Récompenser la collecte d'informations

Mauricio Araya Vincent Thomas
Olivier Buffet François Charpillet

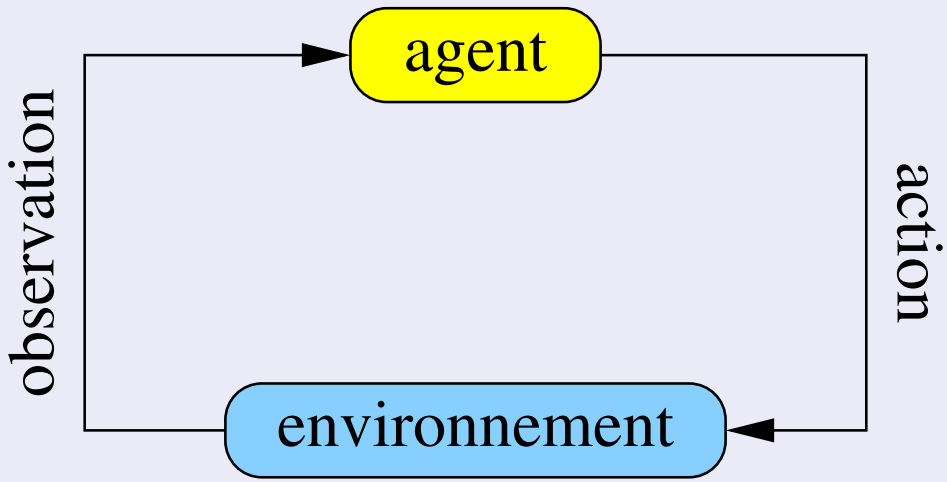
Université de Lorraine / INRIA – LORIA

GDR robotique (GT 8) – 6 septembre 2012

Prise de décision séquentielle dans l'incertain

Problème

Contrôler un système dont l'évolution est incertaine.



Des variantes :

	dynamique	état
MDP	connue	connu
POMDP	connue	caché
RL	cachée	connu
...		

Quand y a-t-il recherche d'information ?

2 cas types :

POMDP : des actions désambiguisent l'état
("*active sensing*")

RL : des actions améliorent le modèle
(exploration vs. exploitation)

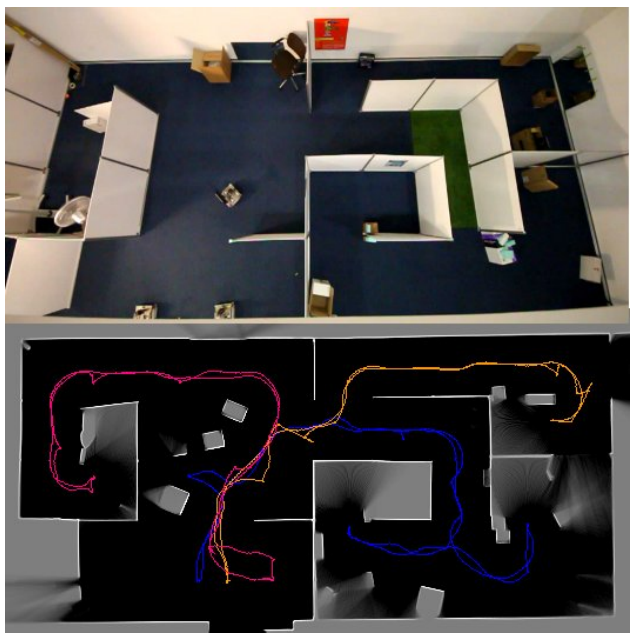
Dans ces 2 cas,

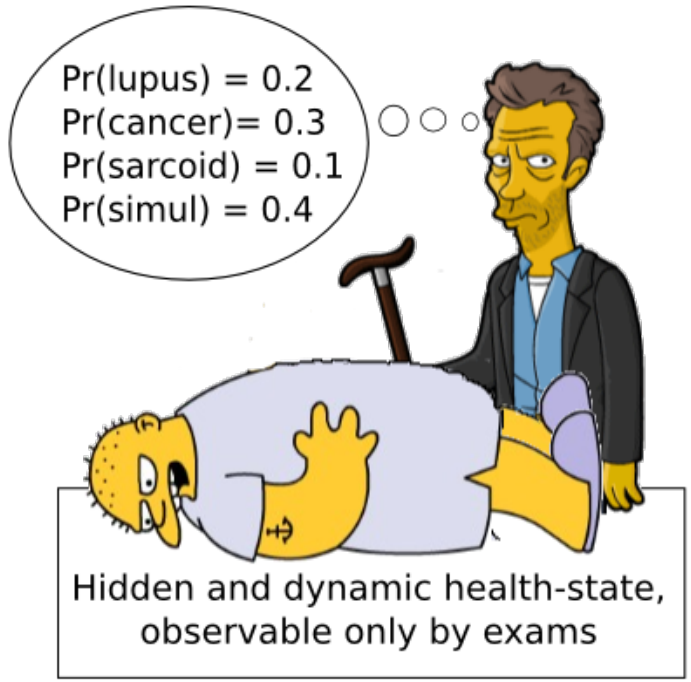
$$\text{acquérir de l'information} = \text{moyen},$$
$$\neq \text{fin}.$$

Ici : but = acquérir de l'information

Ex. :

- diagnostic médical / de panne,
- exploration
  (carte géographique, géologique, ...),
- surveillance d'un lieu
  (localiser entités mobiles ou intrus),
- localisation,
- ...



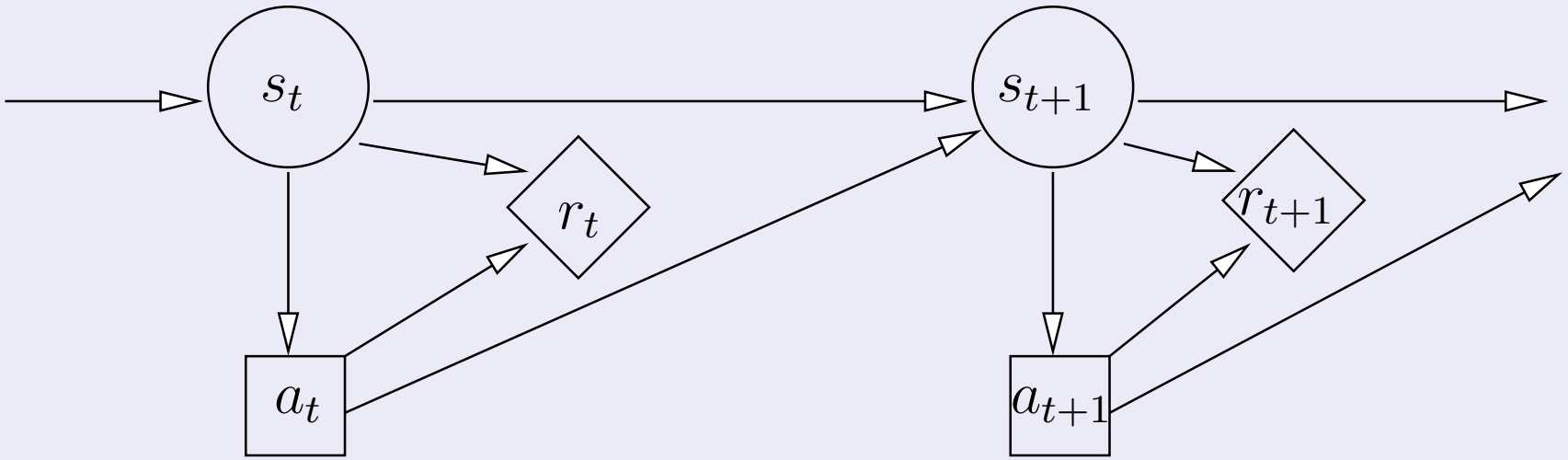


Plan

Plan

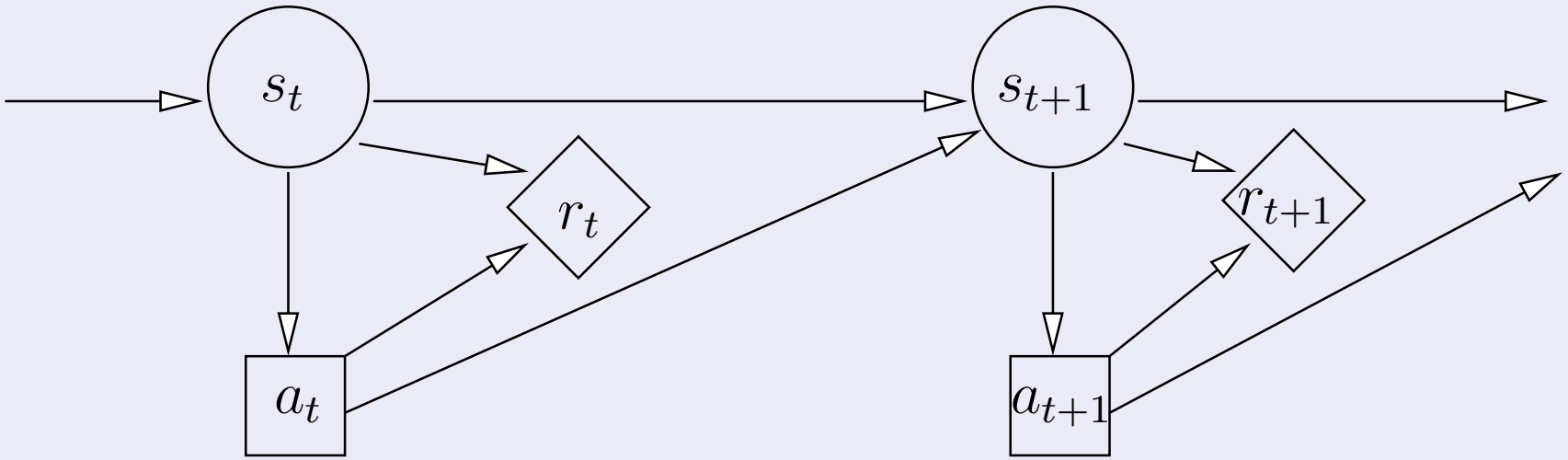
Théorie de la décision

Processus décisionnels de Markov (MDP)



