

# A2DI: Apprentissage de politiques

John Klein

Université de Lille - CRIStAL UMR CNRS 9189





- d'après un document d'Alessandro Lazaric -

Visitez sa [homepage](#)

[Home](#) [Publications](#) [Teaching](#) [Activities](#) [Projects](#)

## Alessandro Lazaric



**Alessandro Lazaric**  
Junior Researcher  
SequeL Team

### Welcome to my site

I received my PhD from the Electronic and Informatics Department of Politecnico di Milano, under the supervision of [Andrea Bonarini](#) and [Marcello Restelli](#).

I'm currently a Junior Researcher (CR1) at INRIA Lille - Nord Europe in the SequeL team led by [Philippe Preux](#) and [Rémi Munos](#).

You can find my (almost) updated CV [here](#).

My main research topics are:

- *Reinforcement Learning*
- *Transfer Learning*
- *Multi-arm Bandit*
- *Online Learning*

I also keep on eye on:

- *Multiagent Learning*
- *Game Theory*
- *Mechanism Design*

Cadre de travail : l'apprentissage par renforcement !

- On a un environnement avec lequel on peut jouer au travers d'actions à choisir
- et une fonction de récompense  $r$  qui nous fournit un retour sur nos actions choisies.

Dans le chapitre précédent, nous avons introduit 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.

→ Apprenons à présent à maximiser cette fonction !

# Plan du chapitre

- 1 Equations de Bellman
- 2 Programmation Dynamique
- 3 Conclusions

## Avertissement

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

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

Problème d'**optimisation** :

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

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



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



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

## Astuce : équation de Bellman

### Proposition

Pour toute politique stationnaire  $\pi = (\pi, \pi, \dots)$ , 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 :*

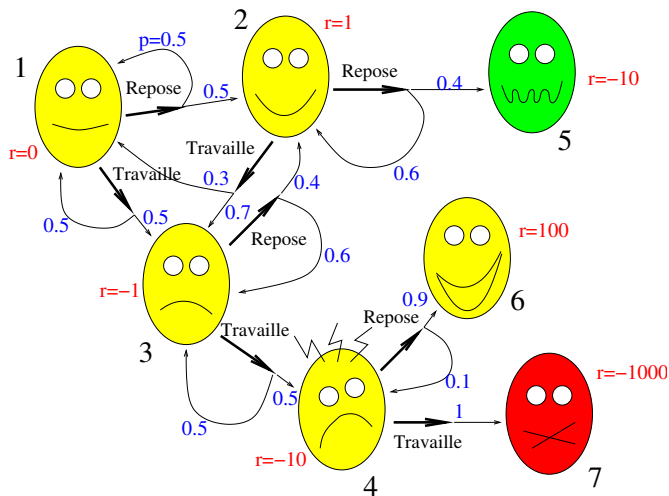
Pour toute politique  $\pi$ , on a

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





## Exemple : le dilemme de l'étudiant



## Exemple : le dilemme de l'étudiant

- **Problème** : horizon infini 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  $x_{t+1}$

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

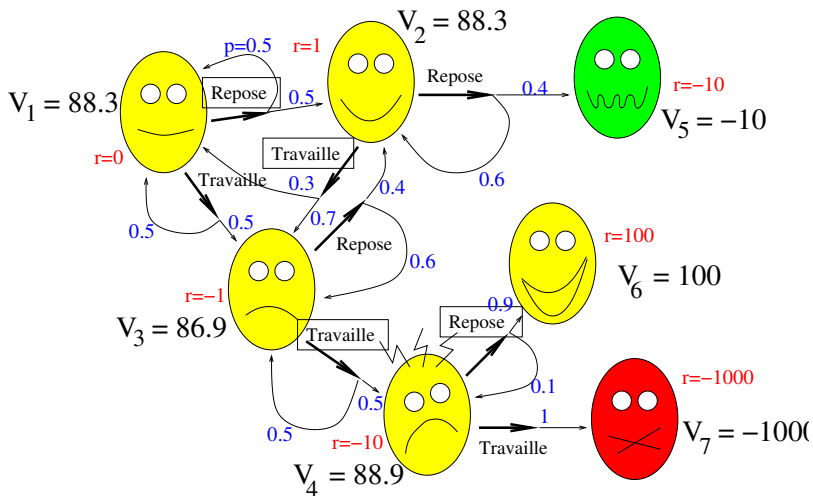
## Exemple : le dilemme de l'étudiant

- 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  $x_{t+1}$

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

- 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  $\pi$ 

Exemple : le dilemme de l'étudiant - calcul incrémental de  $V^\pi$  grâce à Bellman

Notons  $V_i = V^\pi(x_i)$ . 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  $V^\pi$  grâce à Bellman

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

$$V_4 = 88.8$$

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

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

et ainsi de suite pour le reste...

