{t+1} = y | X_t = x; a = 1\right) = \begin{bmatrix} 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0.7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Exemple : le dilemme de l'étudiant

- Espace d'état : $X = \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\}$ et les états x_5, x_6, x_7 sont terminaux .
- Espace d'action : $A = \{\text{repos}\,;\,\text{travail}\} \rightarrow \{0; 1\}$.
- Dynamique Markovienne : a_t et x_t sont des stats. suff. pour x_{t+1}
- Récompense :
  $r : \{x_1; x_2; x_3; x_4; x_5; x_6; x_7\} \rightarrow \{0; 1; -1; -10; -10; 100; -1000\}$

Exemple : le dilemme de l'étudiant : choix d'une politique π



Exemple : le dilemme de l'étudiant : calcul incrémental de V^{π} grâce à Bellman

Calculer V_4 :

$$V_6 = 100$$
$$V_4 = -10 + (0.9V_6 + 0.1V_4)$$

$$\Rightarrow V_4 = \frac{-10 + 0.9V_6}{0.9} = 88.8$$

Exemple : le dilemme de l'étudiant : calcul incrémental de V^π grâce à Bellman

Calculer V_3 : pas besoin d'examiner toutes les trajectoires possibles

$$V_4 = 88.8$$
$$V_3 = -1 + (0.5 V_4 + 0.5 V_3)$$

$$\Rightarrow V_3 = \frac{-1 + 0.5 V_4}{0.5} = 86.8$$

et ainsi de suite pour le reste...

Exemple : le dilemme de l'étudiant : calcul incrémental de V^π grâce à Bellman

Calculer V_3 : pas besoin d'examiner toutes les trajectoires possibles

$$V_4 = 88.8$$
$$V_3 = -1 + (0.5 V_4 + 0.5 V_3)$$

$$\Rightarrow V_3 = \frac{-1 + 0.5 V_4}{0.5} = 86.8$$

et ainsi de suite pour le reste...

Equation de Bellman optimale :

Principe d'Optimalité de Bellman :

> *"Une politque optimale a pour propriété que, quel que soit l'état initial et quelle que soit la décision initiale, les décisions restantes doivent constituer une politique optimale par rapport l'état résultant de la 1ère décision."*

Equation de Bellman optimale :

Proposition

La fonction de valeur optimale V^* (i.e., $V^* = \max_\pi V^\pi$) est la solution à l'équation optimale de Bellman :

$$V^*(x) = \max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \right].$$

et la politique optimale est

$$\pi^*(x) = \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \right].$$

Equation de Bellman optimale :

Preuve

Pour toute politique $\pi = (a, \pi')$ (possiblement non-stationnaire),

$$\begin{aligned}
V^*(x) &\overset{(a)}{=} \max_\pi \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\Big] \\
&\overset{(b)}{=} \max_{(a,\pi')} \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi'}(y) \Big] \\
&\overset{(c)}{=} \max_a \Big[r(x,a) + \gamma \sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y) \Big] \\
&\overset{(d)}{=} \max_a \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \Big].
\end{aligned}$$

■

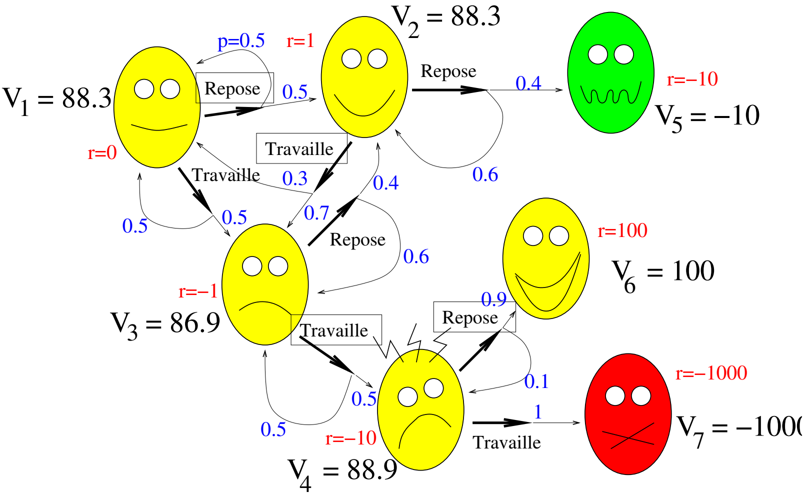
Système d'équations :

