Modèles stochastiques de la prise de décision collective

Alain Dutech

Modèles stochastiques de la prise de décision collective.

Alain Dutech

Equipe MAIA / LORIA
Campus Scientifique, BP 239, 54506 Vandoeuvre les Nancy
Alain.Dutech@loria.fr

Résumé – Nous présentons le cadre mathématique des processus décisionnels de Markov partiellement observables qui permet, dans le domaine de l'intelligence artificielle, de formaliser la notion d'apprentissage par renforcement. En théorie, ces modèles formels sont aussi applicables aux systèmes collectifs composés de plusieurs agents. Néanmoins, par le biais de plusieurs travaux que nous avons effectués, nous montrons certaines limitations de ces modèles, limitations dues aux hypothèses sous-jacentes implicites ou à l'utilisation pratique de ces outils théoriques. Ainsi que nous le discutons, ces limitations peuvent être utiles pour, entre autre, identifier ou délimiter des fonctions cognitives qui seraient nécessaires pour que l'apprentissage par renforcement « artificiel » puisse s'appliquer à des problèmes que peut résoudre l'être humain. Cette question est l'une de celle que nous abordons dans le cadre de notre discussion sur les apports possible de notre domaine aux sciences cognitives.

Mots-Clés : modèles stochastiques, apprentissage par renforcement, systèmes multi-agents, théorie de la décision.

1. Introduction

En intelligence artificielle, l'apprentissage par renforcement désigne un ensemble de mécanismes d'adaptation inspirés des travaux sur le conditionnement (Sutton & Barto, 1998). Ces méthodes ont connu un essor considérable depuis leur apparition dans les années 90, notamment pour des problématiques de contrôle optimal, en s'appuyant sur un cadre mathématique rigoureux. Elles sont aussi particulièrement séduisantes pour concevoir des agents autonomes évoluant dans des environnements dynamiques et incertains.
Le but de cet article est de présenter les grandes lignes de ce formalisme et son utilisation dans un cadre multi-agent (section 2). Nous voulons aussi, au travers de plusieurs travaux que nous avons menés au sein de l'équipe MAIA, mettre en lumières les limites formelles et pratiques de l'apprentissage par renforcement (section 3). Ainsi que nous l'argumentons (section 4), ces limites sont non-seulement utiles pour définir nos futurs travaux de recherche mais aussi pour apporter des éléments de réflexion dans le cadre des sciences cognitives.

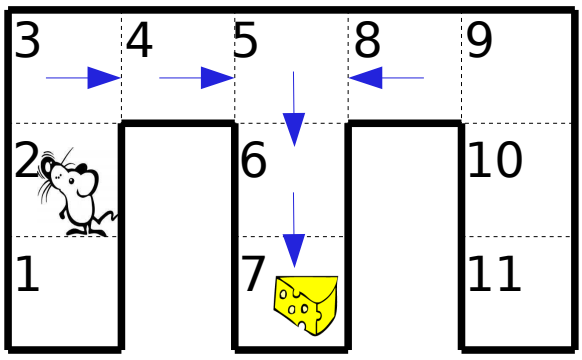
2. Apprentissage par renforcement

Le cadre général de l'apprentissage par renforcement s'intéresse à des agents en interaction avec leur environnement. Un agent est une entité pro-

active guidée par des **buts** qui **agit** sur son environnement en fonction de ses **perceptions** et de son **expérience passée**. La notion de but est matérialisée par le fait que l'agent peut recevoir un signal de récompense (positif ou négatif) après une action, et on considère que l'agent cherche à trouver quel est le comportement qui lui assure les plus grandes récompenses possibles sur le long terme.
Le cadre mathématique des Processus Décisionnels de Markov Partiellement Observé (POMDP[1], voir Cassandra, 1998) permet une approche formelle de l'apprentissage par renforcement. Un POMDP est constitué de :

- un ensemble S des différents états que peut prendre le système constitué de l'agent et de son environnement.
- un ensemble O qui contient les différentes perceptions possibles que l'agent peut avoir sur l'état du système.
- un ensemble A d'actions que l'agent peut effectuer.
- une fonction de transition $T(s,a,s')$ qui, pour chaque état s du système, indique la probabilité que le système se retrouve ensuite dans l'état s' quand l'agent effectue l'action a .
- une fonction de récompense $R(s,a)$ qui indique la récompense (positive, négative ou nulle) reçue par l'agent quand il fait l'action a dans l'état s .
- une fonction d'observation $\Omega(s,o)$ qui indique la probabilité qu'a l'agent de percevoir o quand le système est dans l'état s .