Théorie de la décision

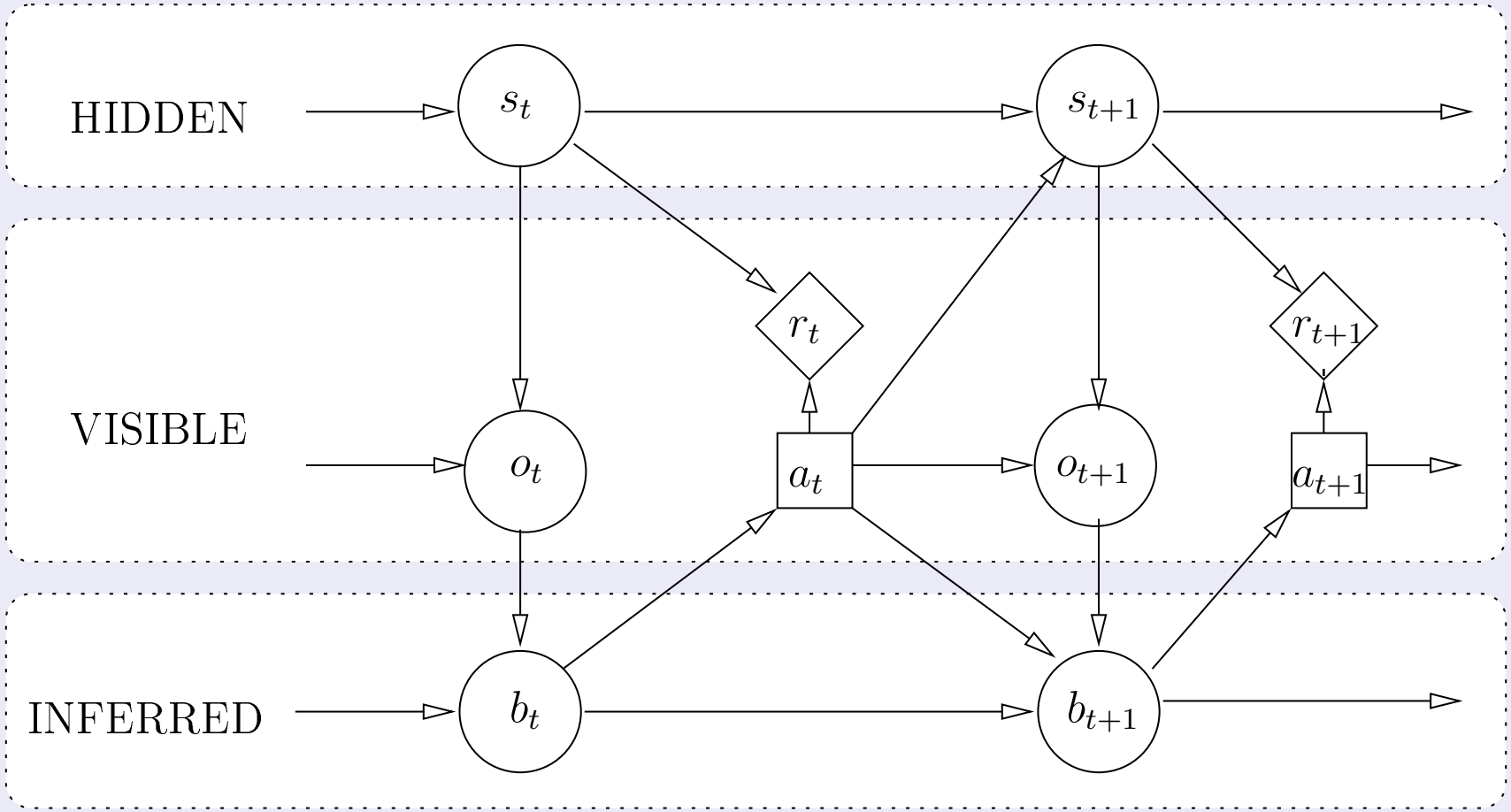
Processus décisionnels de Markov (MDP)



Objectif : trouver une politique optimale.

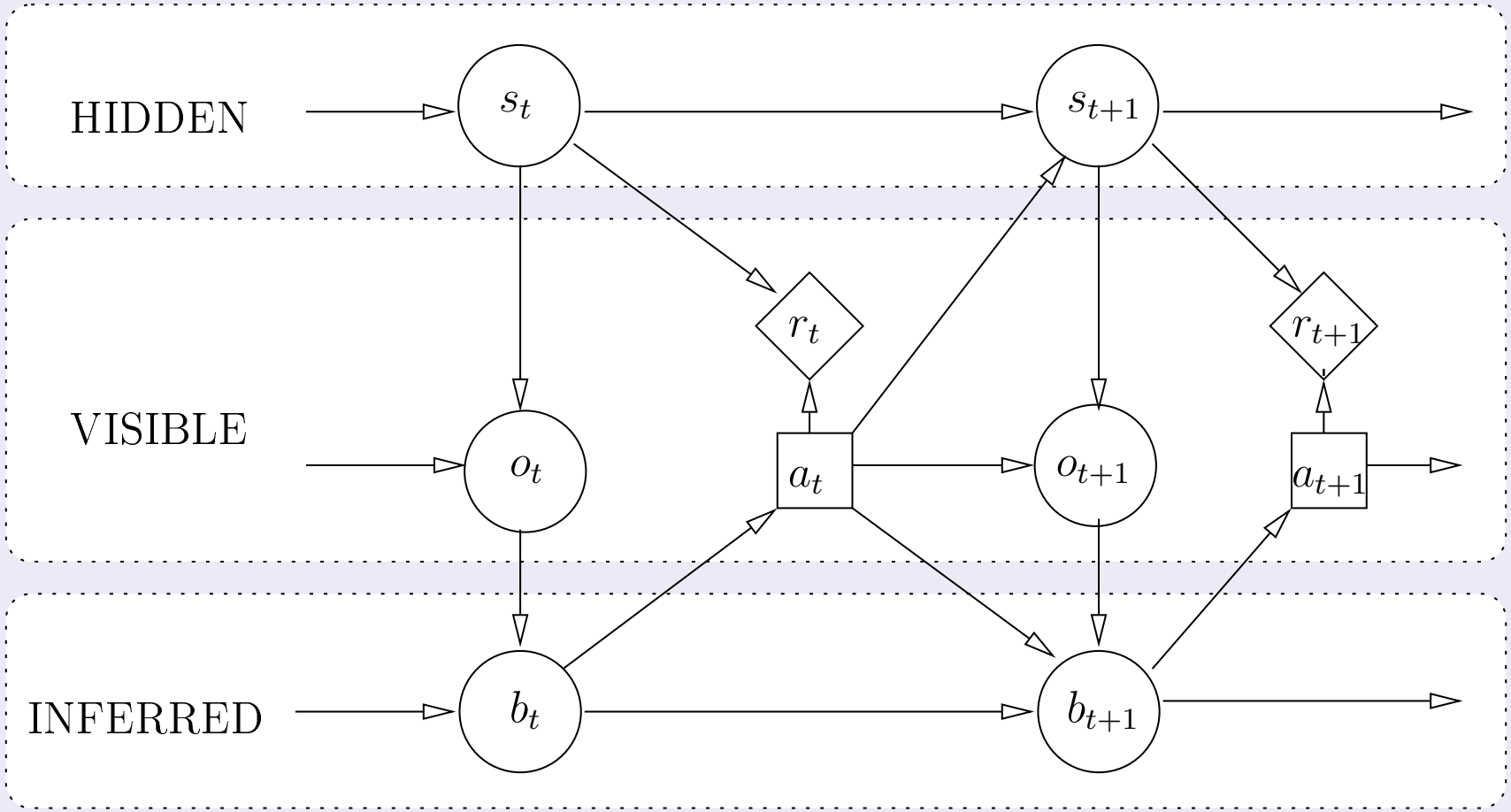
Théorie de la décision (cont)

MDP avec observation partielle (POMDP)



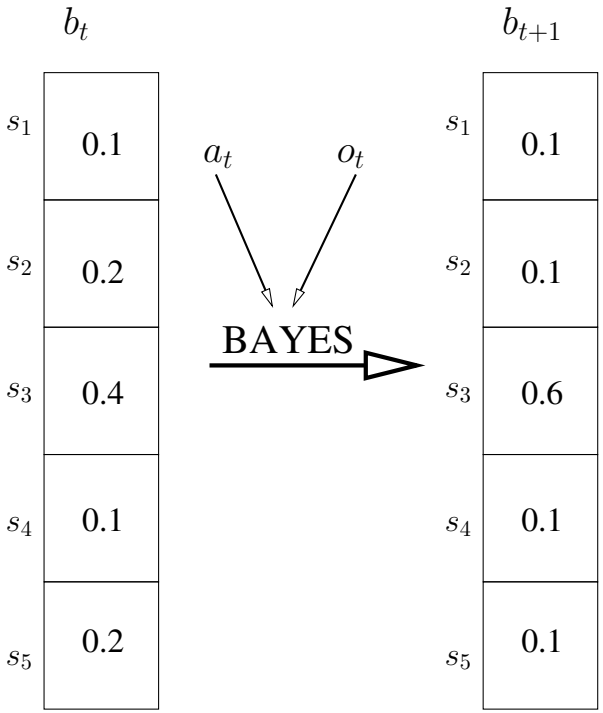
Théorie de la décision (cont)

MDP avec observation partielle (POMDP)



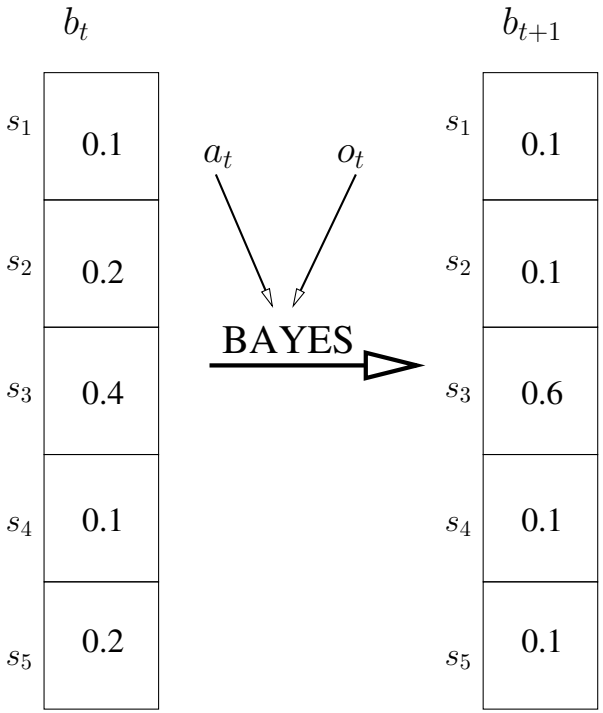
Même objectif, mais en utilisant **l'état de croyance**

L'état de croyance

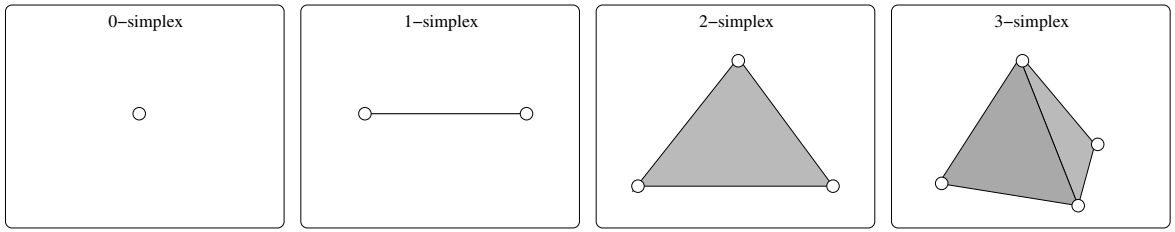


- Est une distribution de probabilité sur les états

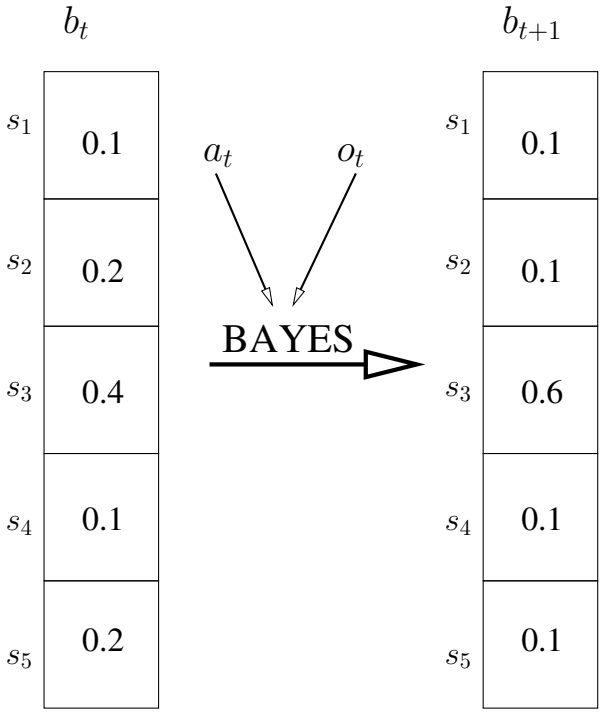
L'état de croyance



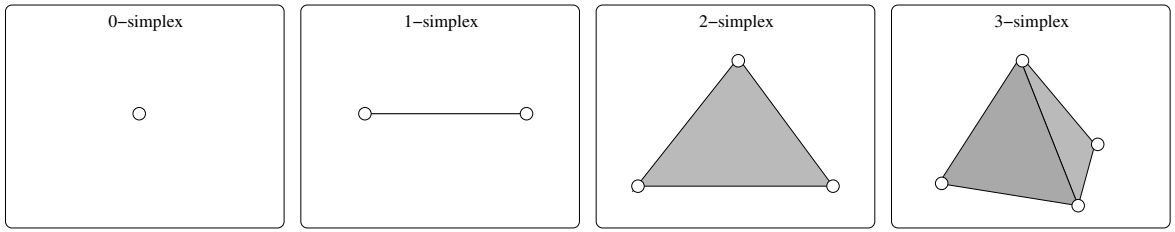
- Est une distribution de probabilité sur les états
- $(b \in)\Delta$ forme un **simplexe** $(\sum_s b(s) = 1)$



