Apprentissage par renforcement bayésien versus recherche directe de politique hors-ligne en utilisant une distribution a priori: comparaison empirique

Michaël Castronovo[1], Damien Ernst[1], Raphaël Fonteneau[1]

Département d'Electricité, Electronique et Informatique
Université de Liège, Belgique
{m.castronovo,dernst,raphael.fonteneau}@ulg.ac.be

Résumé : Cet article aborde le problème de prise de décision séquentielle dans des processus de décision de Markov (MDPs) finis et inconnus. L'absence de connaissance sur le MDP est modélisée sous la forme d'une distribution de probabilité sur un ensemble de MDPs candidats connue a priori. Le critère de performance utilisé est l'espérance de la somme des récompenses actualisées sur une trajectoire infinie. En parallèle du critère d'optimalité, les contraintes liées au temps de calcul sont formalisées rigoureusement. Tout d'abord, une phase « hors-ligne » précédant l'interaction avec le MDP inconnu offre à l'agent la possibilité d'exploiter la distribution a priori pendant un temps limité. Ensuite, durant la phase d'interaction avec le MDP, à chaque pas de temps, l'agent doit prendre une décision dans un laps de temps contraint déterminé. Dans ce contexte, nous comparons deux stratégies de prise de décision : OPPS, une approche récente exploitant essentiellement la phase hors-ligne pour sélectionner une politique dans un ensemble de politiques candidates et BAMCP, une approche récente de planification en-ligne bayésienne.

Nous comparons empiriquement ces approches dans un contexte bayésien, en ce sens que nous évaluons leurs performances sur un large ensemble de problèmes tirés selon une distribution de test. A notre connaissance, il s'agit des premiers tests expérimentaux de ce type en apprentissage par renforcement. Nous étudions plusieurs cas de figure en considérant diverses distributions pouvant être utilisées aussi bien en tant que distribution a priori qu'en tant que distribution de test. Les résultats obtenus suggèrent qu'exploiter une distribution a priori durant une phase d'optimisation hors-ligne est un avantage non-négligeable pour des distributions a priori précises et/ou contraintes à de petits budgets temps en-ligne.

Mots-clés : Apprentissage par Renforcement

1 Introduction

Interagir efficacement dans le cadre d'un processus de décision de Markov (MDP) inconnu afin de maximiser les récompenses collectées sur le long terme reste un défi pour les algorithmes d'Apprentissage par Renforcement (A/R) [2]. La difficulté principale réside dans le compromis Exploration / Exploitation (E/E) : l'agent doit prendre des décisions dans le but d'améliorer sa connaissance du système, afin d'être en mesure de prendre de bonnes décisions sur le long terme, quitte à prendre de mauvaises décisions sur le court terme.

Au cours des quinze dernières années, l'A/R bayésien [4, 19] s'est imposé comme étant une approche sensée pour aborder le compromis E/E. La philosophie du A/R bayésien consiste à fournir à l'agent une connaissance a priori sous forme d'une distribution de probabilités définie sur un ensemble de MDPs candidats (distribution a priori). Durant la phase d'interaction, une distribution a posteriori est maintenue à partir de la distribution a priori et de l'historique des transitions observées. La distribution a posteriori est utilisée afin de calculer une décision bayésienne (quasi-)optimale, définissant la stratégie d'exploration à proprement parler. Différentes approches d'exploration bayésienne dans les MDPs ont déjà été proposées : les méthodes d'A/R bayésien utilisant un modèle maintiennent une distribution a posteriori sur des modèles de transitions (et éventuellement de récompense) [1, 9, 10, 13, 14, 16, 17, 18]. Il existe également des méthodes d'A/R bayésien sans modèle, qui ne conservent pas explicitement une distribution a posteriori sur des modèles de transitions, mais plutôt sur les fonctions de valeurs utilisées par l'agent pour prendre ses décisions (voir par exemple [5, 6, 7, 8, 11, 12]).

Guez et al. [13] ont récemment introduit l'algorithme BAMCP (de l'anglais Bayes-Adaptive Monte Carlo Planning), une approche A/R bayésienne utilisant un modèle combinant le principe de l'algorithme UCT (de l'anglais Upper Confidence bounds for Trees) avec une méthode d'échantillonnage parcimonieuse de la distribution a posteriori. Cet algorithme s'est montré empiriquement performant, atteignant les performances de l'état-de-l'art. Parallèlement, Castronovo et al. [3] ont proposé un algorithme exploitant une distribution a priori durant une phase « hors-ligne » précédant l'interaction avec le MDP inconnu. Cette approche consiste à résoudre un problème de recherche directe parmi un ensemble de stratégies construites à partir de petites formules servant d'index. La stratégie sélectionnée est ensuite appliquée au MDP inconnu durant la phase d'interaction. La contribution de cet article est de comparer ces deux approches dans un contexte bayésien, c'est à dire sur un grand nombre de problèmes de test. Ces problèmes sont tirés selon différentes distributions de probabilités. On réalise plusieurs jeux d'expériences correspondant à différentes associations distribution a priori / distribution de tirage des problèmes de test (les distributions pouvant être similaires ou différentes). Ces expériences sont réalisées en tenant compte du temps de calcul minimal requis pour exécuter les algorithmes. Les résultats montrent que, dans le cas où les contraintes temporelles en-ligne sont serrées, il est avantageux d'utiliser une approche exploitant la distribution a priori durant la phase hors-ligne lorsque la distribution a priori est « informative », tandis que les approches qui maintiennent une distribution a posteriori se montrent généralement plus efficaces dans les autres cas.

Cet article est organisé de la manière suivante : la section 2 formalise le problème. La section 3 présente le protocole expérimental déployé ainsi que les résultats empiriques obtenus. La section 4 discute les résultats obtenus, et la section 5 conclut l'article.

