Représentations prédictives pour la prise de décision dans l'incertain

Abdeslam Boularias

Mercredi, le 28 juillet 2010

UNIVERSITÉ LAVAL

DAMAS
Décision, Adaptation, Multi-AgentS
www.damas.ift.ulaval.ca

Contrôle d'un système dynamique

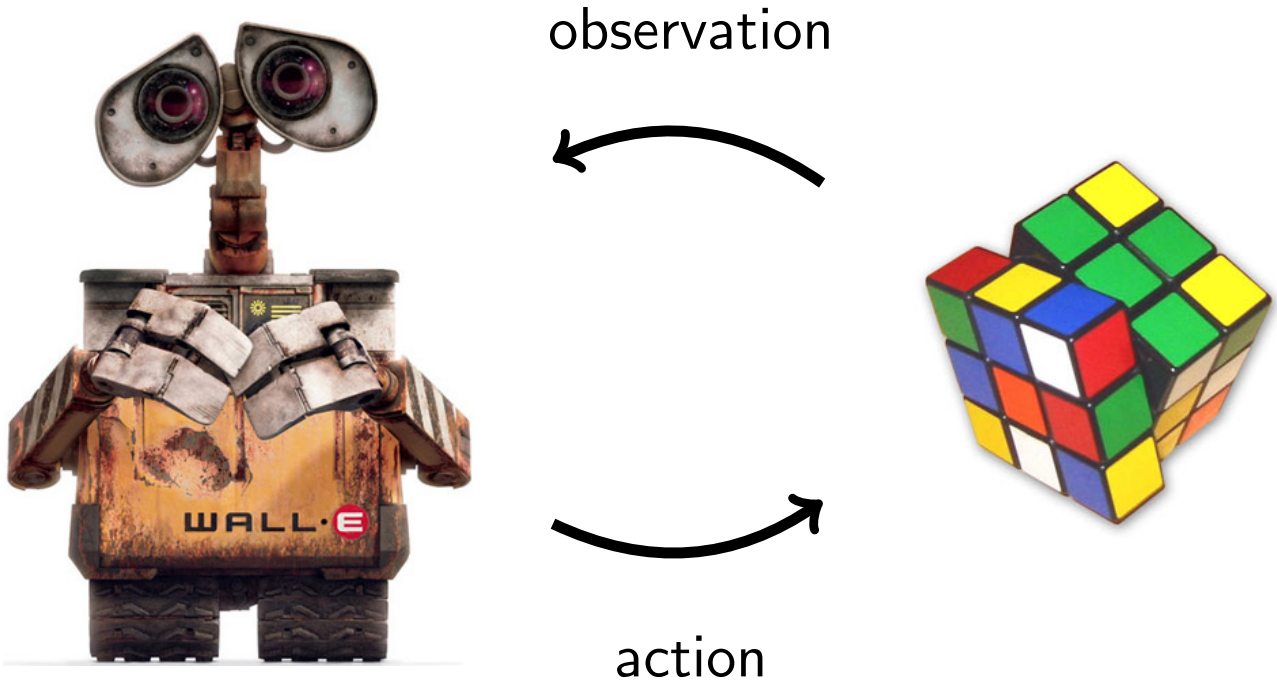


FIG.: L'interaction entre un agent et un système dynamique.

Prise de décision séquentielle

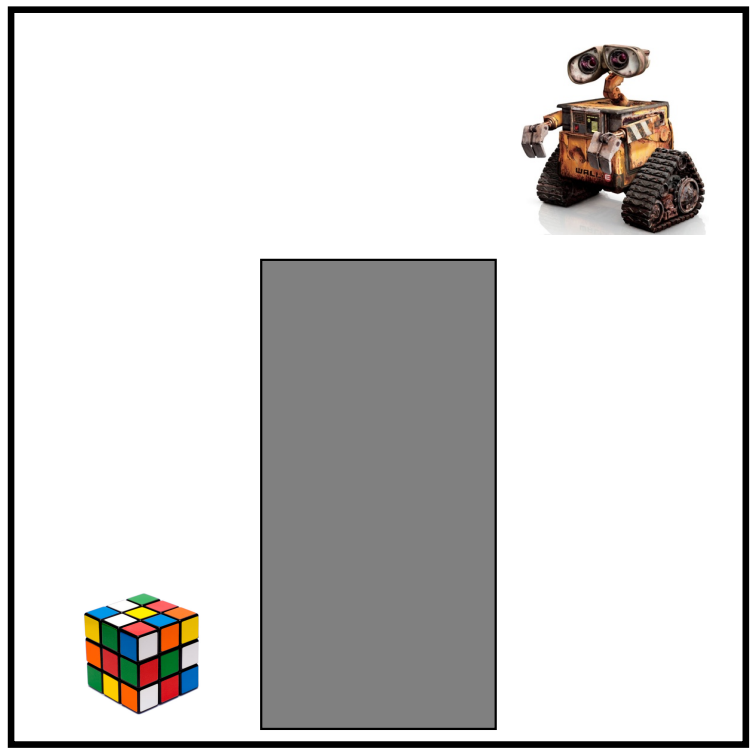


FIG.: Un simple problème de prise de décision séquentielle.

Prise de décision séquentielle

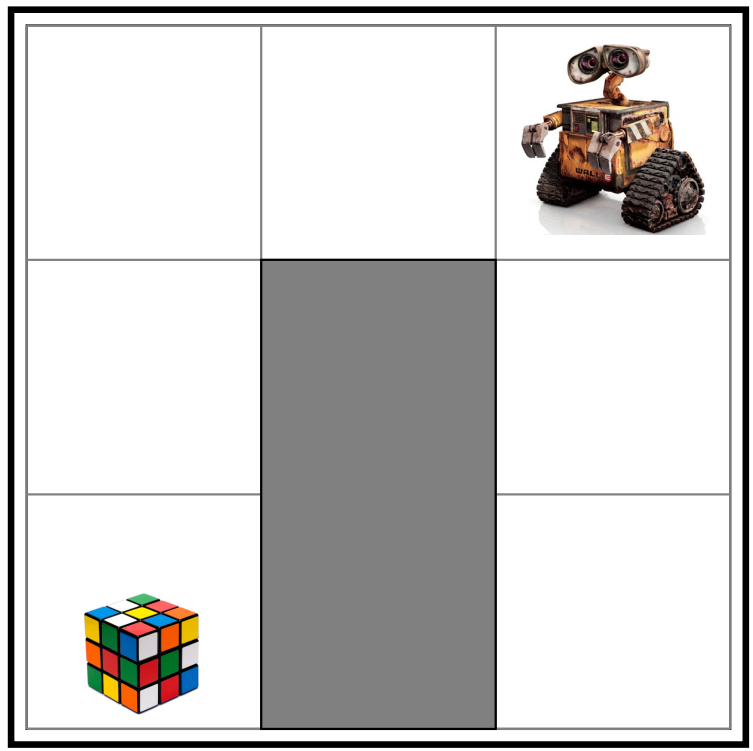


FIG.: Un simple problème de prise de décision séquentielle.

Prise de décision séquentielle

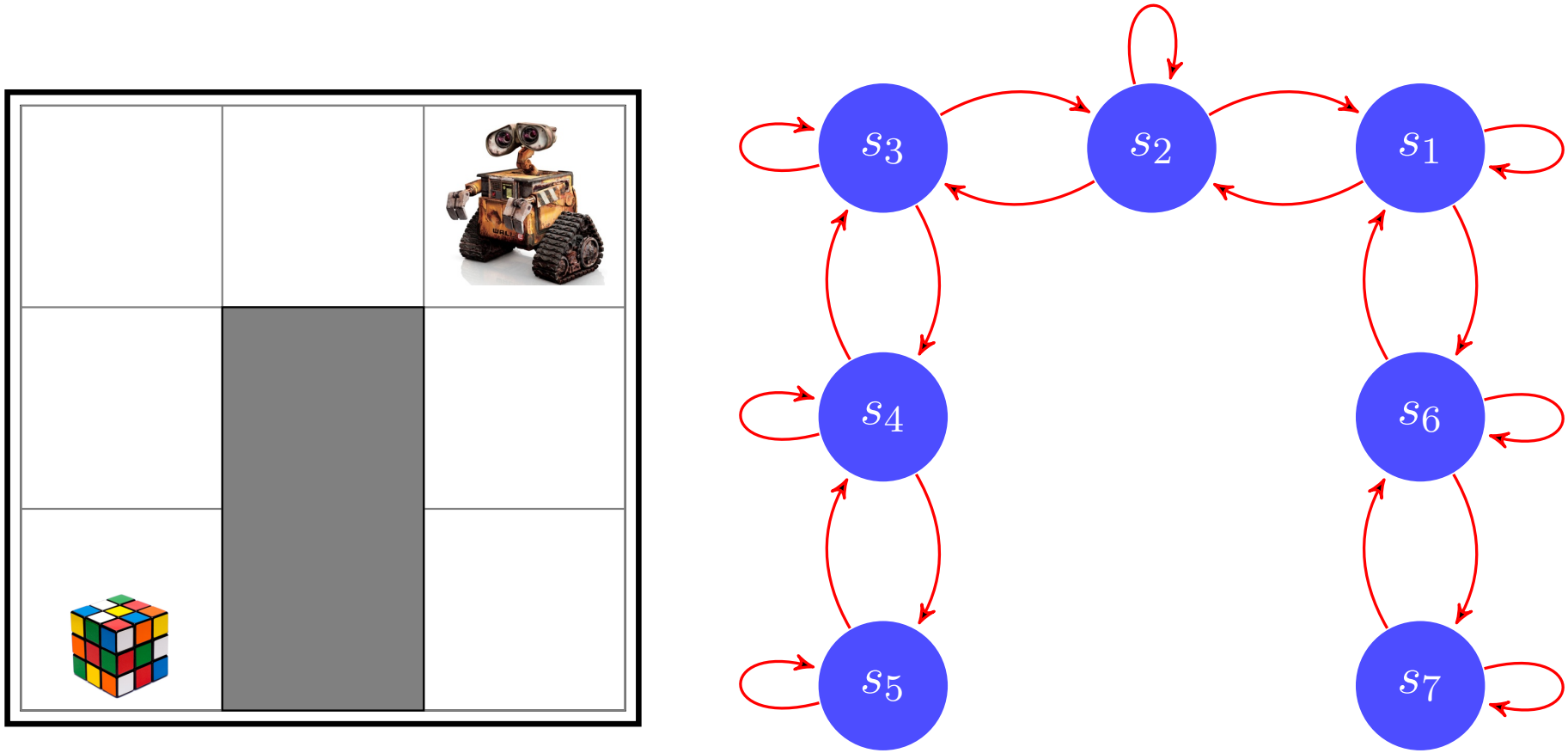


FIG.: Un processus décisionnel de Markov (MDP).

Un processus décisionnel de Markov (MDP) est défini par :

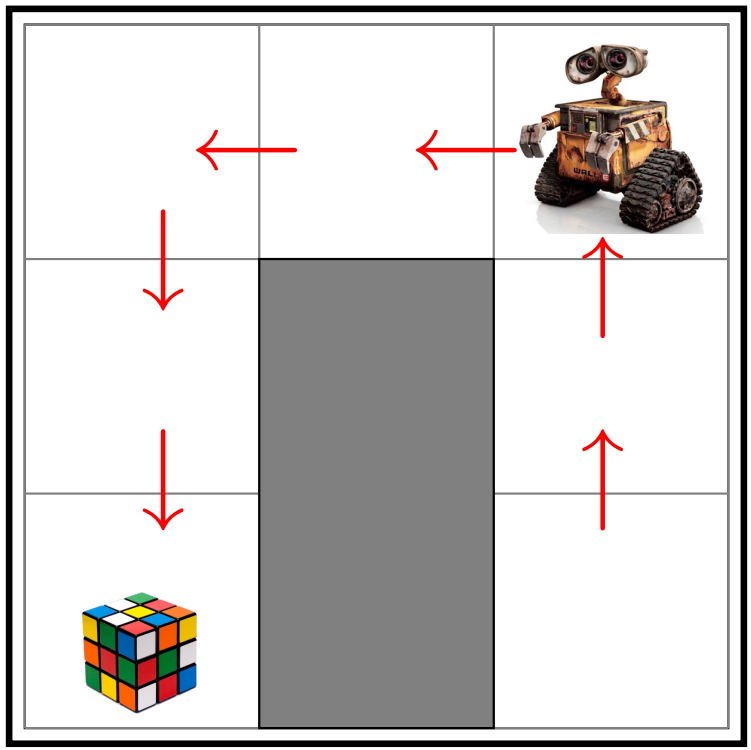
- $\mathcal{S}$: un ensemble des états.
- $\mathcal{A}$: un ensemble des actions.
- T^a : une fonction de transition stochastique pour chaque action $a \in \mathcal{A}$.

$$T^a(s, s') = Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

- R : une fonction de récompense (ou de coût).

Politiques

- Une politique est une fonction qui associe une action à chaque état.



Politiques

- La valeur (désirabilité) d'une politique π est donnée par :

$$V(\pi) = \mathbb{E}_{s_t}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi, T\right]$$

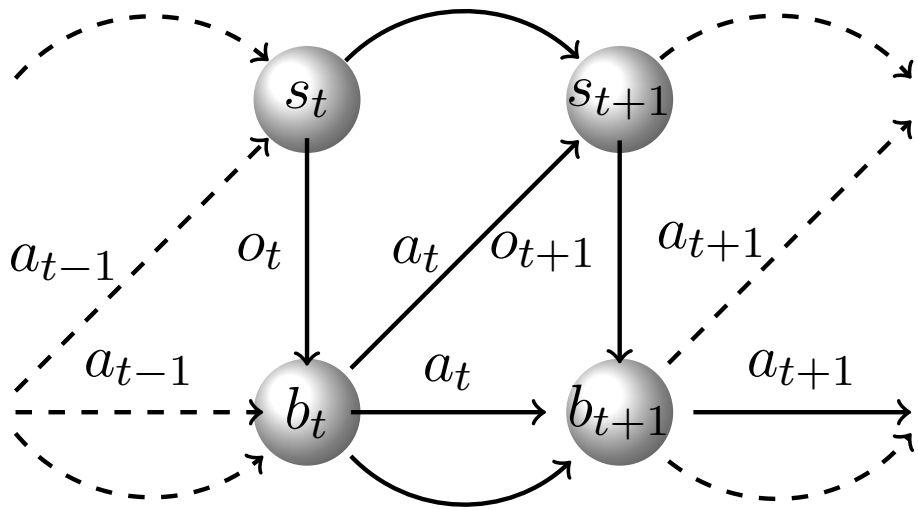
- $\gamma \in [0, 1)$ est un facteur d'escompte (ou d'amortissement).
- Résoudre un MDP $\Rightarrow$ Trouver une politique optimale.

Processus décisionnels de Markov partiellement observables (POMDPs)

- Les observations peuvent êtres partielles et incorrectes.
- Un MDP partiellement observable (POMDP) est un MDP avec un ensemble d'observations $\mathcal{O}$, et une fonction d'observation Z.
- $Z(s, o)$ est la probabilité d'observer o dans l'état s.

États réels (cachés) :

États de croyance :



Exemple d'un POMDP

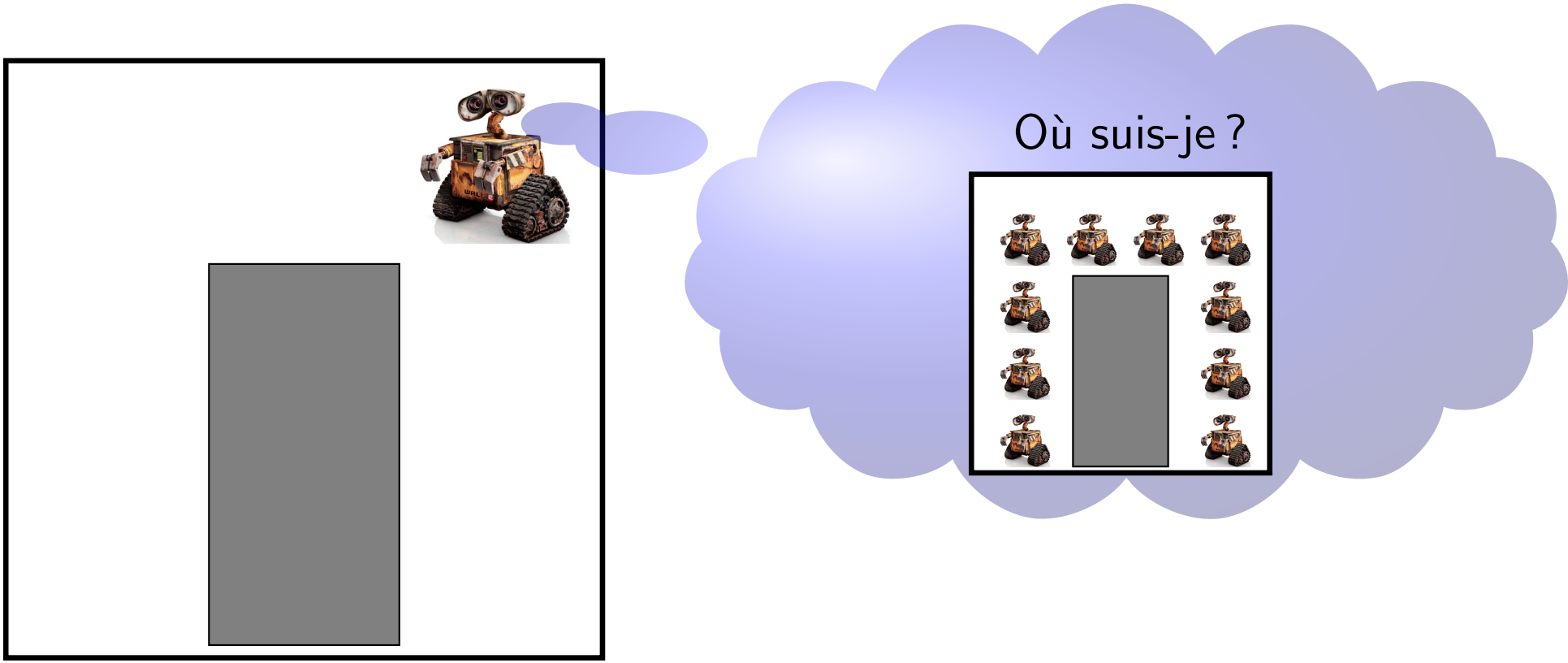


FIG.: État de croyance initial.

Exemple d'un POMDP

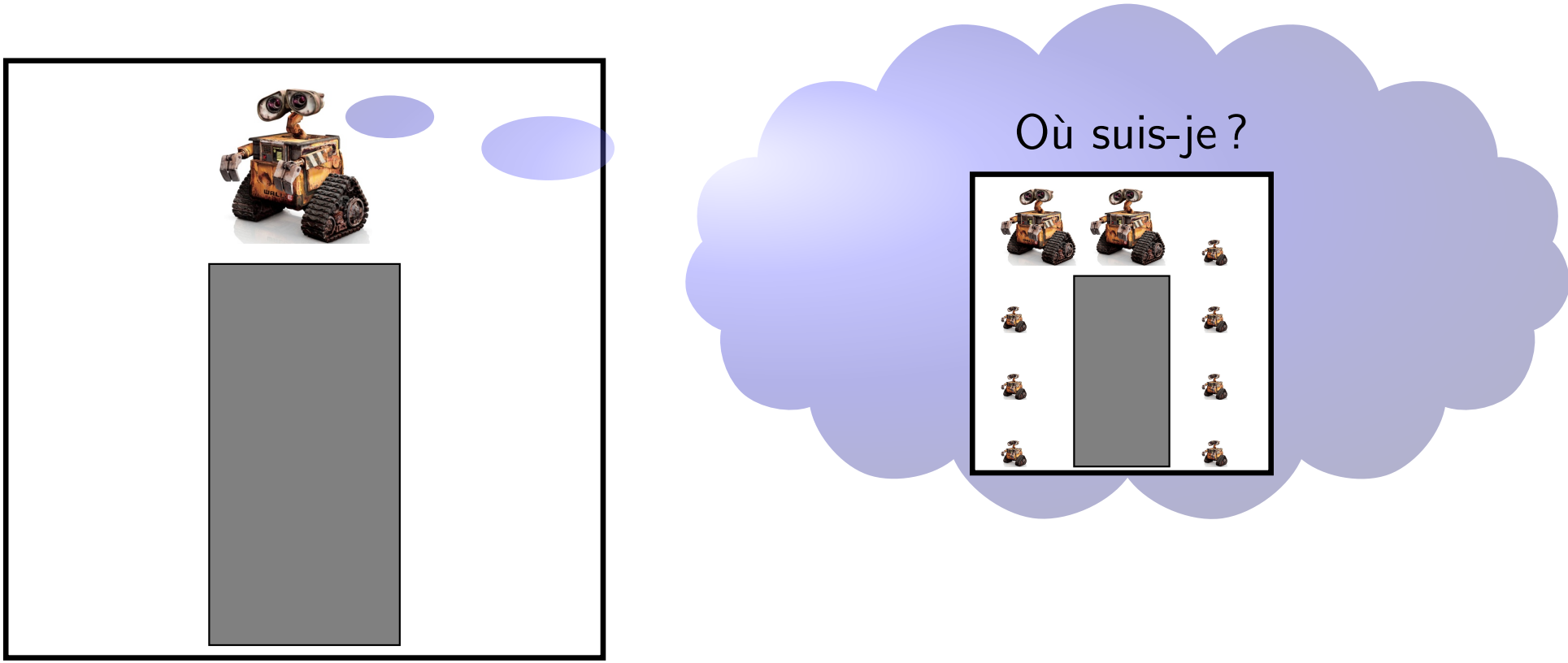


FIG.: État de croyance après avoir exécuté une action.