L'état de croyance

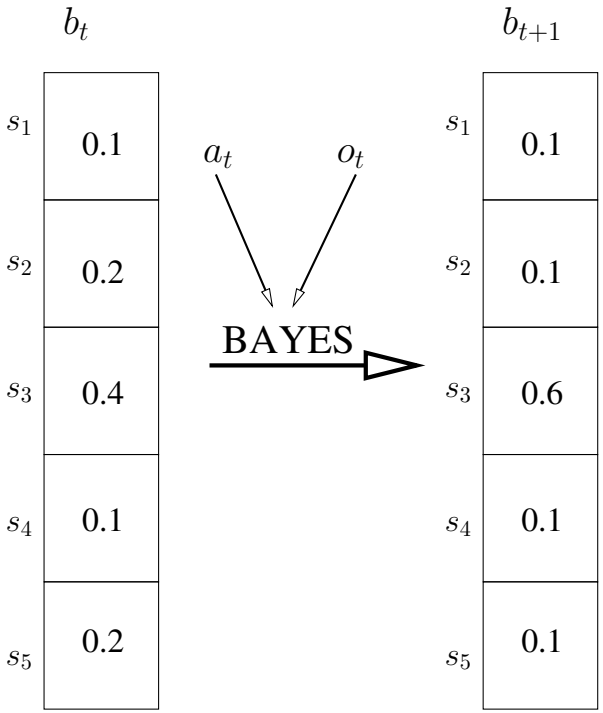


- Est une distribution de probabilité sur les états
- $(b \in)\Delta$ forme un **simplexe** $(\sum_s b(s) = 1)$

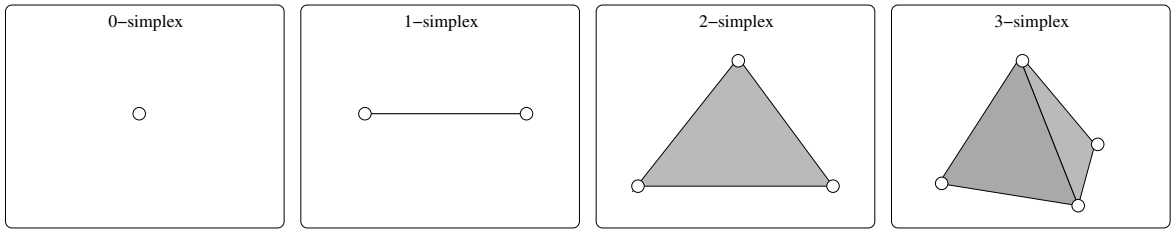


- Une `prior` b_0 est toujours nécessaire

L'état de croyance

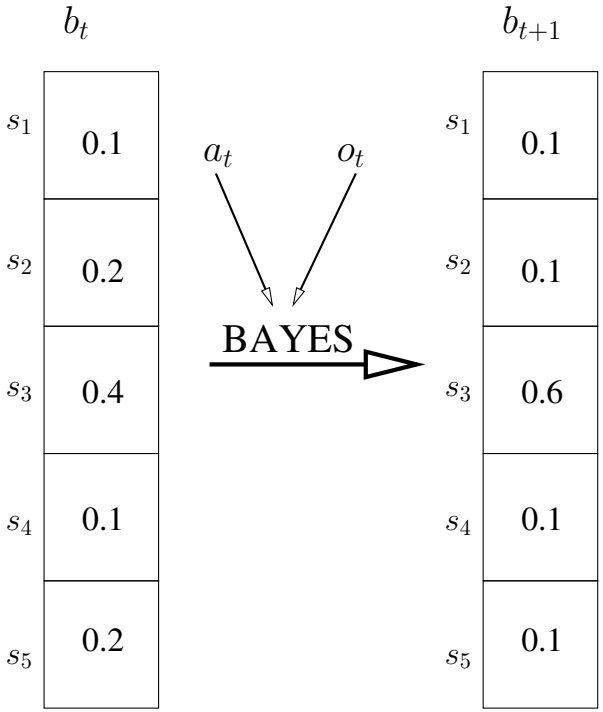


- Est une distribution de probabilité sur les états
- $(b \in)\Delta$ forme un **simplexe** $(\sum_s b(s) = 1)$

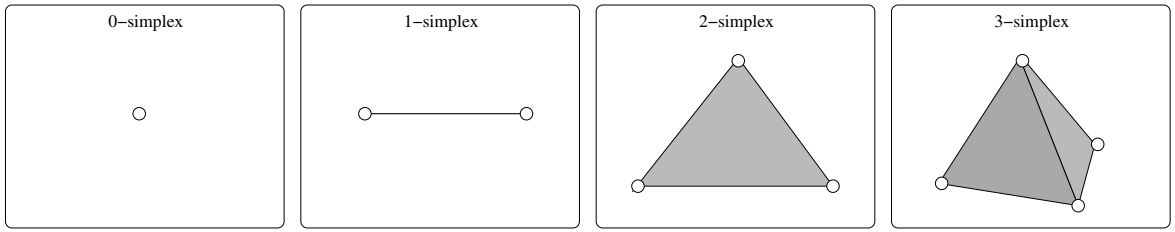


- Une `prior` b_0 est toujours nécessaire
- Est mis à jour en utilisant la règle de Bayes

L'état de croyance



- Est une distribution de probabilité sur les états
- $(b \in)\Delta$ forme un **simplexe** $(\sum_s b(s) = 1)$



- Une `prior` b_0 est toujours nécessaire
- Est mis à jour en utilisant la règle de Bayes

Formellement

$$b^{a,o}(s') \propto Pr(o|a,s') \sum_{s \in \mathcal{S}} Pr(s'|a,s) b(s)$$

MDP sur l'état de croyance

Pour résoudre un POMDP, on le transforme en un **belief MDP**, c'est à dire un MDP défini sur l'état de croyance.

MDP sur l'état de croyance

Pour résoudre un POMDP, on le transforme en un **belief MDP**, c'est à dire un MDP défini sur l'état de croyance.

Formellement

$$V_n(b_0) = \max_{\pi} E\left[\sum_{t=0}^{n} \gamma^t \rho(b_t, \pi(b_t)) \mid b_0\right] \qquad \text{Gain futur}$$

$$V_n(b) = \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \int_{b' \in \Delta} Pr(b'|a,b) V_{n-1}(b') db'\right] \qquad \text{(Bellman, 1954)}$$

$$= \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \sum_{o} Pr(o|a,b) V_{n-1}(b^{a,o})\right] \qquad \text{Avec observations}$$

MDP sur l'état de croyance

Pour résoudre un POMDP, on le transforme en un **belief MDP**, c'est à dire un MDP défini sur l'état de croyance.

Formellement

$$V_n(b_0) = \max_{\pi} E\left[\sum_{t=0}^{n} \gamma^t \rho(b_t, \pi(b_t)) \mid b_0\right] \qquad \text{Gain futur}$$

$$V_n(b) = \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \int_{b' \in \Delta} Pr(b'|a,b) V_{n-1}(b') db'\right] \qquad \text{(Bellman, 1954)}$$

$$= \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \sum_{o} Pr(o|a,b) V_{n-1}(b^{a,o})\right] \qquad \text{Avec observations}$$

La fonction de **récompense** est définie comme l'espérance de $r(s,a)$:

$$\rho(b,a) = E_{s \sim b}[r(s,a)] = \sum_{s} b(s) r(s,a)$$

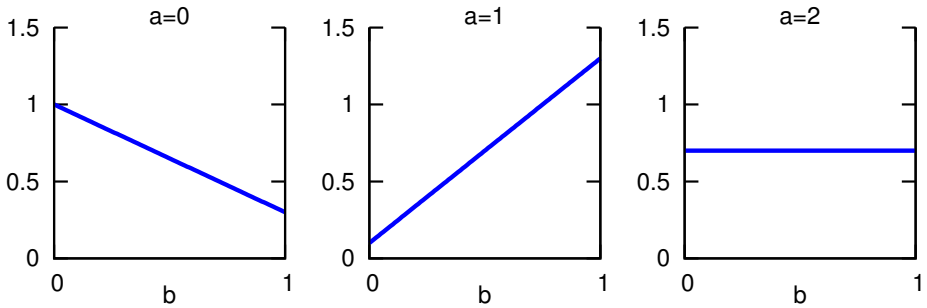
Résolution des POMDP

- La fonction de valeur est convexe et linéaire par morceaux (PWLC)

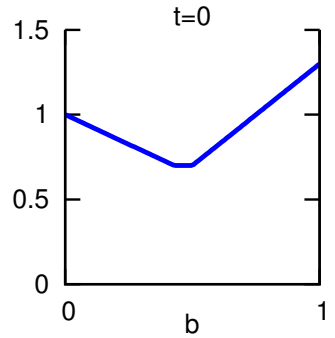
Résolution des POMDP

- La fonction de valeur est convexe et linéaire par morceaux (PWLC)

Fonction de récompense



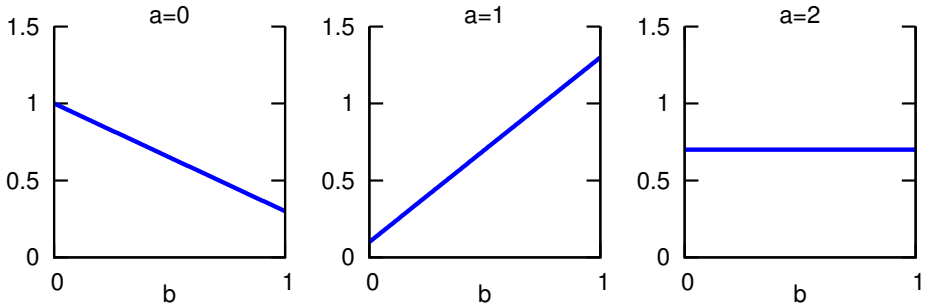
Fonction de valeur



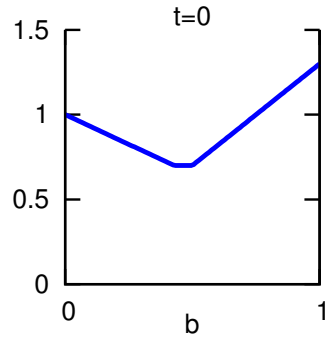
Résolution des POMDP

- La fonction de valeur est convexe et linéaire par morceaux (PWLC)