L'équation de Bellman

$$V^{\pi}(x) = r(x, \pi(x)) + \gamma \sum_{y} p(y|x, \pi(x)) V^{\pi}(y).$$

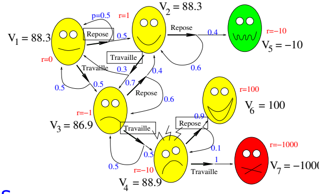
forme un système **linéaire** d'équations avec N inconnues et N contraintes linéaires.

Exemple : le dilemme de l'étudiant (rappel de la politique choisie π)



Exemple : le dilemme de l'étudiant

$$V^{\pi}(x) = r(x,\pi(x)) + \gamma \sum_y p(y|x,\pi(x)) V^{\pi}(y)$$



Système d'équations

$$\begin{cases} V_1 &= 0 + 0.5\,V_1 + 0.5\,V_2 \\ V_2 &= 1 + 0.3\,V_1 + 0.7\,V_3 \\ V_3 &= -1 + 0.5\,V_4 + 0.5\,V_3 \\ V_4 &= -10 + 0.9\,V_6 + 0.1\,V_4 \\ V_5 &= -10 \\ V_6 &= 100 \\ V_7 &= -1000 \end{cases} \Rightarrow$$

$$(V, R \in \mathbb{R}^7,\ P \in \mathbb{R}^{7\times 7})$$

$$V = R + PV$$

$$\Downarrow$$

$$V = (I - P)^{-1} R$$

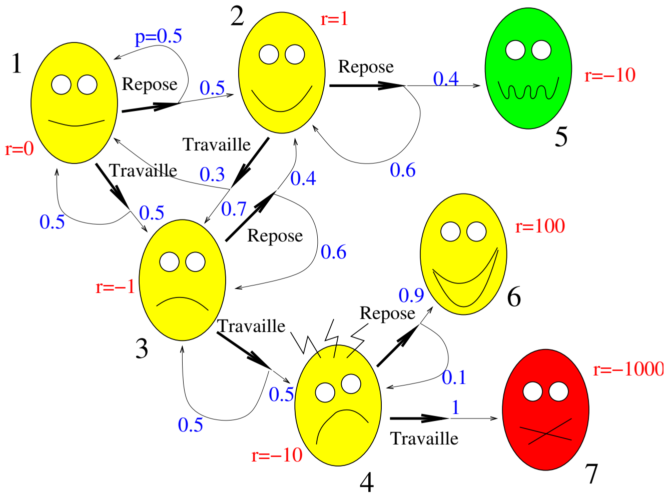
Système d'équations :

L'équation optimale de Bellman

$$V^*(x) = \max_{a \in A}\left[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y)\right].$$

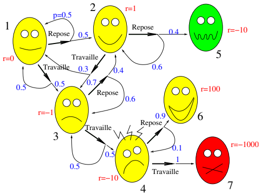
forme un système (hautement) **non-linéaire** d'équations avec N inconnues et N contraintes non-linéaires (l'opérateur max).

Exemple : le dilemme de l'étudiant (politique optimale inconnue)



Exemple : le dilemme de l'étudiant

$$V^*(x) = \max_{a \in A}\left[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y)\right]$$



Système d'équations

$$\begin{cases} V_1 &= \max\left\{0 + 0.5\,V_1 + 0.5\,V_2;\;\; 0 + 0.5\,V_1 + 0.5\,V_3\right\} \\ V_2 &= \max\left\{1 + 0.4\,V_5 + 0.6\,V_2;\;\; 1 + 0.3\,V_1 + 0.7\,V_3\right\} \\ V_3 &= \max\left\{-1 + 0.4\,V_2 + 0.6\,V_3;\;\; -1 + 0.5\,V_4 + 0.5\,V_3\right\} \\ V_4 &= \max\left\{-10 + 0.9\,V_6 + 0.1\,V_4;\;\; -10 + V_7\right\} \\ V_5 &= -10 \\ V_6 &= 100 \\ V_7 &= -1000 \end{cases}$$

$\Rightarrow$ trop compliqué, besoin d'une solution alternative.