2 Enoncé du problème

Le but de cet article est de comparer deux stratégies d'A/R en présence d'une distribution a priori. Nous commençons par décrire le contexte A/R dans la section 2.1. Nous formalisons ensuite l'hypothèse de la connaissance d'une distribution a priori dans la section 2.2, et décrivons brièvement les principes sous-jacents des algorithmes BAMCP et OPPS. La section 2.3 formalise la notion de budget temps définissant les contraintes auxquelles ces méthodes sont soumises, et la section 2.4 présente les spécificités de notre évaluation empirique.

2.1 Apprentissage par renforcement

Soit $M = (X, U, f(\cdot), \rho_M, \rho_{M,0}(\cdot), \gamma)$ un MDP inconnu, où $X = \{x^{(1)}, \ldots, x^{(n_X)}\}$ représente son espace d'état fini et $U = \{u^{(1)}, \ldots, u^{(n_U)}\}$ son espace d'action fini également. Lorsque l'agent se trouve dans l'état $x_t \in X$ à l'instant t et qu'une action $u_t \in U$ est sélectionnée, celui-ci se déplace instantanément vers l'état suivant $x_{t+1} \in X$ avec une probabilité de $P(x_{t+1}|x_t, u_t) = f(x_t, u_t, x_{t+1})$. Une récompense instantanée, déterministe et bornée $r_t = \rho(x_t, u_t, x_{t+1}) \in [R_{\min}, R_{\max}]$ est également observée par l'agent. Dans cet article, nous considérons que la fonction de récompense ρ est parfaitement connue, ce qui est souvent le cas en pratique.

Soit $H_t = (x_0, u_0, r_0, x_1, \cdots, x_{t-1}, u_{t-1}, r_{t-1}, x_t)$ l'historique des transitions observées jusqu'à l'instant t. Une stratégie d'E/E est une politique stochastique h qui, en fonction de l'historique courant H_t, retourne une action $u_t \sim h(H_t)$. Suivant la distribution de probabilité sur les états initiaux $\rho_{M,0}(\cdot)$, l'espérance du retour d'une stratégie d'E/E h pour un MDP donné M se définit comme :

$$J_M^h = \mathop{\mathbb{E}}_{x_0 \sim \rho_{M,0}(\cdot)} [\mathcal{R}_M^h(x_0)],$$

où $\mathcal{R}_M^h(x_0)$ est la somme des récompenses actualisées reçues en appliquant la stratégie d'E/E h, à partir de l'état initial x_0 :

$$\mathcal{R}_M^h(x_0) = \sum_{t=0}^{+\infty} \gamma^t r_t,$$

avec le facteur d ?actualisation γ appartenant à l'intervalle $[0, 1)$. Dans ce contexte, l'objectif est de trouver une politique h^* maximisant J_M^h :

$$h^* \in \arg\max_h \quad J_M^h \ .$$

2.2 Distribution a priori sur un ensemble de modèles

Dans le cas où le MDP est initialement inconnu, les techniques d'A/R bayésien fondées sur un modèle représentent l'incertitude sur le modèle du MDP courant par une distribution de probabilité dont le support est un ensemble de modèles.

Dans cet article, nous supposons que nous connaissons une distribution a priori $p^0_{\mathcal{M}}(\cdot)$ sur un ensemble de MDPs $\mathcal{M}$. De plus, nous faisons les hypothèses suivantes :

— Nous pouvons tirer facilement des modèles de MDPs à partir de $p^0_{\mathcal{M}}(\cdot)$;

— Nous pouvons calculer facilement une distribution a posteriori à partir de $p^0_{\mathcal{M}}(\cdot)$ et de l'historique H_t.

En tenant compte de ces hypothèses, nous avons pour objectif de trouver une stratégie d'E/E h^* maximisant l'espérance de son retour sur un ensemble de modèles de transitions $\mathcal{M}$:

$$h^* \in \arg\max_h \mathop{\mathbb{E}}_{M \sim p^0_{\mathcal{M}}(\cdot)} \left[J^h_M \right] .$$

Dans cet article, nous comparons deux algorithmes tirant parti d'une telle distribution a priori : les algorithmes BAMCP et OPPS.

2.2.1 L'algorithme BAMCP

BAMCP (Bayes-adaptive Monte Carlo planning) est un algorithme A/R bayésien initialement proposé dans [13] et dont les performances atteignent l'état de l'art. Il consiste à appliquer le principe de l'algorithme UCT (Upper Confidence bounds for Trees, voir [15]) dans un MDP bayésien adaptatif [1]. L'algorithme BAMCP est efficace en pratique, notamment grâce à l'emploi d'une technique d'échantillonnage parcimonieuse permettant d'éviter le tirage de modèles à partir des distributions a posteriori dans chaque nœud de l'arbre de planification. En pratique, à partir d'une distribution a priori $p^0_{\mathcal{M}}(.)$ et d'un historique H_t, l'algorithme BAMCP calcule une politique h^{BAMCP}_K, fondée sur la construction d'un arbre de planification comportant exactement K nœuds permettant de déterminer l'action à sélectionner :

$$u_t \sim h^{BAMCP}_K \left(H_t, p^0_{\mathcal{M}}(.) \right) .$$

Notons que, lorsque le nombre de nœuds ajoutés à chaque pas de temps K tend vers l'infini, la décision calculée par l'algorithme BAMCP tend vers l'optimum bayésien.

2.2.2 L'algorithme OPPS

L'algorithme OPPS (Off-line, Prior-based, Policy Search) a été présenté dans [3]. L'approche consiste à (i) construire un ensemble $\mathcal{S}$ de stratégies E/E candidates, et (ii) chercher une politique optimale parmi celles-ci. L'espace des stratégies est construit à partir de stratégies obtenues par maximisation d'indices obtenus à partir de petites formules combinant des informations standards de l'A/R (par exemple, des fonctions de valeurs) via divers opérateurs mathématiques. La recherche d'une stratégie d'E/E optimale est formalisée comme un problème de bandits comportant un bras par stratégie. Tirer un bras revient à tirer un MDP à partir de la distribution a priori, et exécuter une seule fois la stratégie d'E/E candidate correspondant à ce bras. Formellement, l'algorithme OPPS calcule, durant la phase hors-ligne, une politique $h^{OPPS}_{\mathcal{S}}$ dont les décisions sont calculées en-ligne à partir de la distribution a priori $p^0_{\mathcal{M}}(.)$ et de l'historique H_t :

$$u_t \sim h^{OPPS}_{\mathcal{S}}(H_t, p^0_{\mathcal{M}}(.))$$

où

$$h^{OPPS}_{\mathcal{S}} \in \arg\max_{s \in \mathcal{S}} \mathop{\mathbb{E}}_{M \sim p^0_{\mathcal{M}}(\cdot)} \left[J^s_M \right] .$$

Dans cet article, l'ensemble des variables utilisées pour construire les formules diffère légèrement de celui utilisé dans [3]. Cet ensemble est décrit précisément dans l'annexe A.