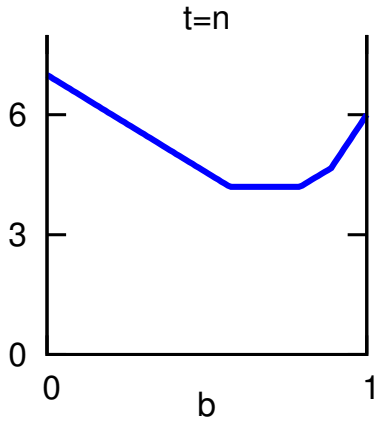
Fonction de récompense

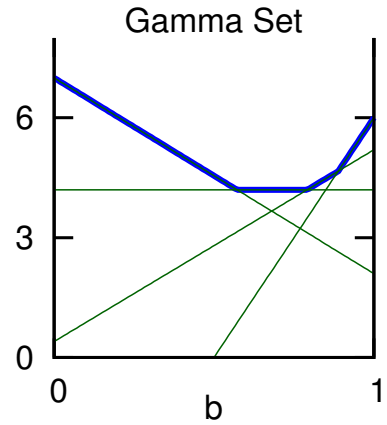


Fonction de valeur



- Prouvé par Smallwood and Sondik (1973)

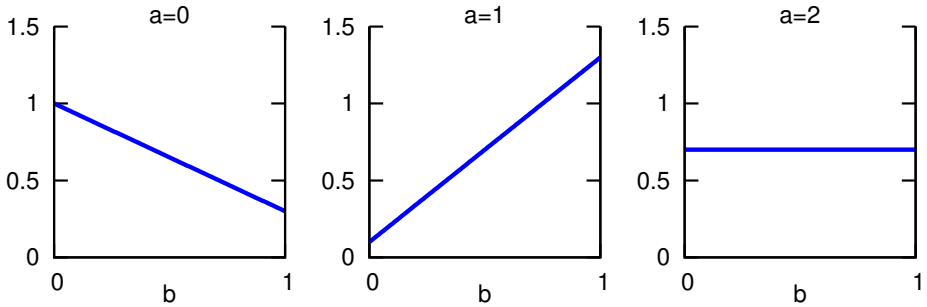




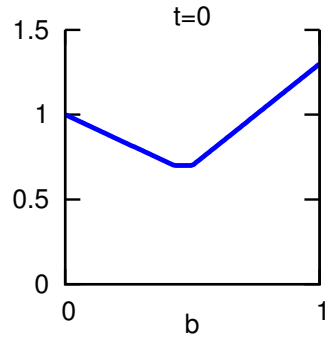
Résolution des POMDP

- La fonction de valeur est convexe et linéaire par morceaux (PWLC)

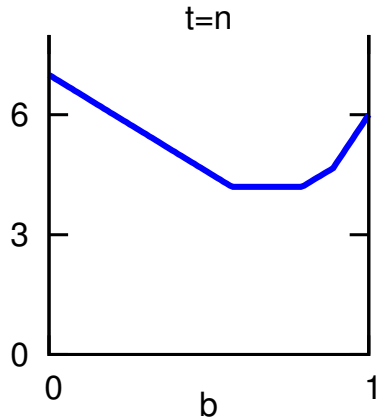
Fonction de récompense

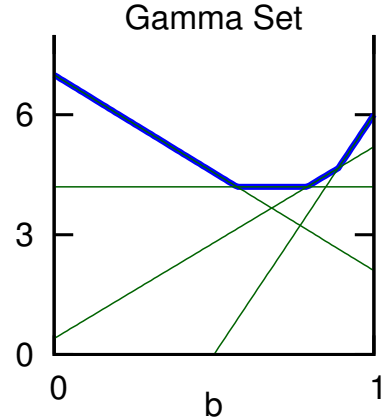


Fonction de valeur



- Prouvé par Smallwood and Sondik (1973)





- Formellement,

$$V_n(b) = \max_{\alpha \in \Gamma_n} \sum_s b(s)\alpha(s),$$

où Γ_n est un ensemble d'hyperplans.

Algorithmes de résolution

On peut résoudre un POMDP

Algorithmes de résolution

On peut résoudre un POMDP

- avec des mises à jour exactes :
  - Batch Enumeration (Monahan, 1982)
  - Incremental Pruning (Cassandra, 1998)

Algorithmes de résolution

On peut résoudre un POMDP

- avec des mises à jour exactes :
  - Batch Enumeration (Monahan, 1982)
  - Incremental Pruning (Cassandra, 1998)
- avec des mises à jour approchées :
  - Point-Based Value Iteration (Pineau et al., 2006)
  - GapMin (Poupart et al., 2011)

Des récompenses dépendant de l'état de croyance

- L'objectif est de trouver le meilleur diagnostic

Des récompenses dépendant de l'état de croyance

- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**

Des récompenses dépendant de l'état de croyance

- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**
- En général, l'objectif est de surveiller des variables

Des récompenses dépendant de l'état de croyance

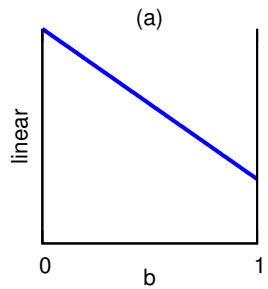
- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**
- En général, l'objectif est de surveiller des variables
  - la surveillance militaire (Hero et al., 2007)
  - la surveillance des enfants (Spaan, 2008)
  - l'exploration d'une zone donnée (Thrun, 2000)

Des récompenses dépendant de l'état de croyance

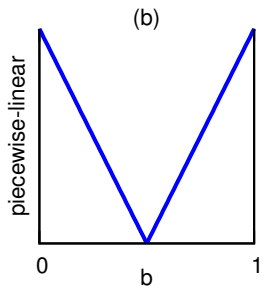
- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**
- En général, l'objectif est de surveiller des variables
  - la surveillance militaire (Hero et al., 2007)
  - la surveillance des enfants (Spaan, 2008)
  - l'exploration d'une zone donnée (Thrun, 2000)
- $\rho(b)$ mesure l'information actuelle dans l'état de croyance

Des récompenses dépendant de l'état de croyance

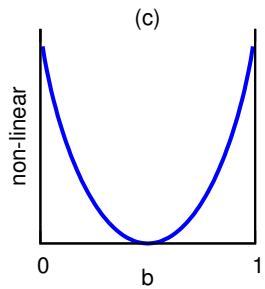
- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**
- En général, l'objectif est de surveiller des variables
  - la surveillance militaire (Hero et al., 2007)
  - la surveillance des enfants (Spaan, 2008)
  - l'exploration d'une zone donnée (Thrun, 2000)
- $\rho(b)$ mesure l'information actuelle dans l'état de croyance
- $\rho(b)$ peut être



$\sum_s r_s b(s)$



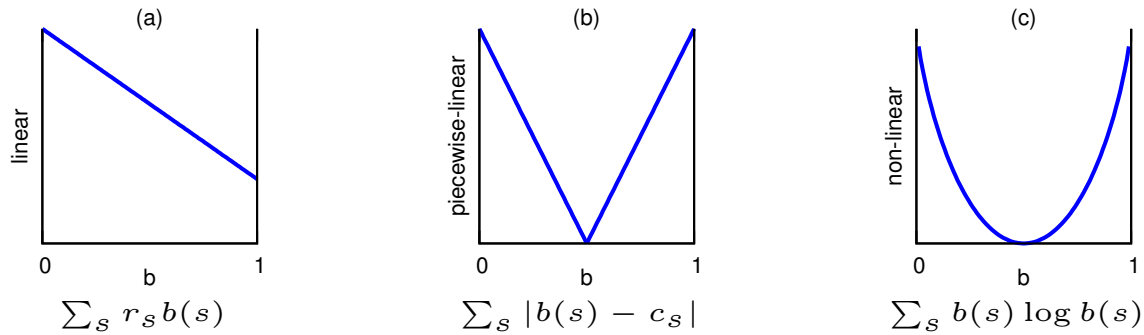
$\sum_s |b(s) - c_s|$



$\sum_s b(s) \log b(s)$

Des récompenses dépendant de l'état de croyance

- L'objectif est de trouver le meilleur diagnostic
  - On n'attribue pas des récompenses aux maladies
  - Mais on attribue des récompenses à la **connaissance**
- En général, l'objectif est de surveiller des variables
  - la surveillance militaire (Hero et al., 2007)
  - la surveillance des enfants (Spaan, 2008)
  - l'exploration d'une zone donnée (Thrun, 2000)
- $\rho(b)$ mesure l'information actuelle dans l'état de croyance
- $\rho(b)$ peut être



- Ces mesures sont généralement convexes

Présentation des ρPOMDP

- Les POMDP ne considèrent pas ce type de récompenses, d'où l'introduction des ρPOMDP :

POMDP	**ρPOMDP**
$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, r, b_0 \rangle$	$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, func(b), b_0 \rangle$
$\downarrow$	$\downarrow$
belief MDP	**belief MDP**
$\langle \Delta, \mathcal{A}, \tau, \rho = \sum_s b(s) r(s,a), b_0 \rangle$	$\langle \Delta, \mathcal{A}, \tau, \rho = func(b), b_0 \rangle$

Présentation des ρPOMDP

- Les POMDP ne considèrent pas ce type de récompenses, d'où l'introduction des ρPOMDP :

POMDP	**ρPOMDP**
$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, r, b_0 \rangle$	$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, func(b), b_0 \rangle$
$\downarrow$	$\downarrow$
belief MDP	**belief MDP**
$\langle \Delta, \mathcal{A}, \tau, \rho = \sum_s b(s) r(s,a), b_0 \rangle$	$\langle \Delta, \mathcal{A}, \tau, \rho = func(b), b_0 \rangle$

- **Problème** : la fonction de valeur n'est plus PWLC.

Présentation des ρPOMDP

- Les POMDP ne considèrent pas ce type de récompenses, d'où l'introduction des ρPOMDP :

POMDP	ρ**POMDP**
$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, r, b_0 \rangle$	$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, func(b), b_0 \rangle$
$\downarrow$	$\downarrow$
belief MDP	**belief MDP**
$\langle \Delta, \mathcal{A}, \tau, \rho = \sum_s b(s) r(s,a), b_0 \rangle$	$\langle \Delta, \mathcal{A}, \tau, \rho = func(b), b_0 \rangle$

- **Problème** : la fonction de valeur n'est plus PWLC.

Contribution :

des outils théoriques permettant de réutiliser les algorithmes existants

Convexité et ρPOMDP

De nombreux algorithmes utilisent la convexité de V.
V reste-t-elle convexe dans les ρPOMDP ?

Convexité et ρPOMDP

De nombreux algorithmes utilisent la convexité de V.
V reste-t-elle convexe dans les ρPOMDP ?

Théorème (Convexité)

Si ρ et V_0 sont des fonctions convexes sur Δ, alors la fonction de valeur V_n du belief MDP est convexe sur Δ à tout pas de temps n.