Les Opérateurs de Bellman :

Rq : sans perte de généralité, on prend le cas d'un espace d'état discret $|X| = N$ et $V^\pi \in \mathbb{R}^N$.

Definition

Pour tout $W \in \mathbb{R}^N$, l' opérateur de Bellman $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$ est

$$\mathcal{T}^\pi W(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) W(y),$$

et l'opérateur optimal de Bellman (ou operateur de programmation dynamique) est

$$\mathcal{T} W(x) = \max_{a \in A} \left[r(x, a) + \gamma \sum_y p(y|x, a) W(y) \right].$$

Les Opérateurs de Bellman :

Proposition

Propriétés des opérateurs de Bellman

1. Monotonie : pour tout $W_1, W_2 \in \mathbb{R}^N$, si $W_1 \leq W_2$ élément par élément, alors

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

2. Décalage : pour tout scalaire $c \in \mathbb{R}$,

$$\begin{aligned} \mathcal{T}^\pi (W + cI_N) &= \mathcal{T}^\pi W + \gamma c I_N, \\ \mathcal{T} (W + cI_N) &= \mathcal{T} W + \gamma c I_N, \end{aligned}$$

Les Opérateurs de Bellman :

Proposition

Propriétés des opérateurs de Bellman

1. Monotonie : pour tout $W_1, W_2 \in \mathbb{R}^N$, si $W_1 \leq W_2$ élément par élément, alors

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

2. Décalage : pour tout scalaire $c \in \mathbb{R}$,

$$\begin{aligned} \mathcal{T}^\pi (W + c I_N) &= \mathcal{T}^\pi W + \gamma c I_N, \\ \mathcal{T} (W + c I_N) &= \mathcal{T} W + \gamma c I_N, \end{aligned}$$

Les Opérateurs de Bellman :

Proposition

3. Contraction pour la norme L_∞ : pour tout $W_1, W_2 \in \mathbb{R}^N$

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4. Point Fixe : pour toute politique π

V^π est l'unique point fixe de $\mathcal{T}^\pi$,

V^* est l'unique point fixe de $\mathcal{T}$.

De plus, pour tout $W \in \mathbb{R}^N$ et toute politique sationnaire π

$$\begin{aligned} \lim_{k \to \infty} (\mathcal{T}^\pi)^k W &= V^\pi, \\ \lim_{k \to \infty} (\mathcal{T})^k W &= V^*. \end{aligned}$$

Les Opérateurs de Bellman :

Proposition

3. Contraction pour la norme L_∞ : pour tout $W_1, W_2 \in \mathbb{R}^N$

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4. Point Fixe : pour toute politique π

V^π est l'unique point fixe de $\mathcal{T}^\pi$,

V^* est l'unique point fixe de $\mathcal{T}$.

De plus, pour tout $W \in \mathbb{R}^N$ et toute politique sationnaire π

$$\begin{aligned} \lim_{k\to\infty} (\mathcal{T}^\pi)^k W &= V^\pi, \\ \lim_{k\to\infty} (\mathcal{T})^k W &= V^*. \end{aligned}$$

Les Opérateurs de Bellman :

Proposition

3. Contraction pour la norme L_∞ : pour tout $W_1, W_2 \in \mathbb{R}^N$

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4. Point Fixe : pour toute politique π

V^π est l'unique point fixe de $\mathcal{T}^\pi$,

V^* est l'unique point fixe de $\mathcal{T}$.