1. Un MDP obtenu en concaténant l'état avec la distribution a posteriori.

2.3 Contraintes de temps

Les algorithmes d'A/R bayésiens peuvent être assez lents durant la phase en-ligne, à cause notamment des mises à jour de la distribution a posteriori durant la phase de planification. Dans cet article, nous formalisons explicitement les budgets temps disponibles à chaque étape. Nous considérons deux types de budgets temps :

— Un budget temps « hors-ligne » B_{-1}, correspondant à la phase durant laquelle l'agent est en mesure d'exploiter la distribution a priori mais ne peut pas encore interagir avec le MDP courant.

— Des budgets temps « en-ligne » $B_0, B_1 \ldots$, où à chaque pas de temps $t \in \mathbb{N}$, est associé un budget temps B_t représentant le temps disponible pour calculer la décision $u_t \in U$ à partir de la distribution a priori $p^0_{\mathcal{M}}(.)$ et de l'historique H_t des transitions observées jusqu'à l'instant t.

2.4 Evaluation bayésienne empirique

Dans cet article, nous proposons une véritable évaluation bayésienne empirique, dans le sens où nous comparons les algorithmes sur un large ensemble de problèmes tirés à partir d'une distribution de probabilité de test. Une telle distribution peut être soit similaire (« précise ») à la distribution a priori, soit différente (« imprécise ») de cette dernière. Formellement, dans chaque expérience, nous avons considéré une distribution a priori $p^0_{\mathcal{M}}(\cdot)$, qui est donnée initialement à chaque algorithme, ainsi qu'une distribution de test $p_{\mathcal{M}}(\cdot)$, utilisée pour tirer les problèmes sur lesquels chaque algorithme est évalué. A notre connaissance, il s'agit de la première fois que des algorithmes d'A/R bayésiens sont comparés en moyenne sur un large ensemble de problèmes plutôt que sur des problèmes uniques.

3 Expériences

Chaque expérience est caractérisée par :

— Une distribution a priori $p^0_{\mathcal{M}}(\cdot)$,

— Une distribution de test $p_{\mathcal{M}}(\cdot)$,

— Un budget temps hors-ligne B_{-1},

— Des budgets temps en-ligne $B_0, B_1, \cdots$ alloués à la prise de décision sur le MDP courant.

Le but de ces expériences est d'identifier l'influence des éléments mentionnés ci-dessus sur les performances des algorithmes et, par conséquent, d'identifier le(s) domaine(s) d'excellence de chaque approche.

La sous-section 3.1 décrit le protocole expérimental utilisé pour comparer les algorithmes décrits dans la section 2.2. La sous-section 3.2 définit précisément les distributions de MDPs considérées dans les expériences présentées dans la sous-section 3.3.

3.1 Protocole expérimental

Pour chaque algorithme,

— Un ensemble de $10\,000$ MDPs est tiré à partir de la distribution $p_{\mathcal{M}}(.)$;

— Nous exécutons une fois l'algorithme sur chaque MDP du groupe ;

— Nous calculons la moyenne empirique des retours escomptés.

Les trajectoires sont tronquées après T décisions, où T est un horizon de troncature défini de la manière suivante :

$$T = \left\lceil \frac{\frac{\epsilon \times (1-\gamma)}{R_{max}}}{\log \gamma} \right\rceil \text{ avec } \epsilon = 0.001.$$

Nous mesurons la moyenne μ et l'écart-type σ de l'ensemble des retours observés. Ces données nous permettent de calculer l'intervalle de confiance à 95% de $J^{h(p^0_{\mathcal{M}}(\cdot))}_{p_{\mathcal{M}}(\cdot)}$:

$$J^{h(p^0_{\mathcal{M}}(\cdot))}_{p_{\mathcal{M}}(\cdot)} \in \left[\mu - \frac{2\sigma}{\sqrt{10\,000}}; \mu + \frac{2\sigma}{\sqrt{10\,000}}\right] \text{ avec une probabilité d'au moins } 95\%.$$

Puisque chaque distribution de MDP décrite ci-dessous peut être utilisée aussi bien comme distribution a priori $p^0_{\mathcal{M}}(\cdot)$ que comme distribution de test $p_{\mathcal{M}}(\cdot)$, le processus est répété pour chaque combinaison possible.

3.2 Distributions de MDPs

Les distributions de MDPs introduites dans cet article sont inspirées du problème « Chain MDP » (MPD chaîné à 5 états) [19]. Pour toutes les distributions de MDPs considérées, l'ensemble des MDPs constituant leur support partagent le même espace d'état X, espace d'action U, fonction de récompense ρ_M, distribution des états initiaux $\rho_{M,0}(\cdot)$ ainsi que le même facteur d'actualisation γ. Dans nos expériences, $X = \{1, 2, 3, 4, 5\}$, $U = \{1, 2, 3\}$, $\gamma = 0.95$ et $x_0 = 1$ avec une probabilité de 1. La fonction de récompense ρ_M est définie de la manière suivante :

$$\begin{aligned} \forall (x, u) \in X \times U, \rho_M(x, u, 1) &= 2.0 \\ \forall (x, u) \in X \times U, \rho_M(x, u, 5) &= 10.0 \\ \forall (x, u) \in X \times U, y \in \{2, 3, 4\}, \rho_M(x, u, y) &= 0.0. \end{aligned}$$