Convexité et ρPOMDP

De nombreux algorithmes utilisent la convexité de V.
V reste-t-elle convexe dans les ρPOMDP ?

Théorème (Convexité)

Si ρ et V_0 sont des fonctions convexes sur Δ, alors la fonction de valeur V_n du belief MDP est convexe sur Δ à tout pas de temps n.

Question : comment mettre en œuvre ces algorithmes ?

Fonctions de récompense PWLC

Cas simple :

- Fonction de récompense du type

$$\rho(b,a) = \max_{\alpha \in \Gamma_\rho^a} \left[\sum_s b(s)\alpha(s) \right].$$

Fonctions de récompense PWLC

Cas simple :

- Fonction de récompense du type

$$\rho(b,a) = \max_{\alpha \in \Gamma_\rho^a} \left[\sum_s b(s)\alpha(s) \right].$$

- On peut propager les fonctions PWLC comme on le fait avec les linéaires

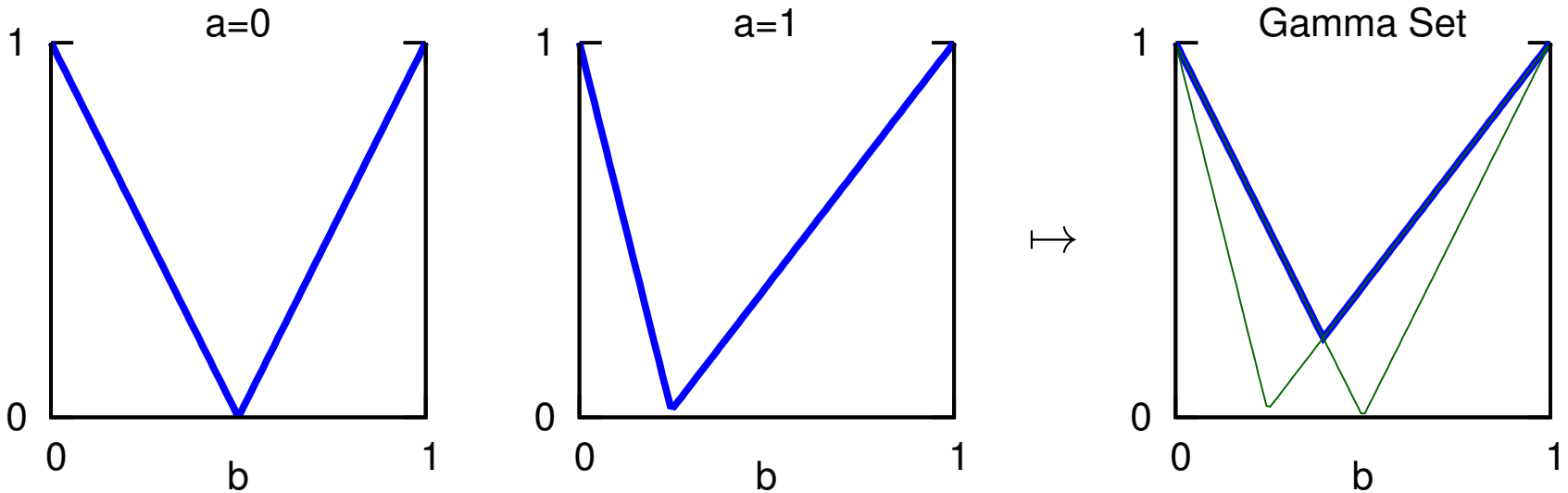
Fonctions de récompense PWLC

Cas simple :

- Fonction de récompense du type

$$\rho(b,a) = \max_{\alpha \in \Gamma_\rho^a} \left[\sum_s b(s)\alpha(s) \right].$$

- On peut propager les fonctions PWLC comme on le fait avec les linéaires

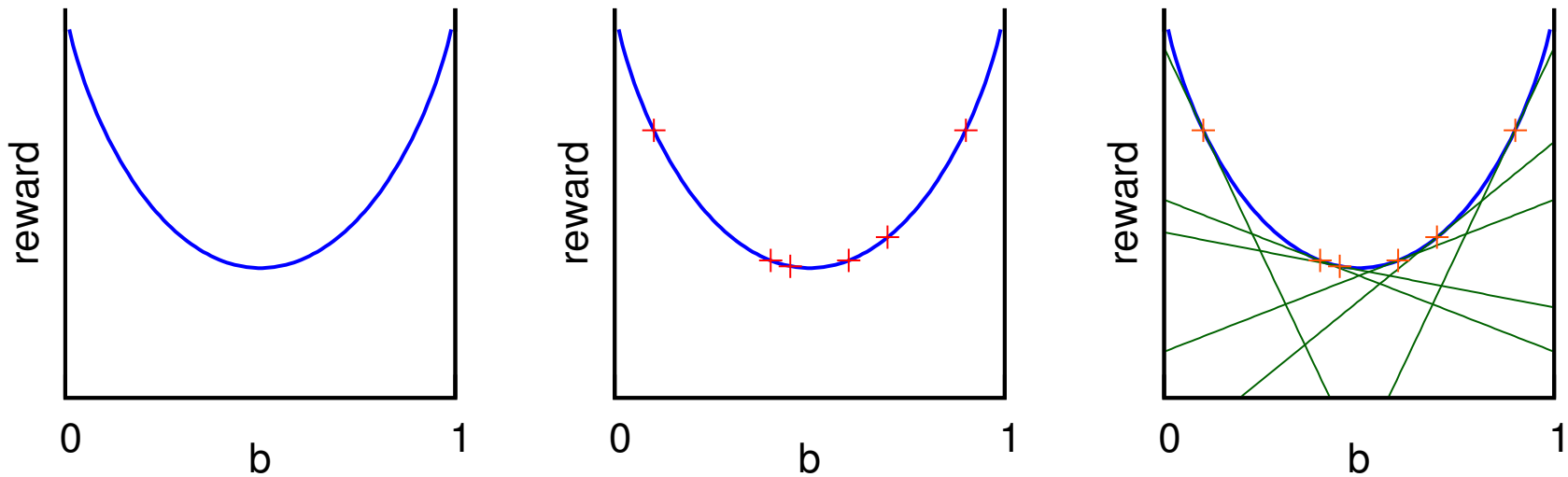


Fonctions de récompense non linéaires

- **Idée :** Utiliser une approximation PWLC de $\rho(b)$ employant un ensemble de points de pivot $b' \in B \subset \Delta$

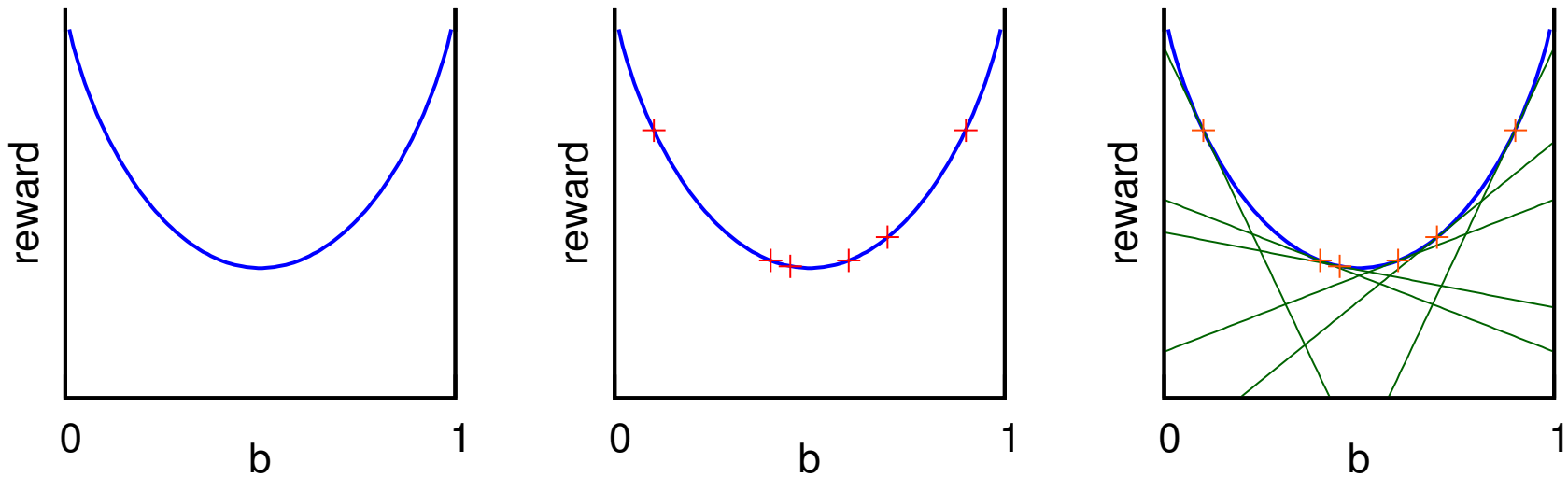
Fonctions de récompense non linéaires

- **Idée :** Utiliser une approximation PWLC de $\rho(b)$ employant un ensemble de points de pivot $b' \in B \subset \Delta$



Fonctions de récompense non linéaires

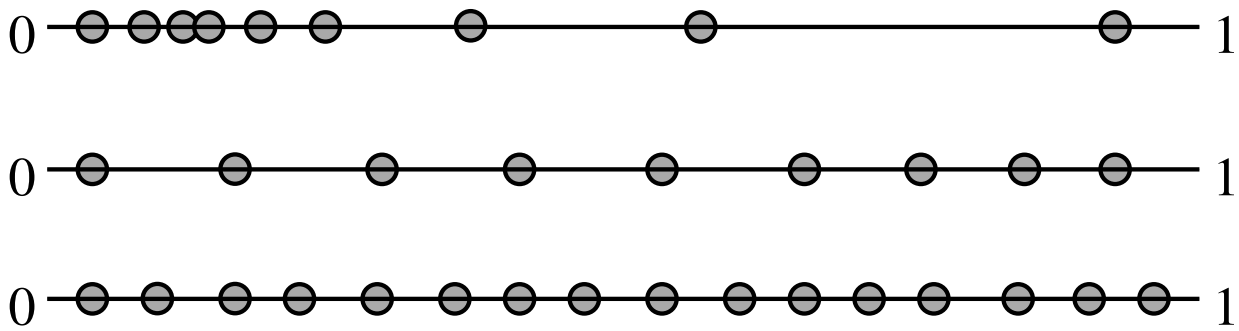
- **Idée :** Utiliser une approximation PWLC de $\rho(b)$ employant un ensemble de points de pivot $b' \in B \subset \Delta$



- **Question :** Cette approximation a-t-elle une erreur bornée ?

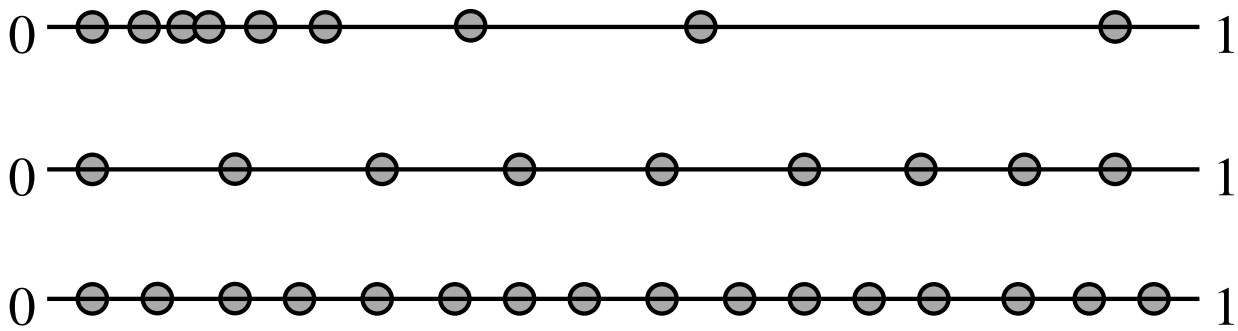
Approximation de $\rho(b)$: l'ensemble des points

- L'erreur dépend de la densité de B



Approximation de $\rho(b)$: l'ensemble des points

- L'erreur dépend de la densité de B



- **Question :** Toute fonction (convexe) peut-elle être approchée avec une erreur bornée ?

