Une règle de conditionnement probabiliste pour le contrôle de robots autonomes

A. Revel & P. Gaussier & C. Joulain, ENSEA ETIS
6 Av du Ponceau, 95014 Cergy Pontoise Cedex
e-mail: gaussier or revel@ensea.fr

1 Introduction

Notre but est de construire une règle d'apprentissage neuronale basée sur des mécanismes de conditionnement complexe, qui permette à un robot autonome réel de "vivre" dans un labyrinthe tel que celui représenté figure 1. Afin de parvenir à trouver une récompense cachée à un endroit dans le labyrinthe, le robot doit apprendre à associer les "bons" mouvements à la reconnaissance de repères visuels. La difficulté de l'apprentissage réside dans le fait que l'ensemble des choix d'associations sensori-motrices effectuées au cours du parcourt n'est renforcé qu'à la sortie du labyrinthe.

De nombreuses techniques de renforcement, telles que le Q-learning par exemple, proposent un cadre théorique à ce type de problèmes et s'appuient sur la décomposition des situations en états distincts. Or la reconnaissance de ces états peut être difficile, voire impossible, étant données les imprécisions inhérentes aux systèmes réels. De plus, différencier les états rend difficile toute tentative de généralisation. Par ailleurs, les justifications biologiques du fonctionnement de ces algorithmes semblent improbables.

Nous avons donc tenté de nous inspirer des mécanismes de conditionnement instrumentaux bien connus des sciences cognitives. Malheureusement, les modèles simulés du conditionnement instrumental, tel que celui proposé par Barto et Sutton, ne permettent pas de gérer les conditionnements impliquant un long délai entre la stimulation et la récompense. Par ailleurs, les implémentations neuronales de ces modèles nécessitent l'ajout de bruit en sortie des neurones afin de permettre de trouver d'autres solutions en cas d'échec. Ce mécanisme peut mener à des situations instables, surtout s'il existe, dans le labyrinthe, une grande disparité entre la fréquence d'apparition d'une situation et d'une autre. Par exemple, dans un labyrinthe, la situation "couloir" est bien plus fréquente que la situation "intersection". Si le robot n'arrive pas à trouver le bon jeu d'associations, il augmentera le niveau du bruit de manière à générer plus de diversité. Cependant, les chances que l'augmentation de ce bruit influe sur le comportement du robot dans un couloir sont beaucoup plus grandes que les chances qu'il intervienne judicieusement pour trouver le bon mouvement à effectuer quand le robot est dans une intersection.

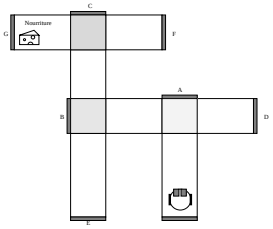

Figure 1: Un exemple de labyrinthe comportant 4 motifs différents (un par intersection et un pour le couloir).

Pour résoudre tous ces problèmes, nous avons donc imaginé une nouvelle règle de conditionnement qui permet au robot de se comporter comme s'il élaborait des hypothèses (ce

que font les animaux ou les hommes [3]). Le principe est basé sur l'utilisation de poids binaires associés à une probabilité entre les neurones dédiés à la reconnaissance et ceux commandant les actions. Cette probabilité peut aussi être considérée comme un facteur de bruit qui serait ramené au niveau des entrées du neurone plutôt qu'en sortie.

2 Algorithme PCR

Le principal problème du conditionnement avec récompense retardée est la difficulté pour le robot de mesurer l'efficacité de son comportement (séquence de perceptions-actions). Il faut donc que ce comportement reste stable pendant un laps de temps suffisament long. Une première solution serait de sélectionner une configuration de poids, de la tester durant un temps donné, de changer de nouveau les poids et de choisir la meilleure configuration. C'est ce que font, par exemple, les algorithmes de programmation dynamique, les méthodes de recuit simulé, le Q-Learning [4] et les algorithmes génétiques.

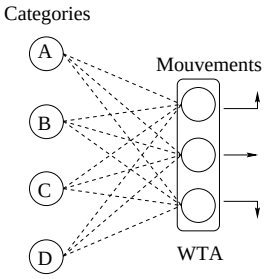

Figure 2: Schéma du réseau de neurones utilisé pour réaliser les associations sensori-motrices

Notre solution offre plus de souplesse. Nous considérerons que l'entrée du réseau de neurones est constituée par un ensemble de neurones qui codent les différentes situations pouvant être rencontrées dans le labyrinthe. La sortie du réseau est formée de neurones associés aux différentes actions qui peuvent être effectuées. Les deux groupes de neurones sont entièrement connectés. Pour simplifier et étant donné que la sortie sur les neurones commandant les actions est en tout ou rien, la valeur des poids est binaire. Le schéma correspondant à un ensemble de 4 catégories associées à 3 actions est donné figure 2.

