Processus Décisionnels de Markov (Partiellement Observables)

(PO)MDP

Alain Dutech

Equipe MAIA - LORIA - INRIA
Nancy, France
Web : http://maia.loria.fr
Mail : Alain.Dutech@loria.fr

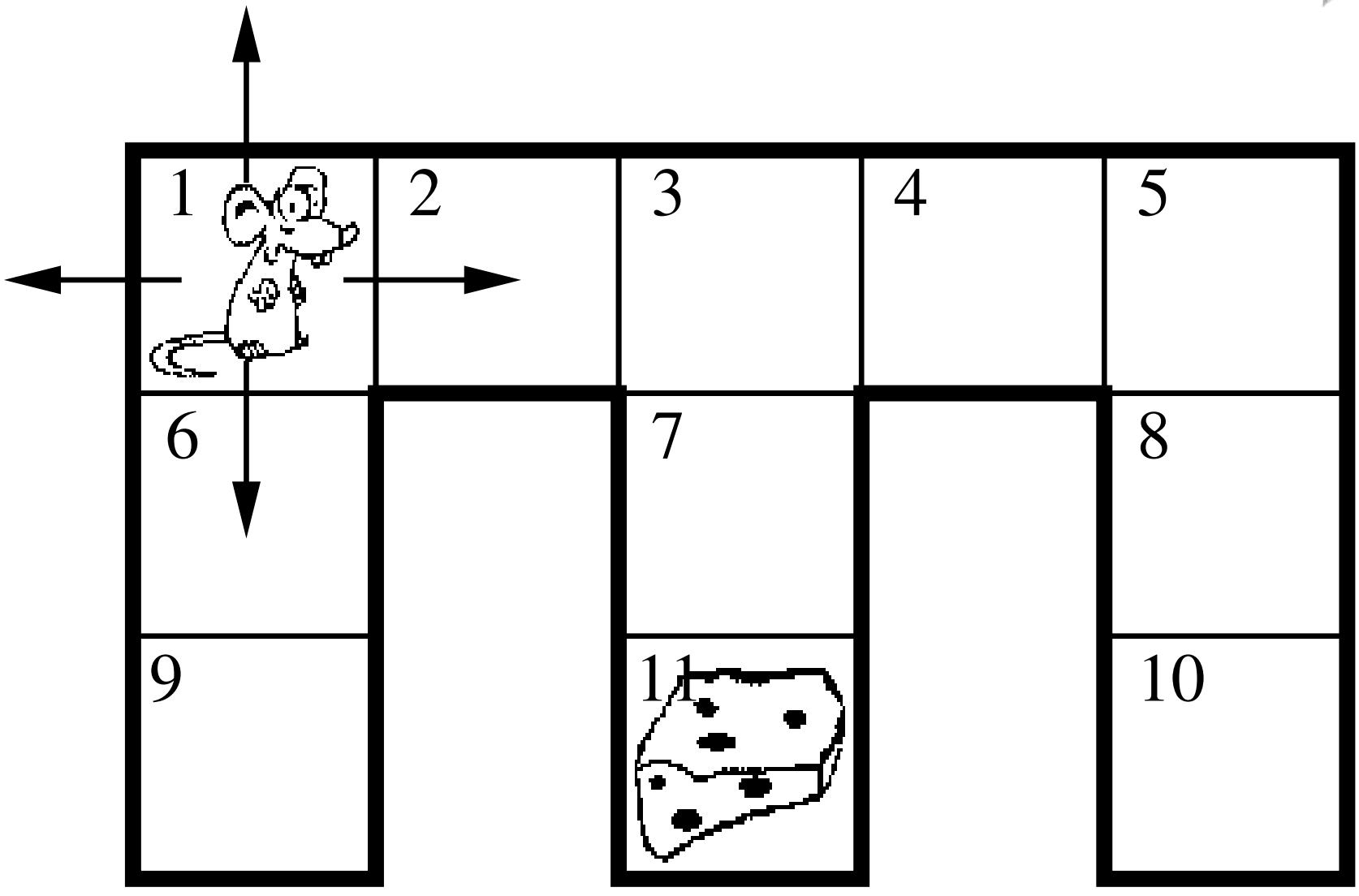
Nancy - 08/07/2014

Plan de l'exposé

Plan de l'exposé

Outline

Décision séquentielle dans l'incertain



Etats 1–11

MDP - Vue Générale (réf : [Bellman, 1957], [Groupe PDMIA, 2008])

Espaces

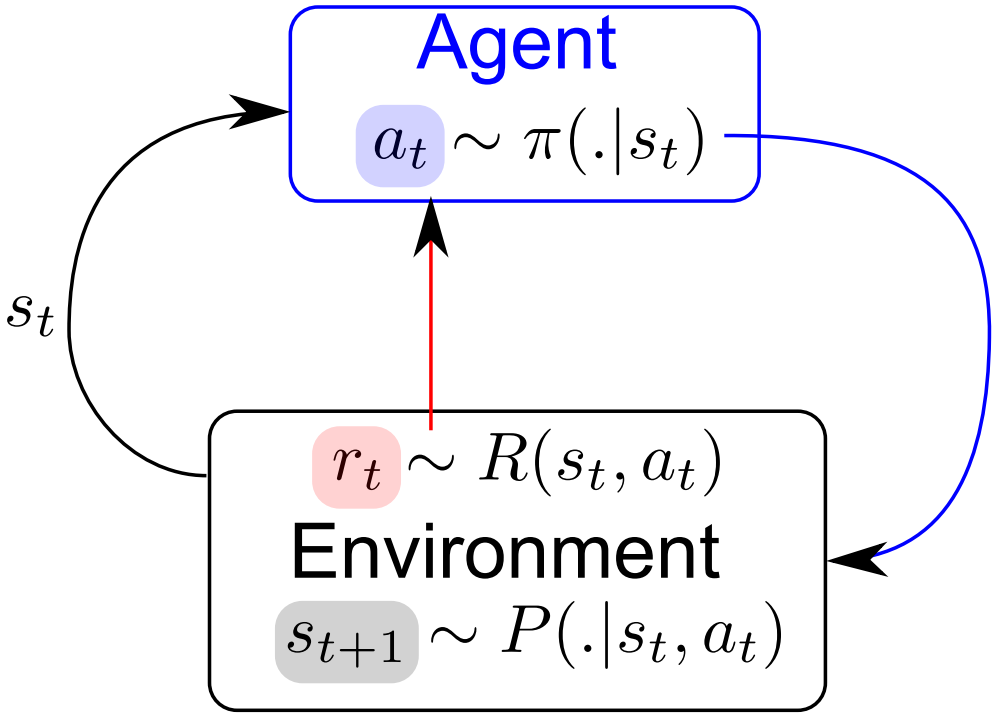
- $\mathcal{S}$: états
- $\mathcal{A}$: actions

Dynamique

- $P(s_{t+1}|s_t, a_t)$: transition
- $R(s, a)$: récompense

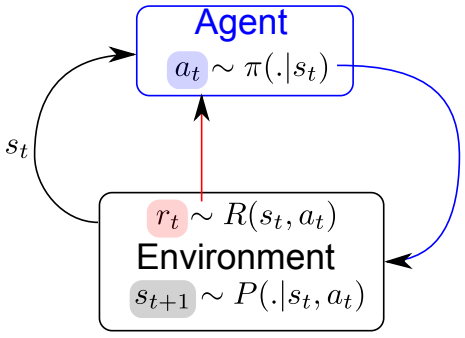
Agent

- $\pi(a_t|s_t)$: politique



Critère $V^\pi(s) = \mathbb{E}_{\sim\pi}\left[\sum_{t=1}^{T} \gamma^t r_t | s_0 = s\right], \qquad \gamma \in [0, 1[$

MDP - Classiquement



- $\mathcal{S}$ fini ($\rightsquigarrow$ grande taille ou continu)
- $\mathcal{A}$ fini ($\rightsquigarrow$ continu moins "facile")
- Critère
  - Horizon infini ou fini
  - ($\rightsquigarrow$ Critère moyen $1/T \sum r_t$)
- Politique déterministe ($\rightsquigarrow$ stochatique peu utile)

Trouver une politique qui optimise le critère

- Il existe des politiques optimales π^*

$$V^{\pi^*}(s) \geq V^{\pi}(s), \quad \forall s \in \mathcal{S}, \forall \pi$$

- Au moins une de ces politiques optimales est déterministe

MDP – Equations de Bellman

Equation de Bellman

$$V(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V(s'), \ \forall s \in \mathcal{S}$$