Approximation de $\rho(b)$: la fonction ρ

- Si $\nabla\rho(b)$ est borné, l'erreur d'approximation est bornée.
- Ex. : Si $\rho(b)$ est **lipschtzienne**, alors $\nabla\rho(b)$ est borné.

$$|f(x) - f(y)| \leq K\|x - y\|_1$$

Approximation de $\rho(b)$: la fonction ρ

- Si $\nabla\rho(b)$ est borné, l'erreur d'approximation est bornée.
- Ex. : Si $\rho(b)$ est **lipschtzienne**, alors $\nabla\rho(b)$ est borné.

$$|f(x) - f(y)| \leq K\|x - y\|_1$$

- Pb. : l'entropie est α-hölderienne, pas lipschtzienne !

$$\exists\alpha \in (0, 1],\ \exists K_\alpha > 0,\ \text{s.t. } |f(x) - f(y)| \leq K_\alpha\|x - y\|_1^\alpha.$$

Approximation de $\rho(b)$: la fonction ρ

- Si $\nabla\rho(b)$ est borné, l'erreur d'approximation est bornée.
- Ex. : Si $\rho(b)$ est **lipschtzienne**, alors $\nabla\rho(b)$ est borné.

$$|f(x) - f(y)| \leq K\|x - y\|_1$$

- Pb. : l'entropie est α-hölderienne, pas lipschtzienne !

$$\exists\alpha \in (0, 1],\ \exists K_\alpha > 0,\ \text{s.t. } |f(x) - f(y)| \leq K_\alpha\|x - y\|_1^\alpha.$$

- Un résultat plus générique peut être prouvé pour les fonctions α-hölderiennes.

Approximation des fonctions α-Hölderiennes

Approximation des fonctions α-Hölderiennes

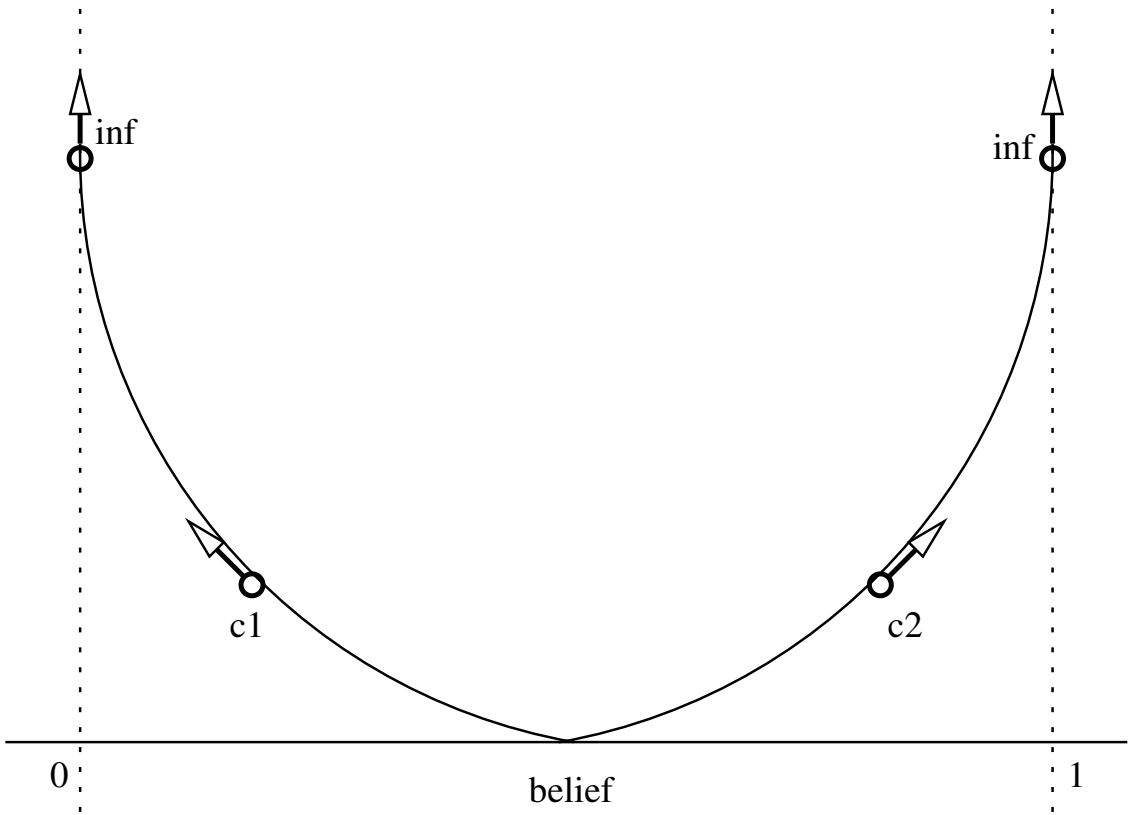
Theoréme (Borne de ρ)

Soit ρ une fonction continue et convexe sur Δ, dérivable partout sur Δ^o (l'intérieur de Δ), et satisfaisant la condition α-Hölder avec la constante K_α. L'erreur d'une approximation ω_B est majorée par $C{\delta_B}^\alpha$, où C est une constante scalaire.

Approximation des fonctions α-Hölderiennes

Theoréme (Borne de ρ)

Soit ρ une fonction continue et convexe sur Δ, dérivable partout sur Δ^o (l'intérieur de Δ), et satisfaisant la condition α-Hölder avec la constante K_α. L'erreur d'une approximation ω_B est majorée par $C{\delta_B}^\alpha$, où C est une constante scalaire.

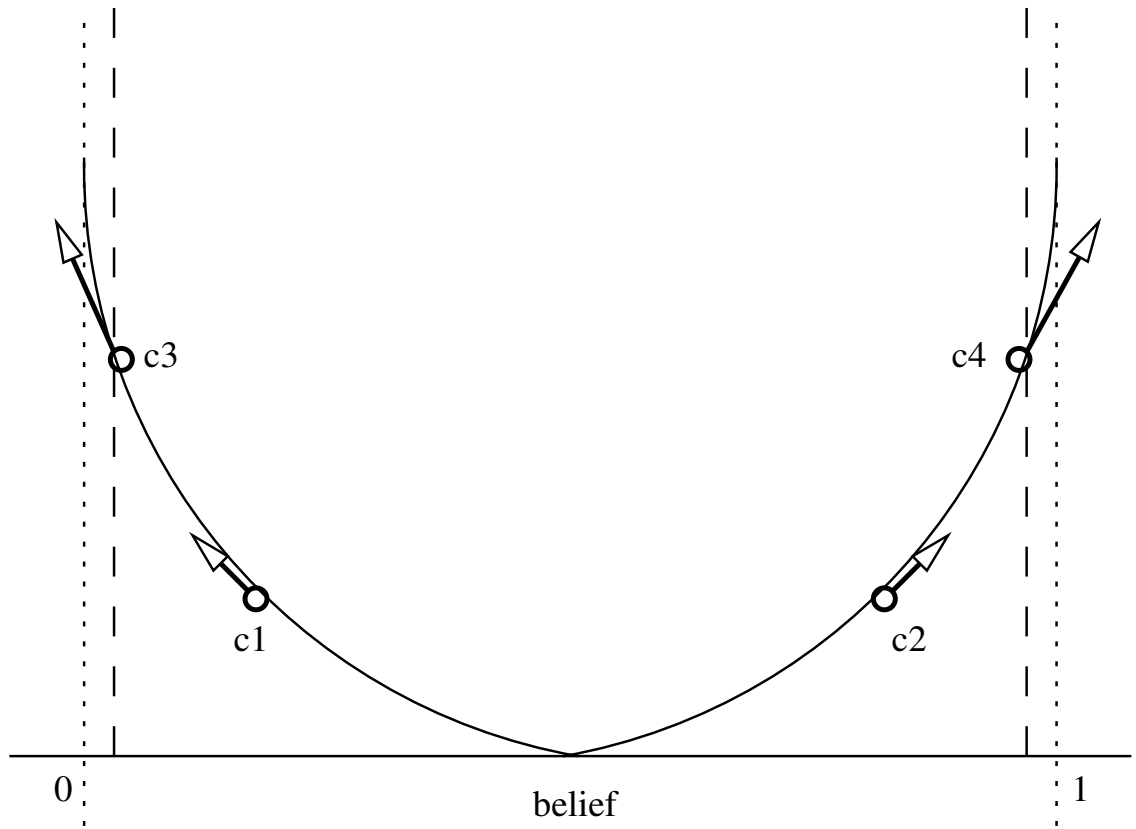


Une fonction α-Hölderienne est lipschitzienne dans un segment loin de la frontière.

Approximation des fonctions α-Hölderiennes

Theoréme (Borne de ρ)

Soit ρ une fonction continue et convexe sur Δ, dérivable partout sur Δ^o (l'intérieur de Δ), et satisfaisant la condition α-Hölder avec la constante K_α. L'erreur d'une approximation ω_B est majorée par $C{\delta_B}^\alpha$, où C est une constante scalaire.



Une fonction α-Hölderienne est lipschitzienne dans un segment loin de la frontière.

Erreur d'approximation dans la résolution

- **Question** : Y a-t-il une borne d'erreur pour les algorithmes proposés ?