Afin de pouvoir changer d'hypothèses, il faut savoir quel crédit accorder à chaque poids. Pour cela, on introduit une probabilité ($p \in [0,1]$) qui détermine le degré de confiance dans la valeur du poids. Lorsqu'une variation du signal de renforcement intervient, elle remet en cause les termes de confiance, mais pas forcement les poids binaires. Ceux-ci ne sont modifiés effectivement qu'après réussite d'un tirage aléatoire dépendant du niveau de confiance accordé aux poids.

De manière à pouvoir modifier le terme de certitude, nous devons stocker une mesure de la corrélation entrée-sortie du neurone. Ce terme est mis à jour à chaque itération. Trois variables à intégration temporelle (voir équation 1 pour le calcul) sont associées respectivement à l'entrée I_i, la sortie O_j et le produit entrée-sortie $input \cdot output$. Les notations sont les suivantes : $\overline{I}$, $\overline{O}$ et $\overline{IO}$. On considère que $I_i \in [0,1]$. La mesure de corrélation est donnée par $\mathcal{C} = \frac{\overline{IO}}{\sqrt{\overline{I}.\overline{O}}}$.

Le groupe associé aux mouvements forme un Winner Take All (WTA). L'activité d'un neurone j du WTA avant compétition est donnée par :

$$Act_j = \sum W_{ij} \cdot I_i + bruit$$

Le bruit est une valeur aléatoire très faible servant à lever les ambiguïtés lorsque deux neurones ont la même activité. Aprés compétition, la sortie O_j est donnée par :

$$O_j = \begin{cases} 1 \text{ si } Act_j = \max_k^n Act_k \\ 0 \text{ sinon} \end{cases}$$

Au début, la valeur des poids est tirée aléatoirement (0 ou 1) et une certitude de $\frac{1}{2}$ leur est associée (équiprobabilité des connexions). Le système commence donc par tester des associations entrée-sortie au hasard. Si l'association est "bonne", un signal de renforcement positif intervient et augmente à son tour la certitude dans les poids considérés. Le système devient plus "sûr" de ses poids.

La probabilité de les modifier décroît. A l'inverse, si l'association est "mauvaise", un signal de renforcement négatif est émis. La confiance dans les poids diminue et les chances d'inverser leur valeur augmentent.

Algorithme PCR

Intégration temporelle

$$\overline{X_j}(t+1) = \frac{\tau \overline{X_j}(t) + X_j(t)}{\tau + 1} \qquad (1)$$

Mise à jour à chaque itération

$\overline{I_i}$,$\overline{O_j}$ et $\overline{IO_{ij}}$ mise à jour en fonction de 1

Si $\left|\frac{\partial P(t)}{\partial t}\right| > \xi$:

Mise à jour des probabilités

$$\begin{aligned} \Delta p_{ij}(t) &= \alpha \cdot \frac{\partial P}{\partial t} \cdot \mathcal{C}_{ij} \cdot (2 \cdot W_{ij} - 1) \\ p_{ij}(t+1) &= p_{ij}(t) + \Delta p_{ij}(t) \end{aligned} \qquad (2)$$

Tirage aléatoire

$$\text{Si} \quad Alea > p_{ij} \quad \text{et } \overline{I} \cdot \overline{O} \neq 0 \quad \text{alors} \quad \begin{cases} W_{ij} = 1 - W_{ij} \\ p_{ij} = 1 - p_{ij} \end{cases} \qquad (3)$$

$P(t)$ est le signal global de renforcement. Il mesure la satisfaction du robot.
α est le coefficient d'apprentissage pour le système à récompense différée
ξ est une constante fixée par l'expérimentateur
Alea est une valeur aléatoire prise dans l'intervale $[0, 1]$

Les signaux de renforcement sont émis par une sorte de mini système limbique qui agit à deux echelles de temps différentes. Si l'action courante amène à une situation susceptible de détériorer le robot (collision), le robot reçoit immédiatement un signal de douleur très important ($\frac{\partial P(t)}{\partial t} \ll -1$), la probabilité du poids associé à l'établissement de cette action tombe à 0 et le poids est forcement changé. Un deuxième mécanisme cherche à imiter les processus de la faim. Aprés un temps donné passé dans le labyrinthe (fonction d'une horloge biologique) sans trouver de nourriture (la récompense à la sortie du labyrinthe), le robot éprouve la sensation de "faim". Cela se traduit par une petite modification de $P(t)$ ($\frac{\partial P(t)}{\partial t} < 0$) qui change les probabilités associées aux poids (la stratégie de recherche dans le labyrinthe est remise en cause).