Dans ce contexte, un MDP est entièrement défini par sa matrice de transition. Nous définissons donc une distribution de probabilité sur les ensembles de matrices de transition possibles via des distributions de Dirichlet Multinomiales Plates (FDM), largement utilisées dans l'A/R bayésien, notamment en raison de la simplicité des mises à jour de Bayes. La densité d_{FDM} d'une telle distribution s'écrit :

$$d_{FDM}(\mu; \boldsymbol{\theta}) = \prod_{x,u} D(\mu_{x,u}; \theta_{x,u})$$

où $D(\cdot; \cdot)$ sont des distributions de Dirichlet indépendantes. Le paramètre $\boldsymbol{\theta}$ représente l'ensemble des compteurs d'observation des transitions $\theta^t_{x,u}$ jusqu'à l'instant t, incluant $\theta^0_{x,u}$, les observations a priori.

La densité de $p_{\mathcal{M}}(\cdot)$ est dès lors définie comme :

$$d_{p_{\mathcal{M}}(\cdot)}(\mu, \boldsymbol{\theta}) = d_{FDM}(\mu; \boldsymbol{\theta})$$

Une distribution de MDP est donc paramétrée par $\boldsymbol{\theta}$, et sera dénotée par $p^{\boldsymbol{\theta}}(\cdot)$ par la suite. Ci-dessous, nous introduisons quatre distributions de MDPs, la distribution « Generalized Chain », la distribution « Optimistic Generalized Chain », la distribution « Pessimistic Generalized Chain » et la distribution « Uniform ».

3.2.1 Distribution « Generalized Chain »

Cette distribution de MDP est une généralisation du « Chain MDP ». Pour chaque action, deux résultats sont possibles :

— L'agent se déplace de l'état x à l'état $x + 1$ (ou reste à l'état x quand $x = 5$) ; ou

— L'agent « glisse » et retourne à l'état initial.

Les probabilités associées à ces deux possibilités sont tirées uniformément. Formellement, $\boldsymbol{\theta}^{GC}$ caractérisant la distribution $p^{\boldsymbol{\theta}^{GC}}(\cdot)$ est défini ainsi :

$$\begin{aligned} \forall u \in U : \theta^{GC}_{1,u} &= [1, 1, 0, 0, 0] \\ \forall u \in U : \theta^{GC}_{2,u} &= [1, 0, 1, 0, 0] \\ \forall u \in U : \theta^{GC}_{3,u} &= [1, 0, 0, 1, 0] \\ \forall u \in U : \theta^{GC}_{4,u} &= [1, 0, 0, 0, 1] \\ \forall u \in U : \theta^{GC}_{5,u} &= [1, 0, 0, 0, 1] \end{aligned}$$

3.2.2 Distribution « Optimistic Generalized Chain »

Cette distribution est une version alternative de « Generalized Chain MDPs » où nous avons utilisé des poids plus élevés pour les transitions permettant à l'agent d'avancer dans la chaîne. Formellement, $\boldsymbol{\theta}^{OGC}$ caractérisant la distribution $p^{\boldsymbol{\theta}^{OGC}}(\cdot)$ est défini ainsi :

$$\begin{aligned} \forall u \in U : \theta^{OGC}_{1,u} &= [1, 5, 0, 0, 0] \\ \forall u \in U : \theta^{OGC}_{2,u} &= [1, 0, 5, 0, 0] \\ \forall u \in U : \theta^{OGC}_{3,u} &= [1, 0, 0, 5, 0] \\ \forall u \in U : \theta^{OGC}_{4,u} &= [1, 0, 0, 0, 5] \\ \forall u \in U : \theta^{OGC}_{5,u} &= [1, 0, 0, 0, 5] \end{aligned}$$

3.2.3 Distribution « Pessimistic Generalized Chain »

Cette distribution est une version alternative de « Generalized Chain MDPs » où nous avons utilisé des poids plus élevés pour les transitions ramenant l'agent à l'état initial. Formellement, $\boldsymbol{\theta}^{PGC}$, caractérisant la distribution $p^{\boldsymbol{\theta}^{PGC}}(\cdot)$ est défini ainsi :

$$\forall u \in U : \theta^{PGC}_{1,u} = [5, 1, 0, 0, 0]$$
$$\forall u \in U : \theta^{PGC}_{2,u} = [5, 0, 1, 0, 0]$$
$$\forall u \in U : \theta^{PGC}_{3,u} = [5, 0, 0, 1, 0]$$
$$\forall u \in U : \theta^{PGC}_{4,u} = [5, 0, 0, 0, 1]$$
$$\forall u \in U : \theta^{PGC}_{5,u} = [5, 0, 0, 0, 1]$$

3.2.4 Distribution « Uniform »

Les probabilités associées à toutes les transitions sont tirées uniformément. Formellement, $\boldsymbol{\theta}^{U}$ caractérisant la distribution $p^{\boldsymbol{\theta}^{U}}(\cdot)$, est défini ainsi :

$$\forall x \in X, u \in U : \theta^{U}_{x,u} = [1, 1, 1, 1, 1]$$

Enfin, notons que, contrairement au « Chain MDP » original, pour lequel l'action 1 est optimale indépendamment de l'état, le comportement optimal pour les MDPs tirés à partir de ces distributions n'est pas défini a priori : celui-ci varie d'un MDP à l'autre.

3.3 Résultats expérimentaux

Nous présentons plusieurs expériences dans lesquelles nous avons considéré différentes associations de distributions a priori /distributions de test.

Pour OPPS, nous avons considéré quatre espaces de stratégies. L'ensemble des variables, des opérateurs et des constantes ont été fixés une fois pour toute. Les quatre espaces de stratégies ne diffèrent que par la taille maximale des formules pouvant être construites à partir de ces éléments. Nous les avons dénotés par $\mathbb{F}_n$ où n représente la taille maximale des formules correspondant à cet espace de stratégies. L'implémentation d'OPPS utilisée dans ces expériences diffère de celle utilisée dans [3] par le choix de l'ensemble des variables. Ces variables sont décrites dans l'annexe A.