Équation de Bellman optimale :

## Principe d'Optimalité de Bellman :

*“Une politique 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 1<sup>ère</sup> décision.”*



Équation 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 :

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

et la politique optimale est

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

## Équation de Bellman optimale :

*Preuve*

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

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



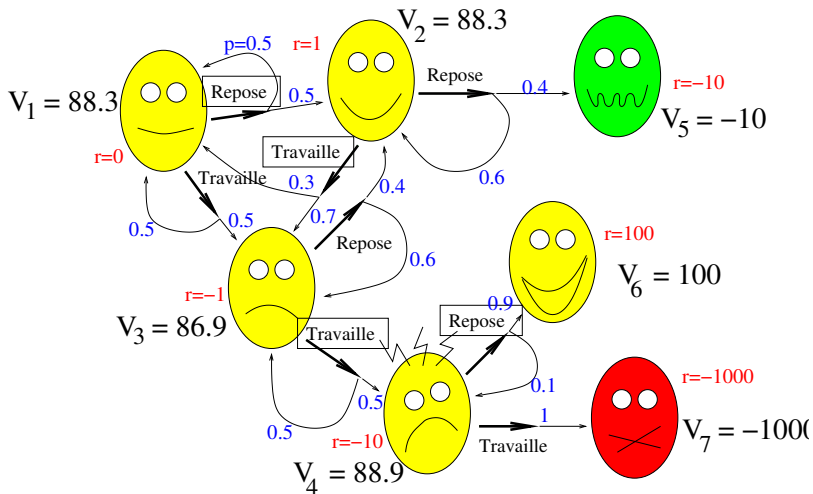
## Système d'équations :

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.

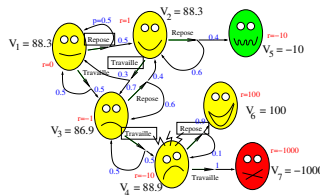
Exemple : le dilemme de l'étudiant (rappel de la politique choisie  $\pi$ )



## 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



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

 $\Rightarrow$ 

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

$$V = R + PV$$

 $\Downarrow$ 

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

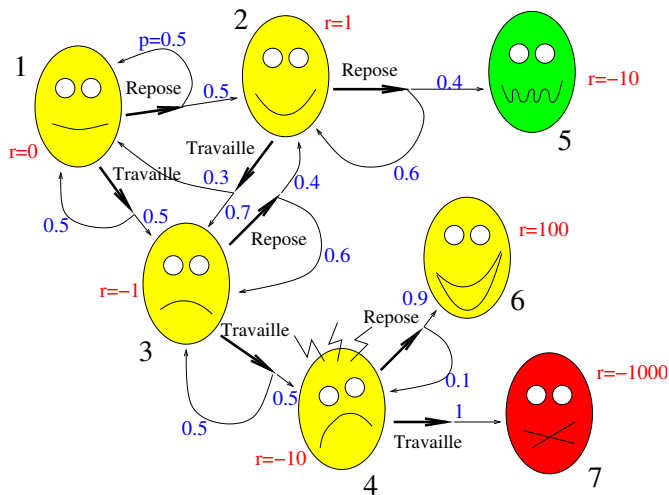
## Système d'équations :

L'équation optimale de Bellman

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

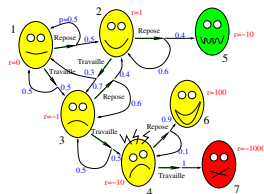
forme un système **non-linéaire** d'équations avec  $N$  inconnues et  $N$  contraintes non-linéaires (l'opérateur **max**).