$\rightsquigarrow$ unique solution est $\mathcal{V}^{\pi}$

Equation d'optimalité de Bellman

$$V(s) = \max_{a \in \mathcal{A}} \left(R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V(s') \right), \ \forall s \in \mathcal{S}$$

$\rightsquigarrow$ unique solution est V^*
$\rightsquigarrow$ politique optimale :

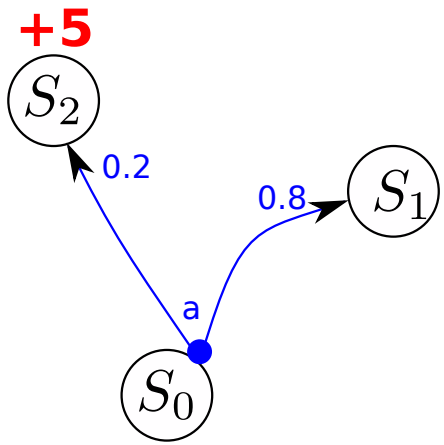
$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \left(R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V^*(s') \right), \ \forall s \in \mathcal{S}$$

MDP - Exemple

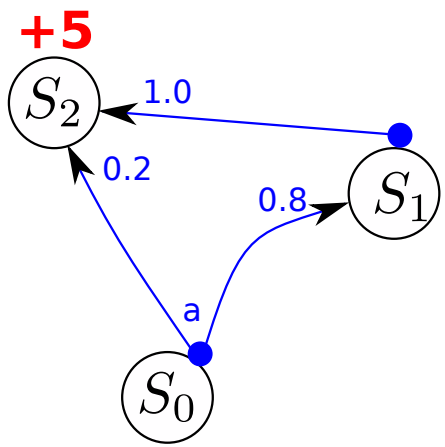
Fonction Qualité

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s')$$

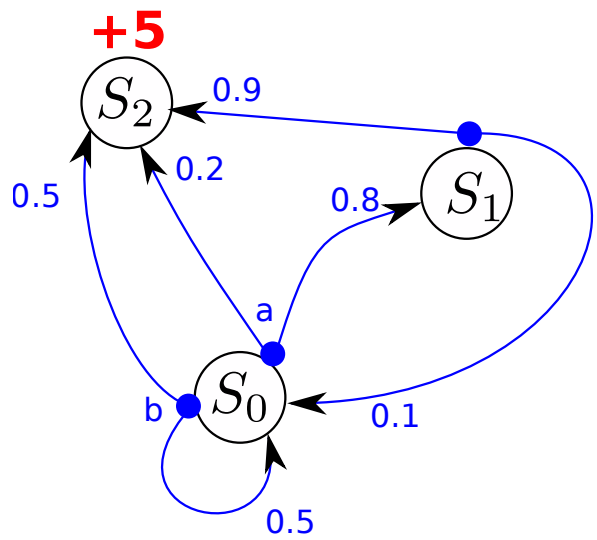
$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$$



- $Q(S_0, a) = 0.9$



- $Q(S_0, a) = 4.14$
- $Q(S_1, .) = 4.5$



- $Q(S_0, a) = 4.08$
- $Q(S_0, b) = 4.09$
- $Q(S_1, .) = 4.41$

Outline

Itération de la Valeur (Value Iteration)

Schéma de calcul itératif

1. $V_0(s) \leftarrow c,\ \forall s \in \mathcal{S}$
2. $V_{i+1}(s) \leftarrow \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V_i(s') \right),\ \forall s \in \mathcal{S}$
3. Stop si $||V_{i+1} - V_i|| < \epsilon$, sinon recommence (2)

Convergence

1. (2) est une contraction : $||V_{n+1} - V_n|| \leq \gamma ||V_n - V_{n-1}||$
2. Théorème de Banach (pt fixe, convergence)
3. $||V^{VI} - V^*|| \leq \epsilon \frac{2\gamma}{1-\gamma}$

avec $||V|| = \max_{s \in \mathcal{S}} V(s)$

Itération de la Politique (Policy Iteration)

Schéma de calcul itératif

1. initialiser $\pi_0(s),\ \forall s \in \mathcal{S}$
2. calculer V^{π_i} $\quad$ ($(I - \gamma P^{\pi_i})^{-1}$ ou Schéma Itératif ou ...)
3. $\pi_{i+1}(s) = \text{argmax}_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi_i}(s') \right),\ \forall s \in \mathcal{S}$
4. Stop si $\pi_{i+1}(s) = \pi_i(s), \forall s \in \mathcal{S}$, sinon (2)

Convergence

- $V^{\pi_{i+1}}(s) \geq V^{\pi_i}(s),\ \forall s \in \mathcal{S}$
- Nombre fini de politiques différentes.

Outline

Apprentissage par Renforcement

Modèle (P et R) **n'est pas connu par agent**

$\rightsquigarrow$ Apprendre valeur/politique avec échantillons (s, a, s', r)

Ex : Q-Learning, approximation stochastique de la valeur

1. Dans un état courant s
2. Agent $\xrightarrow{\pi_{exp}}$ a
3. Environnement $\xrightarrow{P,R}$ r et s'.
4. Mise-à-jour

$$Q(s,a) \longleftarrow Q(s,a) + \alpha \overbrace{[r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)]}^{\Delta}$$

5. Reprendre en (1) avec $s \longleftarrow s'$

Convergence

Visiter chaque (s, a) infinité fois, $\sum \alpha \to \infty$, $\sum \alpha^2 < \infty$.

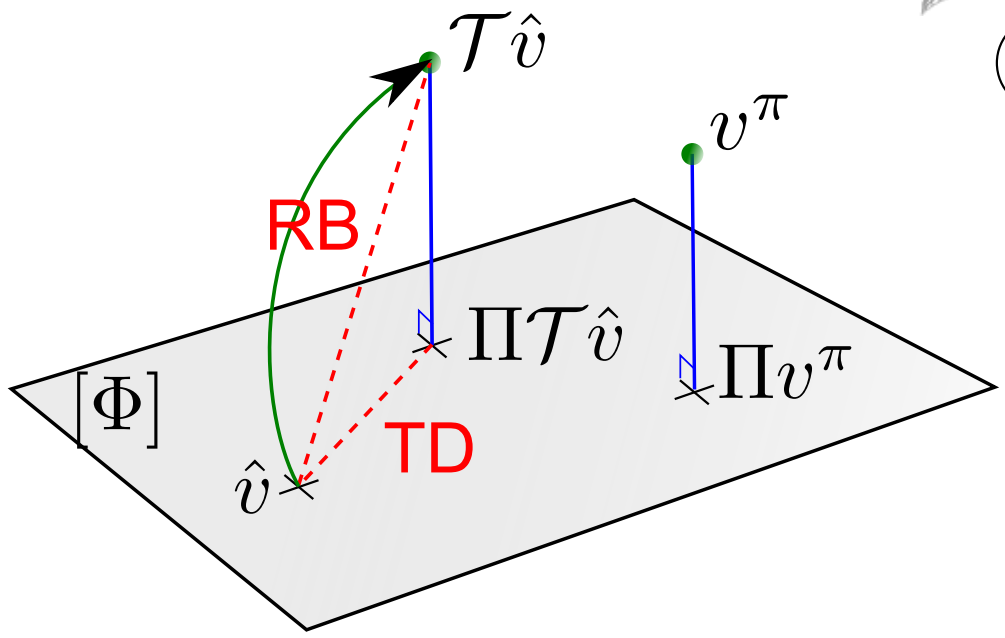
(réf : [Sutton and Barto, 1998], [Szepesvári, 2010])

Outline

Itération de la Valeur Approchée [Scherrer, 2010]

$v = \mathcal{T} v$
où $\mathcal{T} = (I - \gamma P)^{-1}$

$|\mathcal{S}|$ est grand ou $\mathcal{S}$ continu.
Approximation $\hat{v} \in [\Phi]$
Π : Projection sur $[\Phi]$



Approx. Linéaire : $\hat{v} = \sum_k \phi_k w_k = \Phi \mathbf{w}$
Echantillons : (s, a, s', r)

Solution TD

- min $||\Pi\mathcal{T}\hat{v} - \hat{v}||_\xi$
- pas de garantie
- simple échantillonage

Résidu Bellman

- min $||\mathcal{T}\hat{v} - \hat{v}||_\xi$
- garantie :
  $||v^\pi - \hat{v}||_\xi \leq \frac{\sqrt{C(\xi)}}{1-\gamma} ||\mathcal{T}\hat{v} - \hat{v}||_\xi$
- mais double échantillonage

Recherche Directe de Politique

- Politique π est **paramétrée** (par $\mathbf{w}$: $\pi(.|s, \mathbf{w})$
- Associer une valeur à une politique.

$$\bar{J}(\mathbf{w}) = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} R_t\right]$$

- **modifier** les paramètres pour maximiser la valeur.