Quant à BAMCP, nous avons utilisé les paramètres par défaut suggérés par Guez et al. dans [13]. Nous avons construit plusieurs instances de BAMCP en considérant différentes valeurs pour le nombre de nœuds créés à chaque pas de temps. Ce paramètre est noté K.

Nous présentons nos expériences en quatre parties, une pour chaque distribution de test possible. Dans chaque partie, nous présentons un tableau reprenant les résultats expérimentaux obtenus lorsque la distribution a priori et la distribution de test sont identiques, en comparant les différents algorithmes en termes de performances et de budgets temps minimaux requis en hors-ligne et en en-ligne. Nous avons également joint une figure comparant les performances de ces mêmes approches pour d'autres distributions a priori.

3.3.1 Distribution de test « Generalized Chain »

Agent	Budget temps hors-ligne	Budgets temps en-ligne	Score moyen
OPPS ($\mathbb{F}_3$)	$\sim$ 6h	$\sim$ 40ms	42.29 ± 0.45
OPPS ($\mathbb{F}_4$)	$\sim$ 6h	$\sim$ 42ms	41.89 ± 0.41
OPPS ($\mathbb{F}_5$)	$\sim$ 6h	$\sim$ 42ms	41.89 ± 0.41
BAMCP ($K = 1$)	$\sim$ 1ms	$\sim$ 7ms	31.71 ± 0.23
BAMCP ($K = 10$)	$\sim$ 1ms	$\sim$ 54ms	33.23 ± 0.26
BAMCP ($K = 25$)	$\sim$ 1ms	$\sim$ 136ms	33.26 ± 0.26
BAMCP ($K = 50$)	$\sim$ 1ms	$\sim$ 273ms	33.73 ± 0.26
BAMCP ($K = 100$)	$\sim$ 1ms	$\sim$ 549ms	33.99 ± 0.27
BAMCP ($K = 250$)	$\sim$ 1ms	$\sim$ 2s	34.02 ± 0.26
BAMCP ($K = 500$)	$\sim$ 1ms	$\sim$ 3s	34.27 ± 0.26

TABLE 1 – Comparaison avec la distribution a priori « Generalized Chain » sur la distribution de test « Generalized Chain »

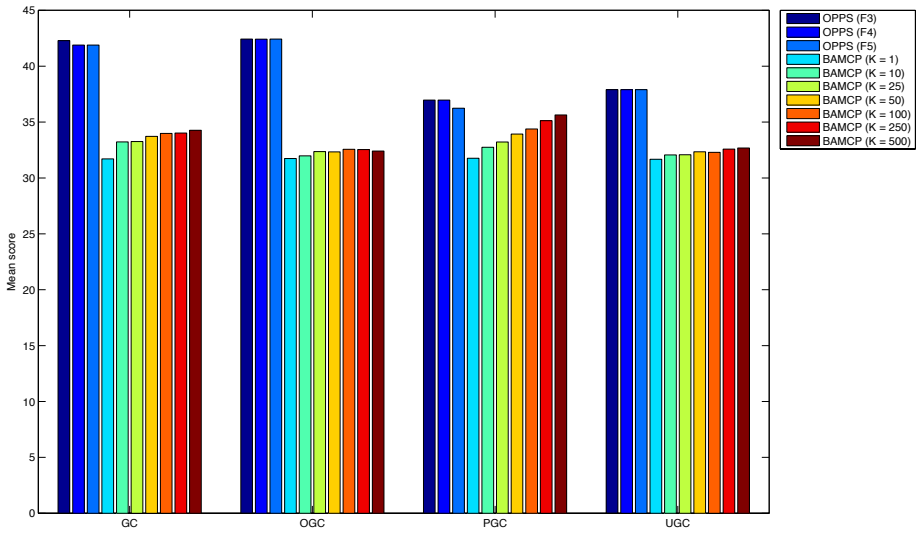

FIGURE 1 – Comparaison sur la distribution de test « Generalized Chain »

La tableau 1 montre que OPPS surclasse BAMCP dans tous les cas de figure, même pour des budgets temps en-ligne très élevés. Le choix de la distribution a priori a un impact significatif sur les performances de OPPS, comme le montre la figure 1. Les performances de OPPS sont similaires pour les distributions a priori « Generalized Chain » et « Optimistic Generalized Chain », mais se dégradent pour la distribution a priori « Uniform ». De son côté, BAMCP est relativement constant sauf lorsqu'il utilise la distribution a priori « Pessimistic Generalized Chain », qui a un impact positif sur ses performances, contrairement à OPPS.

3.3.2 Distribution de test « Optimistic Generalized Chain »

Agent	Budget temps hors-ligne	Budgets temps en-ligne	Score moyen
OPPS ($\mathbb{F}_3$)	$\sim$ 6h	$\sim$ 44ms	110.48 ± 0.61
OPPS ($\mathbb{F}_4$)	$\sim$ 6h	$\sim$ 44ms	110.51 ± 0.61
OPPS ($\mathbb{F}_5$)	$\sim$ 6h	$\sim$ 45ms	110.48 ± 0.61
BAMCP ($K = 1$)	$\sim$ 1ms	$\sim$ 7ms	92.71 ± 0.58
BAMCP ($K = 10$)	$\sim$ 1ms	$\sim$ 56ms	93.97 ± 0.57
BAMCP ($K = 25$)	$\sim$ 1ms	$\sim$ 138ms	94.24 ± 0.58
BAMCP ($K = 50$)	$\sim$ 1ms	$\sim$ 284ms	94.31 ± 0.57
BAMCP ($K = 100$)	$\sim$ 1ms	$\sim$ 555ms	94.59 ± 0.57
BAMCP ($K = 250$)	$\sim$ 1ms	$\sim$ 2s	95.06 ± 0.57
BAMCP ($K = 500$)	$\sim$ 1ms	$\sim$ 3s	95.27 ± 0.58