Figure 1: **Modélisation d'un rat dans un labyrinthe**. Le système est décrit par la position de la souris (1-11), la souris ne connaît que les murs qui l'entourent (observations A,B,C,...). Si on connaît l'état, on peut trouver l'action optimale (flèches), si on ne connaît que l'observation, c'est moins facile...

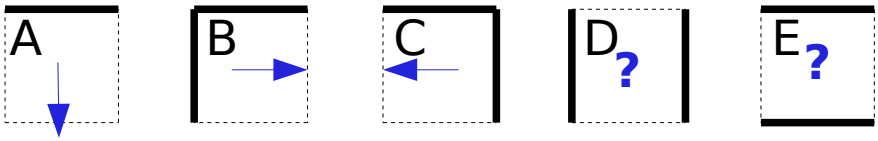


Etats : 1, 2, 3, ...

Actions : flèches

Récompense :
+3 pour le fromage
0 ailleurs

Si on ne connaît que les observations...



Le problème posé est alors pour l'agent d'apprendre un comportement qui

1 De l'anglais « Partially Observable Markov Decision Process ».

lui dit quelle action faire pour maximiser un critère[2], comme par exemple la somme des récompenses r_t reçues par l'agent. Un comportement s'exprime souvent par une politique, c'est à dire une fonction $\pi: O \rightarrow A$ qui, pour chaque observation o indique quelle action a doit faire l'agent.
La figure 1 illustre ce formalisme dans le cas d'une souris qui cherche à trouver du fromage dans un labyrinthe.

2.1. Utiliser un POMDP

Il existe des algorithmes pour construire automatiquement une politique optimale. Ainsi, une fois que l'on a modélisé un problème (il faut définir S, A, O, T, R et Ω), on peut calculer le meilleur comportement pour l'agent. On parle de **planification**. Ces algorithmes sont exacts, c'est-à-dire que l'on peut prouver qu'ils construiront un comportement optimal.
Dans le cadre de l'apprentissage, on peut ne pas connaître la fonction de transition T ou la fonction d'observation Ω ou encore la fonction de récompense R. On dit alors qu'on ne connaît pas le **modèle** du POMDP. Il existe néanmoins des algorithmes qui permettent à un agent d'apprendre un comportement optimal, en s'aidant d'interactions répétées avec l'environnement. L'agent apprend par renforcement à partir de ses expériences passées (Sutton & Barto, 1998). Par contre, les algorithmes ne trouvent un comportement optimal que lorsque le système est **parfaitement observable** par l'agent, c'est-à-dire quand les observations permettent de connaître l'état du système[3].

2.2. En pratique

Le pire ennemi de toutes les méthodes s'appuyant sur les POMDP est la taille des différents espaces de la représentation car la complexité pour trouver une solution dépend directement de ces tailles, et parfois la dépendance est exponentielle. Dans le cas idéal où on a des observations complètes (qui permettent à l'agent de retrouver l'état du système) et un modèle connu, on peut résoudre des problèmes avec quelques dizaine de milliers d'état (parfois au prix de quelques approximations). Les principaux algorithmes de planification sont *policy iteration* et *value iteration* (Puterman, 1994), et dans le cadre de l'apprentissage par renforcement, le *Q-learning*, $TD(\lambda)$ et *actor-critic* (Sutton & Barto, 1998). A l'inverse, dans le cas d'observations partielles, on doit se limiter à quelques états, même quand on connaît le système. On utilise alors des algorithmes comme *Witness* ou *incremental pruning* (Cassandra, 1998). Par contre, comme nous allons le détailler, on ne sait pas apprendre dans le cadre des POMDP, et les problèmes sont encore plus aigus quand on considère plusieurs agents.

---

2 Un critère formel habituellement utilisé est : $U(s) = E\left(\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s\right)$ où $\gamma < 1$

3 par exemple, quand la fonction d'observation est une bijection de S vers O.