Exemple d'un POMDP

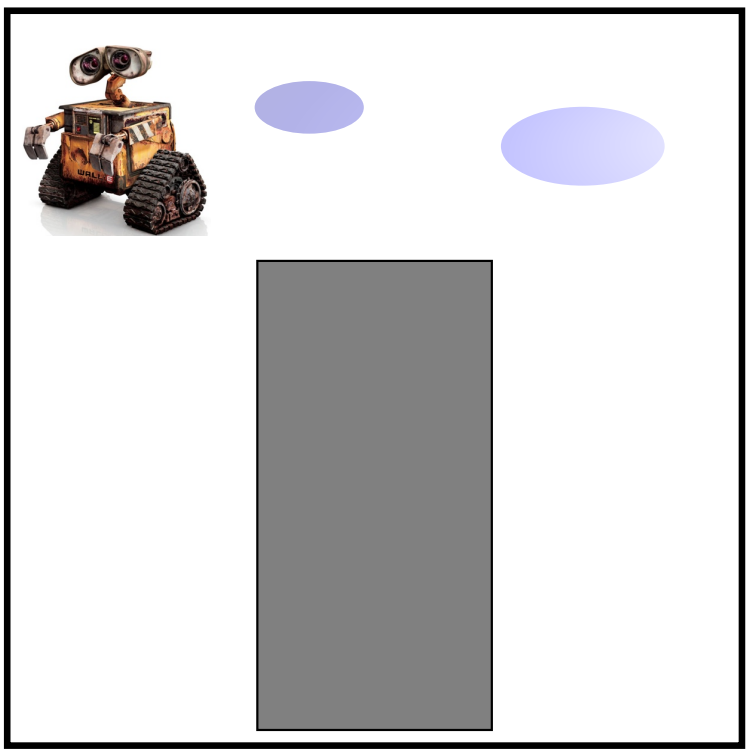
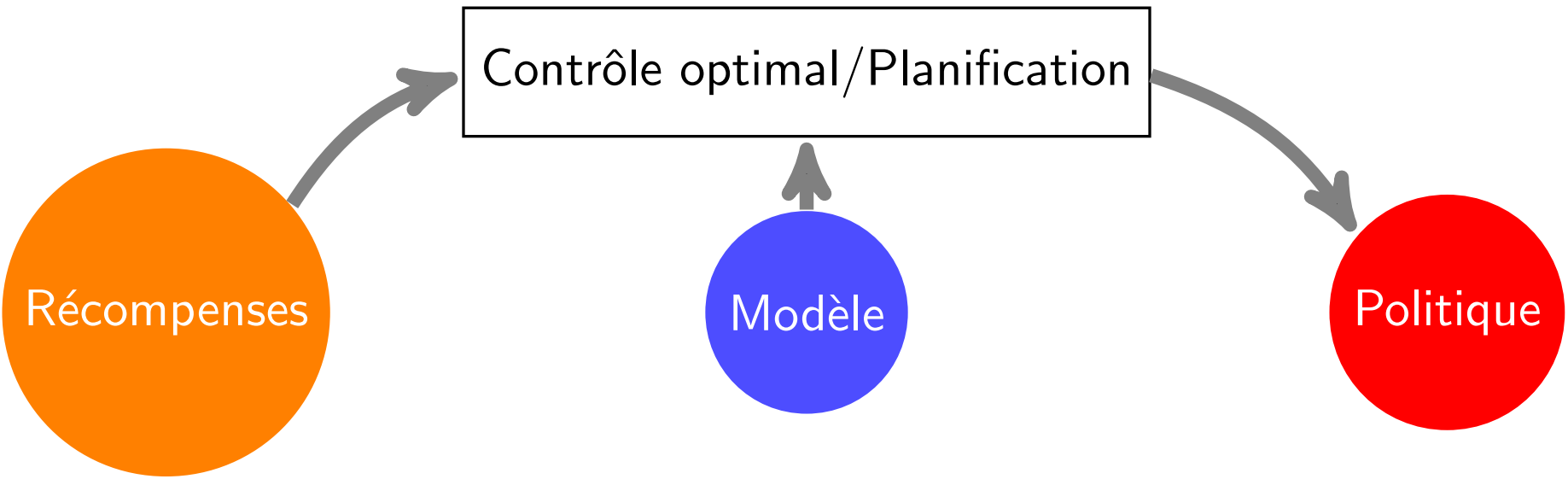


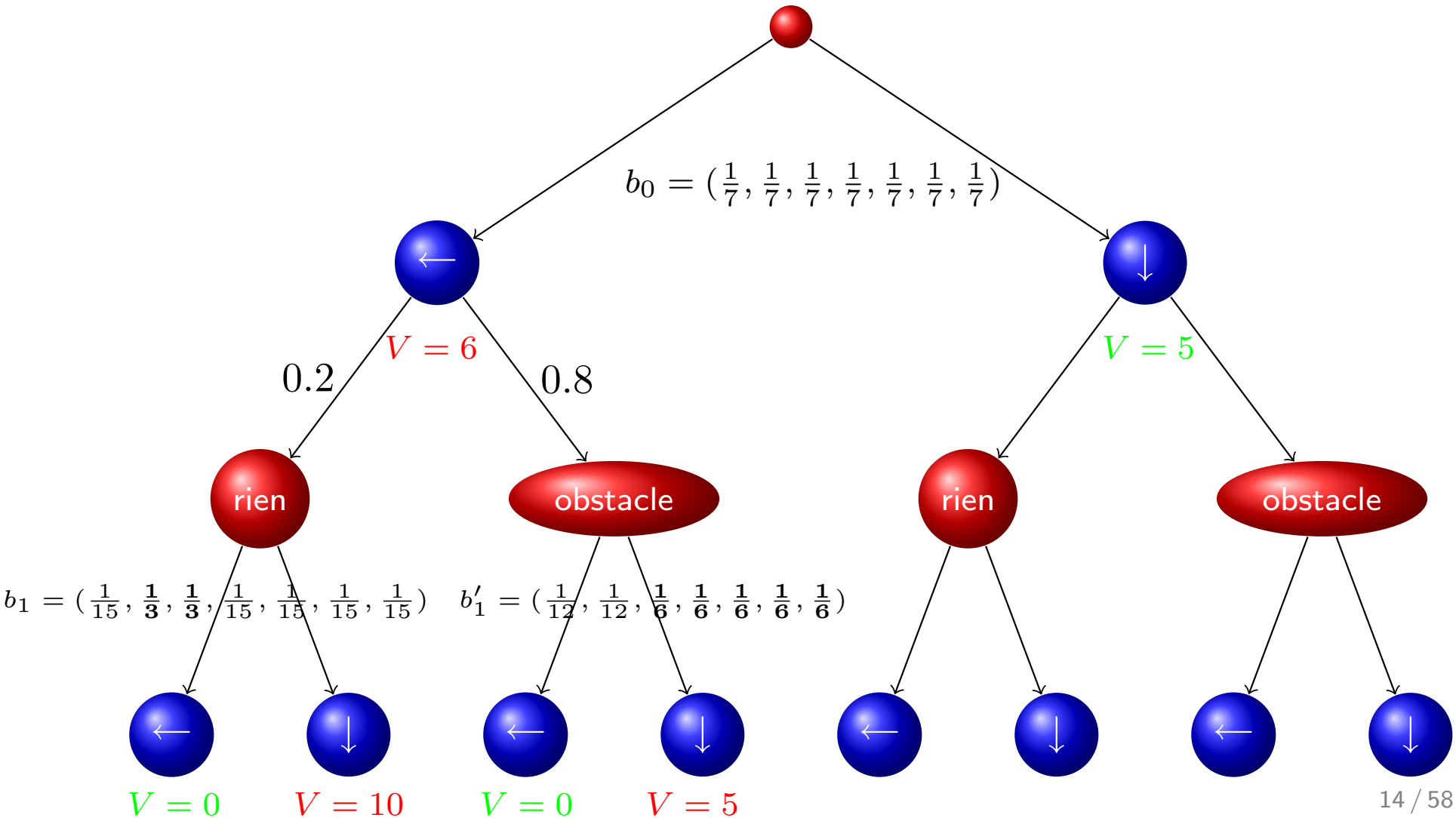
FIG.: État de croyance après avoir exécuté deux actions.

Résoudre un (PO)MDP



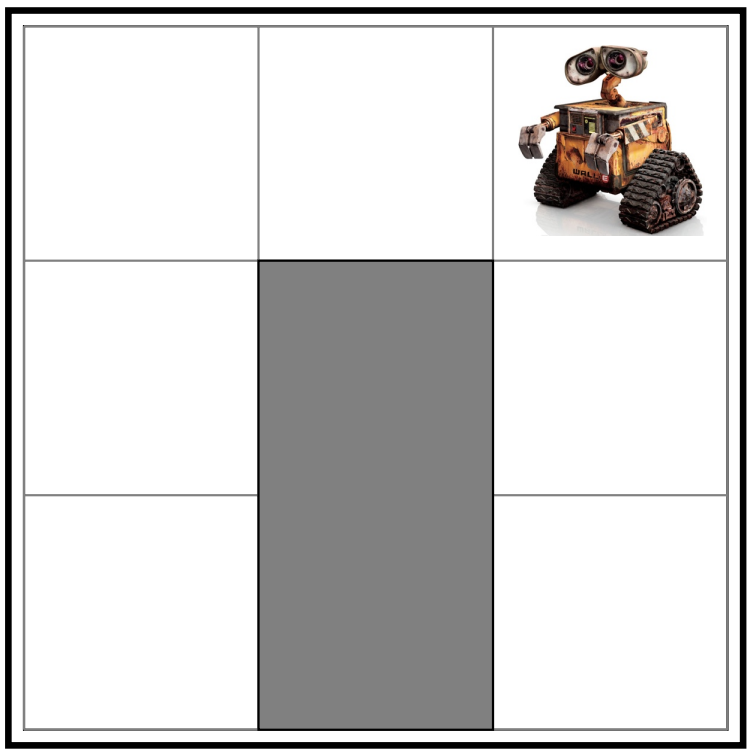
Planification avec les POMDPs

$$V^*(b) = \max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \sum_{o \in \mathcal{O}} Pr(o|b,a) V^*(b_{a,o}) \right]$$

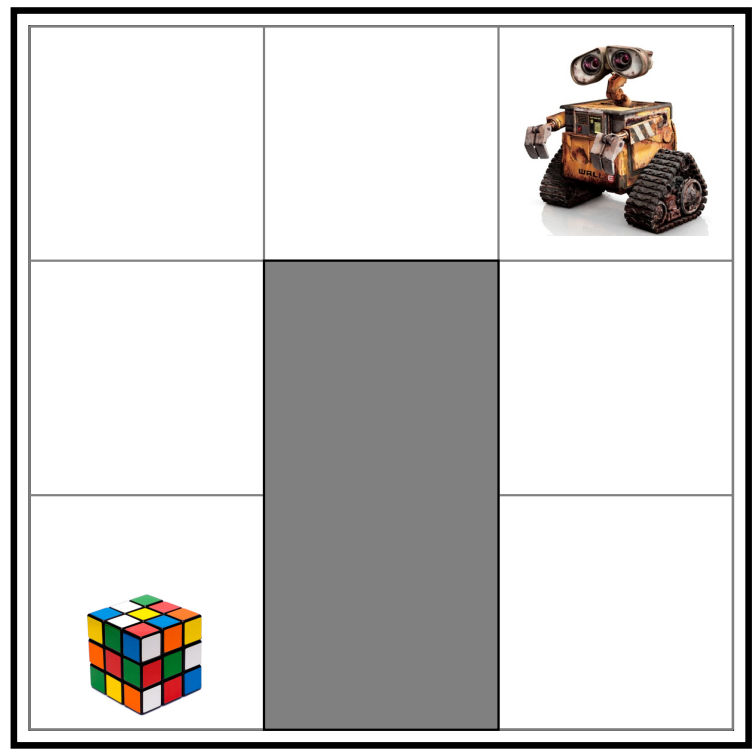


La malédiction de la dimension

La dimension de l'espace des états est exponentielle en nombre de variables qui le définissent.



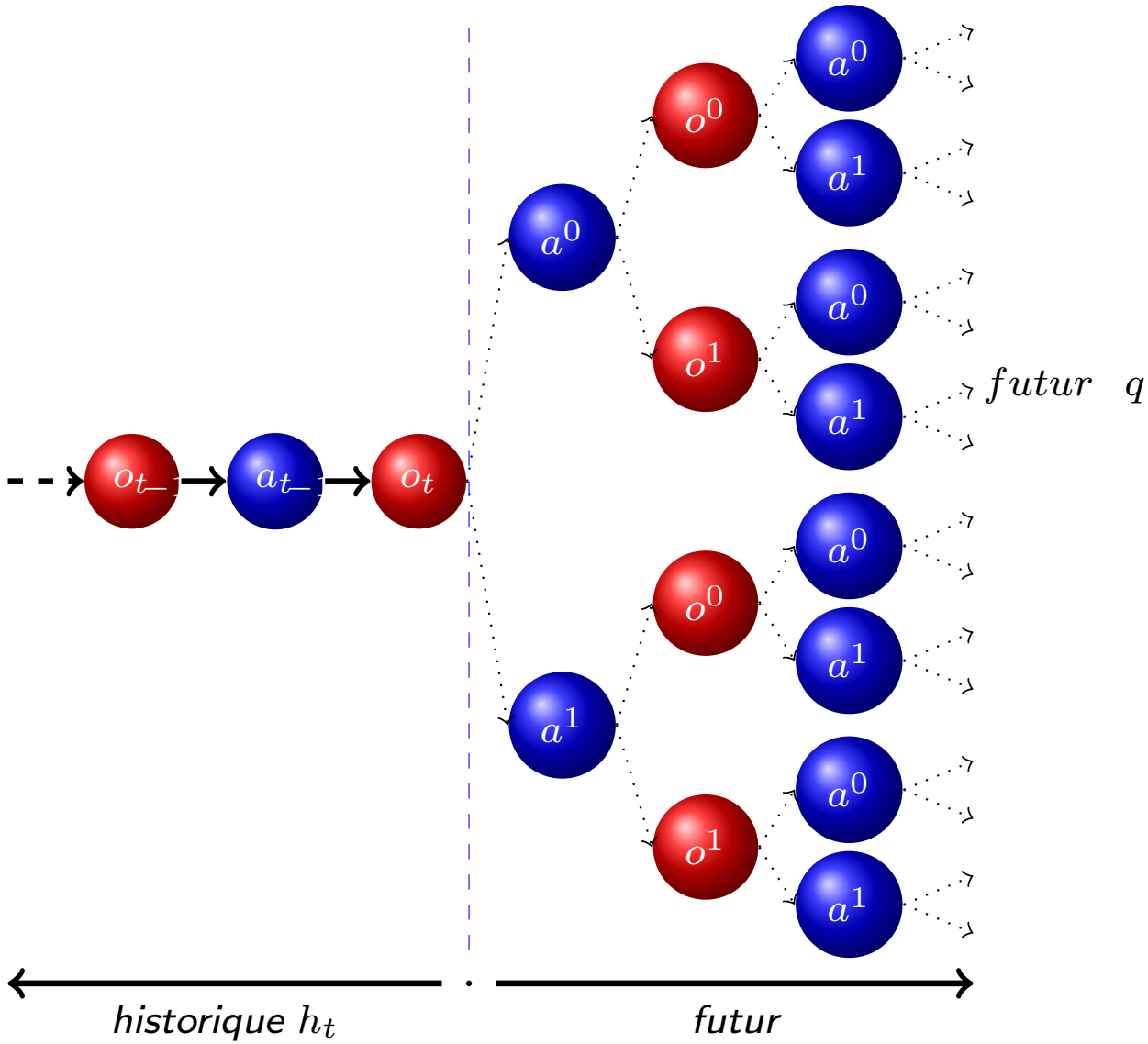
(a) Nombre d'états $= 7$



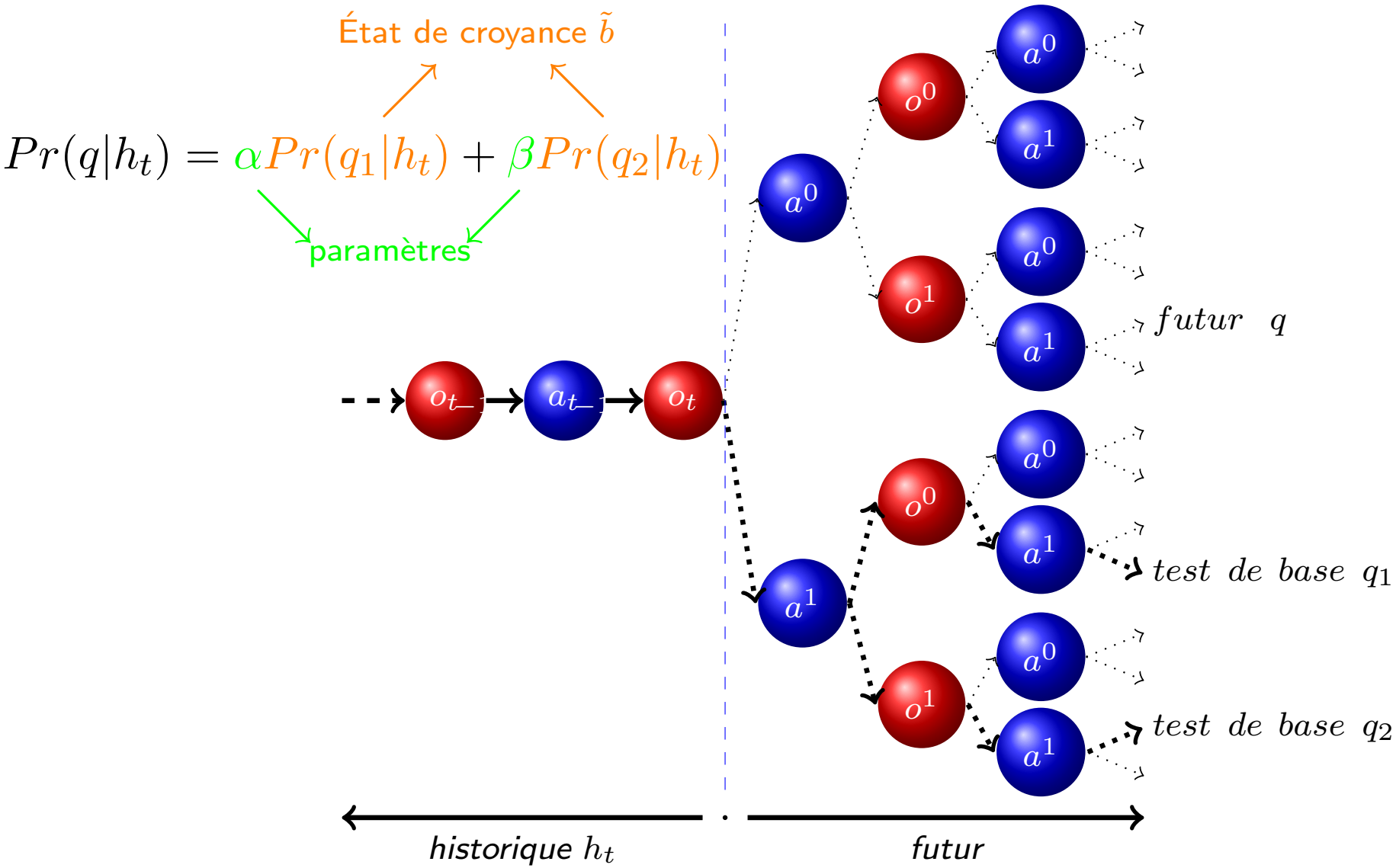
(b) Nombre d'états $= 7 \times 7$

Représentation prédictive des états (PSR) [Littman et al., 2002]