TABLE 2 – Comparaison avec la distribution a priori « Optimistic Generalized Chain » sur la distribution de test « Optimistic Generalized Chain »

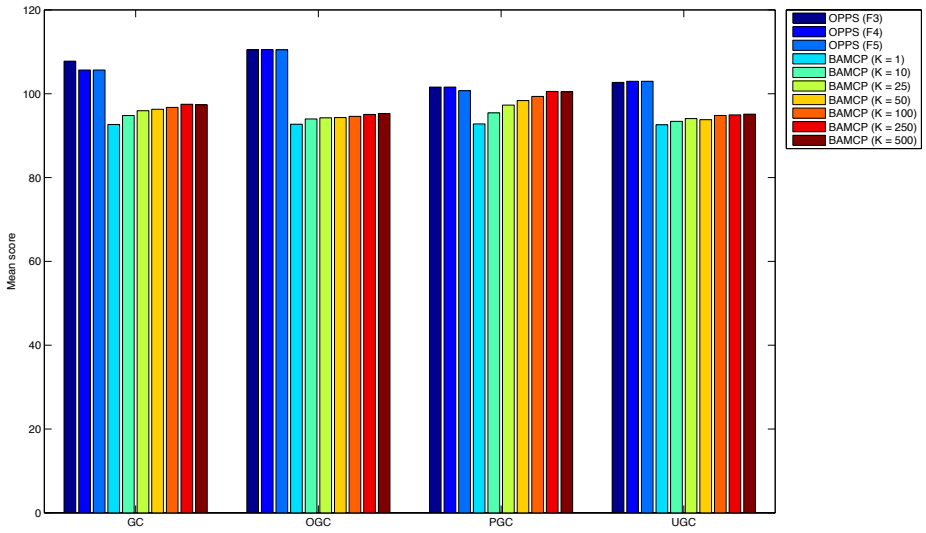

FIGURE 2 – Comparaison sur la distribution de test “Optimistic Generalized Chain"

Le tableau 2 montre que OPPS surclasse clairement BAMCP, même pour des budgets temps en-ligne élevés. Cependant, sur la figure 2, nous pouvons remarquer que BAMCP devient plus compétitif lorsqu'il utilise la distribution a priori « Pessimistic Generalized Chain ». Dans ce cas de figure, les performances de BAMCP sont très proches de celles de OPPS.

3.3.3 Distribution de test « Pessimistic Generalized Chain »

Agent	Budget temps hors-ligne	Budgets temps en-ligne	Score moyen
OPPS ($\mathbb{F}_3$)	$\sim$ 5h	$\sim$ 37ms	35.89 ± 0.06
OPPS ($\mathbb{F}_4$)	$\sim$ 5h	$\sim$ 39ms	35.89 ± 0.06
OPPS ($\mathbb{F}_5$)	$\sim$ 5h	$\sim$ 38ms	35.83 ± 0.06
BAMCP ($K = 1$)	$\sim$ 1ms	$\sim$ 6ms	33.77 ± 0.07
BAMCP ($K = 10$)	$\sim$ 1ms	$\sim$ 54ms	33.97 ± 0.06
BAMCP ($K = 25$)	$\sim$ 1ms	$\sim$ 133ms	34.1 ± 0.06
BAMCP ($K = 50$)	$\sim$ 1ms	$\sim$ 265ms	34.21 ± 0.06
BAMCP ($K = 100$)	$\sim$ 1ms	$\sim$ 536ms	34.37 ± 0.06
BAMCP ($K = 250$)	$\sim$ 1ms	$\sim$ 2s	34.62 ± 0.06
BAMCP ($K = 500$)	$\sim$ 1ms	$\sim$ 3s	34.9 ± 0.06

TABLE 3 – Comparaison avec la distribution a priori « Pessimistic Generalized Chain » sur la distribution de test « Pessimistic Generalized Chain »

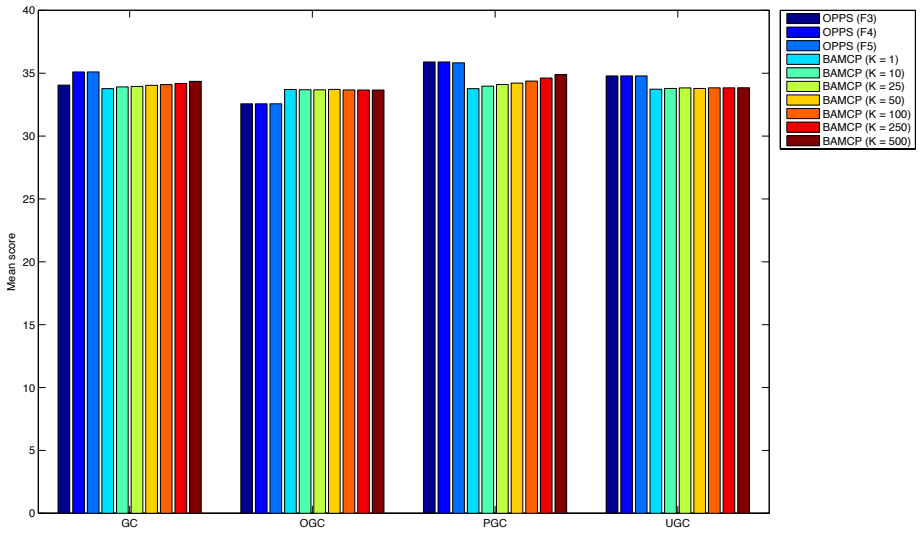

FIGURE 3 – Comparaison sur la distribution de test “Pessimistic Generalized Chain"

Le tableau 3 montre que OPPS et BAMCP ont des performances similaires, même si OPPS reste devant. Néanmoins, BAMCP a besoin de budgets temps en-ligne 80 fois plus importants que OPPS, pour un score légèrement inférieur. Comme le montre la figure 3, BAMCP est systématiquement battu sauf pour la distribution a priori « Optimistic Generalized Chain », où ce dernier surclasse OPPS.