De plus, pour tout $W \in \mathbb{R}^N$ et toute politique sationnaire π

$$\begin{aligned} \lim_{k\to\infty} (\mathcal{T}^\pi)^k W &= V^\pi, \\ \lim_{k\to\infty} (\mathcal{T})^k W &= V^*. \end{aligned}$$

L'équation de Bellman :

Preuve

La propriété de contraction (3) est vérifiée car pour tout $x \in X$ on a

$$
\begin{aligned}
&|\mathcal{T}W_1(x) - \mathcal{T}W_2(x)| \\
&= \left| \max_a \left[r(x,a) + \gamma \sum_y p(y|x,a) W_1(y) \right] - \max_{a'} \left[r(x,a') + \gamma \sum_y p(y|x,a') W_2(y) \right] \right| \\
&\overset{(*)}{\leq} \max_a \left| \left[r(x,a) + \gamma \sum_y p(y|x,a) W_1(y) \right] - \left[r(x,a) + \gamma \sum_y p(y|x,a) W_2(y) \right] \right| \\
&= \gamma \max_a \sum_y p(y|x,a) |W_1(y) - W_2(y)| \\
&\leq \gamma ||W_1 - W_2||_\infty \max_a \sum_y p(y|x,a) = \gamma ||W_1 - W_2||_\infty,
\end{aligned}
$$

where in $(*)$ we used $\max_a f(a) - \max_{a'} g(a') \leq \max_a (f(a) - g(a))$. ■

Plan du chapitre

Question :

Comment calcule-t-on concrètement les fonctions de valeur ? Comment en déduire la politique résolvant notre MDP ?

$\Rightarrow$ avec les *algorithmes d'itération de la Valeur/Politique !*

Système d'équations (Rappel) :

L'équation de Bellman

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y).$$

forme un système **linéaire** d'équations avec N inconnues et N contraintes linéaires.

L'équation optimale de Bellman

$$V^*(x) = \max_{a \in A} \left[r(x, a) + \gamma \sum_y p(y|x, a) V^*(y) \right].$$

forme un système (hautement) **non-linéaire** d'équations avec N inconnues et N contraintes non-linéaires (l'opérateur max).

Système d'équations (Rappel) :

L'équation de Bellman

$$V^{\pi}(x) = r(x, \pi(x)) + \gamma \sum_{y} p(y|x, \pi(x)) V^{\pi}(y).$$

forme un système **linéaire** d'équations avec N inconnues et N contraintes linéaires.

L'équation optimale de Bellman

$$V^{*}(x) = \max_{a \in A} \left[r(x, a) + \gamma \sum_{y} p(y|x, a) V^{*}(y) \right].$$

forme un système (hautement) **non-linéaire** d'équations avec N inconnues et N contraintes non-linéaires (l'opérateur max).

Idée n°1 : Itération par valeurs

1. Soit V_0 un vecteur dans R^N quelconque.
2. A chaque itération $k = 1, 2, \ldots, K$
   - Calculer $V_{k+1} = \mathcal{T} V_k$
3. Retourner la politique *gloutonne*

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \right].$$

Idée n°1 : Itération par valeurs

1. Soit V_0 un vecteur dans R^N quelconque.
2. A chaque itération $k = 1, 2, \ldots, K$
   - Calculer $V_{k+1} = \mathcal{T} V_k$
3. Retourner la politique *gloutonne*

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \right].$$

Idée n°1 : Itération par valeurs

1. Soit V_0 un vecteur dans R^N quelconque.
2. A chaque itération $k = 1, 2, \ldots, K$
   - Calculer $V_{k+1} = \mathcal{T} V_k$
3. Retourner la politique *gloutonne*

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \right].$$

Idée n°1 : Itération par valeurs

1. Soit V_0 un vecteur dans R^N quelconque.
2. A chaque itération $k = 1, 2, \ldots, K$
   - Calculer $V_{k+1} = \mathcal{T} V_k$
3. Retourner la politique *gloutonne*

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \right].$$

Idée n°1 : Itération par valeurs - Compléxité

Complexité *Calculatoire*

- Chaque itération avec le calcul de la politique gloutonne prend $O(N^2|A|)$ opérations.

$$V_{k+1}(x) = \mathcal{T} V_k(x) = \max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_k(y) \right]$$

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \right]$$

- La complexité temporelle totale est $O(KN^2|A|)$

Complexité *Mémoire*

- Occupation mémoire d'un MDP : dynamiques $O(N^2|A|)$ et récompense $O(N|A|)$.
- Occupation mémoire de la fonction de valeur et de la politique optimale $O(N)$.