$Pr(q|h_t) = ?$



Représentation prédictive des états (PSR) [Littman et al., 2002]



Représentation prédictive des états (PSR) [Littman et al., 2002]

$$U(q|s) = Pr(q|s)$$

$$U = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} q_1 & q_2 & q_3 & q_4 & q_5 & q_6 & q_7 \end{array} \\ \begin{pmatrix} 0.2 & 0.5 & 1 & 0 & 0.2 & 0.5 & 1 & \dots \\ 0.5 & 0.8 & 0.3 & 0.7 & 0.5 & 0.8 & 0.3 & \dots \\ 0.5 & 0.9 & 0.1 & 0.3 & 0.5 & 0.9 & 0.1 & \dots \\ 1 & 0 & 1 & 0.2 & 1 & 0 & 1 & \dots \\ 0.5 & 0.3 & 0.8 & 0.9 & 0.5 & 0.3 & 0.8 & \dots \\ 0.9 & 0.6 & 0.7 & 0.3 & 0.9 & 0.6 & 0.7 & \dots \\ 0.9 & 0.9 & 0.5 & 0 & 0.9 & 0.9 & 0.5 & \dots \end{pmatrix} \end{array}$$

Représentation prédictive des états (PSR) [Littman et al., 2002]

$$U(q|s) = Pr(q|s)$$

$$U = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \end{array} \begin{array}{c} \begin{array}{cccc} q_1 & q_2 & q_3 & q_4 \end{array} \\ \left(\begin{array}{cccc} 0.2 & 0.5 & 1 & 0 \\ 0.5 & 0.8 & 0.3 & 0.7 \\ 0.5 & 0.9 & 0.1 & 0.3 \\ 1 & 0 & 1 & 0.2 \\ 0.5 & 0.3 & 0.8 & 0.9 \\ 0.9 & 0.6 & 0.7 & 0.3 \\ 0.9 & 0.9 & 0.5 & 0 \end{array}\right. \\ F \end{array} \begin{array}{ccc} q_5 & q_6 & q_7 \\ \left.\begin{array}{ccc} 0.2 & 0.5 & 1 \\ 0.5 & 0.8 & 0.3 \\ 0.5 & 0.9 & 0.1 \\ 1 & 0 & 1 \\ 0.5 & 0.3 & 0.8 \\ 0.9 & 0.6 & 0.7 \\ 0.9 & 0.9 & 0.5 \end{array}\right) \end{array} \begin{array}{c} \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \end{array}$$

$$Pr(q|s) = aPr(q_1|s) + bPr(q_2|s) + cPr(q_3|s) + dPr(q_4|s)$$

Représentation prédictive des états (PSR) [Littman et al., 2002]

$$U(q|s) = Pr(q|s)$$

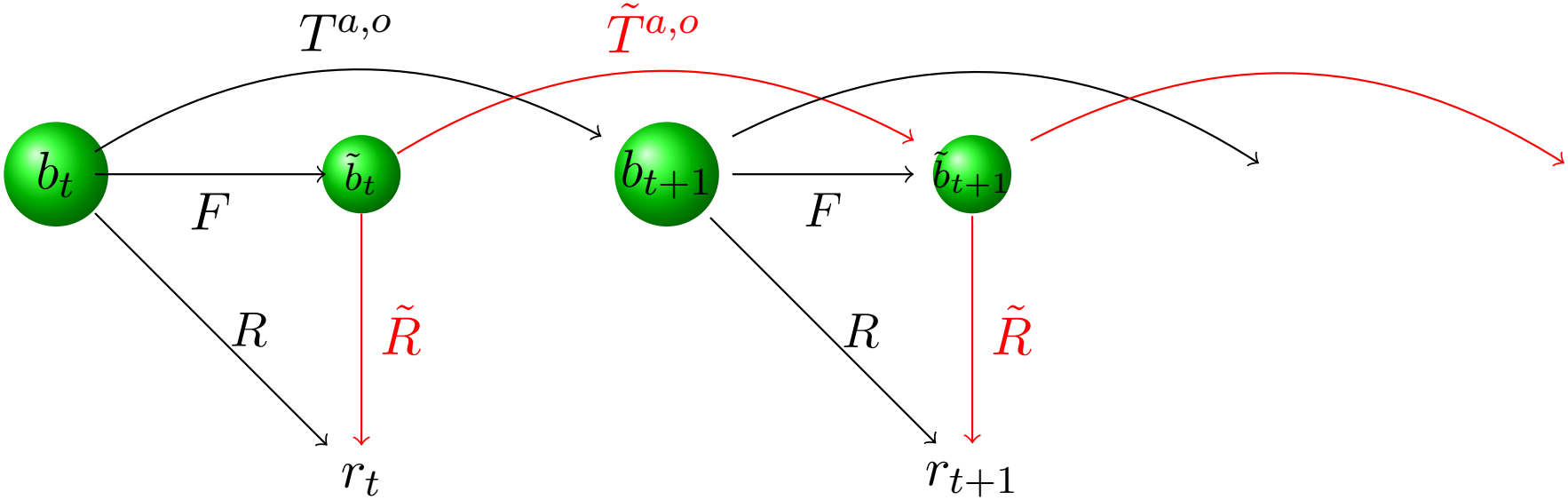
$$U = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} q_1 & q_2 & q_3 & q_4 & q_5 & q_6 & q_7 \end{array} \\ \left(\begin{array}{ccccccc} 0.2 & 0.5 & 1 & 0 & 0.2 & 0.5 & 1 \\ 0.5 & 0.8 & 0.3 & 0.7 & 0.5 & 0.8 & 0.3 \\ 0.5 & 0.9 & 0.1 & 0.3 & 0.5 & 0.9 & 0.1 \\ 1 & 0 & 1 & 0.2 & 1 & 0 & 1 \\ 0.5 & 0.3 & 0.8 & 0.9 & 0.5 & 0.3 & 0.8 \\ 0.9 & 0.6 & 0.7 & 0.3 & 0.9 & 0.6 & 0.7 \\ 0.9 & 0.9 & 0.5 & 0 & 0.9 & 0.9 & 0.5 \end{array} \right. \begin{array}{c} \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \end{array} \\ \left. \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \right) \end{array}$$

F

$$Pr(q|s) \approx a' Pr(q_1|s) + b' Pr(q_2|s) + c' Pr(q_3|s) \cancel{+ d Pr(q_4|s)}$$

Compression de l'espace des états

Minimiser $\|F\tilde{T}^{a,o} - T^{a,o}F\|_\infty$, $\|F\tilde{R} - R\|_\infty$, et $\|F\tilde{Z} - Z\|_\infty$.



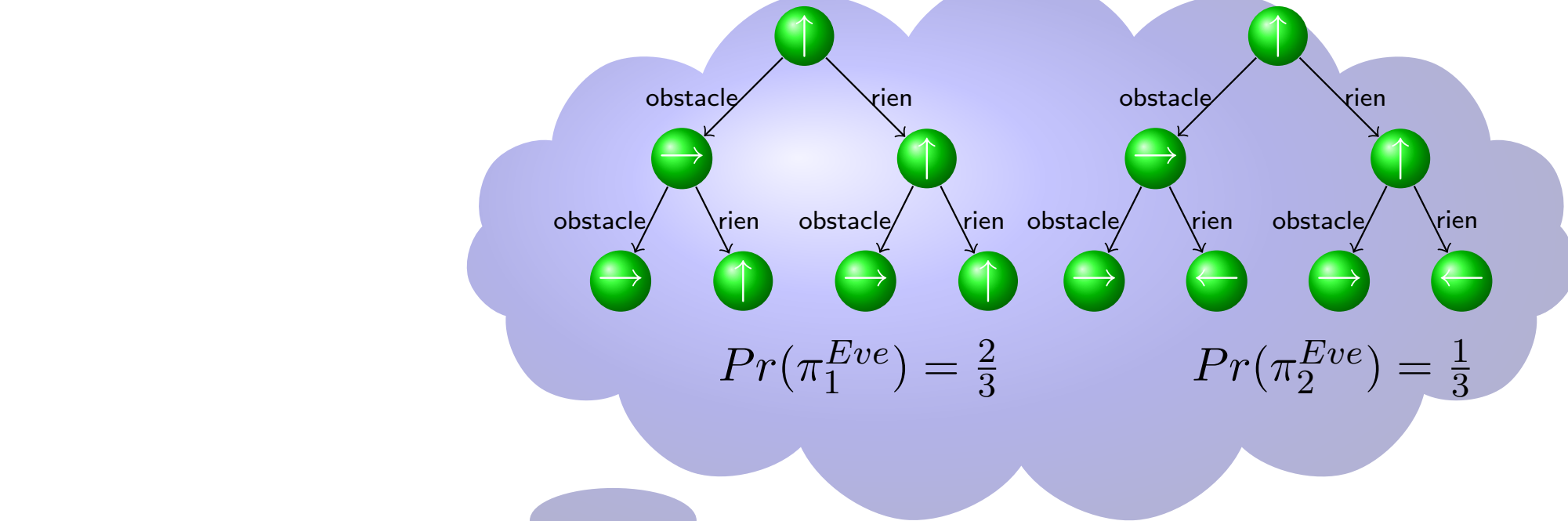
Ce problème peut être formalisé comme un programme linéaire.

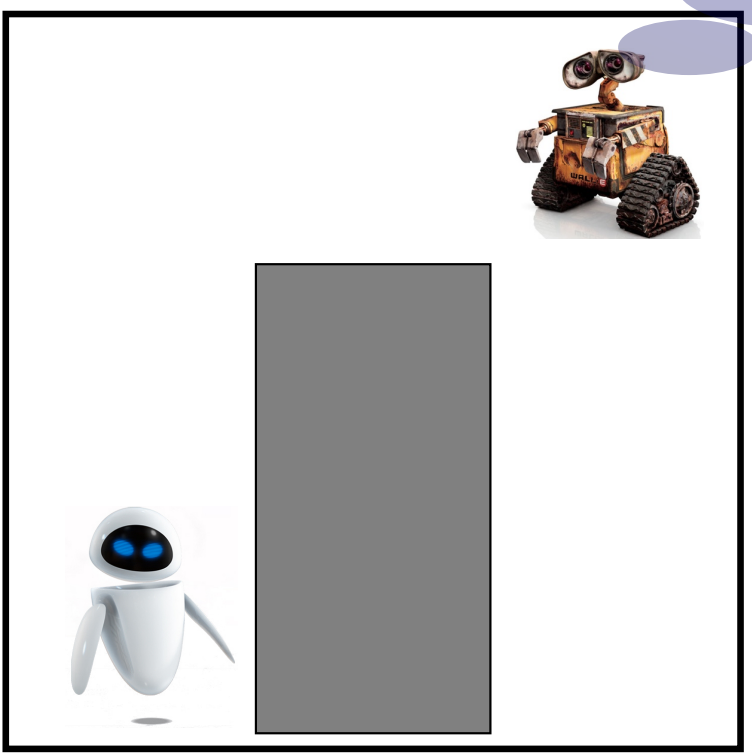
Résultats empiriques [Boularias et al., ICMLA 2008]

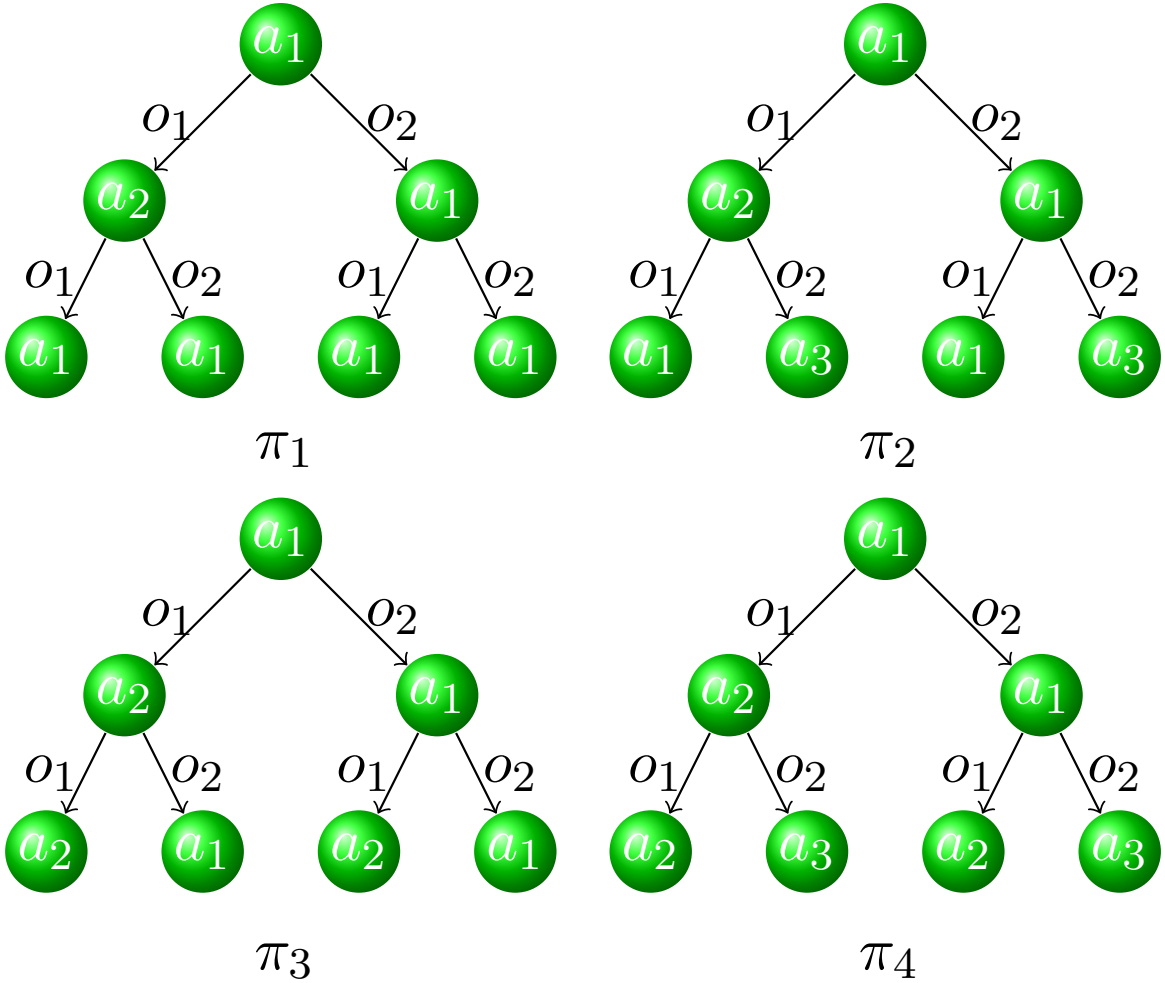
	PSR approximé			POMDP	
Domaine	Compression	Temps	Récompense	Temps	Récompense
Hallway				27.58	0.06
	33%	15.89	0.00		
	66%	20.70	0.06		
Hallway2				80.20	0.02
	22%	1.56	0.00		
	50%	5.54	0.01		

TAB.: Temps moyen, en millisecondes, et récompense par étape.

POMDPs décentralisés


obstacle
rien
obstacle
rien
obstacle
rien
obstacle
rien
obstacle
rien
obstacle
rien
$Pr(\pi_1^{Eve}) = \frac{2}{3}$
$Pr(\pi_2^{Eve}) = \frac{1}{3}$





La malédiction de la dimension : La dimension de l'espace des politiques est exponentielle en nombre d'observations et doublement exponentielle en horizon de planification.

Représentation prédictive des politiques (PPRs)

$$U = \begin{array}{c} \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{array} \begin{array}{c} \begin{array}{cccccccc} a_1o_1a_2o_1a_1 & a_1o_1a_2o_1a_2 & a_1o_1a_2o_2a_1 & a_1o_1a_2o_2a_3 & a_1o_2a_1o_1a_1 & a_1o_2a_1o_1a_2 & a_1o_2a_1o_2a_1 & a_1o_2a_1o_2a_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \text{dimension} = 4$$

Représentation prédictive des politiques (PPRs)

$$
U = \begin{array}{c}
\begin{array}{ccccccccc}
q_1 & q_2 & q_3 & & & & & & \\
\mathbf{a_1o_1a_2o_1a_1} & \mathbf{a_1o_1a_2o_1a_2} & \mathbf{a_1o_1a_2o_2a_1} & a_1o_1a_2o_2a_3 & a_1o_2a_1o_1a_1 & a_1o_2a_1o_1a_2 & a_1o_2a_1o_2a_1 & a_1o_2a_1o_2a_3 &
\end{array} \\
\begin{pmatrix}
\pi_1 & \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 1 & 0 & 1 & 0 \\
\pi_2 & \mathbf{1} & \mathbf{0} & \mathbf{0} & 1 & 1 & 0 & 0 & 1 \\
\pi_3 & \mathbf{0} & \mathbf{1} & \mathbf{1} & 0 & 0 & 1 & 1 & 0 \\
\pi_4 & \mathbf{0} & \mathbf{1} & \mathbf{0} & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
\end{array}
\quad \text{dimension} = 4
$$

Représentation prédictive des politiques (PPRs)

$$
U = \begin{array}{c}
\begin{array}{ccc} q_1 & q_2 & q_3 \end{array} \\
\begin{array}{c} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{array}
\overbrace{\begin{pmatrix}
\mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 1 & 0 & 1 & 0 \\
\mathbf{1} & \mathbf{0} & \mathbf{0} & 1 & 1 & 0 & 0 & 1 \\
\mathbf{0} & \mathbf{1} & \mathbf{1} & 0 & 0 & 1 & 1 & 0 \\
\mathbf{0} & \mathbf{1} & \mathbf{0} & 1 & 0 & 1 & 0 & 1
\end{pmatrix}}^{\begin{array}{c} \mathbf{a_1o_1a_2o_1a_1} \\ \mathbf{a_1o_1a_2o_1a_2} \\ \mathbf{a_1o_1a_2o_2a_1} \\ a_1o_1a_2o_2a_3 \\ a_1o_2a_1o_1a_1 \\ a_1o_2a_1o_1a_2 \\ a_1o_2a_1o_2a_1 \\ a_1o_2a_1o_2a_3 \end{array}}
\end{array}
\text{ dimension} = 4
$$

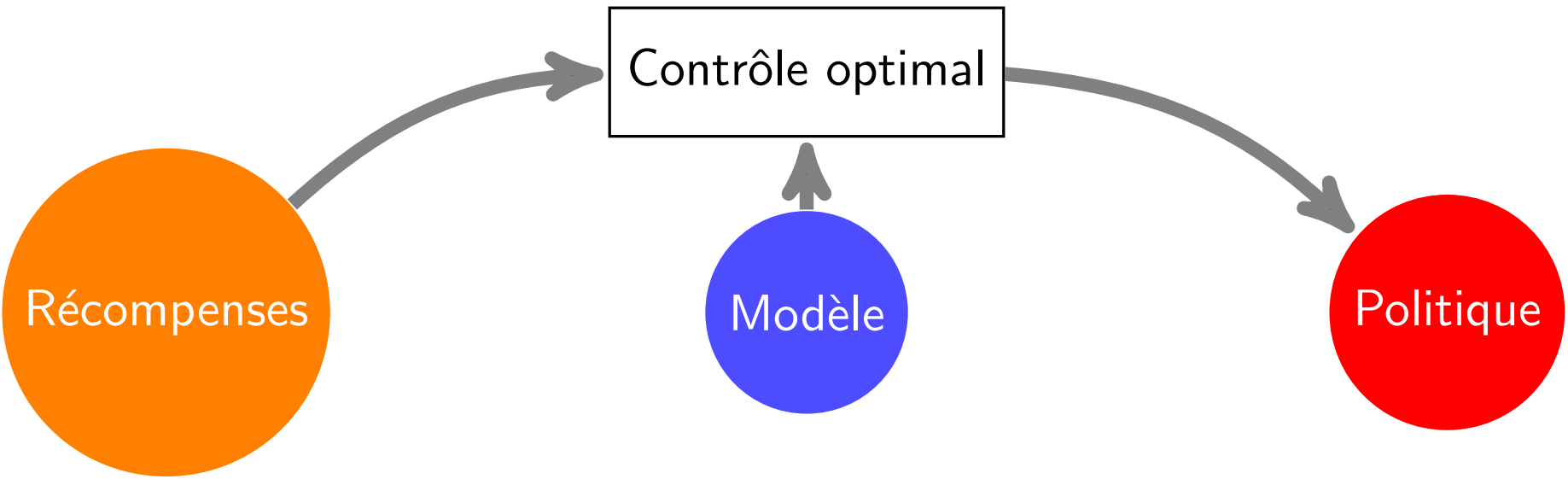
$$
= \underbrace{\begin{pmatrix}
\mathbf{1} & \mathbf{0} & \mathbf{1} \\
\mathbf{1} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{1} & \mathbf{1} \\
\mathbf{0} & \mathbf{1} & \mathbf{0}
\end{pmatrix}}_{F} \times \begin{pmatrix}
1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & -1 & 0 & 0 & 1 & -1
\end{pmatrix} \text{ dimension} = 3
$$