3.3.4 Distribution de test « Uniform »

Agent	Budget temps hors-ligne	Budgets temps en-ligne	Score moyen
OPPS ($\mathbb{F}_3$)	$\sim$ 8h	$\sim$ 52ms	57.37 ± 0.38
OPPS ($\mathbb{F}_4$)	$\sim$ 8h	$\sim$ 53ms	57.37 ± 0.38
OPPS ($\mathbb{F}_5$), UGC)	$\sim$ 8h	$\sim$ 51ms	57.37 ± 0.38
BAMCP ($K = 1$)	$\sim$ 1ms	$\sim$ 6ms	47.92 ± 0.29
BAMCP ($K = 10$)	$\sim$ 1ms	$\sim$ 52ms	48.81 ± 0.3
BAMCP ($K = 25$)	$\sim$ 1ms	$\sim$ 132ms	48.95 ± 0.3
BAMCP ($K = 50$)	$\sim$ 1ms	$\sim$ 256ms	49.3 ± 0.3
BAMCP ($K = 100$)	$\sim$ 1ms	$\sim$ 521ms	49.39 ± 0.31
BAMCP ($K = 250$)	$\sim$ 1ms	$\sim$ 2s	50.08 ± 0.31
BAMCP ($K = 500$)	$\sim$ 1ms	$\sim$ 3s	50.06 ± 0.31

TABLE 4 – Comparaison avec la distribution a priori « Uniform » sur la distribution de test « Uniform »

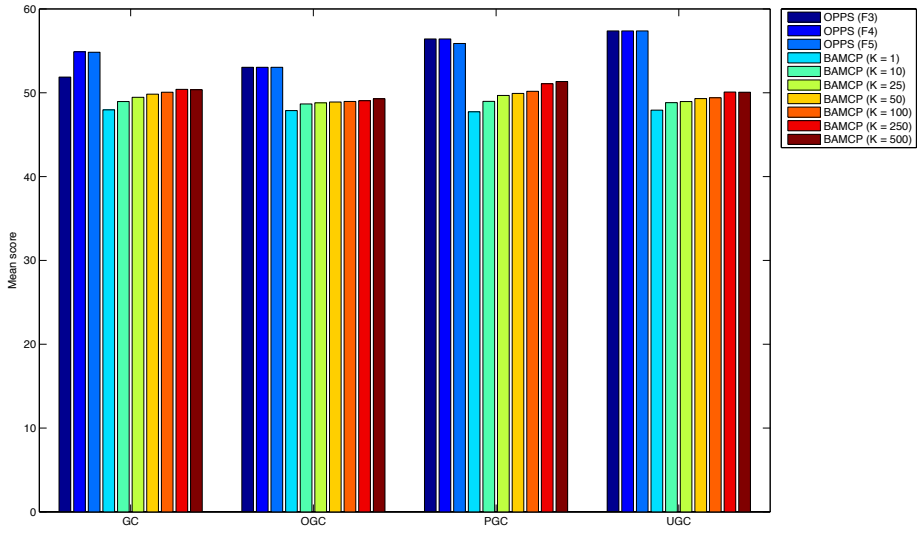

FIGURE 4 – Comparaison sur la distribution de test “Uniform"

Le tableau 4 et la figure 4 montrent une claire victoire de OPPS pour n’importe quelle distribution a priori, même avec des budgets temps en-ligne élevés. Nous pouvons également remarquer que OPPS est plus efficace lorsqu’il utilise la bonne distribution a priori.

4 Discussion

De manière générale, nous constatons que OPPS est meilleur que BAMCP, même pour des budgets temps en-ligne élevés, au prix de plusieurs heures de calculs hors-ligne. Cependant, nous pouvons remarquer que BAMCP lui a tenu tête à plusieurs reprises lorsque nous utilisions la distribution a priori « Pessimistic Generalized Chain ».

Quant à la précision de la distribution a priori, il apparait au regard des résultats obtenus qu'utiliser une distribution a priori différente de la distribution de test a un impact négatif sur les performances de OPPS, ce qui n'est pas étonnant dans la mesure où OPPS effectue une recherche de politique en se basant sur celle-ci. Cet impact est d'autant plus important dans le cas de distributions de test serrées (« Optimistic Generalized Chain » et « Pessimistic Generalized Chain »). De son côté, BAMCP semble moins affecté par l'imprécision de la distribution a priori, probablement grâce à la mise à jour de sa distribution a posteriori.

5 Conclusion

Nous avons présenté une comparaison expérimentale entre deux approches bayésiennes différentes, exploitant intensivement la distribution a priori durant soit la phase hors-ligne (OPPS), soit la phase en-ligne (BAMCP) pour interagir efficacement avec un MDP inconnu. Nos expériences suggèrent que : (i) exploiter une distribution a priori durant une phase hors-ligne n'est jamais une mauvaise idée, tandis que (ii) maintenir une distribution a posteriori peut réduire l'impact des erreurs de précision de la distribution a priori sur les performances de l'agent.

Remerciements

Michaël Castronovo est financé par le F.R.S.-FNRS (FRIA - Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture) et remercie le CECI [2]. Raphaël Fonteneau est Chargé de Recherches F.R.S.-FNRS. Cet article présente des résultats obtenus grâce au Pôle d'Attraction Interuniversitaire (PAI) belge DYSCO (Dynamical Systems, Control and Optimization).

Références

[1] ASMUTH J., LI L., LITTMAN M., NOURI A. & WINGATE D. (2009). A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, p. 19–26 : AUAI Press.

[2] BUŞONIU L., BABUŠKA R., DE SCHUTTER B. & ERNST D. (2010). *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, Florida : CRC Press.

[3] CASTRONOVO M., MAES F., FONTENEAU R. & ERNST D. (2012). Learning exploration/exploitation strategies for single trajectory reinforcement learning. *Journal of Machine Learning Research (JMLR)*, **24**, 1–9.