2.3. Comment apprendre dans les POMDP

Quand on ne connaît pas le modèle (transition et récompense) et que les observations sont incomplètes, il n'existe pas d'algorithme exact d'apprentissage par renforcement pour les POMDP. La principale difficulté vient du fait que les observations ne donnent pas assez d'information pour apprendre les causalités nécessaires à garantir que la solution trouvée est optimale. Des méthodes approchées permettent de trouver des politiques localement optimale (il n'y a pas de meilleure politique proche) en faisant l'hypothèse que les observations sont complètes (Jaakkola, Singh, & Jordan, 1994).
Pour notre part, nous avons adopté une approche plus cognitive qui part du principe que si l'observation seule n'est pas suffisante, une certaine mémoire du passé peut permettre d'obtenir suffisamment d'information pour apprendre de meilleure solution (Dutech & Samuelides, 2003). Là encore, on ne peut pas garantir l'optimalité. Ces travaux nous ont par contre amené à nous pencher sur un problème qui est encore, dans notre communauté, très ouvert : comment savoir quels sont les éléments du passé qui sont pertinents à mémoriser pour pouvoir apprendre par renforcement (Dutech & Scherrer, 2001).

3. Dans un cadre collectif

Le cadre formel et abstrait des POMDP s'étend de manière assez naturelle aux cas multi-agents. Si l'état du système reste décrit par l'ensemble S, chaque agent i possède sa propre fonction d'observation Ω^i, son propre ensemble d'action A^i et même sa propre fonction de récompense R^i. Les transitions entre les différents états du système dépendent maintenant des actions de **tous** les agents, c'est une sorte de méta-action composée des actions individuelles, que l'on peut noter $a=\left(a^1,\ldots,a^n\right)$.
Si la problématique de l'apprentissage dans un cadre multi-agent coopératif est mathématiquement résolue, elle l'est dans un cas assez éloigné de la réalité et des systèmes cognitifs. Notre optique est d'analyser les problèmes soulevés par la mise en oeuvre de l'apprentissage par renforcement avec deux contraintes supplémentaires sur les agents : ils doivent être **indépendants** et pratiquement réalisables du point de vue informatique. Si la première contrainte concerne naturellement les systèmes cognitifs, nous verrons que la deuxième, pourtant éloignée des réalités biologiques, peut apporter un éclairage intéressant sur la cognition.
Nous allons maintenant lister les problématiques qui ont émergé des travaux de notre équipe.

3.1. Coopération et agents individualistes

3.1.1. Problématique

Même avec seulement 2 agents, Hart & Mas-Colell (2003) ont montré qu'il

est impossible de trouver un algorithme général qui construit des comportements optimaux dans tous les cas où les agents apprennent indépendamment l'un de l'autre (pas de communication ou de couplage). En particulier, des agents indépendants n'ont pas forcément les mêmes objectifs en terme de récompense et il est alors difficile pour les agents d'apprendre à coopérer. Le plus fréquemment, le comportement global appris par les agents se stabilise sur une solution qui se contente d'assurer aux agents des pertes minimales[4].

3.1.2.Couplage entre les agents

La question que nous nous sommes posée est alors de savoir quelles étaient les contraintes minimales à rajouter aux agents pour que, chaque agent apprenant égoïstement de manière indépendante, le comportement global collectif résultat soit néanmoins satisfaisant. Dans (Aras, Dutech & Charpillet, 06), nous avons ainsi proposé un couplage assez léger entre les agents en leur permettant de passer des pseudo-contrats entre eux. En s'échangeant virtuellement des récompenses, les agents peuvent parfois s'auto-inciter à choisir des actions moins égoïstes et ainsi apprendre à collaborer plus efficacement.

3.2. Co-évolution

3.2.1. Problématique

Quand plusieurs agents apprennent en même temps on parle de co-évolution. Chaque agent se voit finalement confronté à un environnement évolutif car composé, entre autre, des autres agents. On dit alors que l'agent qui apprend est placé dans un environnement non-stationnaire. Ce problème spécifique de la non-stationarité est des plus complexe à résoudre avec des agents indépendants. Sa solution théorique - qui est d'inclure explicitement le temps dans l'état du système - n'est pas appliquable en pratique.