Résultats empiriques [Boularias & Chaib-draa, ICAPS 2008]

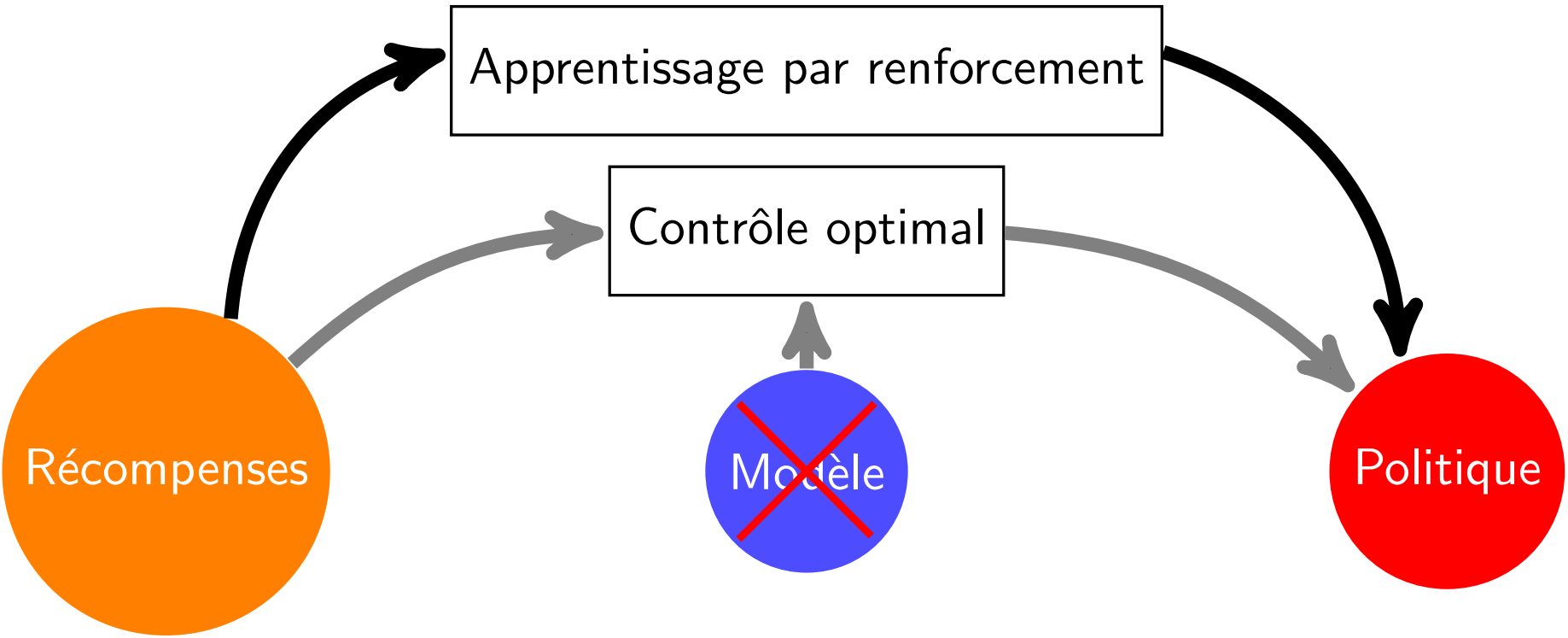
		Arbres de décisions		**Représentations prédictives**	
Problème	T	temps	arbres	temps	séquences
Multiagent Tiger	2	0.20	(27,27)	0.17	(18,18)
	3	2.29	(675,675)	1.79	(90,90)
	4	-	-	534.90	(540,540)
Multiagent Broadcasting Channel	2	0.12	(8,8)	0.14	(8,8)
	3	0.46	(72,72)	0.36	(24,24)
	4	17.59	(1800,1458)	4.59	(80,80)

TAB.: Temps d'exécution (en secondes) et nombre de politiques et de séquences retenues par l'algorithme de programmation dynamique.

Résoudre un (PO)MDP



Résoudre un (PO)MDP



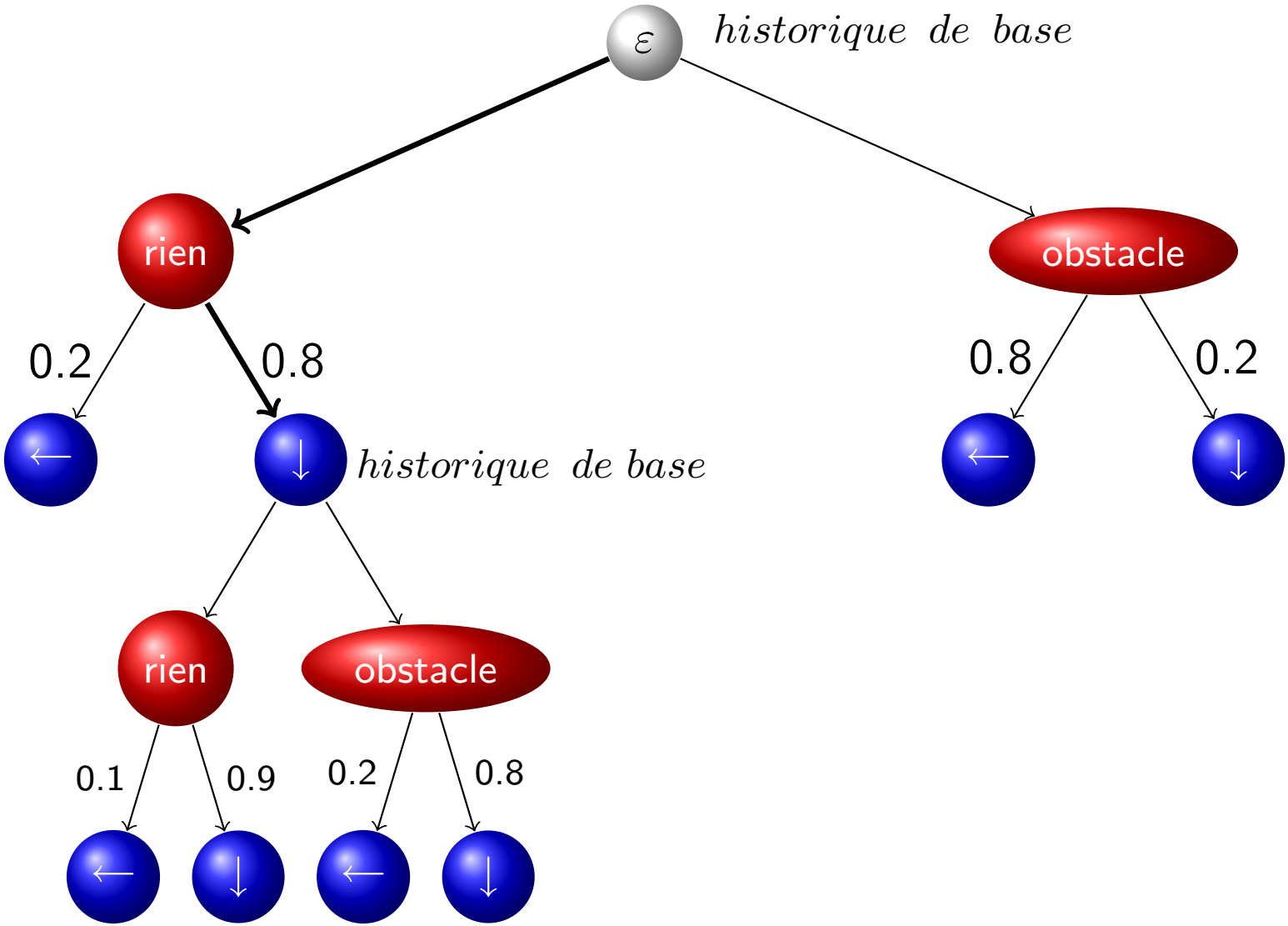
Gradient de la politique

Le gradient de la fonction de valeur d'une politique définie par le vecteur de paramètres θ est donné par :

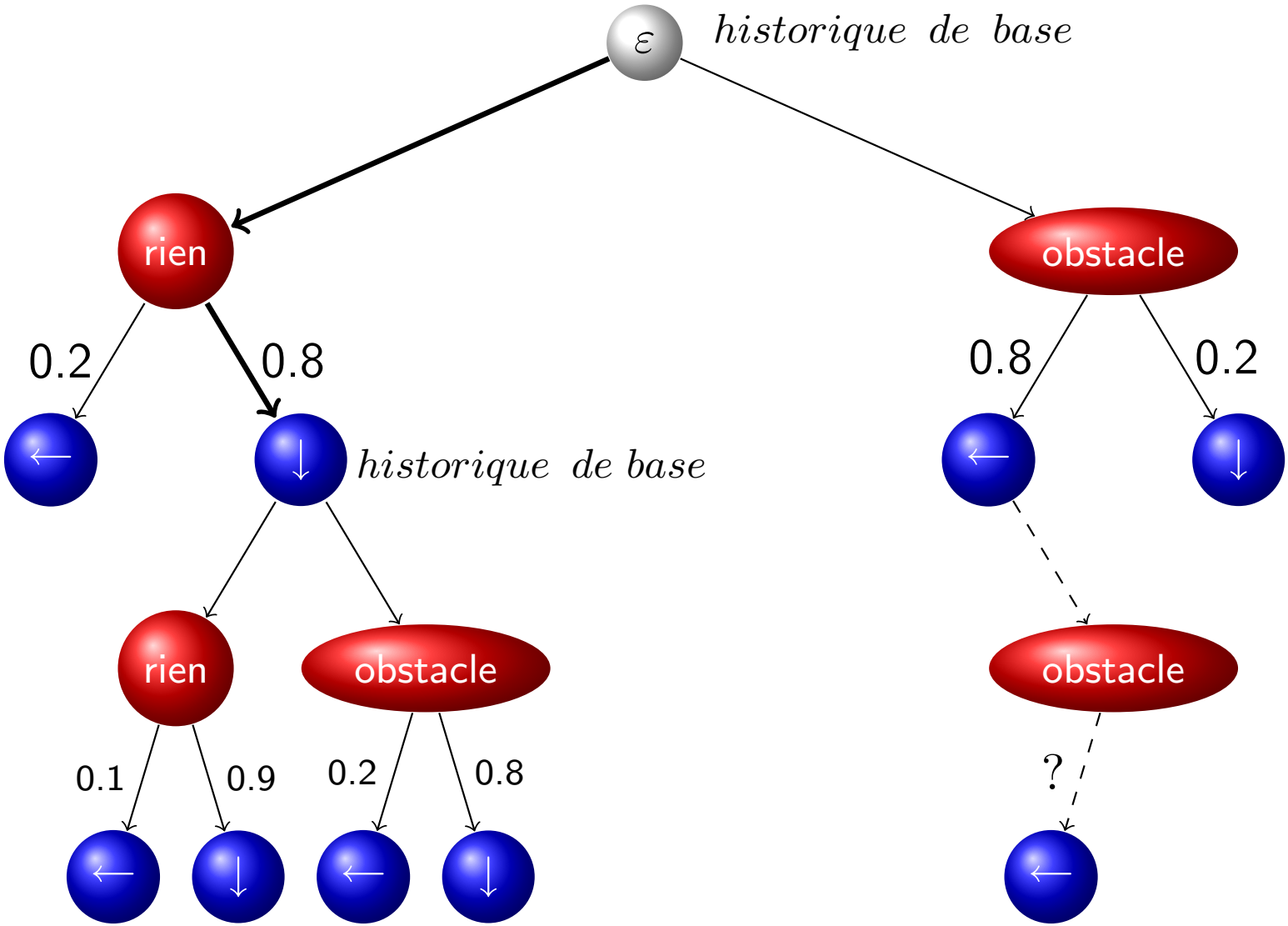
$$\frac{\partial V(\theta)}{\partial \theta_i} = \sum_{t=0}^{\infty} \gamma^t \sum_{a_1 o_1 \ldots a_t o_t a_{t+1}} \underbrace{\prod_{i=0}^{t} Pr(o_i | a_1 o_1 \ldots a_i, \theta) \mathbb{E}(r_t)}_{environnement} \times \underbrace{\frac{\partial \prod_{i=0}^{t} Pr(a_{i+1} | a_1 o_1 \ldots a_i o_i, \theta)}{\partial \theta_i}}_{politique}$$

Le terme $\prod_{i=0}^{t} Pr(o_i | a_1 o_1 \ldots a_i, \theta) \mathbb{E}(r_t)$ peut être estimé à partir des trajectoires en utilisant l'échantillonnage préférentiel (Importance Sampling).

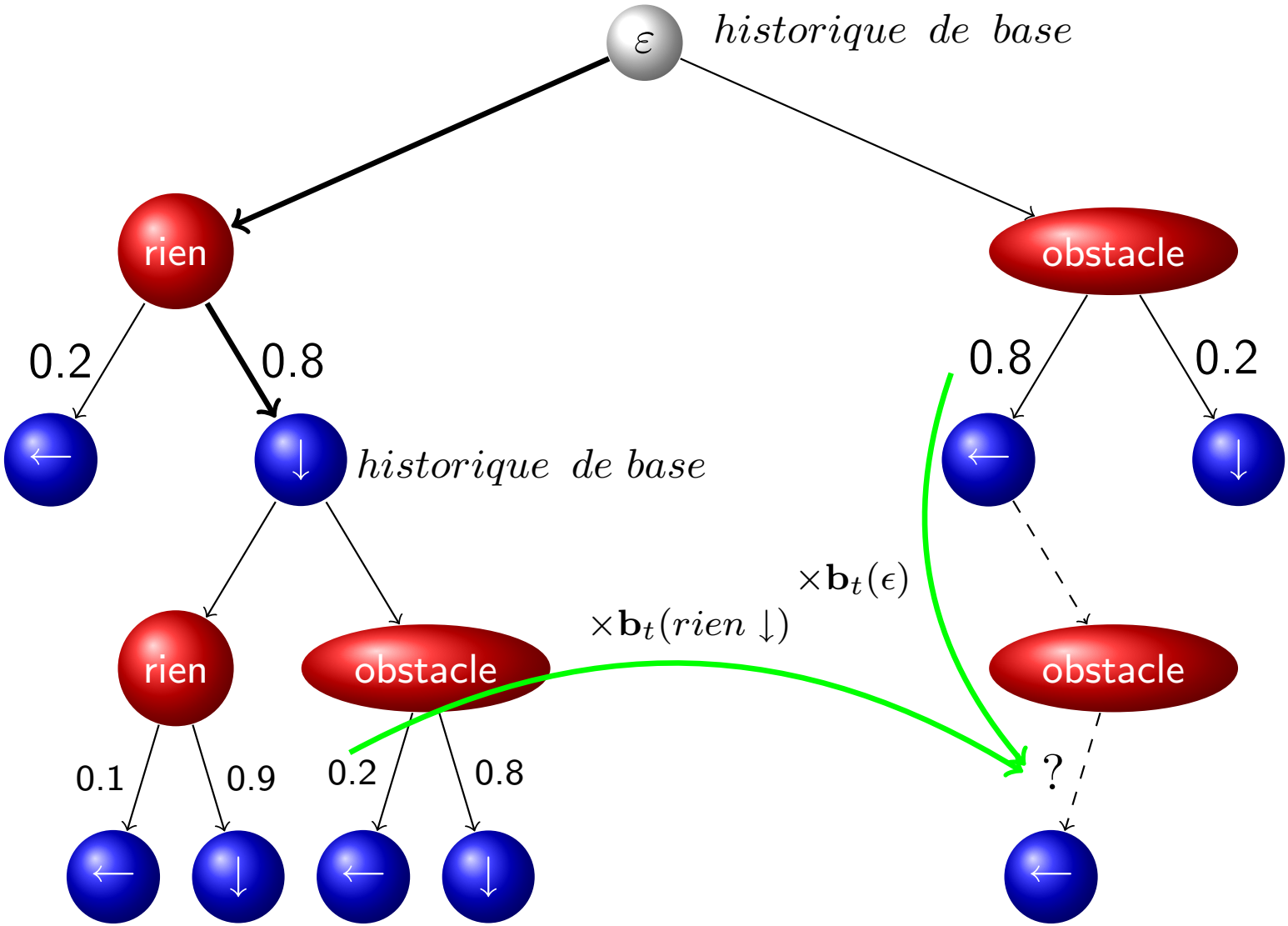
Représentations prédictives des politiques (PPRs)



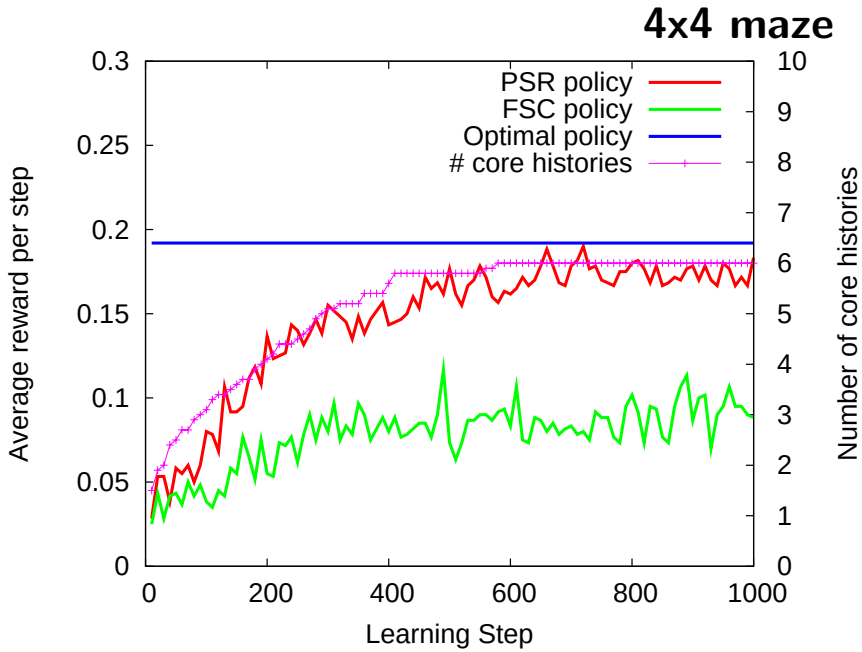
Représentations prédictives des politiques (PPRs)

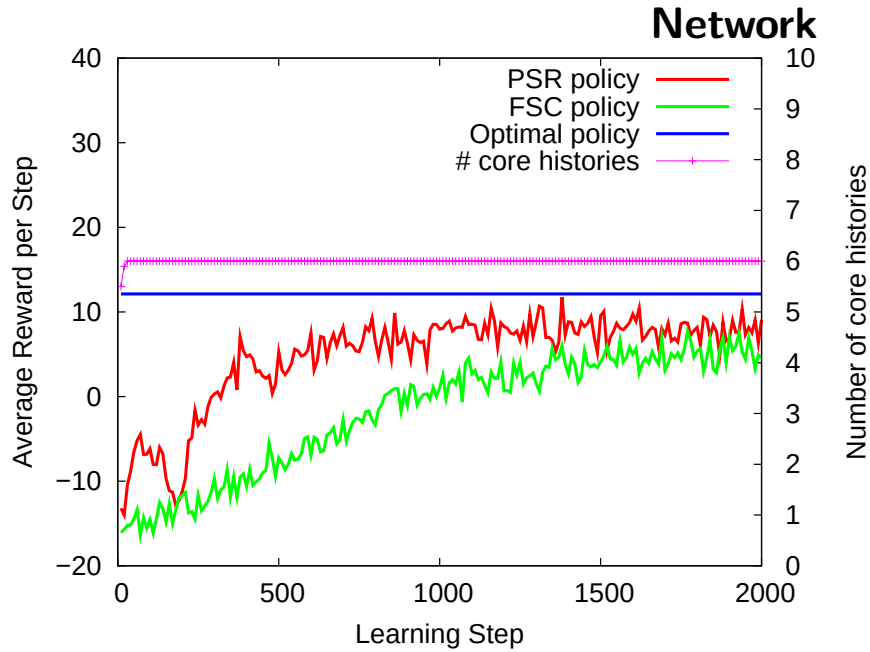


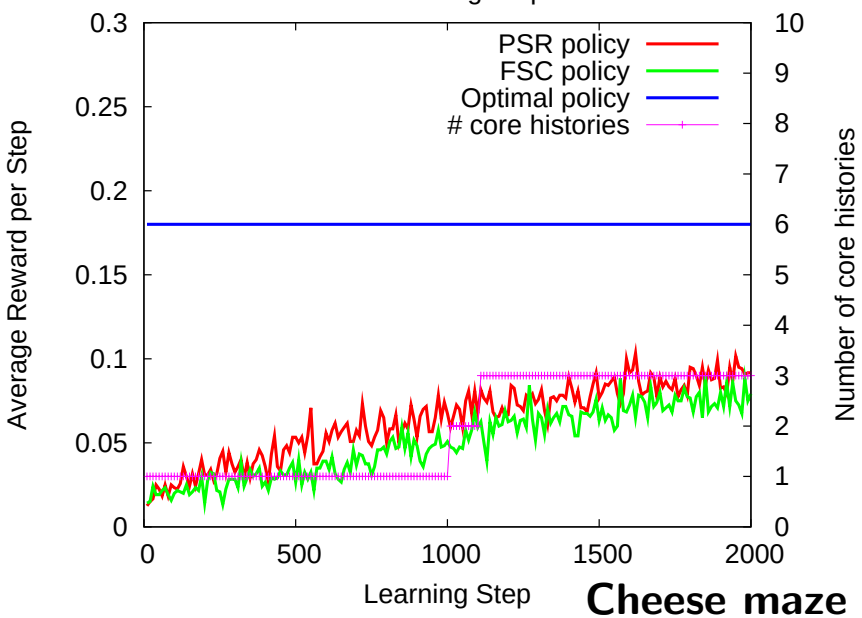
Représentations prédictives des politiques (PPRs)