[4] DEARDEN R., FRIEDMAN N. & ANDRE D. (1999). Model based Bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, p. 150–159 : Morgan Kaufmann.

[5] DEARDEN R., FRIEDMAN N. & RUSSELL S. (1998). Bayesian Q-learning. In *Proceedings of Fifteenth National Conference on Artificial Intelligence (AAAI)*, p. 761–768 : AAAI Press.

[6] ENGEL Y., MANNOR S. & MEIR R. (2003). Bayes meets Bellman : the Gaussian process approach to temporal difference learning. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, p. 154–161.

[7] ENGEL Y., MANNOR S. & MEIR R. (2005a). Reinforcement learning with Gaussian processes. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, p. 201–208.

[8] ENGEL Y., SZABO P. & VOLKINSHTEIN D. (2005b). Learning to control an octopus arm with Gaussian process temporal difference methods. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, p. 347–354 : MIT Press.

2. CECI est l'acronyme de 'Consortium des Equipements de Calcul Intensif' ; un consortium de centres de calculs haute performance de UCL, ULB, ULg, UMons et UNamur.

[9] FARD M. M. & PINEAU J. (2010). PAC-Bayesian model selection for reinforcement learning. In *Neural Information Processing Systems (NIPS)*.

[10] FONTENEAU R., BUSONIU L. & MUNOS R. (2013). Optimistic planning for belief-augmented Markov decision processes. In *International Symposium on Adaptive Dynamic Programming And Reinforcement Learning (IEEE ADPRL)*, p. 77–84 : IEEE.

[11] GHAVAMZADEH M. & ENGEL Y. (2006). Bayesian policy gradient algorithms. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)* : MIT Press.

[12] GHAVAMZADEH M. & ENGEL Y. (2007). Bayesian actor-critic algorithms. In *Proceedings of the Twenty-Fourth International Conference on Machine Learning (ICML)*.

[13] GUEZ A., SILVER D. & DAYAN P. (2012). Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Neural Information Processing Systems (NIPS)*.

[14] HENNIG P., STERN D. & GRAEPEL T. (2009). Bayesian quadratic reinforcement learning. In *Neural Information Processing Systems (NIPS)*.

[15] KOCSIS L. & SZEPESVÁRI C. (2006). Bandit based Monte-Carlo planning. *European Conference on Machine Learning (ECML)*, p. 282–293.

[16] POUPART P. (2008). Model-based Bayesian reinforcement learning in partially observable domains. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.

[17] ROSS S. & PINEAU J. (2008). Model-based Bayesian reinforcement learning in large structured domains. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)* : AUAI Press.

[18] ROSS S., PINEAU J., CHAIB-DRAA B. & KREITMANN P. (2011). A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research (JMLR)*, **12**, 1729–1770.

[19] STRENS M. (2000). A Bayesian framework for reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, p. 943–950 : ICML.

A Description de l'ensemble des variables utilisées dans l'algorithme OPPS

L'implémentation de OPPS utilisée dans cet article diffère de celle introduite dans [3] de part l'ensemble des variables utilisées pour construire les formules. L'ensemble des variables considéré dans cet article est composé de trois variables. Ces trois variables correspondent à trois Q-fonctions calculées par itération sur les valeurs à partir de trois modèles différents. Formellement, à partir de l'ensemble des transitions observées au cours d'un historique h, on note $N_h(x,u)$ le nombre d'occurrences d'une transition prenant son origine à partir de la paire état-action (x,u) dans h, et $N_h(x,u,y)$ le nombre d'occurrences de la transition (x,u,y) dans h. On définit les trois fonctions de transition $f_{mean}, f_{uniform}, f_{self}$ ainsi :

1. f_{mean} correspond à l'espérance de la fonction de transition selon la distribution a posteriori courante. Si $\boldsymbol{\theta}^0$ dénote la liste des compteurs des transitions observées de la distribution a priori $p^0_{\mathcal{M}}(\cdot)$, f_{mean} est défini comme :
$$\forall x,u,y : \theta^h_{x,u}(y) = \theta^0_{x,u}(y) + N_h(x,u,y)$$
Formellement, le modèle de transitions moyen est défini comme :
$$\forall x,u,y : f_{mean}(x,u,y) = \frac{\theta^h_{x,u}(y)}{\sum_{y'} \theta^h_{x,u}(y')}$$

2. $f_{uniform}$ correspond à l'espérance de la fonction de transition selon la distribution a posteriori obtenue à partir de l'historique courant partant d'une distribution a priori uniforme. Formellement, le modèle de transitions uniforme est défini comme :
$$\forall x,u,y : f_{uniform}(x,u,y) = \frac{1 + N_h(x,u,y)}{|U| + N_h(x,u)}$$

3. f_{self} correspond à l'espérance de la function de transition selon la distribution a posteriori courante obtenue à partir de l'historique courant partant avec une initialisation des compteurs correspondant

à une impulsion de Dirac centrée autour d'un MDP déterministe où chaque état n'est atteignable que depuis lui-même (pour toute action). Formellement, la fonction de transition « self » est définie comme :

$$\forall x, u : f_{self}(x, u, x) = \frac{1 + N_h(x, u, x)}{1 + N_h(x, u)}$$
$$\forall x, u, y \neq x : f_{self}(x, u, x) = \frac{N_h(x, u, x)}{1 + N_h(x, u)}.$$

B Erratum

Nous avons commis une erreur dans les expériences que nous avons présentées dans la première version de cet article. Cela nous a conduit à surestimer les performances de l'algorithme BAMCP.

Cette erreur vient du fait que la fonction de récompense utilisée par BAMCP n'était pas égale à $R(x, u, y')$ mais bien à l'espérance de cette fonction de récompense, que nous avons appelée $R'(x, u)$:

$$R'(x, u) = \sum_{y' \in X} P(y'|x, u)\ R(x, u, y')$$

Dans ce contexte, BAMCP fonctionne nettement mieux, sans doute parce que la fonction $R'(x, u)$ donne à l'agent certaines informations sur la matrice de transition.