3.2.2. Apprentissage alterné

Un solution qui a été proposée au sein de notre équipe (Chades, 2003) est d'alterner les groupes d'agents qui apprennent. On gèle ainsi le comportement de tous les agents - ils n'apprennent pas et ne modifient pas leur comportement - sauf un. L'agent apprend donc dans un environnement stationnaire. On passe ensuite à un autre agent, toujours en gelant le comportement des autres agent, et ainsi de suite.
Cette manière de procéder, bien que s'appuyant sur un apprentissage individuel, garanti qu'une certaine organisation collective va se mettre en place puisqu'il a été montré que ces algorithmes de co-évolution se stabilisent dans des positions d'équilibre où aucun agent n'a intérêt à changer de comportement ou de cycle de comportements[5].

4 les agents vont converger vers un équilibre de Nash (Myerson, 1991).
5 ce sont encore une fois des équilibres de Nash.

3.3. Structuration par les interactions

3.3.1. Problématique

Quand le nombre d'agent augmente, la complexité du POMDP permettant de représenter le problème croît de manière exponentielle. Comme la résolution elle même est d'une complexité qui est pire qu'exponentielle en la taille de l'état, on arrive à un problème qui est doublement exponentiel. En pratique, il devient rapidement utopique de vouloir résoudre le problème directement.

3.3.2. Prise en compte explicite des interactions

La solution proposée par Thomas, V., Bourjot C., & Chevrier, V. (2006) est élégante à plusieurs points de vue. L'idée principale est de modéliser explicitement les interactions possibles entre les agents dans le modèle au lieu de les représenter de manière implicite (par exemple, une interaction pourrait être le résultat de certaines actions conjointes dans des situations bien particulières). Ce modèle, appelé *Interac-Dec-POMDP*, permet de réduire la complexité de résolution en décomposant finalement le problème en trois sous problèmes :

- apprendre à agir au mieux individuellement entre deux interactions.
- apprendre quand et avec qui déclencher des interactions.
- les ensembles d'agents qui sont impliqués dans des interactions doivent apprendre à résoudre au mieux leur interaction. C'est uniquement dans cette phase que l'on a besoin de prendre en compte plusieurs agents pour la résolution.

De plus, cette approche permet de mettre l'accent non seulement sur les actions et les perceptions mais aussi, et surtout, sur les interactions dans un problème multi-agent. Cela permet aussi, du point de vue de l'agent, de tenir compte des autres agents sans avoir besoin de les modéliser, les problèmes collectifs (coordination, coopération, concurrence) étant concentrés dans les interactions. Cette approche permet de travailler avec des agents ayant une architecture simple et qui n'ont pas besoin de mettre en œuvre des fonctions cognitives complexes.

3.3.3. Du shaping à l'apprentissage incrémental.

Nous avons aussi abordé le problème de la taille du problème avec comme fil directeur de procéder de manière incrémentale (Dutech, Buffet, & Charpillet, 2001). Alors que le but est d'apprendre à de nombreux agents des comportements leur permettant de résoudre une tâche collective, nous avons proposé de n'utiliser d'abord que 2 agents (le minimum requis). De plus, ces deux agents n'essaient pas de résoudre directement la tâche finale, mais sont progressivement placés devant des tâches de plus en plus complexes les amenant à apprendre à résoudre la tâche finale. Ensuite, on augmente progressivement le nombre d'agents dans l'environnement (les nouveaux bénéficiant de l'expérience des agents plus aguerris), les agents continuant naturellement à apprendre. Cette démarche s'inspire des

travaux sur le shaping[6] en psychologie comportementale (Skinner, 1938).

4. Conclusion

Nous avons présenté le cadre formel dans lequel la communauté de l'intelligence artificielle étudie l'apprentissage par renforcement. Le **cadre mathématique** des POMDP, qui permet d'asseoir formellement la notion d'apprentissage par renforcement, permet en théorie de modéliser tout type de système. Cependant, la théorie nous enseigne que l'apprentissage n'est possible que dans des cas idéaux et assez éloignés de la réalité : états complets du système connus, un seul agent, environnement stationnaire. Ainsi, quand on essaie d'utiliser ces outils dans un cadre multi-agent, même avec des hypothèses minimalistes de plausibilité (agents indépendants), on se trouve confronté à de nouveaux problèmes, qui sont autant de questions ouvertes.