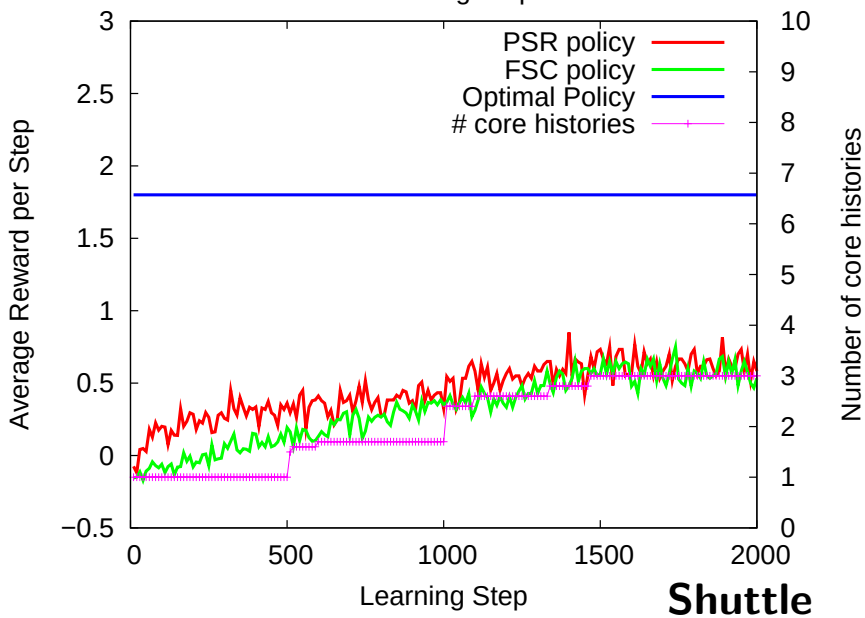


Résultats empiriques [Boularias & Chaib-draa, ICML 2009]

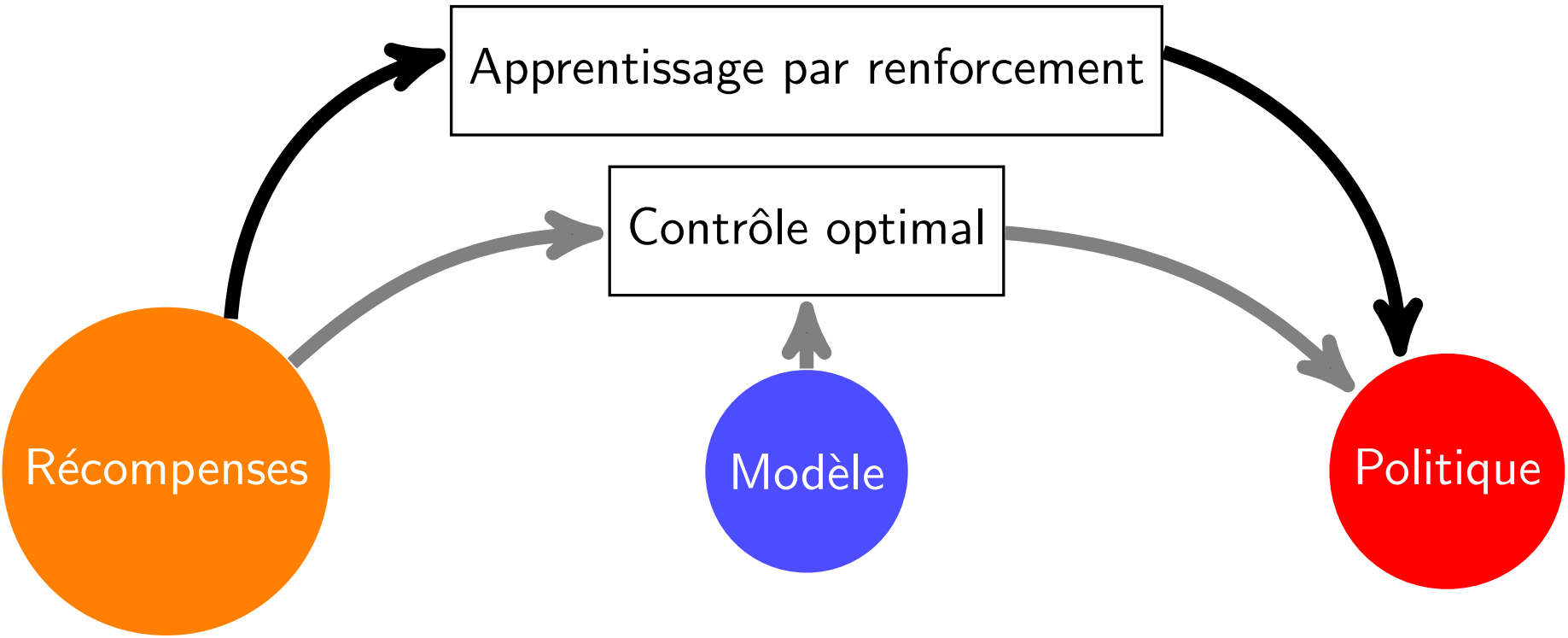




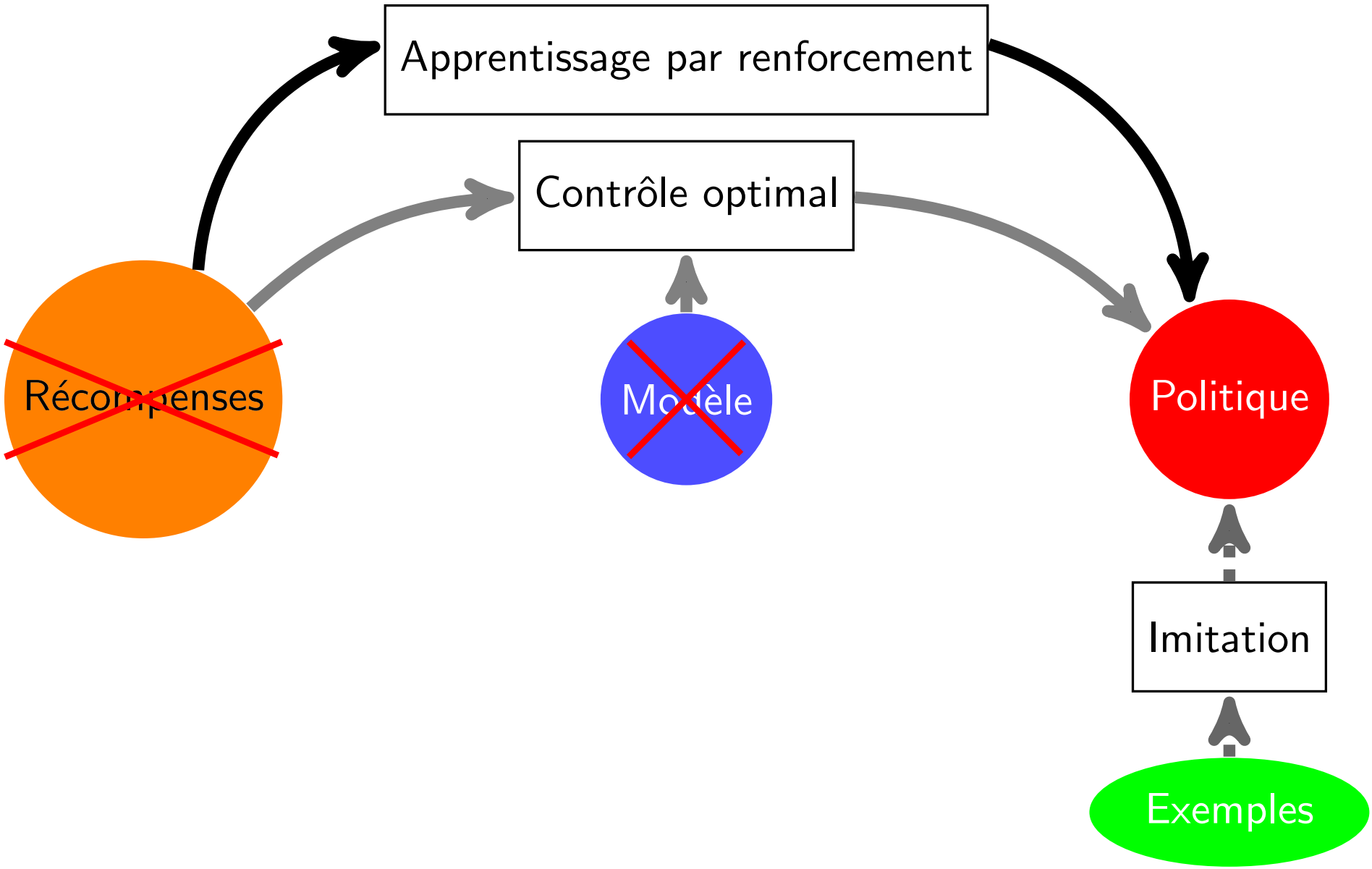




Résoudre un (PO)MDP



Résoudre un (PO)MDP



Apprentissage par imitation direct (Behavioral cloning)

✗ Ce problème a été peu étudié dans le context d'un environnement partiellement observable.

✓ L'expert peut être modélisé comme un système dynamique
$\Rightarrow$ Représentation prédictive de la politique de l'expert.

Apprentissage par imitation avec les représentations prédictives

$$U(q_i|h_j) = Pr(q_i|h_j)$$

$$U = \begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} q_1 & q_2 & q_3 & q_4 & q_5 & q_6 & q_7 \end{array} \\ \left(\begin{array}{ccccccc} 0.2 & 0.5 & 1 & 0 & 0.2 & 0.5 & 1 \\ 0.5 & 0.8 & 0.3 & 0.7 & 0.5 & 0.8 & 0.3 \\ 0.5 & 0.9 & 0.1 & 0.3 & 0.5 & 0.9 & 0.1 \\ 1 & 0 & 1 & 0.2 & 1 & 0 & 1 \\ 0.5 & 0.3 & 0.8 & 0.9 & 0.5 & 0.3 & 0.8 \\ 0.9 & 0.6 & 0.7 & 0.3 & 0.9 & 0.6 & 0.7 \\ 0.9 & 0.9 & 0.5 & 0 & 0.9 & 0.9 & 0.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \right. \end{array} \begin{array}{c} \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \end{array}$$

Apprentissage par imitation avec les représentations prédictives

$$U(q_i|h_j) = Pr(q_i|h_j)$$

$$U = \begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ \vdots \end{array} \begin{pmatrix} q_1 & q_2 & q_3 & q_4 & q_5 & q_6 & q_7 & \\ 0.2 & 0.5 & 1 & 0 & 0.2 & 0.5 & 1 & \dots \\ 0.5 & 0.8 & 0.3 & 0.7 & 0.5 & 0.8 & 0.3 & \dots \\ 0.5 & 0.9 & 0.1 & 0.3 & 0.5 & 0.9 & 0.1 & \dots \\ 1 & 0 & 1 & 0.2 & 1 & 0 & 1 & \dots \\ 0.5 & 0.3 & 0.8 & 0.9 & 0.5 & 0.3 & 0.8 & \dots \\ 0.9 & 0.6 & 0.7 & 0.3 & 0.9 & 0.6 & 0.7 & \dots \\ 0.9 & 0.9 & 0.5 & 0 & 0.9 & 0.9 & 0.5 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

F

Résultats empiriques [Boularias, ICMLA 2008]

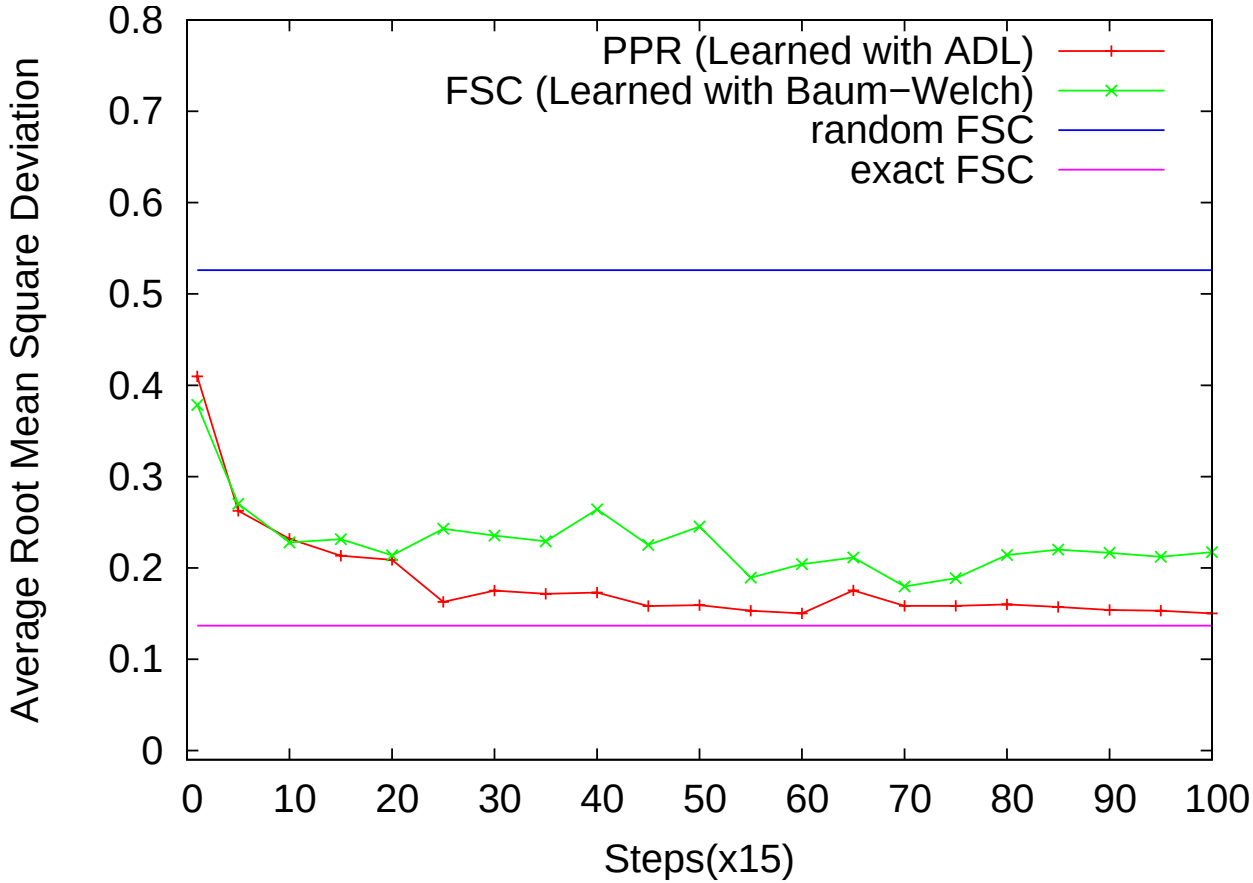


FIG.: Erreur quadratique moyenne de la prédiction des actions de l'expert.

Apprentissage par imitation direct (Behavioral cloning)

- L'optimalité d'une action dans un état donné dépend de :
  1. les caractéristiques de l'état.
  2. les caractéristiques des états voisins.
  3. la dynamique du système.
- L'apprentissage par imitation est un problème de prédiction de sortie structurée (structured output prediction).

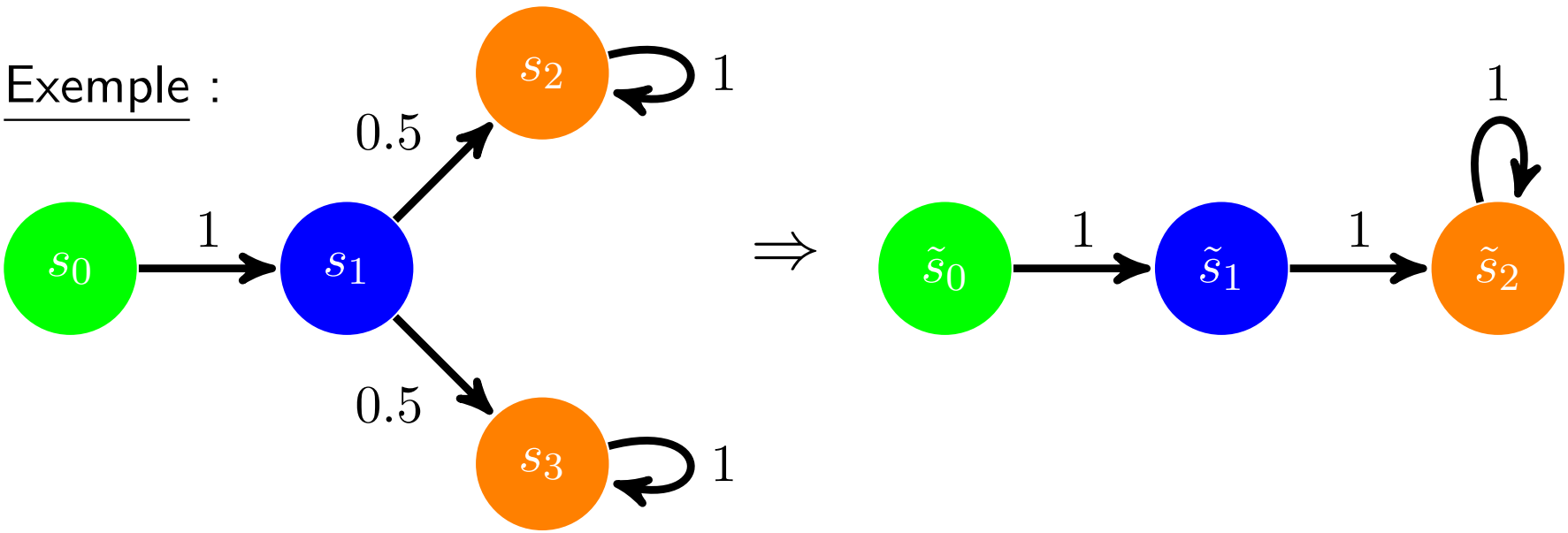
Morphisme des processus décisionnels de Markov [Ravindran, 2004]

Un morphisme entre deux MDPs $M(\mathcal{S}, \mathcal{A}, \{T^a\})$ et $\tilde{M}(\tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \{\tilde{T}^a\})$ est une fonction surjective f définie par :

$$f : \mathcal{S} \to \tilde{\mathcal{S}}$$

$$\tilde{T}^a(f(s), \tilde{s}) = \sum_{s' \in \mathcal{S}, f(s') = \tilde{s}} T^a(s, s')$$

Exemple :



Morphismes souples [Sorg & Singh, 2009]

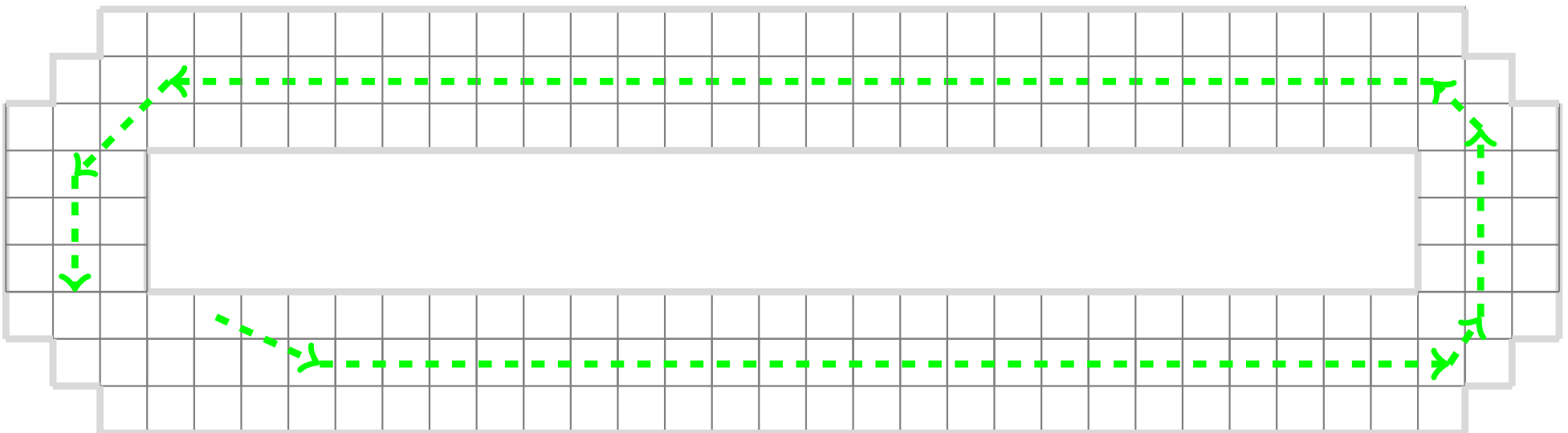
Un morphsime souple entre deux MDPs M et $\tilde{M}$ est une fonction f définie par :

$$f : \mathcal{S} \times \tilde{\mathcal{S}} \rightarrow [0, 1]$$
$$\| \sum_{\tilde{s} \in \tilde{\mathcal{S}}} f(s, \tilde{s}) \tilde{T}^a(\tilde{s}, \tilde{s}') - \sum_{s' \in \mathcal{S}} T^a(s, s') f(s, \tilde{s}') \| \leq \epsilon$$

✓ Le problème de trouver un morphisme souple peut être résolu à l'aide d'un programme linéaire.