Nouvel outil : Fonction de valeur état-action / Q-fonction

Definition

Dans les problèmes à horizon infinie avec affaiblissement, pour tout politique π, la fonction de valeur état-action (ou Q-fonction) est $Q^\pi : X \times A \mapsto \mathbb{R}$ is

$$Q^\pi(x, a) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a, a_t = \pi(x_t), \forall t \geq 1\Big],$$

et la Q-fonction optimale correspondante est

$$Q^*(x, a) = \max_\pi Q^\pi(x, a).$$

Q-fonction

Les relations entre la V-fonction et la Q-fonction sont :

$$\begin{aligned} Q^{\pi}(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) V^{\pi}(y) \\ V^{\pi}(x) &= Q^{\pi}(x, \pi(x)) \\ Q^{*}(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) V^{*}(y) \\ V^{*}(x) &= Q^{*}(x, \pi^{*}(x)) = \max_{a \in A} Q^{*}(x,a). \end{aligned}$$

Idée n°1 : Itération par valeurs - Implémentations alternatives

Q-iteration.

1. Soit Q_0 une Q-fonction quelconque
2. A chaque itération $k = 1, 2, \ldots, K$
   - Calculer $Q_{k+1} = \mathcal{T} Q_k$
3. Retourner la politique gloutonne

$$\pi_K(x) \in \arg\max_{a \in A} Q(x,a)$$

Comparaison

- Complexité calculatoire et occupation mémoire augmentée à $O(N|A|)$ et $O(N^2|A|^2)$
- Complexité calculatoire de la politique gloutonne diminuée à $O(N|A|)$

Idée n°1 : Itération par valeurs - Implémentations alternatives

Asynchronous VI.

1. Soit V_0 un vecteur quelconque dans R^N
2. A chaque itération $k = 1, 2, \ldots, K$
   - ***Choisir un état*** x_k
   - Calculer $V_{k+1}(x_k) = \mathcal{T} V_k(x_k)$
3. Retourner la politique gloutonne

$$\pi_K(x) \in \arg\max_{a \in A} \left[r(x, a) + \gamma \sum_y p(y|x, a) V_K(y) \right].$$

Comparison

- Complexité calculatoire réduite à $O(N|A|)$.
- Nombre d'itération maximum redaugmenté à $O(KN)$ mais possiblement bien plus faible en pratique les états sont *priorisés* correctement.
- Convergence garantie.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A}\left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y)\right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques

1. Soit π_0 une politique stationnaire *quelconque*
2. A chaque itération $k = 1, 2, \ldots, K$
   - *Evaluation de la politique* sachant π_k, calculer V^{π_k}.
   - *Amélioration de la politique* : calculer la politique gloutonne

$$\pi_{k+1}(x) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y) \right].$$

3. Retourner la dernière politique π_K

Rq : habituellement K est le plus petit entier k tel que $V^{\pi_k} = V^{\pi_{k+1}}$.

Idée n°2 : Itération par politiques - Garanties

Proposition

L'algorithme par itération de politiques génère une suite de politiques avec des performances *croissantes* :

$$V^{\pi_{k+1}} \geq V^{\pi_k},$$

et il converge vers π^* en un nombre *fini* d'itérations.

Idée n°2 : Itération par politiques - Garanties

Preuve

D'après la définition des opérateurs de Bellman et de la politique gloutonne π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

d'après la monotonie de $\mathcal{T}^{\pi_{k+1}}$, il s'en suit que

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

En mettant bout à bout toutes les inégalités, on obtient

$$V^{\pi_k} \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Alors $(V^{\pi_k})_k$ est une suite croissante.