MAIS

- Quel algorithme pour modifier les paramètres ? (**Gradient** ou **EM**)
- Besoin d'un *critic* (pour estimer V) ?
- Minimum local, (espaces de grande taille ?).

Montée de Gradient [Kimura et al., 1997], [Baxter and Bartlett, 2001]

Le Gradient de la fonction objectif $\nabla_{\mathbf{w}} \bar{J} = R(s) \frac{\nabla_{\mathbf{w}} \mu(s;\pi,w)}{\mu(s;\pi,w)}$

où μ est la distribution stationnaire

Estimation **non-biasée** si processus régénérant ou avec épisodes.

- A chaque étape : $z_{t+1} = z_t + \frac{\nabla_{\mathbf{w}} \pi(s_t,a_t,\mathbf{w})}{\pi(s_t,a_t,\mathbf{w})}$, $v_{t+1} = v_t + R_t$
- Fin d'un épisode : $\Delta_{j+1} = \Delta_j + v_t z_t / \text{length}$
- résultat Δ_N / N

Estimation **biaisée** en-ligne

- A chaque étape : $z_{t+1} = \beta z_t + \frac{\nabla_{\mathbf{w}} \pi(s_t,a_t,\mathbf{w})}{\pi(s_t,a_t,\mathbf{w})}$, $w_{t+1} = w_t + \alpha_t . R_t . z_t$
- résultat w_T

$\rightsquigarrow$ eNAC, ou EM (mais Primitive Dynamique de Mvt, Importance Sampling), "faible" nb de DoF ?

Outline

Références I

Baxter, J. and Bartlett, P. (2001).

Infinite-horizon policy-gradient estimation.
Journal of Artificial Intelligence Research, 15 :319–350.

Bellman, R. (1957).

Dynamic programming.
Princeton University Press, Princeton, New-Jersey.

Groupe PDMIA (2008).

Processus Décisionnels de Markov en Intelligence Artificielle. (Edité par Olivier Buffet et Olivier Sigaud), volume 1 & 2.
Lavoisier - Hermes Science Publications.

Kimura, H., Miyazaki, K., and Kobayashi, K. (1997).

Reinforcement learning in POMDPs with function approximation.
In *Proc. of the Fourteenth Int. Conf. on Machine Learning (ICML'97)*, pages 152–160.

Scherrer, B. (2010).

Should one compute the Temporal Difference fix point or minimize the Bellman Residual ? The unified oblique projection view.
In *Proc. of the 27th Int. Conf. on Machine Learning (ICML'2010)*.

Sutton, R. and Barto, A. (1998).

Reinforcement Learning.
Bradford Book, MIT Press, Cambridge, MA.

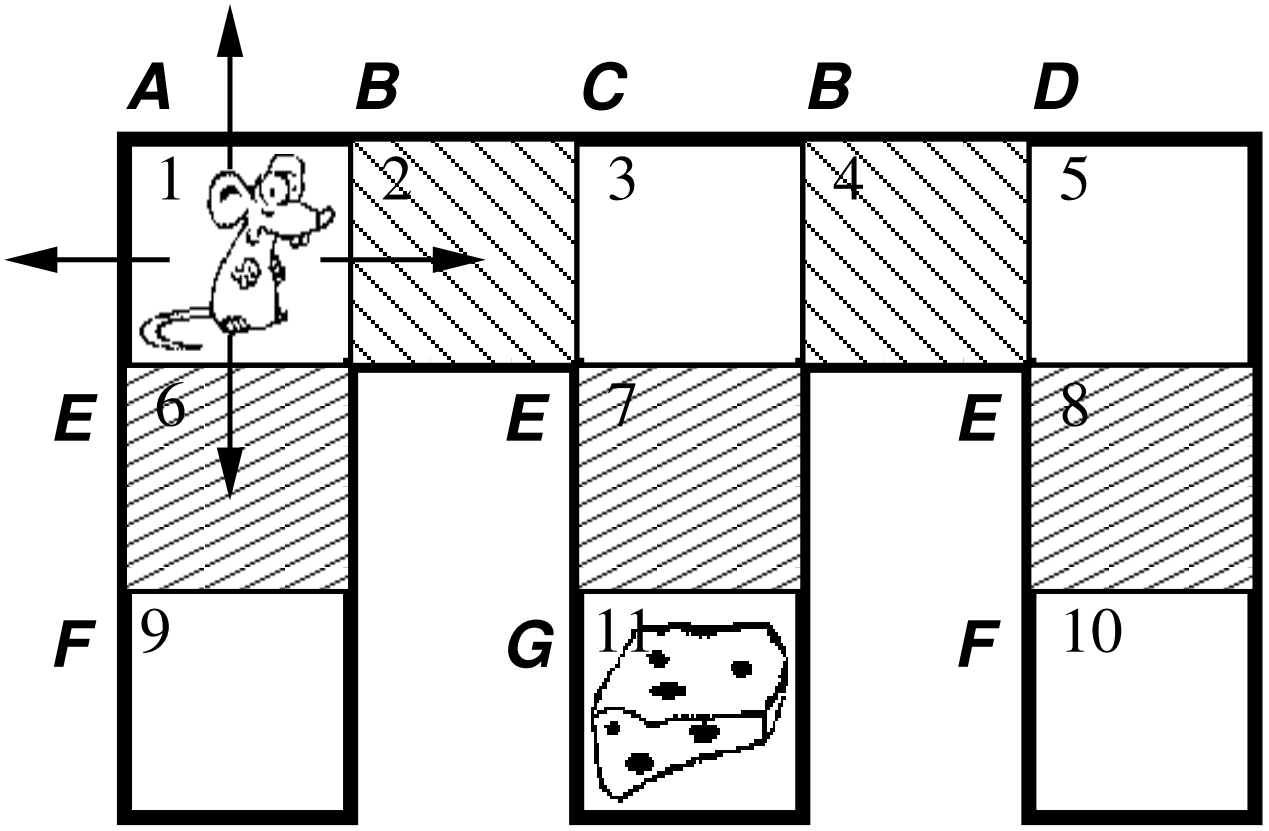
Szepesvári, C. (2010).

Algorithms for Reinforcement Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning).
Morgan and Claypool.

Outline

Partiellement Observable

Perceptions : dans la case 1, la souris ne perçoit que les murs à gauche et en haut.



Etats 1–11 **Observations A–G**

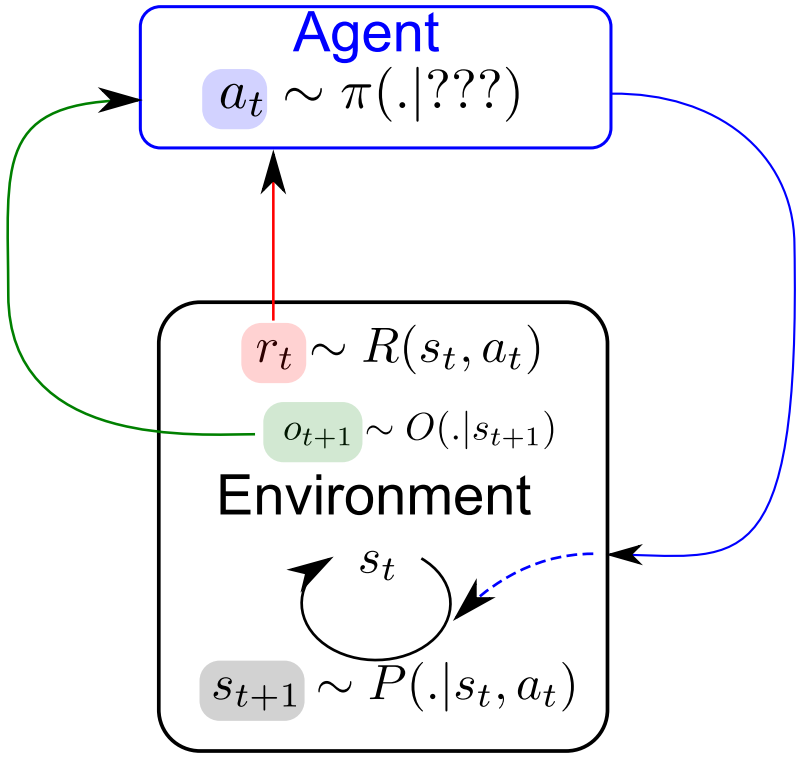
POMDP - Vue Générale [Groupe PDMIA, 2008]