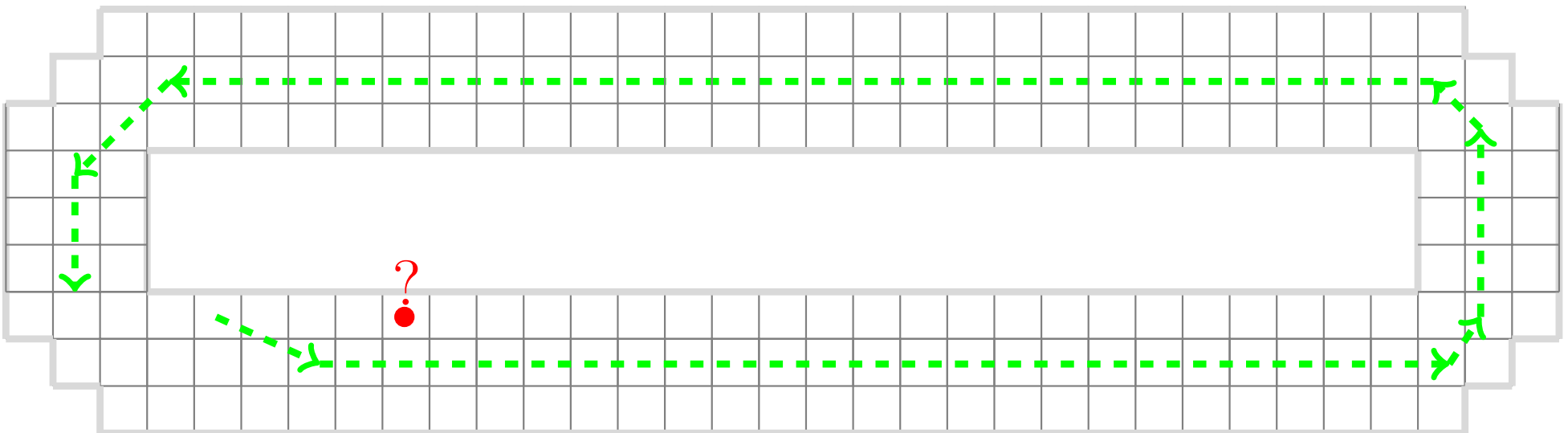
L'exemple du circuit

trajectoire d'un expert

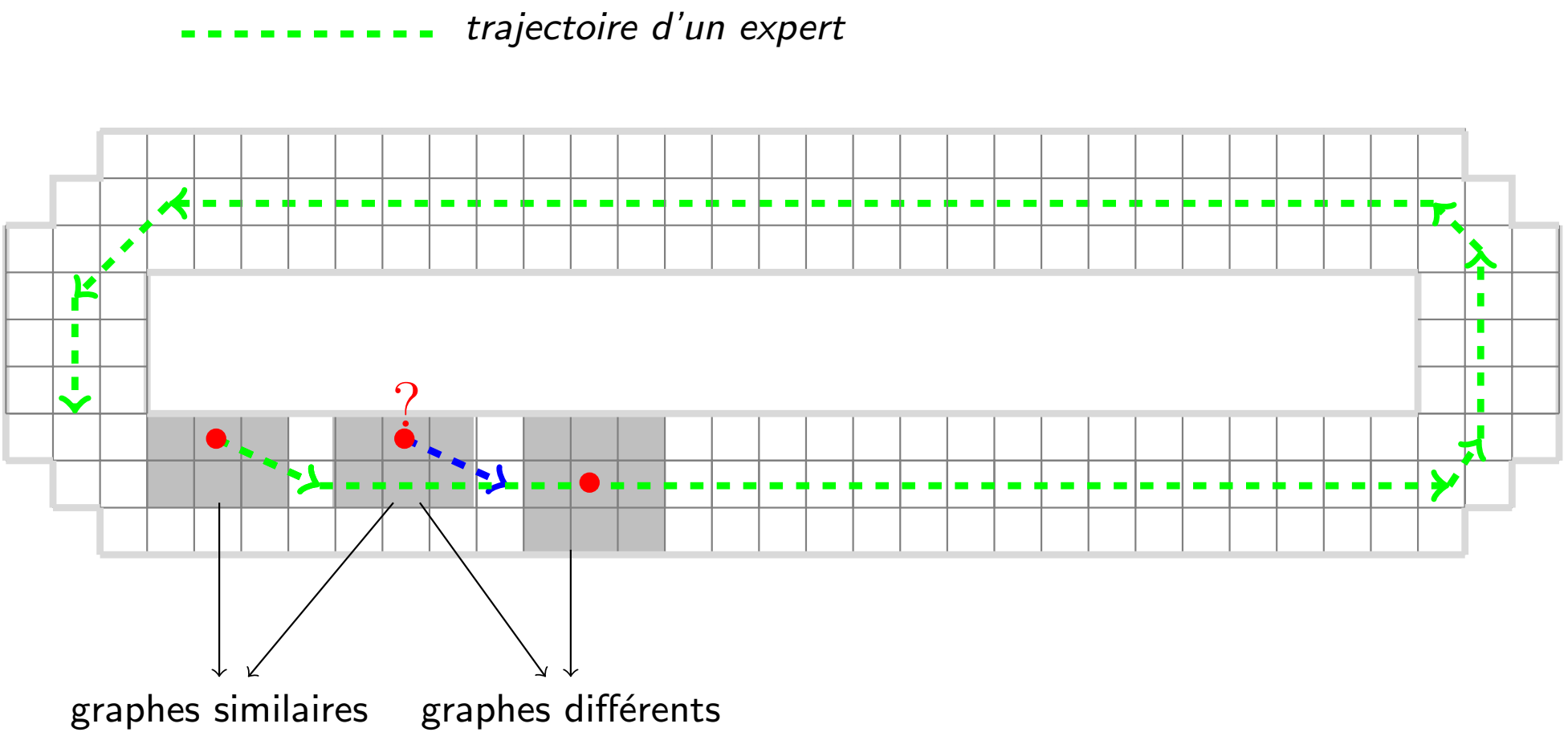


L'exemple du circuit

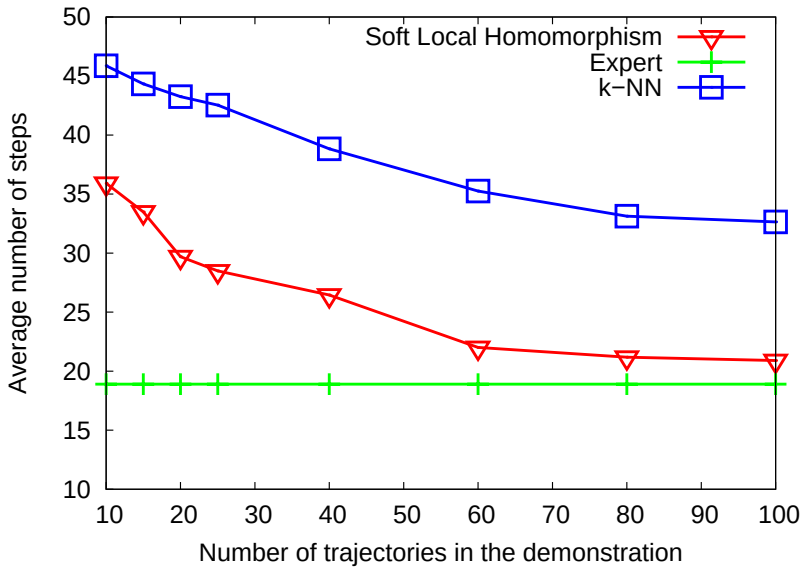
trajectoire d'un expert



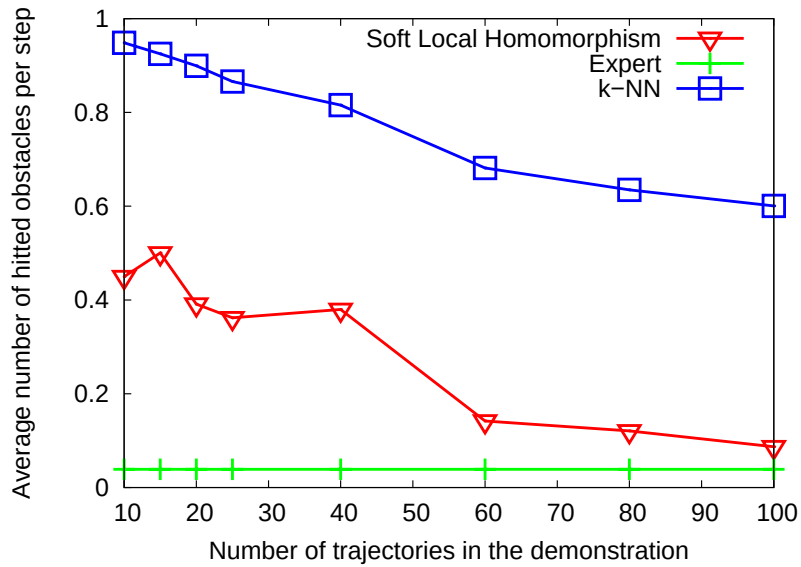
L'exemple du circuit



Résultats empiriques [Boularias & Chaib-draa, ICRA 2010]

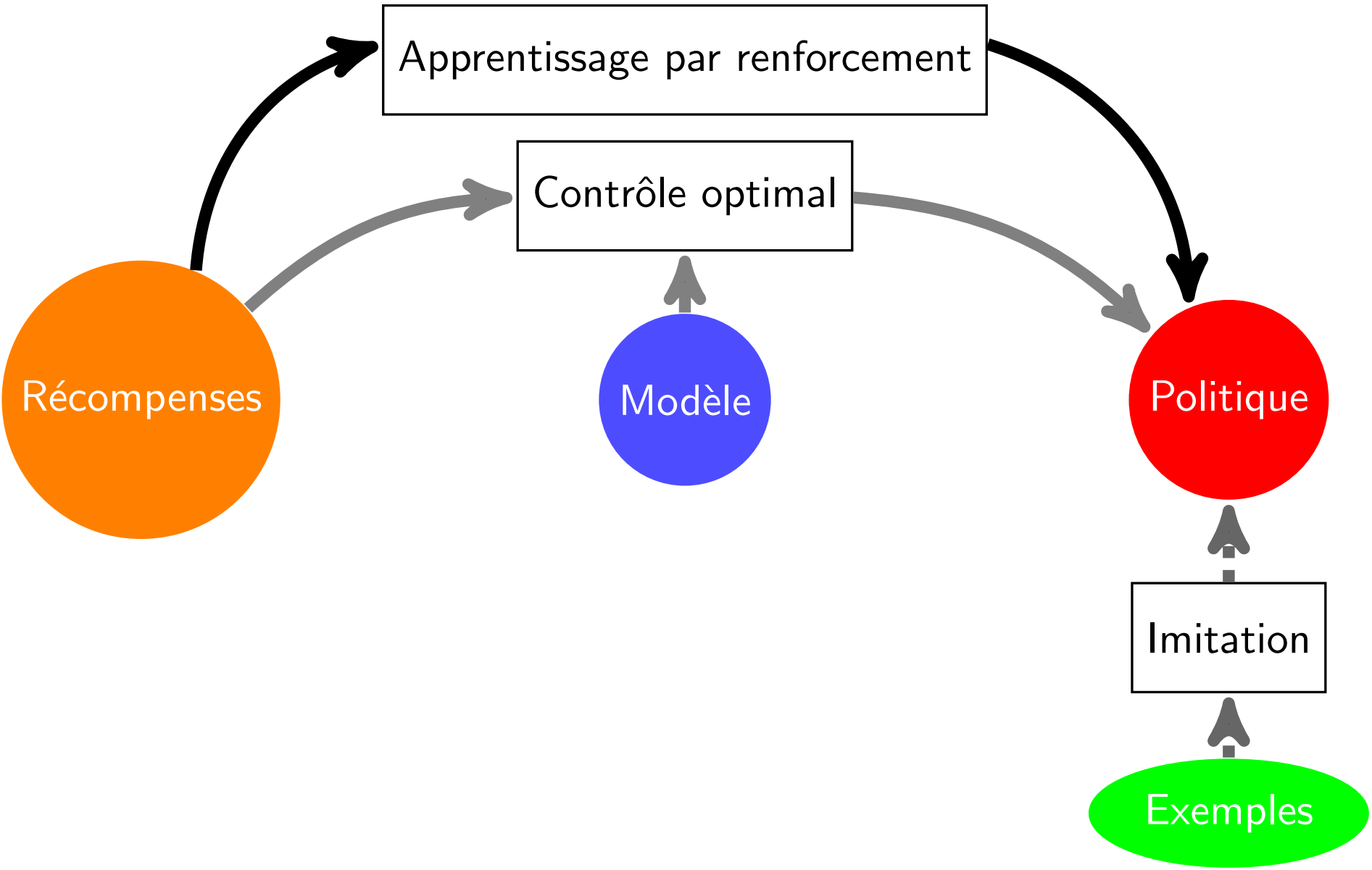


(a) Durée moyenne d'une course

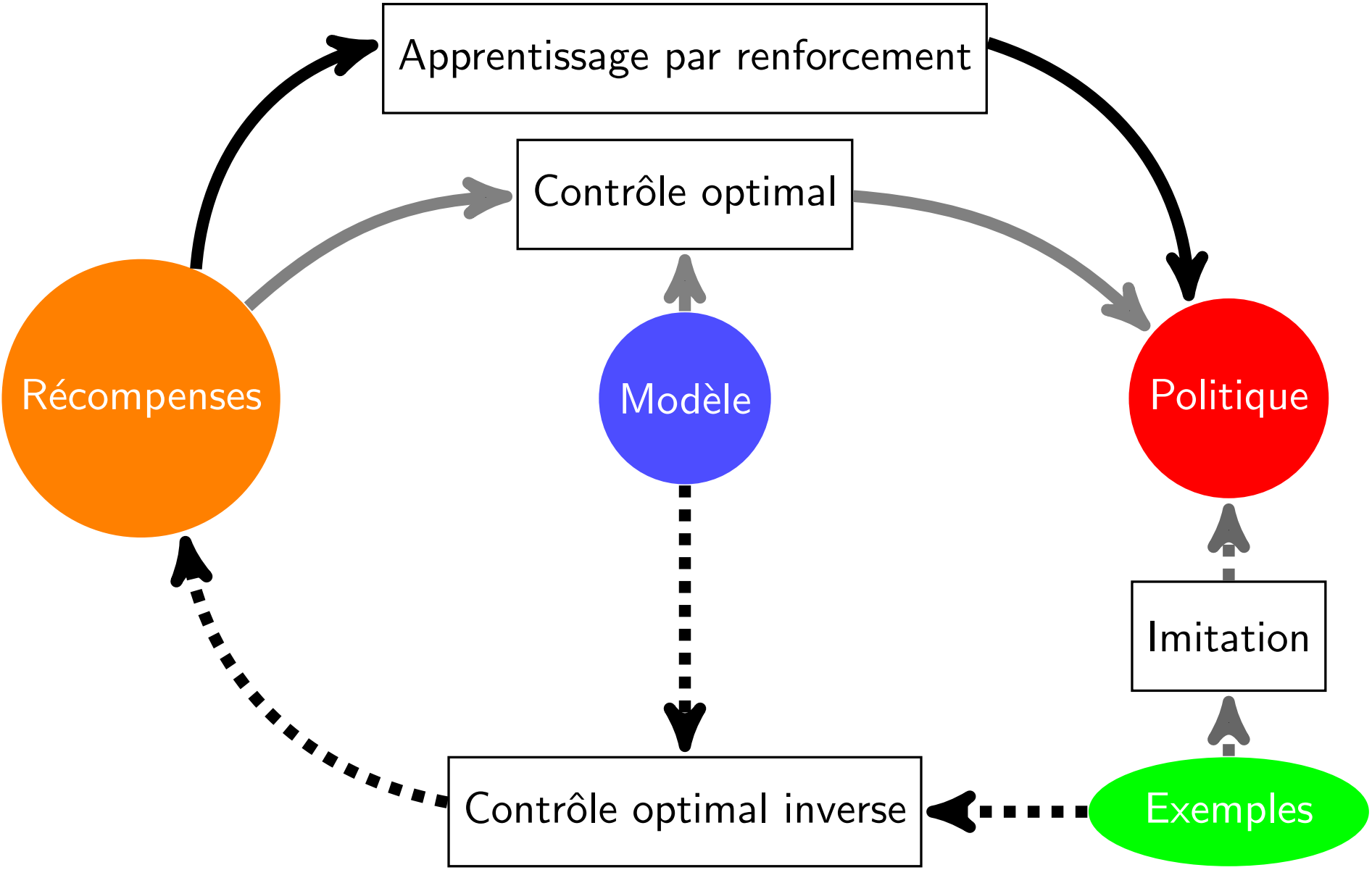


(b) Fréquence des sorties de piste

Résoudre un (PO)MDP



Résoudre un (PO)MDP



Apprentissage par renforcement inverse

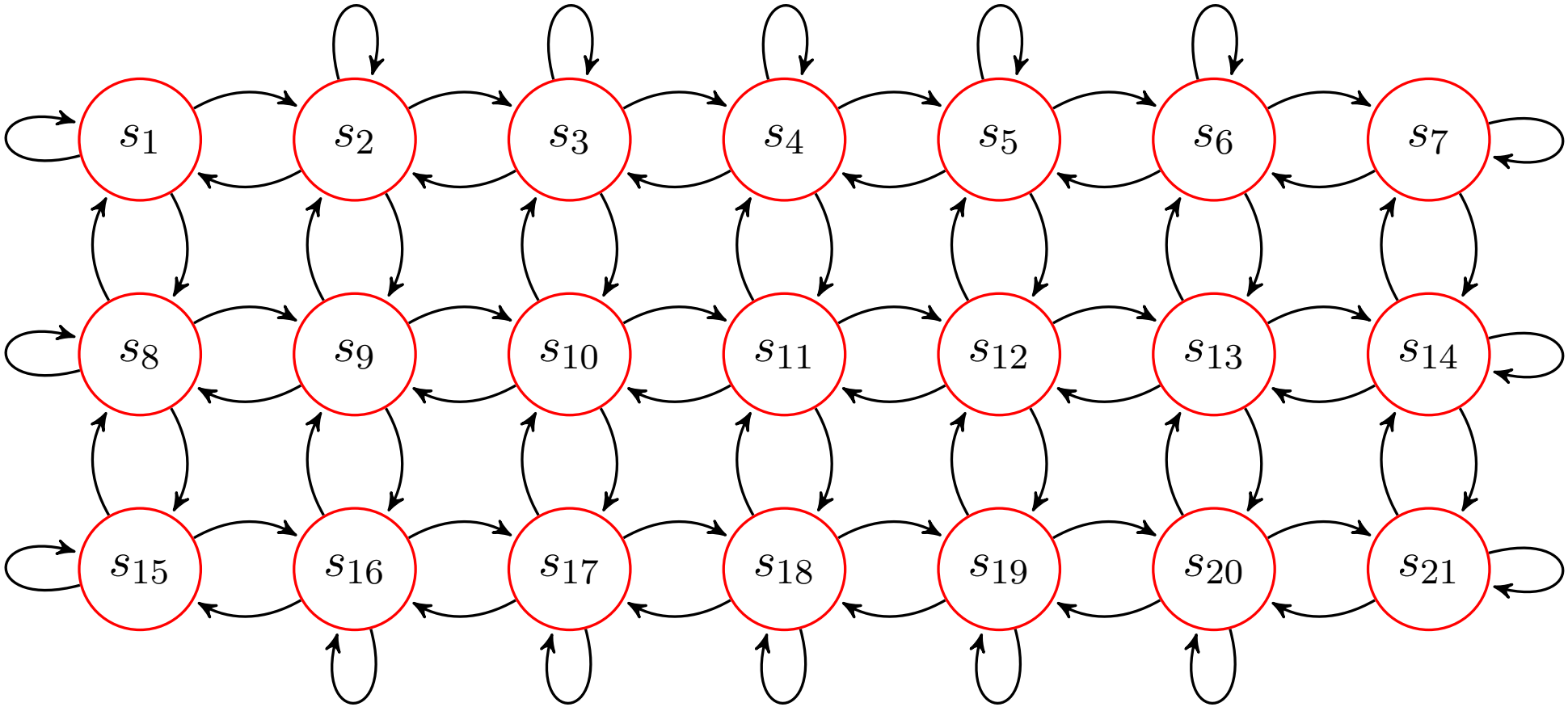
Étant donnés des exemples d'une politique π^* de l'expert, trouver un vecteur de récompenses R tel que :

$$R^T \mu_{\pi^*} \geq \max_{\pi \in \Pi(M)} R^T \mu_\pi$$

Deux problèmes principaux :