Idée n°2 : Itération par politiques - Garanties

Preuve
D'après la définition des opérateurs de Bellman et de la politique gloutonne π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

d'après la monotonie de $\mathcal{T}^{\pi_{k+1}}$, il s'en suit que

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

En mettant bout à bout toutes les inégalités, on obtient

$$V^{\pi_k} \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Alors $(V^{\pi_k})_k$ est une suite croissante.

Idée n°2 : Itération par politiques - Garanties

Preuve
D'après la définition des opérateurs de Bellman et de la politique gloutonne π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

d'après la monotonie de $\mathcal{T}^{\pi_{k+1}}$, il s'en suit que

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

En mettant bout à bout toutes les inégalités, on obtient

$$V^{\pi_k} \leq \lim_{n\to\infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Alors $(V^{\pi_k})_k$ est une suite croissante.

Idée n°2 : Itération par politiques - Garanties

Preuve
D'après la définition des opérateurs de Bellman et de la politique gloutonne π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

d'après la monotonie de $\mathcal{T}^{\pi_{k+1}}$, il s'en suit que

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

En mettant bout à bout toutes les inégalités, on obtient

$$V^{\pi_k} \leq \lim_{n\to\infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Alors $(V^{\pi_k})_k$ est une suite croissante.

Idée n°2 : Itération par politiques - Garanties

Preuve (suite)
Puisqu'un MDP fini admet un nombre fini de politiques, la condition d'arrêt est finallement vérifiée pour un certain k.

Ainsi la relation (1) est vraie avec une égalité et on obtient

$$V^{\pi_k} = \mathcal{T} V^{\pi_k}$$

et $V^{\pi_k} = V^*$ impliquant que π_k est une politique optimale. ∎

Idée n°2 : Itération par politiques

Notations supplémentaires :

- Pour toute politique π, le *vecteur* de récompense est noté $r^{\pi}(x) = r(x, \pi(x))$,
- La *matrice* de transition est notée $[P^{\pi}]_{x,y} = p(y|x, \pi(x))$.

Idée n°2 : Itération par politiques - Etape d'évaluation de la politique

- *Par calcul direct.* Pour toute politique π calculer

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Complexité : $O(N^3)$ (peut être ramenée à $O(N^{2.807})$).

- *Iterativement.* Pour toute politique π

$$\lim_{n \to \infty} \mathcal{T}^\pi V_0 = V^\pi.$$

Complexité : Une approximation de V^π à ϵ près nécessite $O(N^2 \frac{\log 1/\epsilon}{\log 1/\gamma})$ étapes.

- *Par simulation type Monte-Carlo.* Dans chaque état x, simuler n trajectoires $((x_t^i)_{t\geq 0,})_{1\leq i \leq n}$ en suivant la politique π et calculer

$$\hat{V}^\pi(x) \simeq \frac{1}{n} \sum_{i=1}^{n} \sum_{t \geq 0} \gamma^t r(x_t^i, \pi(x_t^i)).$$

Complexité : En chaque état, l'erreur d'approximation est $O(1/\sqrt{n})$.

Idée n°2 : Itération par politiques - Etape d'évaluation de la politique

- *Par calcul direct.* Pour toute politique π calculer
$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$
Complexité : $O(N^3)$ (peut être ramenée à $O(N^{2.807})$).
- *Iterativement.* Pour toute politique π
$$\lim_{n\to\infty} \mathcal{T}^\pi V_0 = V^\pi.$$
Complexité : Une approximation de V^π à ϵ près nécessite $O(N^2 \frac{\log 1/\epsilon}{\log 1/\gamma})$ étapes.
- *Par simulation type Monte-Carlo.* Dans chaque état x, simuler n trajectoires $((x_t^i)_{t\geq 0,})_{1\leq i\leq n}$ en suivant la politique π et calculer
$$\hat{V}^\pi(x) \simeq \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geq 0}\gamma^t r(x_t^i, \pi(x_t^i)).$$
Complexité : En chaque état, l'erreur d'approximation est $O(1/\sqrt{n})$.