Espaces

- $\mathcal{S}$: états
- $\mathcal{A}$: actions
- Ω : observations

Dynamique

- $P(s_{t+1}|s_t, a_t)$: transition
- $R(s, a)$: récompense
- $O(o_t|s_t)$: f. observation
- $b_0 = P(s_0)$: initial



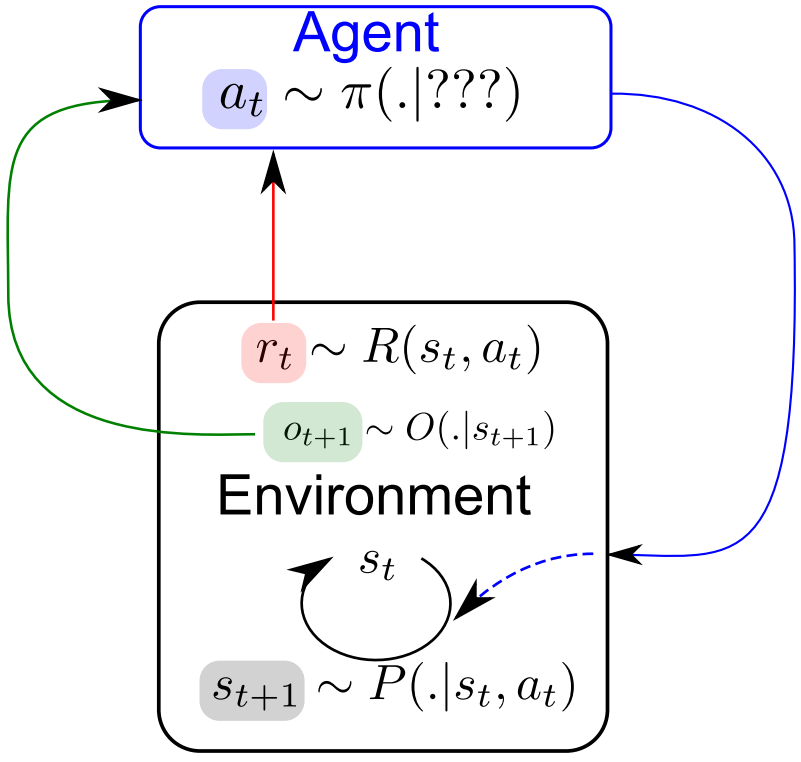
POMDP - Vue Générale [Groupe PDMIA, 2008]

Espaces

- $\mathcal{S}$: états
- $\mathcal{A}$: actions
- Ω : observations

Dynamique

- $P(s_{t+1}|s_t, a_t)$: transition
- $R(s, a)$: récompense
- $O(o_t|s_t)$: f. observation
- $b_0 = P(s_0)$: initial



Problème

Trouver une politique optimale (maximise $\mathbb{E}_{\sim\pi}\left[\sum_{t=1}^{T} \gamma^t r_t | s_0 = s\right]$).

- non-markovien : existence d'une *Fonction de Valeur* ?
- état d'information : quel est le support de la politique ? $\pi(a_t|???)$