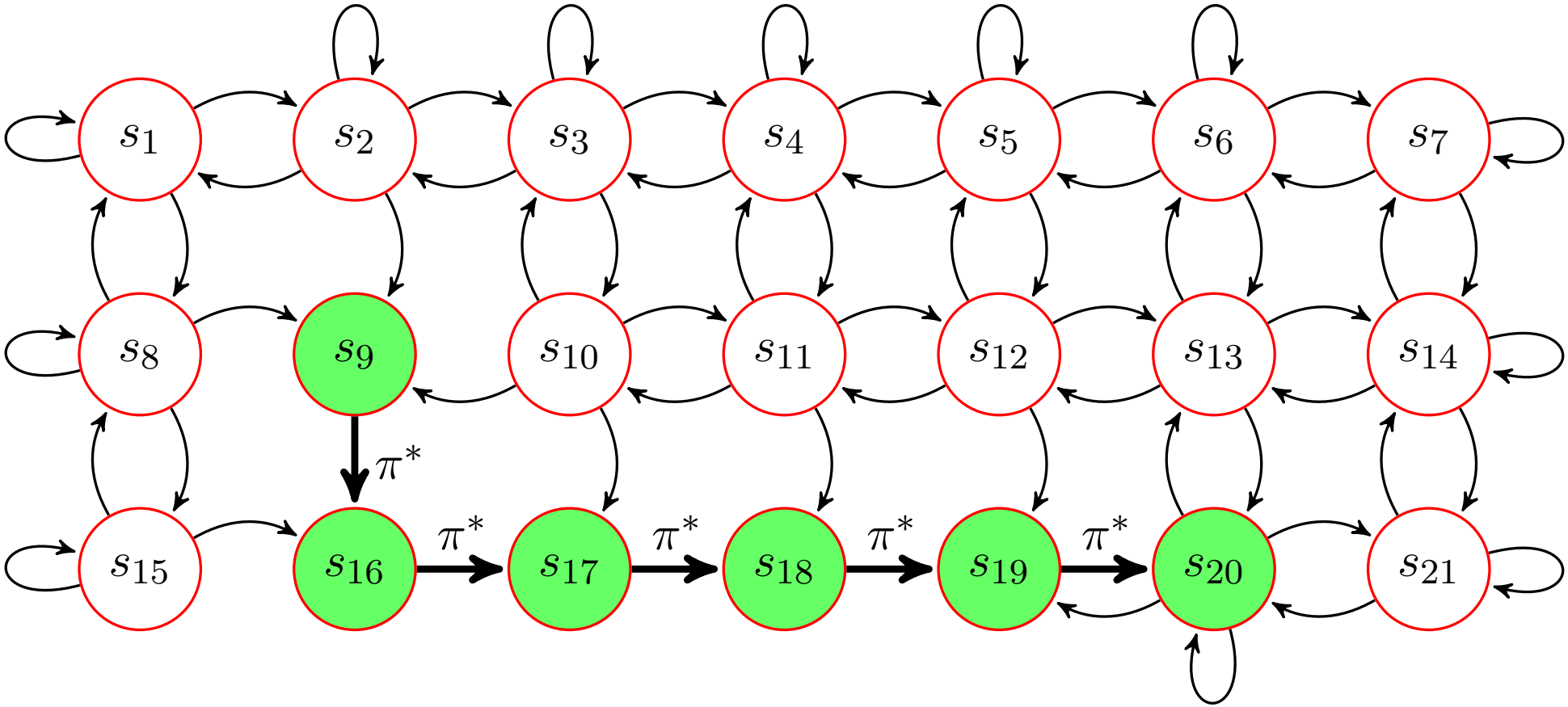
- ✗ L'espace des solutions est infini (par exemple, $R = (0, 0, \ldots, 0)$ est une solution triviale. Cela est un problème de régularisation.
- ✗ La politique de l'expert π^* n'est pas complètement connue, la distribution stationnaire μ_{π^*} ne peut pas être calculée analytiquement.

Bootstrapping des exemples de l'expert



Un processus décisionnel de Markov M.

Bootstrapping des exemples de l'expert



Un processus décisionnel de Markov modifié M'.

$$\max_{\pi \in \Pi(M')} R^T \mu_\pi \geq \max_{\pi \in \Pi(M)} R^T \mu_\pi$$

Résultats empiriques

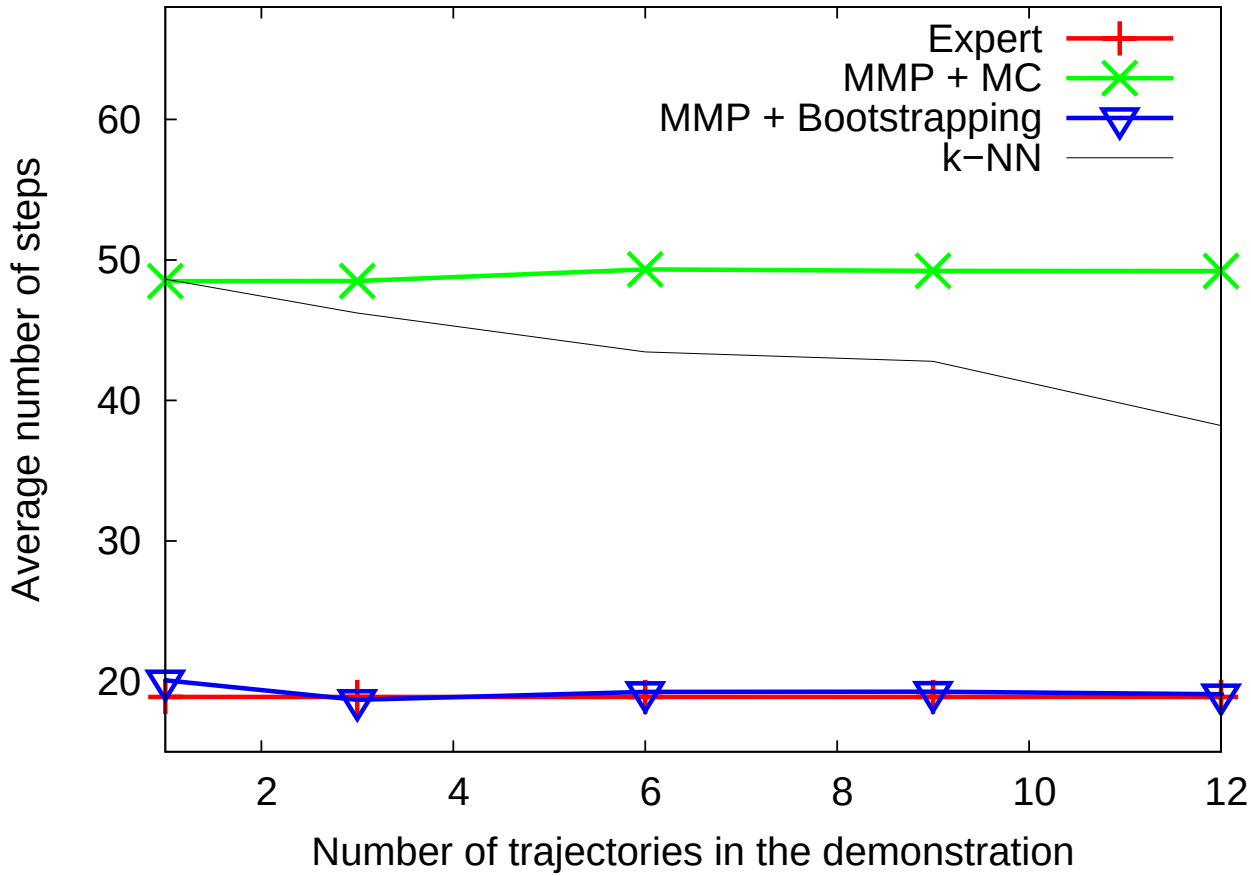


FIG.: Durée moyenne d'une course.

Conclusion

- ✓ Les représentations prédictives des états et des politiques sont une solution prometteuse aux problèmes de contrôle et d'apprentissage dans les systèmes dynamiques.
- ✓ Les représentations prédictives réduisent la complexité des structures définissant les états et les politiques.

- ✗ Les paramètres des représentations prédictives n'ont pas une interpretation intuitive.
- ✗ Les états de croyance soufrent d'une instabilité numérique.
- ✗ L'apprentissage des tests et des historiques de base reste un problème ouvert.

Quelques références :


Boularias, A. (2008).
A Predictive Model for Imitation Learning in Partially Observable Environments.
In *ICMLA'08*, pages 83–90.


Boularias, A., Izadi, M., and Chaib-draa, B. (2008).
Prediction-directed Compression of POMDPs.
In *ICMLA'08*, pages 99–105.


Boularias, A. and Chaib-draa, B. (2008a).
Exact Dynamic Programming for Decentralized POMDPs with Lossless Policy Compression.
In *ICAPS'08*, pages 20–27.


Boularias, A. and Chaib-draa, B. (2009).
Predictive Representations for Policy Gradient in POMDPs.
In *ICML'09*, pages 65–72.


Boularias, A. and Chaib-draa, B. (2010).
Apprenticeship Learning via Soft Local Homomorphisms.
In *ICRA'10*.

Questions ?

Quelques références :


Jaeger, H. (2000).
Observable Operator Models for Discrete Stochastic Time Series.
Neural Computation, 12(6) :1371–1398.


Aberdeen, D., & Baxter, J. (2002).
Scaling Internal-State Policy-Gradient Methods for POMDPs.
Proc. 19th Int. Conf. Machine Learning (pp. 3–10).


Littman, M., Sutton, R., & Singh, S. (2002).
Predictive Representations of State.
Advances in Neural Information Processing Systems 14 (pp. 1555–1561).


Peters, J. and Schaal, S. (2008).
Natural Actor-Critic.
Neurocomputing, 71 (pp. 1180–1190).


Sutton, R. S., Mcallester, D., Singh, S., & Mansour, Y. (2000).
Policy Gradient Methods for Reinforcement Learning with Function Approximation.
In Advances in Neural Information Processing Systems 12 (NIPS'00) (pp. 1057–1063).

Automates finis stochastiques (FSM)

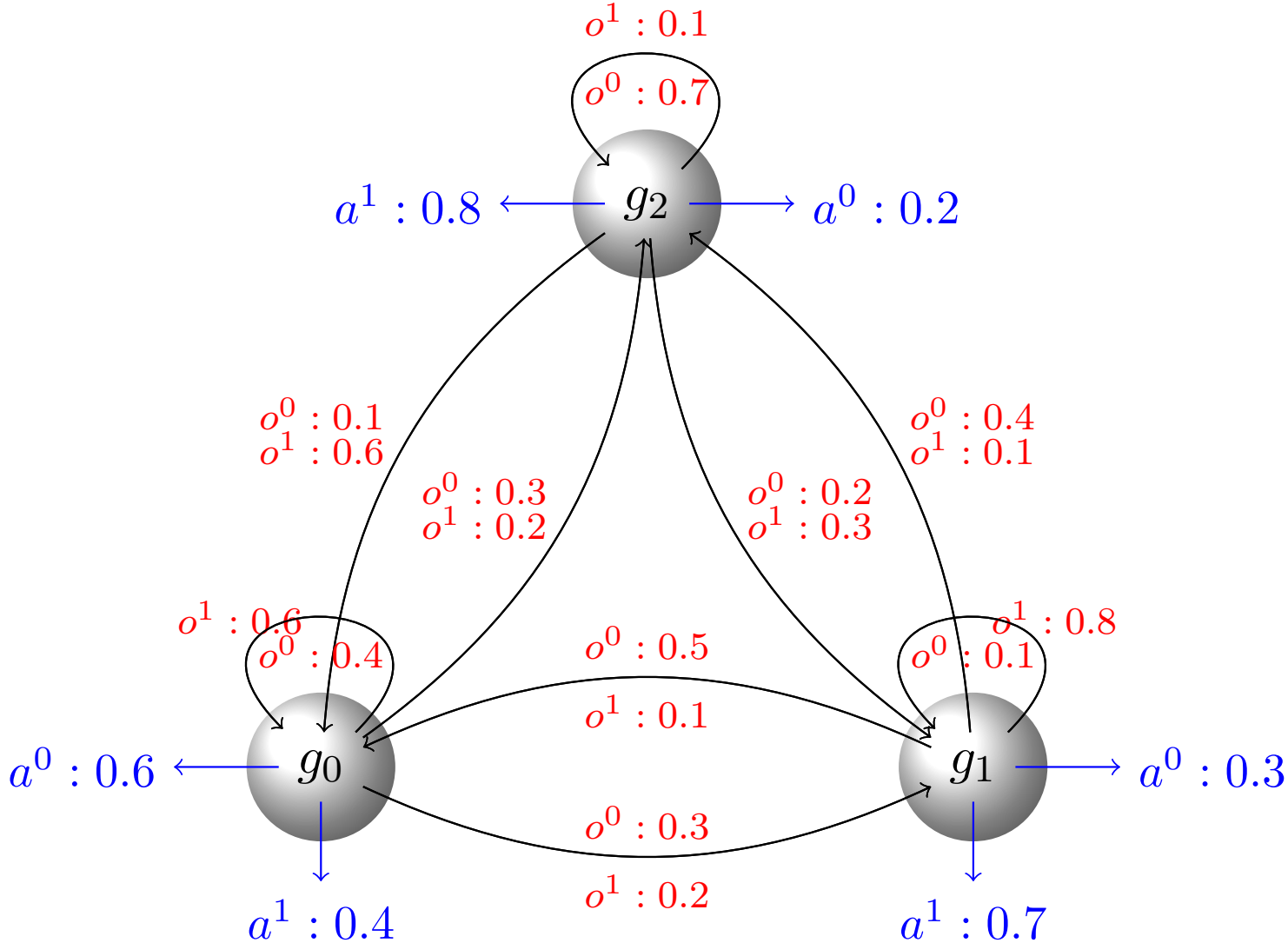


FIG.: Un automate fini stochastique.

Apprentissage par imitation direct (Behavioral cloning)

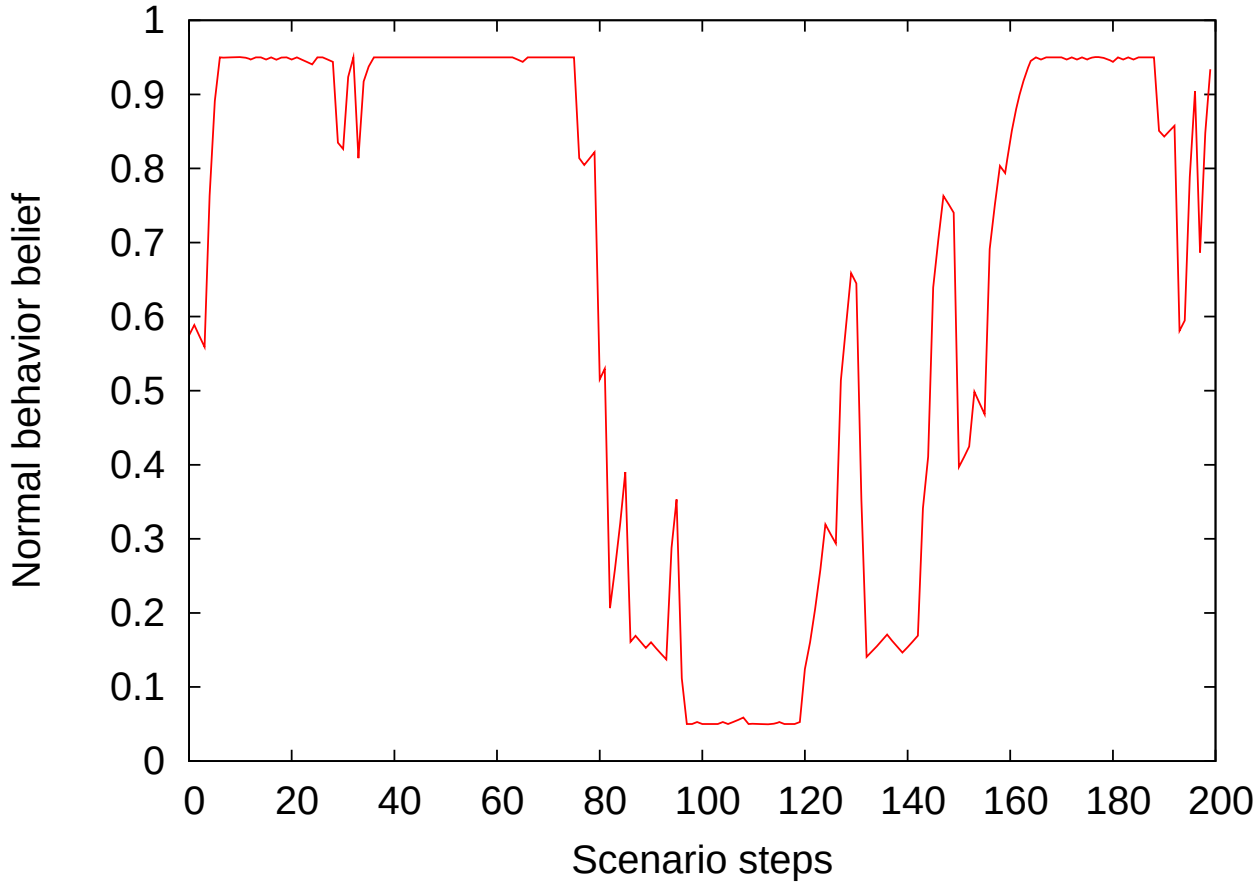
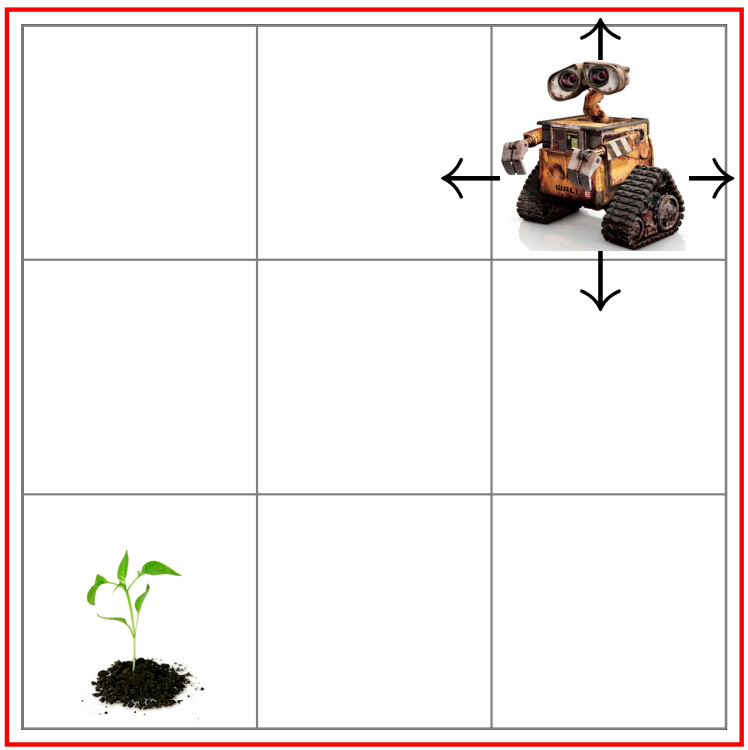