Idée n°2 : Itération par politiques - Etape d'évaluation de la politique

- *Par calcul direct.* Pour toute politique π calculer
$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$
Complexité : $O(N^3)$ (peut être ramenée à $O(N^{2.807})$).
- *Iterativement.* Pour toute politique π
$$\lim_{n\to\infty} \mathcal{T}^\pi V_0 = V^\pi.$$
Complexité : Une approximation de V^π à ϵ près nécessite $O(N^2 \frac{\log 1/\epsilon}{\log 1/\gamma})$ étapes.
- *Par simulation type Monte-Carlo.* Dans chaque état x, simuler n trajectoires $((x_t^i)_{t\geq 0,})_{1\leq i\leq n}$ en suivant la politique π et calculer
$$\hat{V}^\pi(x) \simeq \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geq 0}\gamma^t r(x_t^i, \pi(x_t^i)).$$
Complexité : En chaque état, l'erreur d'approximation est $O(1/\sqrt{n})$.

Idée n°2 : Itération par politiques - Etape d'amélioration de la politique

- Si la politique est évaluée avec V, alors l'amélioration de la politique a une complexité $O(N|A|)$ (calcul d'une espérance).

- Si la politique est évaluée avec Q, alors l'amélioration de la politique a une complexité $O(|A|)$ correspondant à

$$\pi_{k+1}(x) \in \arg\max_{a \in A} Q(x, a),$$

Idée n°2 : Itération par politiques - Etape d'amélioration de la politique

- Si la politique est évaluée avec V, alors l'amélioration de la politique a une complexité $O(N|A|)$ (calcul d'une espérance).
- Si la politique est évaluée avec Q, alors l'amélioration de la politique a une complexité $O(|A|)$ correspondant à

$$\pi_{k+1}(x) \in \arg\max_{a \in A} Q(x, a),$$

Idée n°2 : Itération par politiques - Nombre d'itérations

- Au pire $O(\frac{N|A|}{1-\gamma}\log(\frac{1}{1-\gamma}))$

Idée n°1 vs Idée n°2 / Itération par valeurs vs Itération par politiques

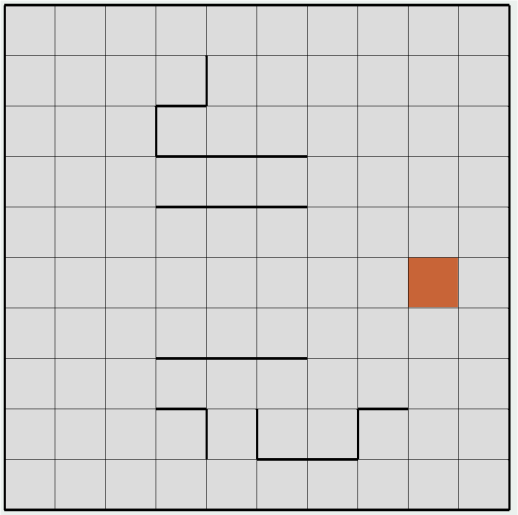
Itération par valeur

- $+$: chaque itération est très efficace d'un point de vue calculatoire.
- $-$ *:* la convergence est seulement *asymptotique*.

Itération par politiques

- $+$: converge en un nombre fini d'itérations (souvent petit en pratique).
- $-$ *:* chaque itération nécessite une *évaluation de la politique* complète potentiellement coûteuse.

Exemple : le problème *Grid-World*



Plan du chapitre

Messages importants du chapitre :

- Les équations de Bellman offre une formulation compacte et inductive des fonctions de valeur.
- La programmation dynamique permet de déterminer la politique optimale pour un problème de type MDP.
- Deux alternatives pour la mise en oeuvre : itération par valeur ou par politique.

Messages importants du chapitre :

- Les équations de Bellman offre une formulation compacte et inductive des fonctions de valeur.

- La programmation dynamique permet de déterminer la politique optimale pour un problème de type MDP.

- Deux alternatives pour la mise en oeuvre : itération par valeur ou par politique.

Messages importants du chapitre :

- Les équations de Bellman offre une formulation compacte et inductive des fonctions de valeur.
- La programmation dynamique permet de déterminer la politique optimale pour un problème de type MDP.
- Deux alternatives pour la mise en oeuvre : itération par valeur ou par politique.