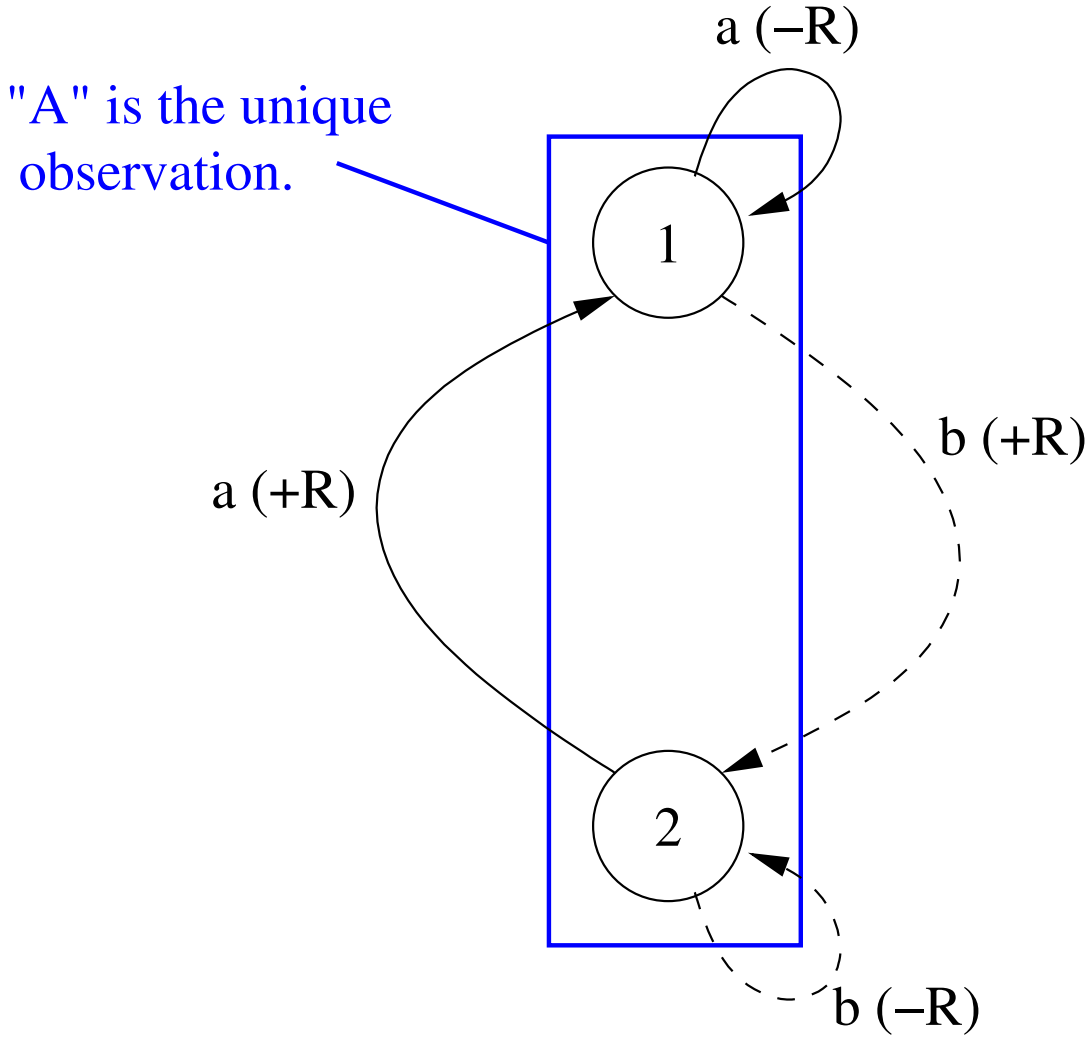
Outline

Politique Adaptée Stochastique

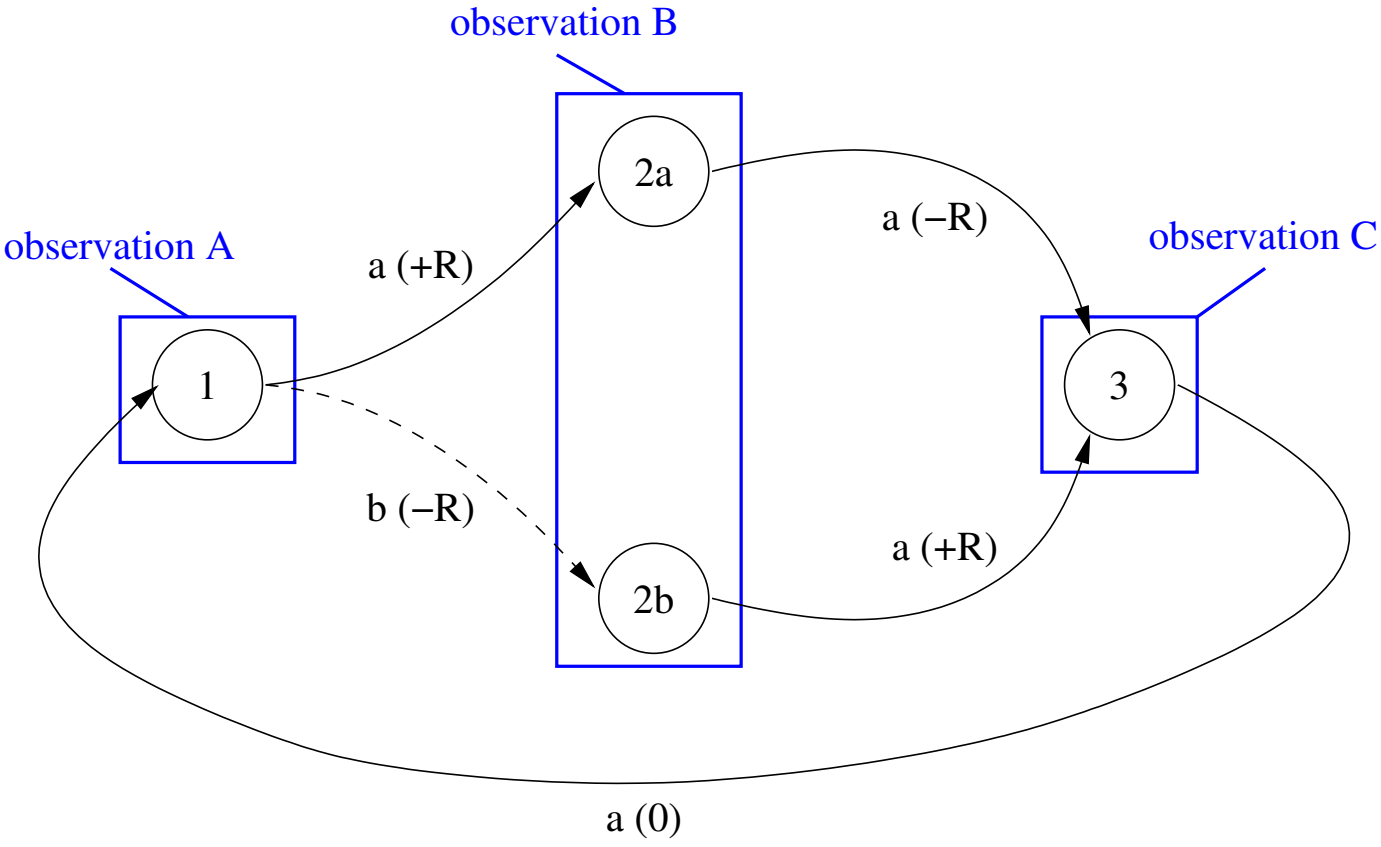
Politique adaptée : sans-mémoire

$\pi : \Omega \longrightarrow \mathcal{A}$

Une politique **stochastique** sans-mémoire peut être **arbitrairement meilleure** qu'une politique sans mémoire déterministe. [Singh et al., 1994]



Pas de Politique sans-Mémoire Optimale



Il n'existe pas de politique sans-mémoire qui optimise la fonction de valeur "*adaptée*" [Singh et al., 1994]

$$\mathcal{V}^{\pi}(o) = \sum_{s \in \mathcal{S}} P^{\pi}(s|o) V^{\pi}(s)$$

Convergence des algorithmes “classiques” [Jaakkola et al., 1994]

- TD(0)

$$\forall o \in \Omega,\ \mathcal{V}(o) = \sum_{s \in \mathcal{S}} P^{\pi}(s|o) \left[R(s) + \gamma \sum_{o' \in \Omega} P^{\pi}(s, o') \mathcal{V}(o') \right],$$

où $P^{\pi}(s, o') = \sum_{s' \in \mathcal{S}} P^{\pi}(s'|s) O(o'|s')$.

- Q-Learning

$$\mathcal{Q}(o, a) = \sum_{s \in \mathcal{S}} P^{\pi_{\exp}}(s|o, a) \left[R(s, a) + \gamma \sum_{o' \in \Omega} P_a(s, o') max_{a' \in \mathcal{A}} \mathcal{Q}(o', a') \right],$$

où $P^{\pi_{\exp}}(s|o, a)$ est la distribution de probabilité asymptotique d'occupation et où $P_a(s, o') = \sum_{s' \in \mathcal{S}} P(s'|s, a) O(o'|s')$.

Outline

"*Belief States*" : statistique suffisante du passé

"Belief States"

distribution sur les états : $b_t(s) = P(s_t = s)$

- Statistique suffisante : $b_t(s) = P(s_t | a_t, s_{t-1}, \ldots, s_0)$

$\rightsquigarrow$ Etat d'Information complet.

- Mise à jour bayésienne :

$$b_o^a(s') = P(s' | b, a, o)$$
$$= \frac{O(o|s') \sum_{s \in \mathcal{S}} P(s'|s,a) b(s)}{\sum_{s \in \mathcal{S}} \sum_{s'' \in \mathcal{S}} O(o|s'') P(s''|s,a) b(s)}.$$

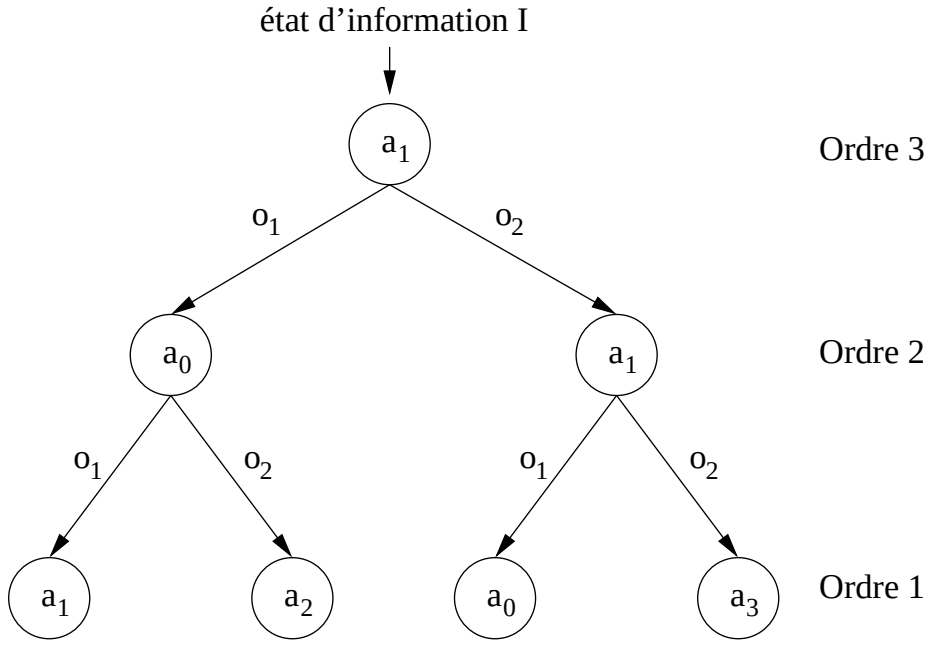
$\rightsquigarrow$ définit un MDP sur un espace d'état continu, qui peut être résolu [Aström, 1965].

Politique en Arbre

Si les supports sont les *belief states*

- Politique optimale Déterministe
- Sorte de Plan Conditionnel.