Erreur d'approximation dans la résolution

- **Question** : Y a-t-il une borne d'erreur pour les algorithmes proposés ? Comment se propage l'erreur d'approximation de la fonction de récompense ?
  - L'erreur de propagation pour les **algorithmes exacts** est

$$\|V_t - V_t^*\|_\infty \leq \frac{C\delta_B^\alpha}{1-\gamma}$$

  - Pour les **algorithmes à base de points**, l'erreur d'approximation commune est

$$\|\hat{V}_t - V_t^*\|_\infty \leq \frac{(R_{max} - R_{min} + C\delta_B^\alpha)\delta_B}{1-\gamma} + \frac{C\delta_B^\alpha}{1-\gamma}$$

Erreur d'approximation dans la résolution

- **Question** : Y a-t-il une borne d'erreur pour les algorithmes proposés ? Comment se propage l'erreur d'approximation de la fonction de récompense ?
  - L'erreur de propagation pour les **algorithmes exacts** est

$$\|V_t - V_t^*\|_\infty \leq \frac{C\delta_B^\alpha}{1-\gamma}$$

  - Pour les **algorithmes à base de points**, l'erreur d'approximation commune est

$$\|\hat{V}_t - V_t^*\|_\infty \leq \frac{(R_{max} - R_{min} + C\delta_B^\alpha)\delta_B}{1-\gamma} + \frac{C\delta_B^\alpha}{1-\gamma}$$

- Si on diminue la densité, on peut approcher la fonction optimale d'aussi près que l'on veut

Illustration expérimentale

Camera Clean (surveillance)

- N zones disposées circulairement
- 1 cible tourne au hasard (sens horaire)
- 1 caméra cherche cette cible

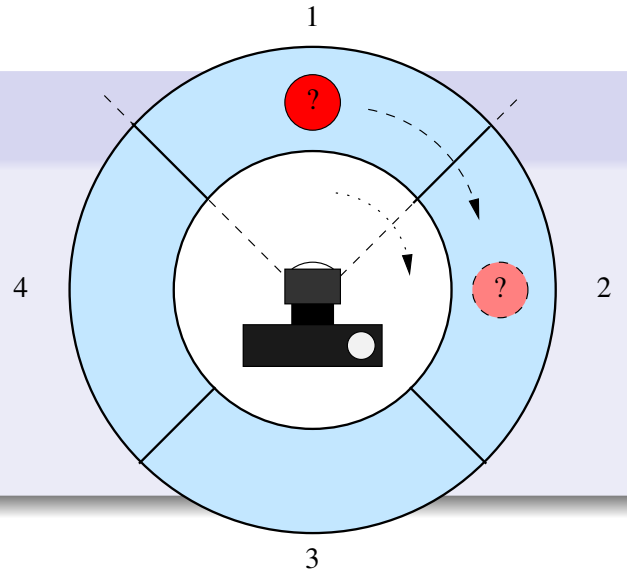
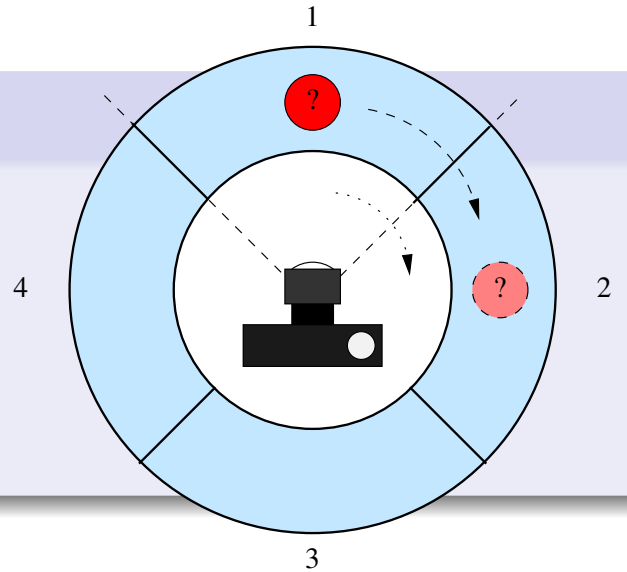


Illustration expérimentale

Camera Clean (surveillance)

- N zones disposées circulairement
- 1 cible tourne au hasard (sens horaire)
- 1 caméra cherche cette cible



- Cible :

$$\begin{aligned} P(move) &= 0,05 \\ P(stop) &= 0,95 \\ \bar{T}_{wait} &= 20 \end{aligned}$$

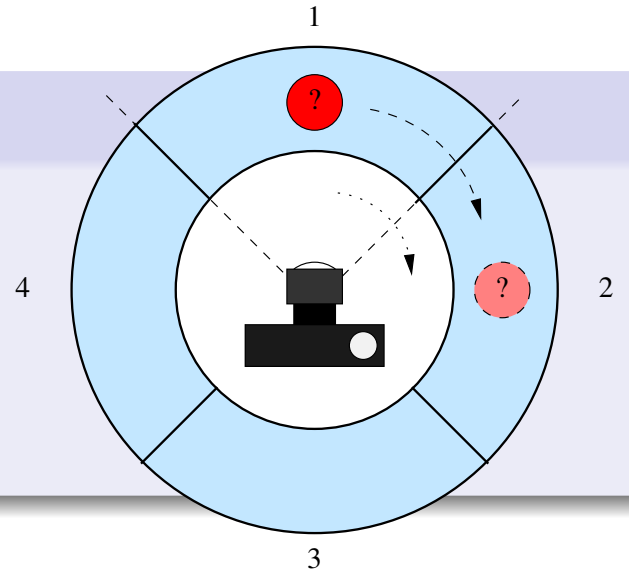
- Actions :
  - $clean$: nettoie l'objectif
  - $move$: pointe sur la zone suivante
  - $shoot$: prend une photo

$$P(good\ classification|clean) = 0,80$$
$$P(good\ classification|dirty) = 0,55$$

Illustration expérimentale

Camera Clean (surveillance)

- N zones disposées circulairement
- 1 cible tourne au hasard (sens horaire)
- 1 caméra cherche cette cible



- Cible :

$$\begin{aligned} P(move) &= 0,05 \\ P(stop) &= 0,95 \\ \bar{T}_{wait} &= 20 \end{aligned}$$

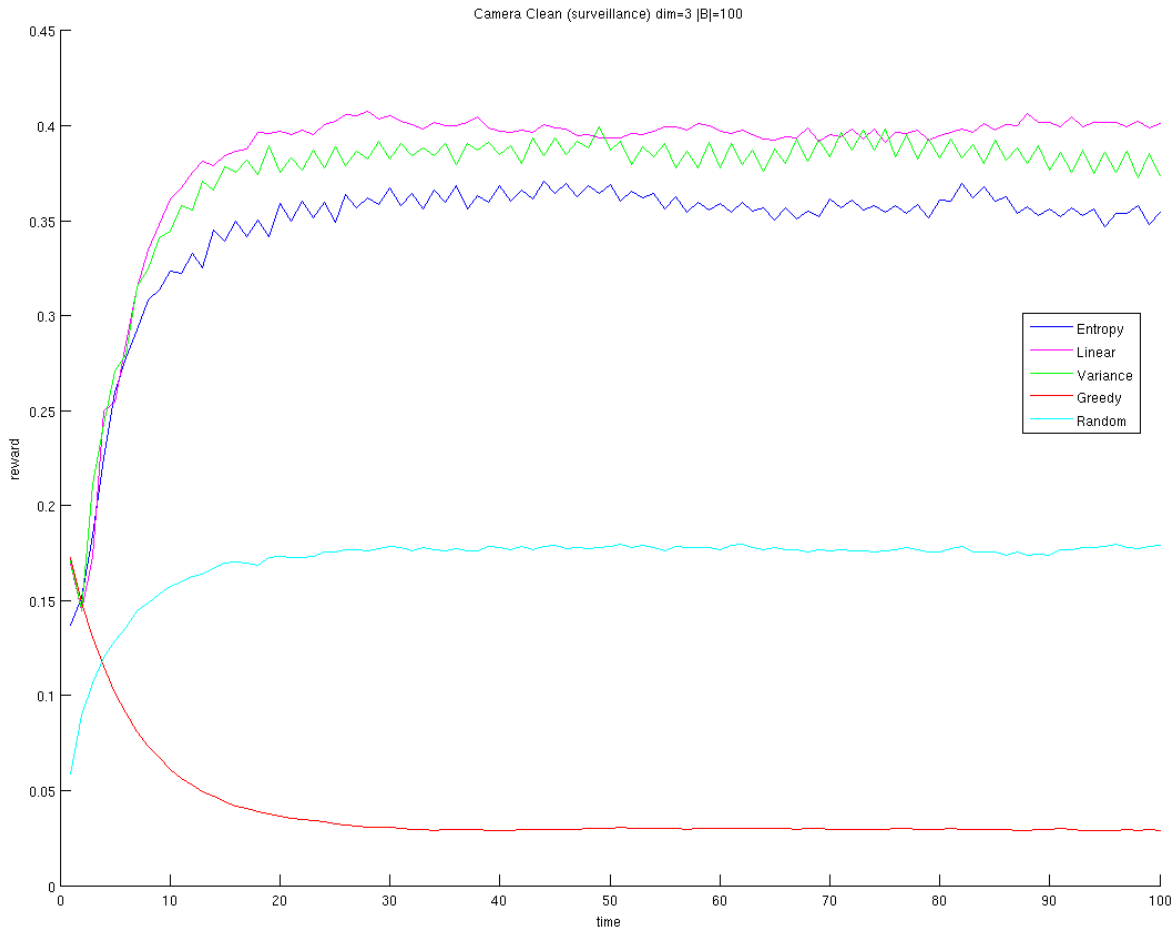
- Actions :
  - $clean$: nettoie l'objectif
  - $move$: pointe sur la zone suivante
  - $shoot$: prend une photo

$$\begin{aligned} P(good\ classification|clean) &= 0,80 \\ P(good\ classification|dirty) &= 0,55 \end{aligned}$$

- 5 algorithmes (**random**, **greedy**, ρ =**linear**, ρ =**quad.**, ρ =**-entropy**)
- $|B| = 100$ (par défaut)
- chaque algorithme exécuté 10 fois
- chaque exécution évaluée sur 500 trajectoires
- chaque trajectoire dure 100 pas de temps
- $\gamma = 0,95$