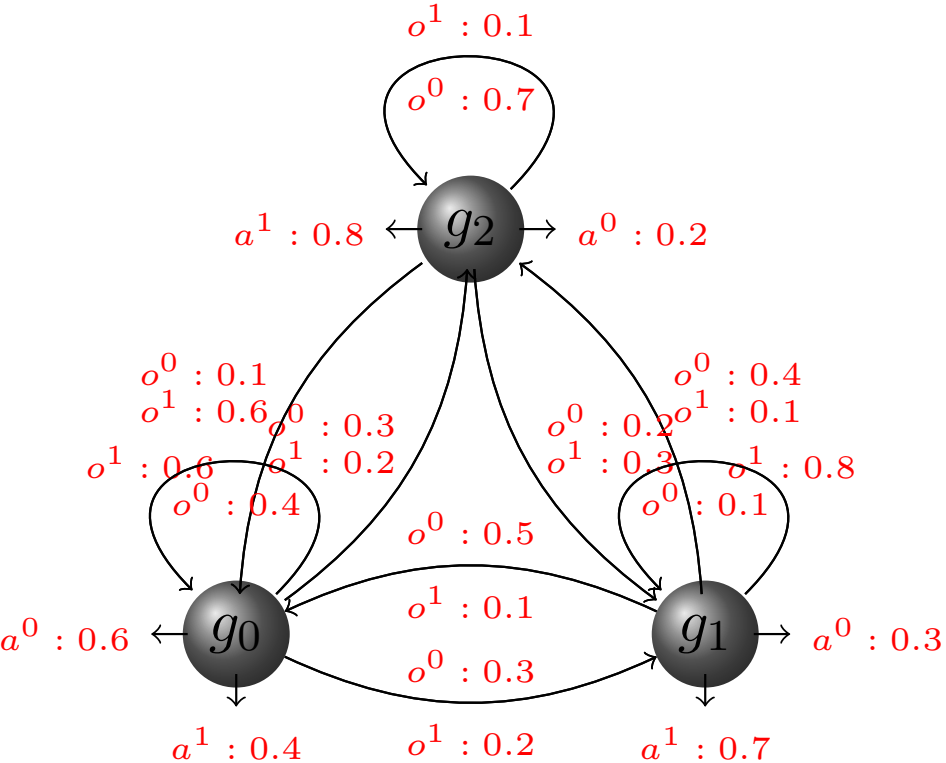
FIG.: Détection d'une déviation du comportement normal.

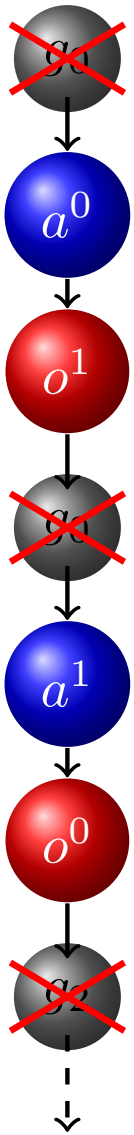
Predictive State Representations (PSR)- a simple example



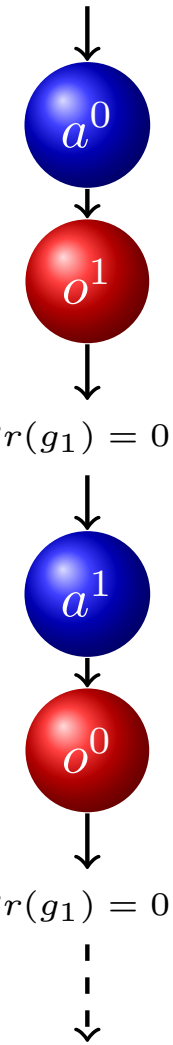
- A POMDP belief state is a probability distribution over the 9 hidden states.
- A PSR belief state is the probabilities of the tests :
  1. $\text{action} : \uparrow, \text{observation} : \text{wall}$
  2. $\text{action} : \downarrow, \text{observation} : \text{wall}$
  3. $\text{action} : \leftarrow, \text{observation} : \text{wall}$
  4. $\text{action} : \rightarrow, \text{observation} : \text{wall}$
- The current belief state of the robot is :
  1. $Pr(\text{wall}|\uparrow) = 1$
  2. $Pr(\text{wall}|\downarrow) = 0$
  3. $Pr(\text{wall}|\leftarrow) = 0$
  4. $Pr(\text{wall}|\rightarrow) = 1$

FSC with Internal Belief States [Aberdeen & Baxter, 2002]





$Pr(g_0) = 1, Pr(g_1) = 0, Pr(g_2) = 0$

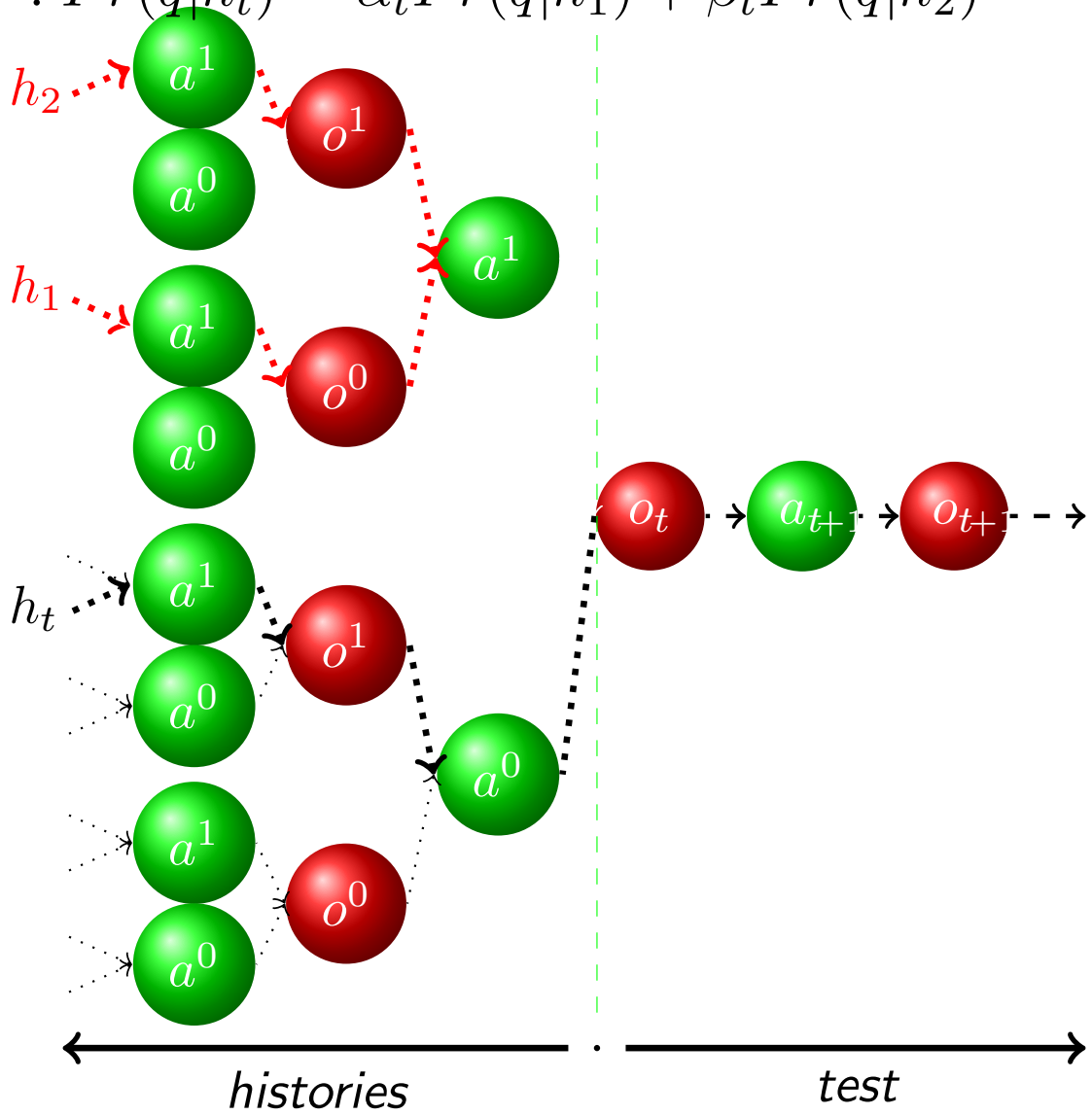


$Pr(g_0) = 0.6, Pr(g_1) = 0.2, Pr(g_2) = 0.2$

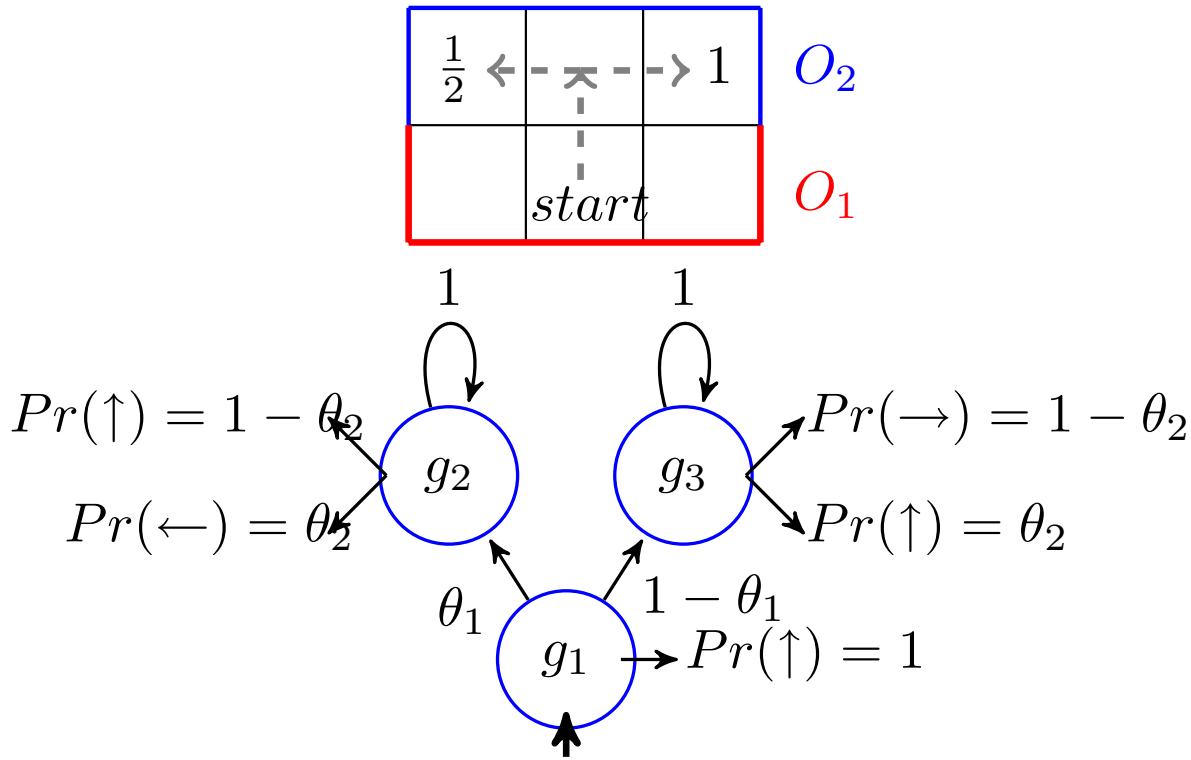
$Pr(g_0) = 0.36, Pr(g_1) = 0.24, Pr(g_2) = 0.4$

Predictive State Representations (PSRs) - core histories

$$\forall q \in \{\mathcal{A} \times \mathcal{O}\}^* : Pr(q|h_t) = \alpha_t Pr(q|h_1) + \beta_t Pr(q|h_2)$$

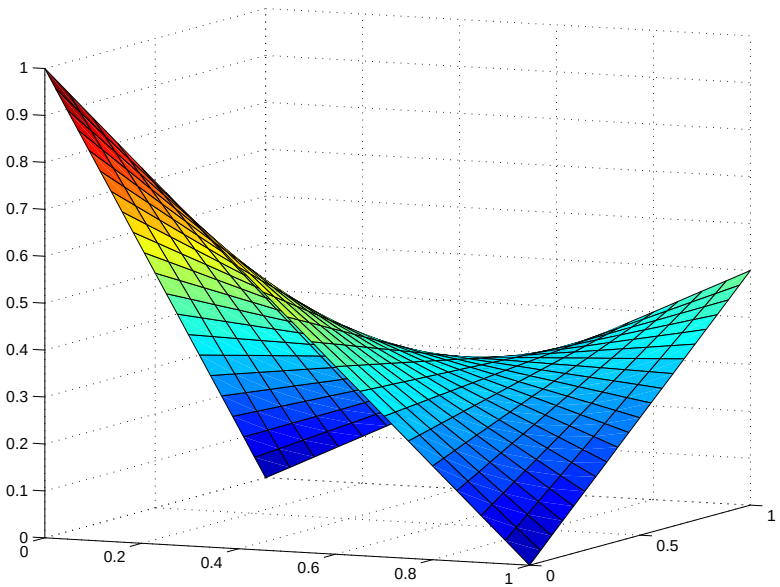


Small example

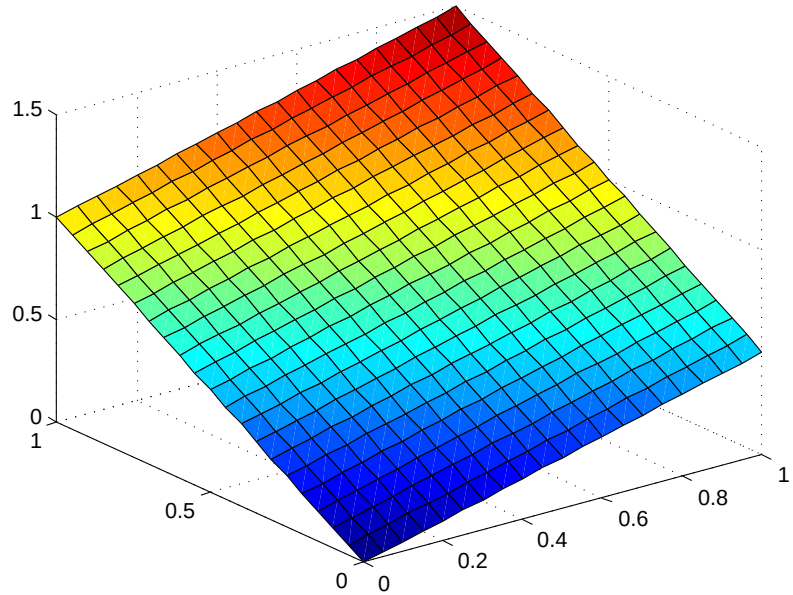


- The equivalent PSR representation has one core history $h_0 = \emptyset$, therefore $b_t(\emptyset) = 1, \forall t \geq 0$.
- The parameters of the PSR representation are : $Pr(\leftarrow |O_2, \emptyset) \stackrel{def}{=} \theta_1', Pr(\rightarrow |O_2, \emptyset) \stackrel{def}{=} \theta_2', Pr(\uparrow |O_2, \emptyset) \stackrel{def}{=} \theta_3'$

Example



$$V(\pi_\theta^{FSC}) = \frac{1}{2}\theta_1\theta_2 + (1-\theta_2)(1-\theta_1)$$



$$V(\pi_{\theta'}^{PSR}) = \frac{1}{2}{\theta_1}' + {\theta_2}'$$
$$\text{s.t. } {\theta_1}' + {\theta_2}' \leq 1$$

The value functions of an FSC and its equivalent PSR policy.

Predictive Reward Representation

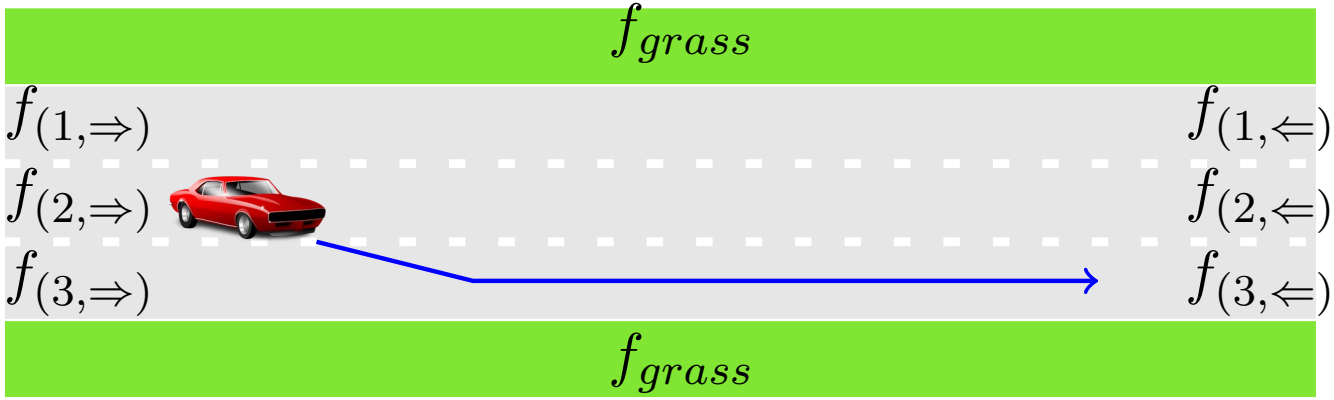


FIG.: A car driving simulation and the corresponding reward features.

- Actions : steer right, steer left, accelerate, decelerate, nothing.
- A high reward is given for driving on the right side of the road.
- However, the feature $f_{(1,\Leftarrow)}$ does not appear in the demonstration !
- This problem can be solved by adding the feature $\tilde{f}_{(steer\,right,\ f_{grass})}$

Predictive Reward Representation

The set of extended features is denoted by $\tilde{F}$.

1. $\tilde{F}^0$ is initialized with linearly independent base features F ;
2. For $t > 0$, $\tilde{F}^{t+1} = \tilde{F}^t \cup \{T^a \tilde{f} | a \in \mathcal{A}, \tilde{f} \in \tilde{F}^t\}$;
3. At the end of each iteration t, only independent vectors are kept in $\tilde{F}^t$;

This process stops when $\tilde{F}^{t+1} = \tilde{F}^t$.

Results 2

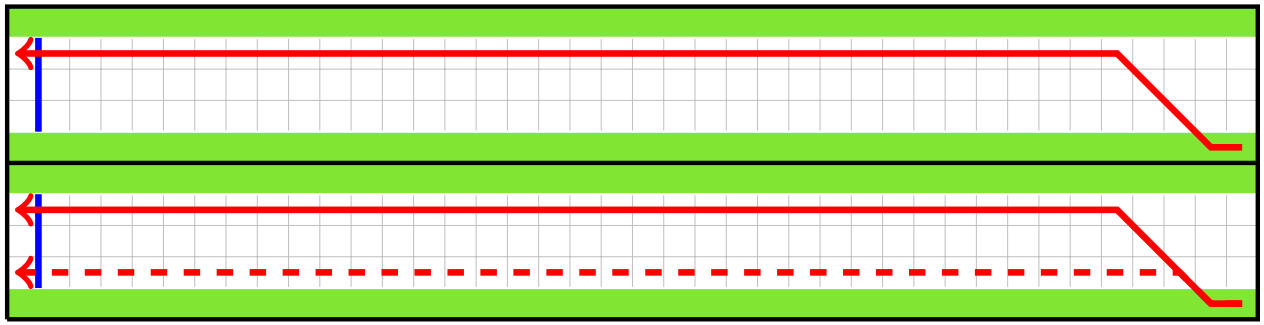


FIG.: The trajectories of the expert cover only the upper half of the road. In the lower half, the policy learned using extended features was able to keep the car on the right side, contrary to the policy learned using base features (dashed line).

- How to determine if a feature is important.
- Computational cost of searching for latent features.
- Other types of latent features (using images as input).
- Nonparametric bayesian method for infinite latent features.