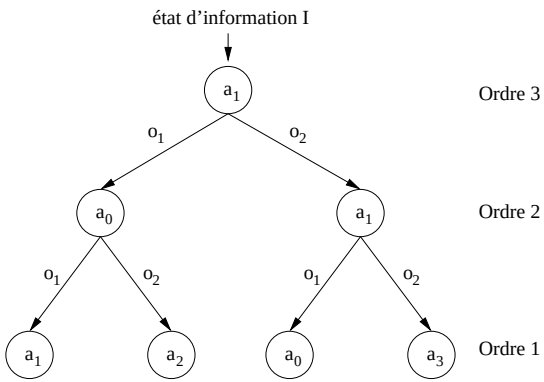
⇝ représentée par un arbre

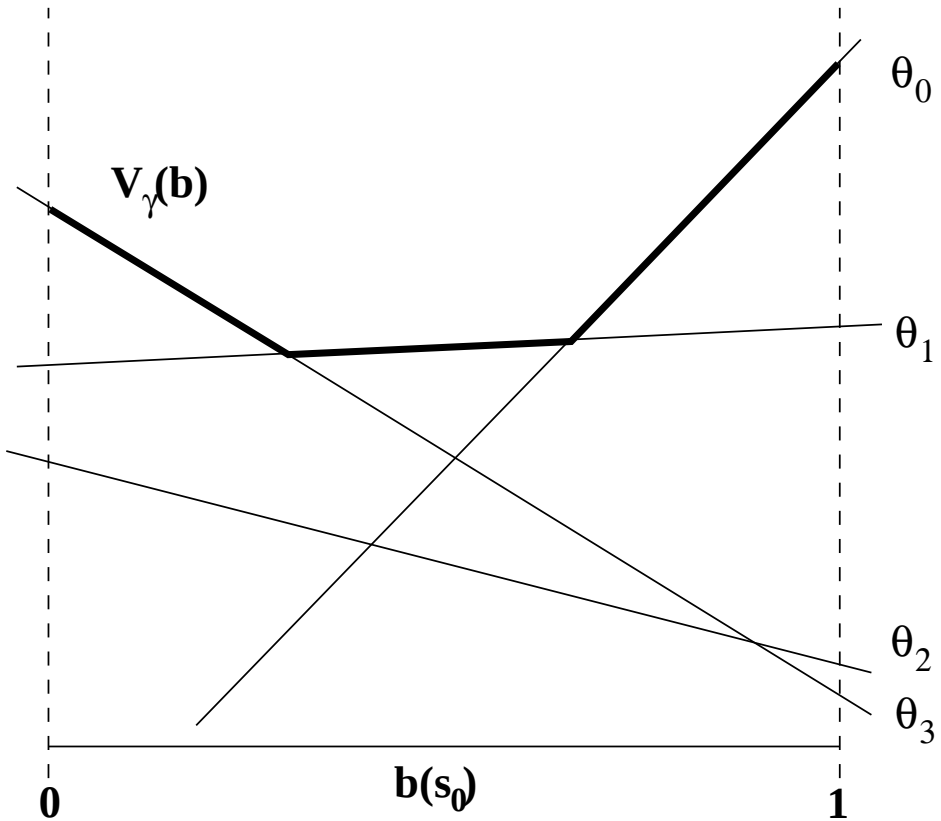


Fonction de Valeur Linéaire Par Morceau Convexe (LPMC)

$$V_n^*(b) = \max_{a \in \mathcal{A}} \left[R(b,a) + \gamma \sum_{o \in \Omega} P(o|b,a) V_{n-1}^*(b_o^a)) \right].$$

- Horizon fini n
- Vecteur θ_i = une politique...





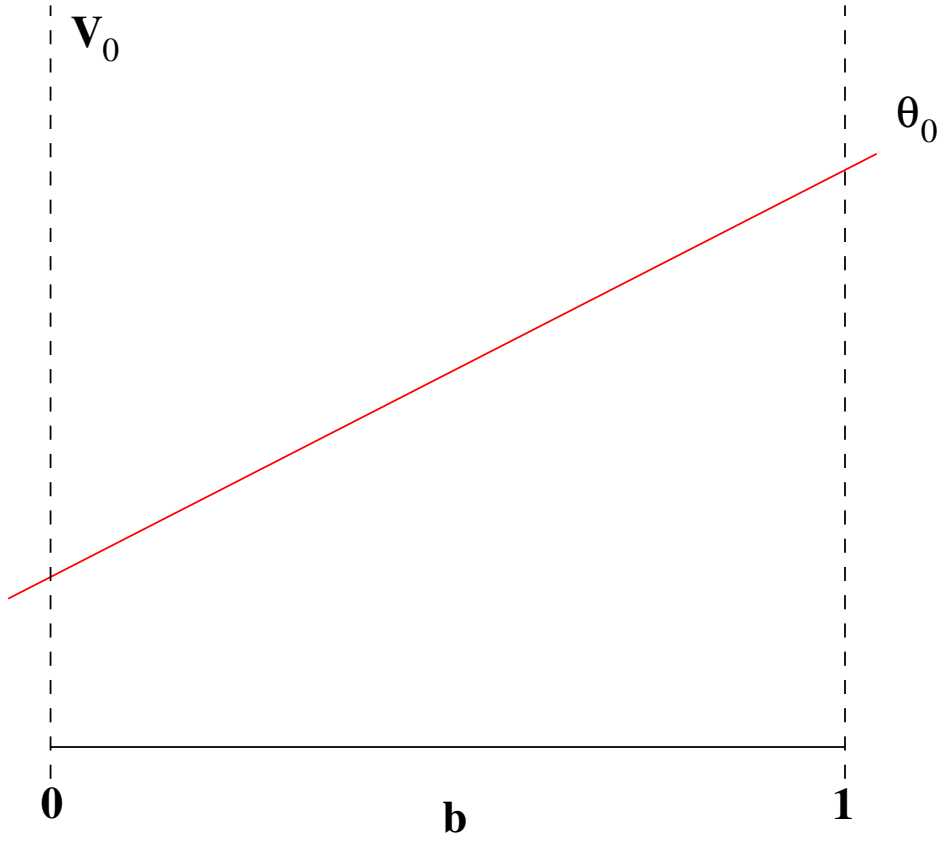
Outline

Prog. Dynamique et Fonction de Valeur LPMC (0)

Horizon de 0

"*belief state*" : b
Fonction de valeur linéaire en b.
Trivialement convexe

$$R(b) = \sum_s b(s)R(s)$$

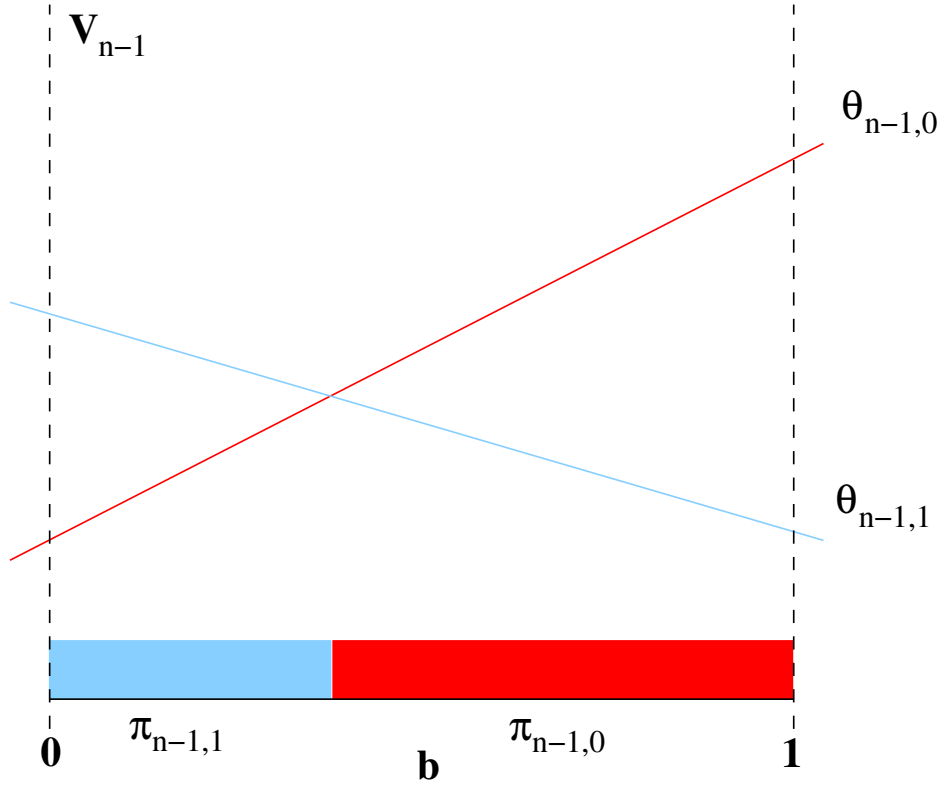


Prog. Dynamique et Fonction de Valeur LPMC (1)