- comment apprendre par renforcement quand les perceptions immédiates de l'agent ne lui permettent pas d'obtenir des informations suffisantes sur l'état du système (section 2.3). La solution « terre à terre » que nous avons commencé à explorer pose le problème de l'apprentissage de causalités ou de contingences pertinentes.
- quelle forme minimale de couplage (par la communication directe ou indirecte) est nécessaire pour que des agents puissent apprendre de manière collective (section 3.1). Il semble qu'on ne puisse apprendre ensemble en faisant abstraction de l'autre.
- comment apprendre ensemble sans se trouver confronté à une tâche qui soit trop dynamique du fait que tous les agents apprennent en même temps (section 3.2). Les solutions actuelles, peu satisfaisantes, reposent sur un contrôle central répartissant le temps d'apprentissage sur les agents.

D'un point de vue informatique, la faisabilité pratique des agents ajoute la question suivante :

- comment rendre faisable en pratique l'utilisation de l'apprentissage par renforcement dans les problèmes multi-agents alors que, par nature, la complexité des algorithmes est exponentielle. Une solution est de structurer le problème par l'utilisation explicite d'interaction entre les agents (section 3.3), ce qui pose alors le problème de pouvoir différencier les contingences individuelles et collectives. Une autre est de guider les agents dans leur apprentissage (section 3.4), mais il est alors nécessaire de pouvoir quantifier la difficulté des tâches que l'on soumet aux agents.

Au vu de toutes ces difficultés, on peut légitimement se demander quel est l'utilité de ce type de modèle et donc de notre approche. Essayons d'aborder cette question du point de vue des sciences cognitives.

D'une part, il est clair que le formalisme de l'apprentissage par renforcement ne peut prétendre vouloir modéliser ni la cognition dans son ensemble, ni l'ensemble des mécanismes d'apprentissage. Au départ, ces méthodes se veulent un simple modèle abstrait du conditionnement, c'est-

6 que l'on pourrait traduire par « façonnage ».

à-dire déconnecté des mécanismes biologiques sous-jacents. De fait, le formalisme de l'apprentissage par renforcement ne semble pouvoir s'appliquer que pour des comportements cognitivement simples, quasiment « **réactifs** ». D'ailleurs, il a été utilisé par certains auteurs pour montrer, entre autre, qu'il était possible d'expliquer des formes d'organisation collective sans faire appel à des fonction cognitives complexes (Cotel et. al., 2005).

De plus, il est important de signaler que le modèle formel de l'apprentissage par renforcement et des POMDP suppose des apprenants **rationnels**, c'est-à-dire qu'ils recherchent des solutions optimales. Or, ce n'est clairement pas le cas des êtres humains. Nous pensons néanmoins que les problèmes que nous avons soulevés précédemment, bien que mis en évidence sous cette hypothèse de rationalité, soulèvent des questions valides dans un cadre plus général. Ils se retrouvent par exemple dans les problèmes de « viabilité » (Aubin & *al*., 2005) où, pourtant, on ne cherche pas de solution optimale.

Sans nous attarder sur les apports possibles pour d'autres communautés (théorie de la décision, automatique, robotique), nous argumentons que, du point de vue des sciences cognitives, l'utilité des POMDP se trouve justement dans les **questions soulevées par ces modèles**.

En premier lieu, dans le cas de fonctions cognitives qui peuvent être modélisée par l'apprentissage par renforcement, le fait de mieux connaître les possibilités et les limites du modèle peut permettre de mieux explorer ou rechercher les mécanismes biologiques qui pourraient mettre en oeuvre ce genre d'apprentissage.

De plus, explorer les limites de l'apprentissage par renforcement est nécessaire pour identifier les fonctions cognitives que ce formalisme pourrait contribuer à modéliser. Ce travail peut permettre de mieux définir les hypothèses, les pré-requis et les limites des fonctions cognitives ainsi étudiées.

En particulier, on peut se demander quelles sont les fonctions cognitives nécessaires pour que des agents apprennent de manière collective et si l'apprentissage par renforcement peut y avoir une place. Les carences, que nous avons fait ressortir au cours de nos études, permettent justement de mettre le doigt sur les fonctions cognitives qu'il semble nécessaire de mettre en oeuvre pour apprendre dans ce cadre :

- comment apprendre les liens de cause à effet pertinents de manière à pallier la pauvreté des perceptions immédiates de l'environnement.
- comment gérer un monde changeant et évolutif?
- comment anticiper en présence d'autres agents?