3 Résultats

Afin de valider notre algorithme nous avons testé ses performances sur différents types de labyrinthes. De plus, pour pouvoir mettre en avant ses qualités nous avons réalisé les mêmes expériences en utilisant le Q-learning qui est un algorithme classique de renforcement.

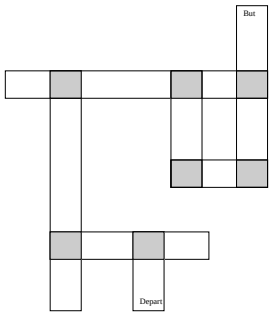

Figure 3: Labyrinthe "long" comportant une boucle.

Les labyrinthes utilisés ont été construits pour mettre en avant les paramètres influençant la convergence de PCR ou du Q-learning. Nous utilisons un labyrinthe "court" comportant 30 cases et dont le chemin optimal est de 20 cases (voir figure 1). Nous donnons aussi les résultats pour un labyrinthe "long" (65 cases) avec une chemin optimal de 40 cases. Enfin nous avons testé les résultats sur un labyrinthe comportant une boucle (voir figure 3.

Par ailleurs, un deuxième paramètre jouant un rôle dans le temps d'apprentissage est le nombre de motifs différents pouvant être utilisés pour réaliser les associations. Dans un premier cas on choisit d'utiliser un motif pour le couloir, un motif pour tourner à droite et un motif pour tourner à gauche. Dans un second cas, chaque intersection possède un motif diffèrents.

Les résultats sont présentés de manière synthétique dans le tableau ci-dessous :

PCR		
	Nombre de catégories	
Type	**3**	**4**
Court	17 ± 16	30 ± 34
Long	13 ± 10	24 ± 33
A boucle	13 ± 10	24 ± 33

Q-learning		
	Nombre de catégories	
Type	**3**	**4**
Court	35 ± 3	35 ± 3
Long	67 ± 4	67 ± 4
A boucle	?	?

Tout d'abord, comme on le voit, dans tous les cas notre algorithme est plus performant que le Q-learning. Cependant il faut noter que les différences de temps de convergence ne dépendent pas des mêmes paramètres. Dans le cas du Q-learning, c'est le nombre de cases du labyrinthe qui conditionne principalement la vitesse de convergence. En fait cela vient de la séparation des situations en états distincts, ce qui interdit les mécanismes de généralisation. Dans notre algorithme, on voit que la taille du labyrinthe n'influence en aucune manière les performances. Par contre le nombre de formes pouvant être associées influence grandement le temps d'apprentissage.

Un atout important de notre algorithme est qu'il continue à fonctionner même dans des situations délicates telles que les boucles à l'intérieur de labyrinthe. Le Q-learning ne peut pas fonctionner dans pareil cas car cet algorithme crée un état pour chaque instant de la simulation. Il est donc incapable de repérer qu'il tourne en rond, il peut repasser deux fois au même endroit, mais considére que cela représente deux états distincts. Notre algorithme arrive lui à résoudre ce problème grâce aux capacités de généralisation de la carte PTM utilisée pour catégoriser les entrées, mais aussi à cause du mécanisme simulant le système limbique qui lui permet d'éviter que le robot tourne en rond.

4 Conclusion

Nous avons présenté une règle de conditionnement probabiliste permettant de créer des associations pertinentes entre des catégories visuelles et des actions et bien que le signal de renforcement ne soit pas disponible immédiatement après que la bonne action a été effectuée. Pour l'étude de l'algorithme nous avons considéré que le problème de la catégorisation était déjà résolu. Cependant, dans une expérience avec un robot réel, la catégorisation est une étape décisive car elle permet de réduire drastiquement la compléxité d'analyse d'une scène. Dans d'autres travaux [1, 2] nous montrons comment PCR peut s'insérer dans une architecture hiérarchique pour gérer la navigation de notre robot dans un labyrinthe réel. Cependant, nous mettons en évidence qu'une telle architecture nécessite que l'apprentissage se fasse selon une technique de "Shaping" (apprentissage graduel des problèmes les plus simples vers les problèmes les plus compliqués).

References

[1] P. Gaussier, A. Revel, C. Joulain, and B. Gas. Living in a partially structured environment: How to bypass the limitation of classical reinforcement techniques. *submitted to Robotics and Autonomous Systems*, 1996.

[2] C. Joulain, P. Gaussier, and A. Revel. Apprentissage de catégories sensori-motrices par un robot autonome. In *NSI*, 1996.

[3] Marvin Levine. Hypothesis theory and nonlearning despite ideal s-r-reinforcement contingencies. *Psychological Review*, 78(2):130–140, 1971.

[4] C.J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, Psychology Department, Cambridge University, Cambridge, England, 1989.