LPMC V_{n-1} à l'étape $n-1$

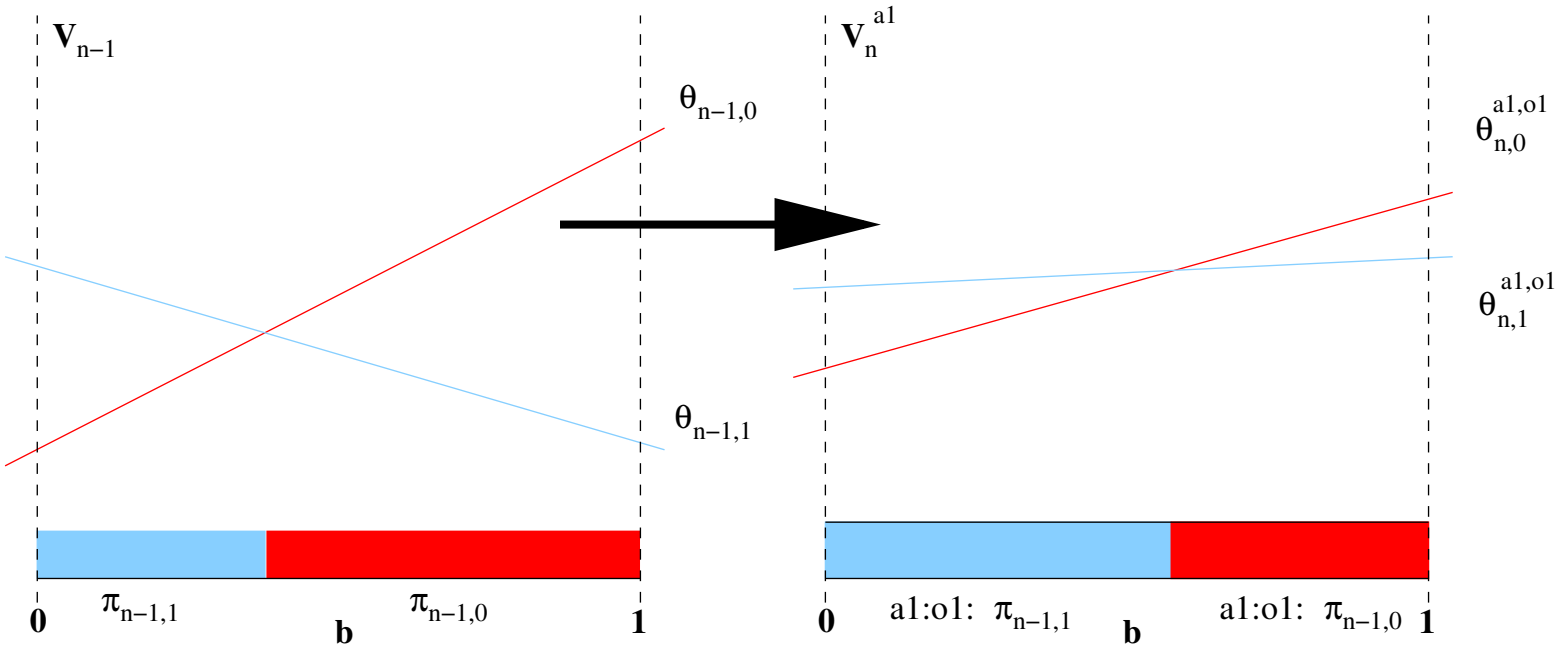
$$V_{n-1}(b) = \max_{\theta \in \Theta_{n-1}} b.\theta$$

un θ est associé à une politique



Prog. Dynamique et Fonction de Valeur LPMC (2)

LPMC V_n à l'étape n pour la première action $a1$ et l'observation $o1$



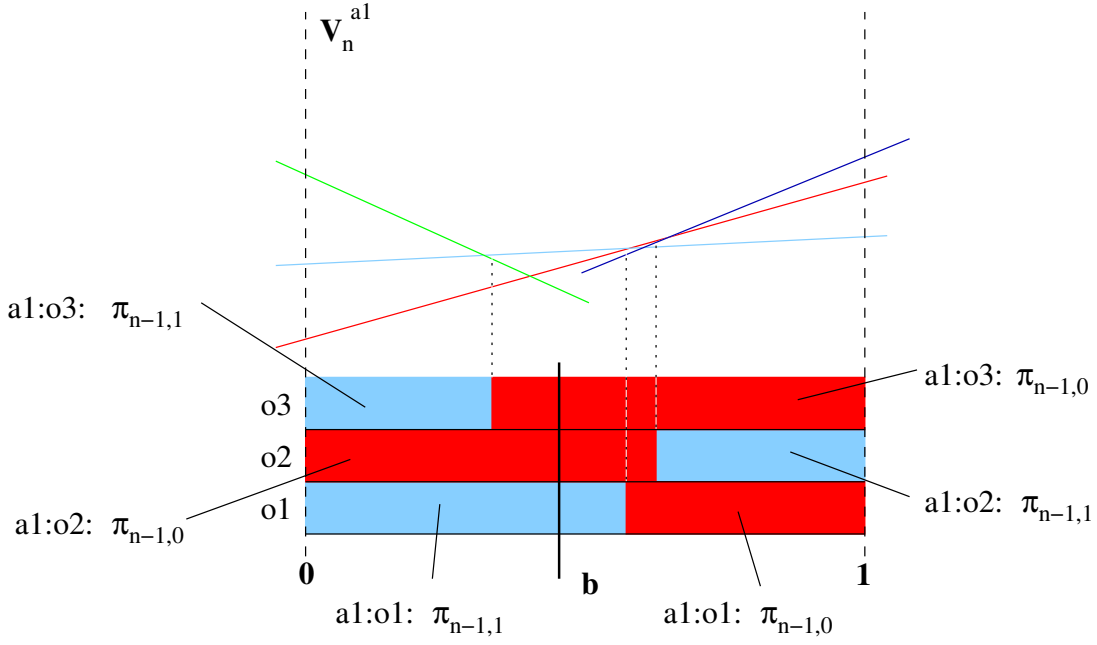
$$\theta_n^{a1,o1}(b,s) = \frac{R(s,a1)}{|\Omega|} + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a1) O(o1|s') \theta_{n-1}^{a1,o1}(b^{a1,o1},s).$$

Prog. Dynamique et Fonction de Valeur LPMC (3)

LPMC V_n à l'étape n pour la première action $a1$

$$\theta_n^{a1}(b) = \sum_{o \in \Omega} \theta_n^{a1,o}(b).$$

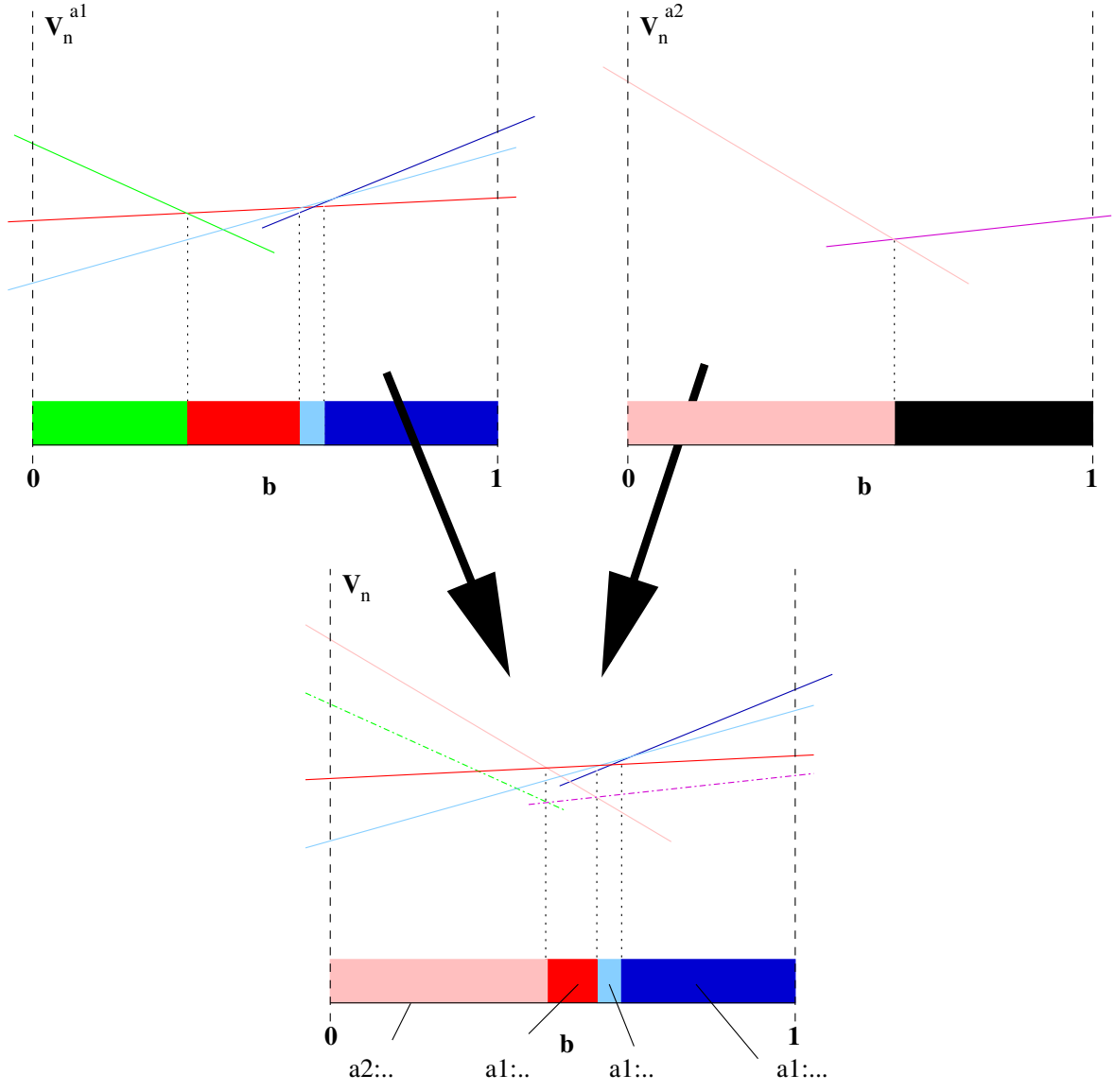
Il y a au plus $|\Theta_{n-1}|^{|\Omega|}$ vecteurs