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



## Exemple : le dilemme de l'étudiant

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



## Système d'équations

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

⇒ 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  $\mathcal{T}^\pi : \mathbb{R}^N \rightarrow \mathbb{R}^N$  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 opérateur de programmation dynamique) est

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

# 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  $c \in \mathbb{R}$ ,

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

## Les Opérateurs de Bellman :

### Proposition

3. **Contraction pour la norme  $L_\infty$**  : 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  $\pi$

$$\begin{aligned} V^\pi &\text{ est l'unique point fixe de } \mathcal{T}^\pi, \\ V^* &\text{ est l'unique point fixe de } \mathcal{T}. \end{aligned}$$

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

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

## L'équation de Bellman :

### Preuve

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

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

où dans  $(*)$  on utilise  $|\max_a f(a) - \max_{a'} g(a')| \leq \max_a |f(a) - g(a)|$ . ■

# Plan du chapitre

- 1 Equations de Bellman
- 2 Programmation Dynamique
- 3 Conclusions

## Question :

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

⇒ 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

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

forme un système **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  $k = 1, 2, \dots, K$ 
  - 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 - Complexité

## 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 infini avec affaiblissement, pour tout politique  $\pi$ , la *fonction de valeur état-action* (ou Q-fonction) est

$Q^\pi : X \times A \mapsto \mathbb{R}$  is

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

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 :

$$Q^\pi(x, a) = r(x, a) + \gamma \sum_{y \in X} p(y|x, a) V^\pi(y)$$

$$V^\pi(x) = Q^\pi(x, \pi(x))$$

$$Q^*(x, a) = r(x, a) + \gamma \sum_{y \in X} p(y|x, a) V^*(y)$$

$$V^*(x) = Q^*(x, \pi^*(x)) = \max_{a \in A} Q^*(x, a).$$

## Idée n°1 : Itération par valeurs - Implémentations alternatives

### *Q-iteration.*

- ❶ Soit  $Q_0$  une Q-fonction quelconque
- ❷ A chaque itération  $k = 1, 2, \dots, K$ 
  - Calculer  $Q_{k+1} = \mathcal{T}Q_k$
- ❸ 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.

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

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

## Comparaison

- Complexité calculatoire **réduite** à  $O(N|A|)$ .
- Nombre d'itération maximum **augmenté** 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  $\pi_0$  une politique stationnaire *quelconque*
- ② A chaque itération  $k = 1, 2, \dots, K$ 
  - *Évaluation de la politique* sachant  $\pi_k$ , calculer  $V^{\pi_k}$ .
  - *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  $\pi_K$

*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  $\pi^*$  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  $\pi_{k+1}$

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

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

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

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 finalement 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  $\pi_k$  est une politique optimale. ■

## Idée n°2 : Itération par politiques

### Notations supplémentaires :

- Pour toute politique  $\pi$ , 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 - Étape d'évaluation de la politique

- *Par calcul direct.* Pour toute politique  $\pi$  calculer

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

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

- *Itérativement.* Pour toute politique  $\pi$

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

*Complexité* : Une approximation de  $V^\pi$  à  $\epsilon$  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  $\pi$  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 - Étape 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  $O\left(\frac{N|A|}{1-\gamma} \log\left(\frac{1}{1-\gamma}\right)\right)$

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

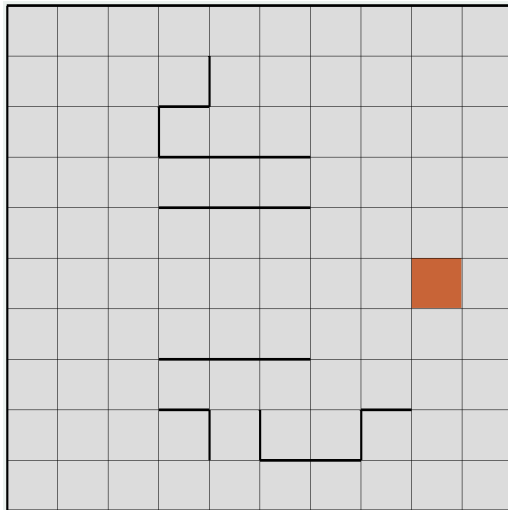
### *Itération par valeur*

- + : 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

- 1 Equations de Bellman
- 2 Programmation Dynamique
- 3 Conclusions**



## Messages importants du chapitre :

- Les équations de **Bellman** offrent 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**.