Chacune de ces question peut amener à son tour d'autres questions. Ainsi, une possibilité pour anticiper en présence d'autre agent est de se doter d'un modèle du comportement des autres agents, donc de savoir reconnaître et différencier les autres agents, puis de savoir remettre en cause son modèle, *etc*.

Si l'on se place d'un point de vue plus global, il apparaît clairement que de nombreuses tâches apprises par les êtres humains, voire même par les animaux, ne peuvent être exprimés que dans des modèles pour lesquels les méthodes d'apprentissage par renforcement existantes sont inopérantes. De nombreux autres formalismes d'apprentissages existent (réseaux de

neurones récurrents, cartes auto-organisatrices, réseaux bayésiens, apprentissage statistique, *etc.*) mais aucun ne peut modéliser l'apprentissage humain dans son entier. Doit-on les combiner? Comment les combiner? Sont-ils suffisants pour tout expliquer en les combinant? Comment sont-ils mis en oeuvre? Sont-ils plausibles? Les réponses à ces questions passent par l'identification des capacités, du champ d'application et des limites de ces formalismes. C'est ce à quoi nous nous attachons dans le cadre de l'apprentissage par renforcement.
Les nombreuses questions qui parsèment cet articles, qui dépassent souvent le cadre strict de l'apprentissage par renforcement, peuvent néanmoins être abordées dans ce cadre, fournissant ainsi des pistes et des axes de recherches qui ne demandent qu'à être croisés avec d'autres thèmes de recherche et d'autres disciplines. C'est en ce sens que nous nous y intéressons.

5. Références

Aubin, J.-P., Bayen, A., & Bonneuil N. & Saint-Pierre, P. (2005). *Viability, Control and Games: Regulation of Complex Evolutionary Systems Under Uncertainty and Viability Constraints*, Springer-Verlag

Aras, R., Dutech, A., & Charpillet, F. (2006). Efficient Learning in Games. *Proceedings de la huitème Conférence Francophone sur l'Apprentissage (CAp'06)*, Trégastel, France.

Cassandra, A. (1998). Exact and approximate algorithms for partially observable Markov decision processes. *PhD Thesis, Brown University, Providence, RI*.

Chades, I. (2003). Planification distribuée dans les systèmes multi-agents à l'aide de processus décisionnels de Markov. *Thèse de l'Université de Nancy I*.

Cotel, M.-C., Thomas, V., Bourjot, C., Desor, D., Chevrier, V., & Schroeder, H. (2005) Processus cognitifs et differenciation sociale de groupes de rats : interet de la modelisation multi-agents. *6e colloque des jeunes chercheurs en sciences cognitives.*

Dutech, A., Buffet, O.,& Charpillet, F. (2001). Multi-Agent systems by incremental gradient reinforcement learning. *Sevtenteenth International Joint Conference on Artificial Intelligence,* IJCAI-01, Seattle, USA.

Dutech, A., & Scherrer, B. (2001). Learning to use contextual information for solving partially observable Markov decision problems. *Proc. of the 5th European Workshop on Reinforcement Learning, EWRL-5, Utrecht.*

Dutech, A., & Samuelides, M. (2003). Apprentissage par renforcement pour les processus décisionnels de Markov partiellement observés. *Revue d'Intelligence Artificielle, RIA, volume 17(4)*.

Hart, S., & Mas-Colell, A. (2003). Uncoupled dynamics do not lead to Nash equilibrium, *American Economic Review*, pages 1830-1836.

Myerson, R. (1991). *Game theory: Analysis of conflicts*. Harvard University Press.

Puterman, M. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons Inc., New York.

Skinner, B. (1938). *The Behavior of Organisms: An Experimental Analysis*. Prentice Hall, New-Jersey.

Sutton, R., & Barto, A. (1998). *Reinforcement Learning*. Bradford Book, MIT Press, Cambridge, MA.

Thomas, V., Bourjot C., & Chevrier, V. (2006). Heuristique pour l'apprentissage décentralisé d'interaction dans les systèmes multi-agents réactifs. *Reconnaissance des Forme et Intelligence Artificielle*, RFIA'06, Tours.