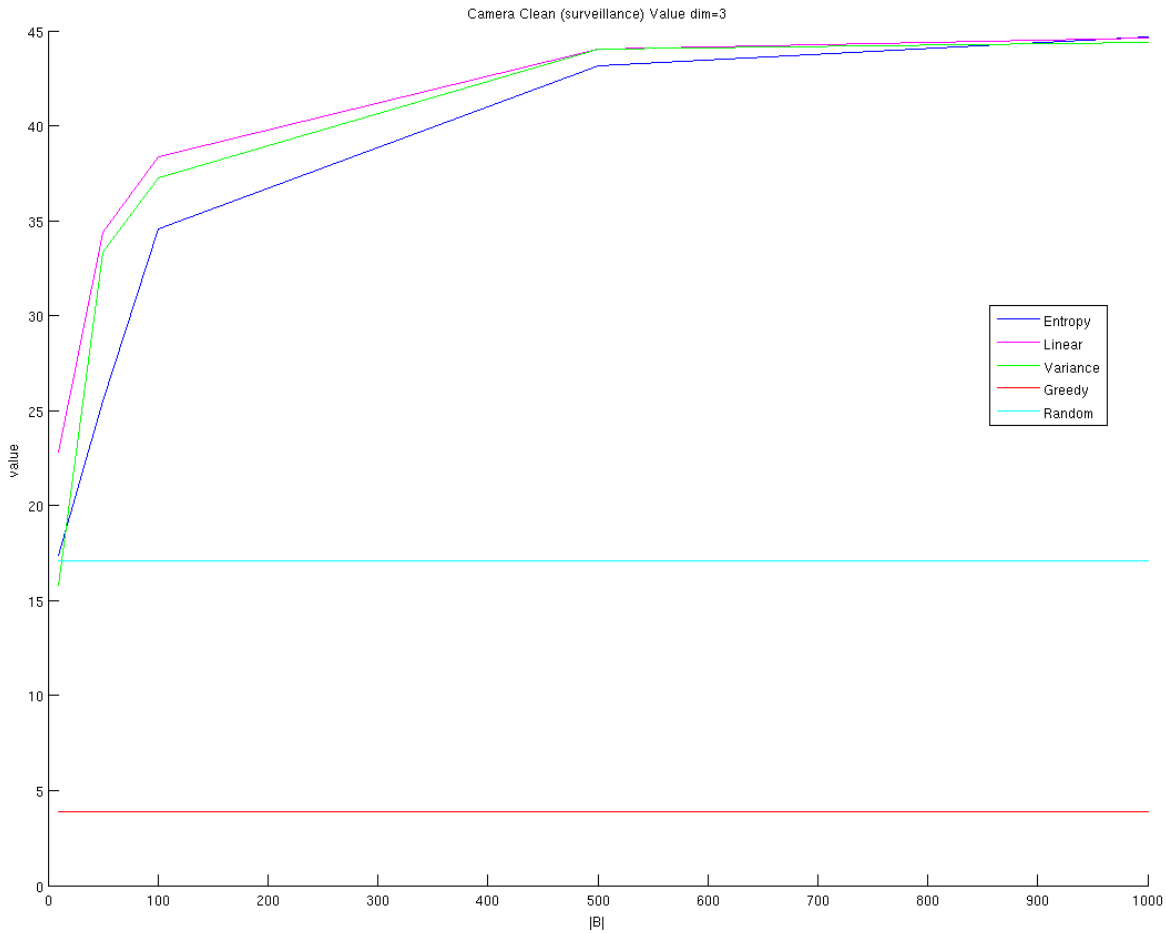
$r = -$Entropie moyenne $= f(t)$

$[N = 3]$



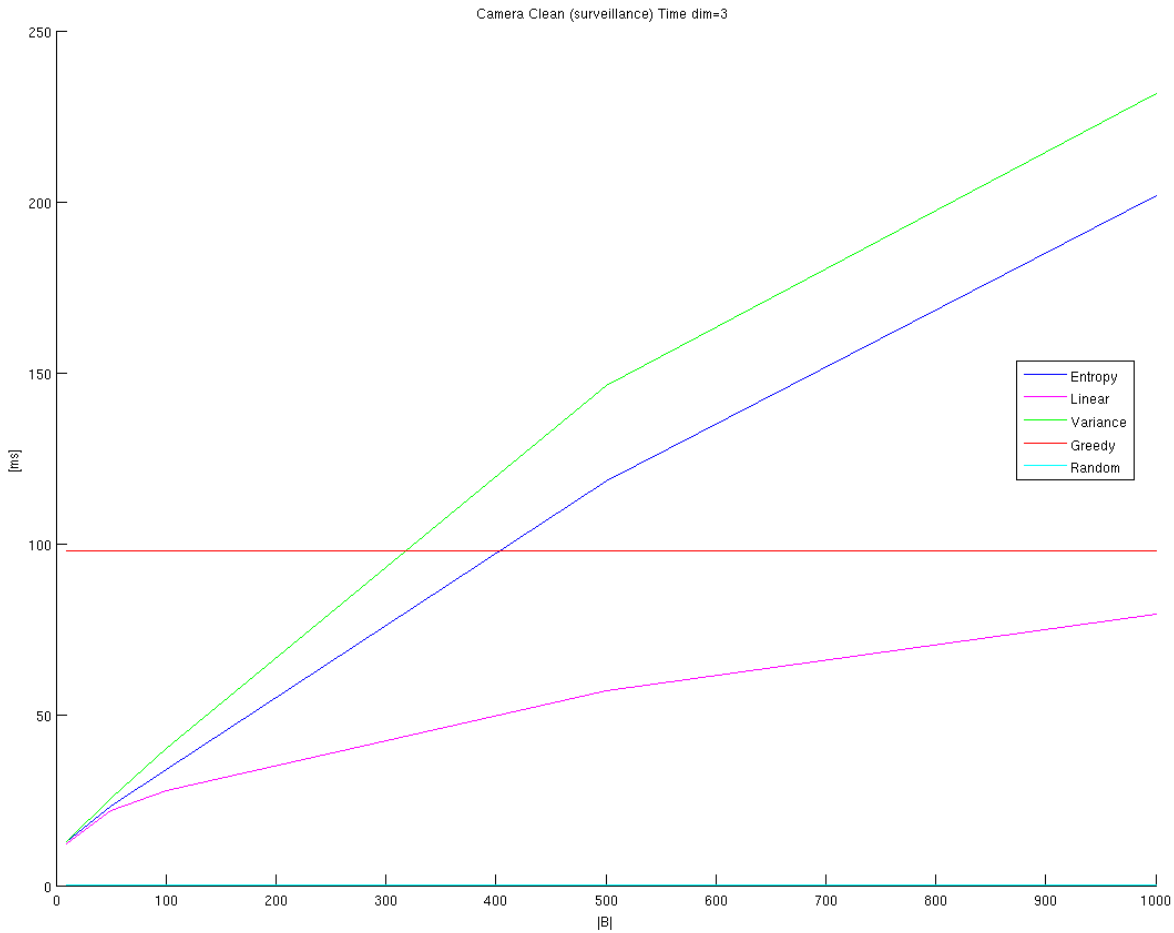
Récompense cumulée $= f(|B|)$

$[N = 3]$



Temps de calcul $= f(|B|)$

$[N = 3]$



Bilan

- Les ρPOMDP étendent les POMDP en permettant de récompenser le gain d'information.

- Pour les résoudre :
  - les approximations PWLC fonctionnent toujours ;
  - les algorithmes travaillant sur le bMDP (sans approx. PWLC) sont inchangés (RTDP-bel, POMCP, ...) ;
  - nombre de bornes usuelles (Q-MDP, ...) sont peu utiles.

Plan

Heuristiques gourmandes

Une heuristique souvent efficace :

maximiser l'espérance du gain immédiat d'information.
Ex. :

- Mastermind, pendu, ...
- échantillonnage dans un champ de Markov (Bonneau et al., 2012)
- suivi d'une cible

Heuristiques gourmandes

Une heuristique souvent efficace :

maximiser l'espérance du gain immédiat d'information.
Ex. :

- Mastermind, pendu, ...
- échantillonnage dans un champ de Markov (Bonneau et al., 2012)
- suivi d'une cible

C'est en particulier le cas (Williams et al., 2007) si

1. on souhaite maximiser le gain en information mutuelle, et
2. il n'y a pas d'état "interne" à contrôler.

Apprentissage actif de modèle

RL bayésien avec modèle

RL bayésien avec modèle (ici $\mu = T$)

- On maintient une croyance (belief) $\boldsymbol{b}_t(\boldsymbol{\mu})$
- Un cas simple :

$$\boldsymbol{\theta}_{s,a,s'} = \text{nbre de transitions } s \xrightarrow{a} s' \text{ observées} + 1$$

- On se ramène à nouveau à résoudre un belief MDP.
  (Mais Pb bcp + dur.)

Apprentissage actif de modèle

Comment apprendre T activement ?

3 étapes

1. Choisir un critère de performance
2. En dériver r (ou plutôt ρ)
3. Trouver π

Apprentissage actif de modèle

Comment apprendre T activement ?

3 étapes

1. Choisir un critère de performance
2. En dériver r (ou plutôt ρ)
3. Trouver π

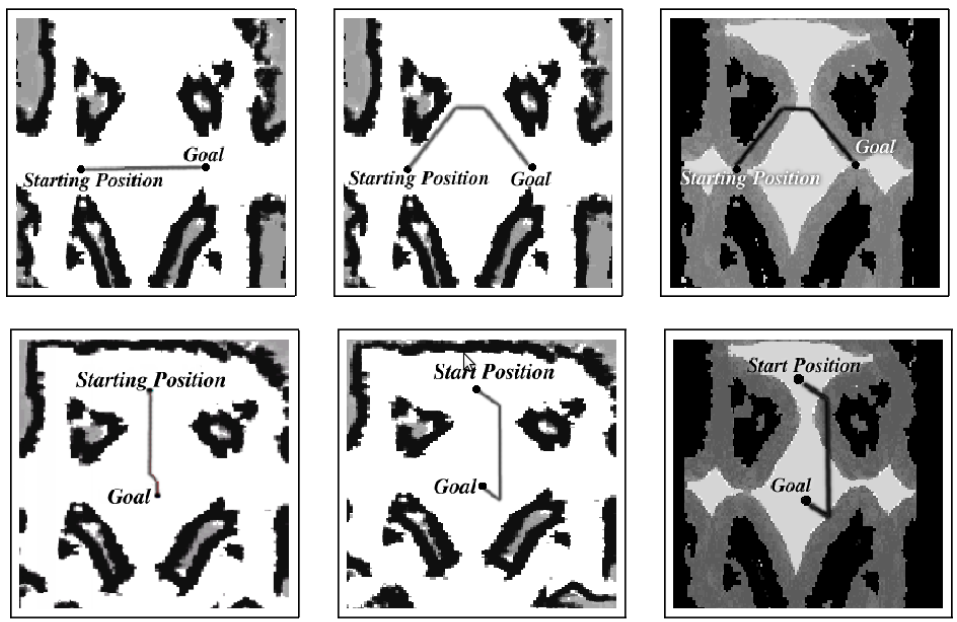
Note : La fonction de récompense force une exploration.

[Plus de détails dans (Araya-López et al., 2011).]

Heuristiques pour la planification ?

Ajouter des récompenses dépendant de l'état de croyance
$=^{?}$ bonne heuristique pour :

- compenser une mauvaise connaissance de b_0 (Teichteil et al.)
- guider des algorithmes vers une première politique (Thrun, 2000)
  (Intuition : Mieux vaut toujours savoir assez bien où l'on est.)



(a) Conventional (b) Coastal (c) Sensor Map

Plan

Conclusion

- Formalisation possible de problèmes de recherche d'information dans un cadre bayésien (ρPOMDP, ρBRL, ...).
- Choix du critère pas évident (problème mal posé ?).
- Une façon de modéliser la curiosité ? (s'intéresse à des variables cachées, au modèles, pas aux capacités)
- Résolution similaire aux problèmes classiques (plus coûteuse, mais même ordre de grandeur).
- Des heuristiques souvent efficaces.

Richard Bellman. The theory of dynamic programming. Bull. Amer. Math. Soc., 60 :503–516, 1954.

R. Smallwood and E. Sondik. The optimal control of partially observable Markov decision processes over a finite horizon. Operation Research, 21 :1071–1088, 1973.

G. Monahan. A survey of partially observable Markov decision processes. Management Science, 28 :1–16, 1982.

A. Cassandra. Exact and approximate algorithms for partially observable Markov decision processes. PhD thesis, Brown University, Providence, RI, USA, 1998.

J. Pineau, G. Gordon, and S. Thrun. Anytime point-based approximations for large POMDPs. Journal of Artificial Intelligence Research (JAIR), 27 :335–380, 2006.

P. Poupart, Kee-Eung Kim, and Dongho Kim. Closing the gap : Improved bounds on optimal POMDP solutions. In Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS), 2011.

A. Hero, D. Castan, D. Cochran, and K. Kastella. Foundations and Applications of Sensor Management. Springer Publishing Company, Incorporated, 2007. ISBN 0387278923, 9780387278926.

M. Spaan. Cooperative active perception using POMDPs. In AAAI 2008 Workshop on Advancements in POMDP Solvers, July 2008.

S. Thrun. Probabilistic algorithms in robotics. AI Magazine, 21(4) :93–109, 2000.

M. Bonneau, N. Peyrard, and R. Sabbadin. A reinforcement-learning algorithm for sampling design in Markov random fields. In Proceedings of the Twentieth European Conference on Artificial Intelligence (ECAI'12), 2012.

J.L. Williams, J.W. Fisher III, and A.S. Willsky. Performance guarantees for information theoretic active inference. In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2007.

M. Araya-López, O. Buffet, V. Thomas, and F. Charpillet. Active learning of MDP models. In Proceedings of the Ninth European Workshop on Reinforcement Learning (EWRL-11), 2011.