Prog. Dynamique et Fonction de Valeur LPMC (4)

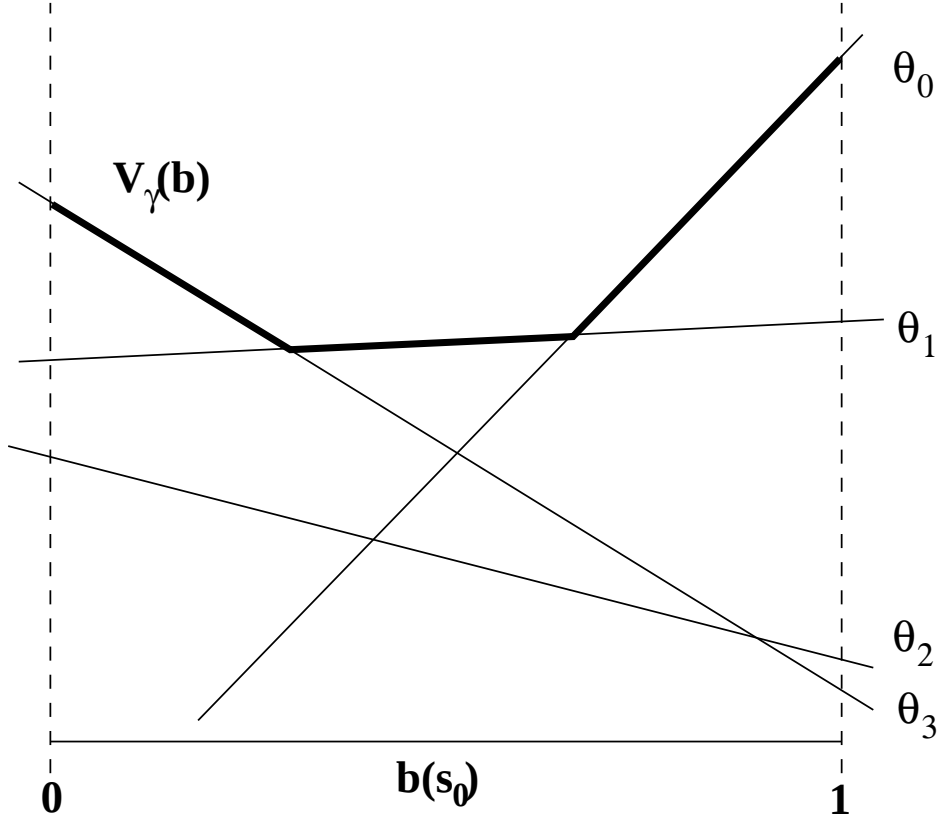
LPMC V_n à l'étape n

$$\theta_n(b) = \max_{a \in \mathcal{A}} \theta_n^a(b)$$



Fonction LPMC avec "Belief States"

- POMDP à Horizon Fini
  - La Fonction de Valeur optimale est LPMC [Smallwood and Sondik, 1973]
  - $$V_n(b) = \max_{\theta \in \Theta_n} b.\theta$$
- POMDP à Horizon Infini
  - Il existe des Fonction de Valeur LPMC ϵ-optimales
  - LPMC optimale seulemnt pour des POMDP *transients* [Sondik, 1971]



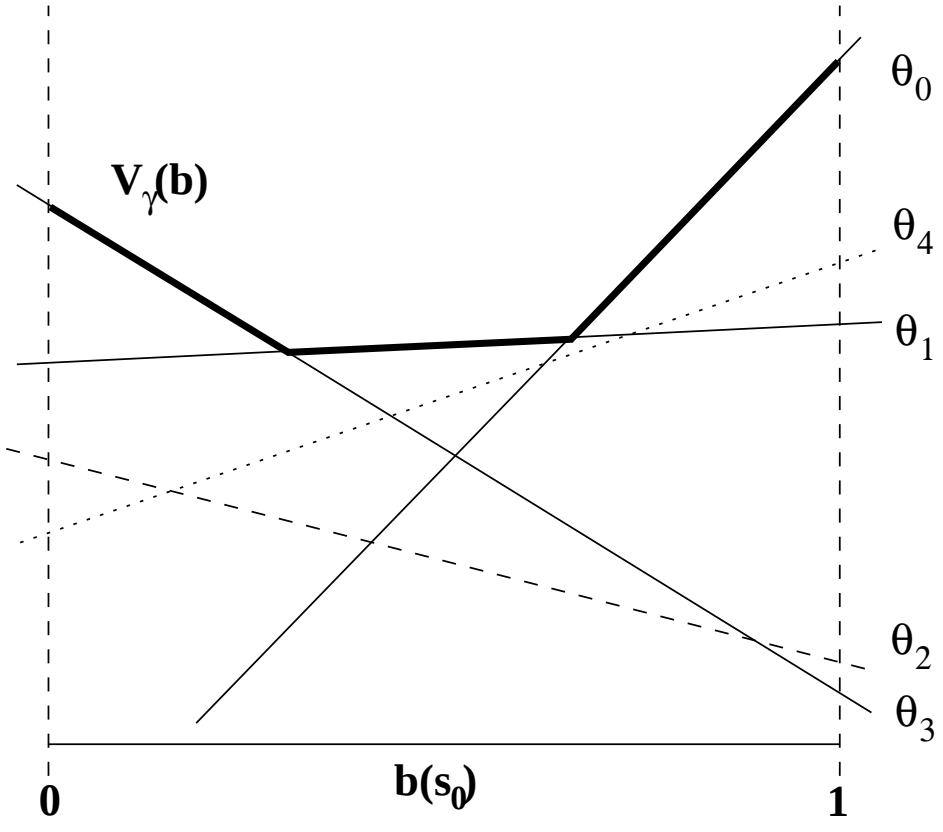
$\rightsquigarrow$ le vrai problème c'est la taille de l'espace des vecteurs Θ.

Représentation Parcimonieuse

θ de Θ est dominé : $b.\theta \leq \max_{\theta' \in \Theta} b.\theta'$.

- Il existe une représentation minimale
  [Littman and Szepesvári, 1996]

- θ_2 : entièrement dominé
- θ_4 : avec algo PRUNING



Outline

Références I

Aström, K. (1965).

Optimal control of Markov decision processes with incomplete state estimation.
Journal of Mathematical Analysis and Applications, 10 :174–205.

Cassandra, A. (1998).

Exact and Approximate Algorithms for Partially Observable Markov Decision Processes.
PhD thesis, Brown University, Department of Computer Science, Providence, RI.

Groupe PDMIA (2008).

Processus Décisionnels de Markov en Intelligence Artificielle. (Edité par Olivier Buffet et Olivier Sigaud), volume 1 & 2.
Lavoisier - Hermes Science Publications.

Hansen, E. (1998).

Solving POMDPs by searching in policy space.
In *Proc. of the Fourteenth Conf. on Uncertainty in Artificial Intelligence (UAI'98*.

Jaakkola, T., Singh, S., and Jordan, M. (1994).

Reinforcement learning algorithm for partially observable markov decision problems.
In Tesauro, G., Touretsky, D., and Leen, T., editors, *Advances in neural information processing systems*, volume 7. MIT Press, Cambridge, Massachusetts.

Références II

Littman, M. and Szepesvári, C. (1996).

A generalized reinforcement-learning model : Convergence and applications.
In *Proc. of the Thirteenth Int. Conf. on Machine Learning (ICML'96)*.

Singh, S., Jaakkola, T., and Jordan, M. (1994).

Learning without state estimation in partially observable markovian decision processes.
In *Proceedings of the Eleventh International Conference on Machine Learning*.

Smallwood, R. D. and Sondik, E. J. (1973).

The optimal control of partially observable Markov processes over a finite horizon.
Operations Research, 21 :1071–1088.

Sondik, E. (1971).

The optimal control of partially observable markov decision processes.
PhD thesis, Stanford University, California.