HAL
open science

BRL Quasi-Optimal à l'aide de Transitions Locales Optimistes

Mauricio Araya, Vincent Thomas, Olivier Buffet

To cite this version:

Mauricio Araya, Vincent Thomas, Olivier Buffet. BRL Quasi-Optimal à l'aide de Transitions Locales Optimistes. Journées Francophones sur la planification, la décision et l'apprentissage pour le contrôle des systèmes - JFPDA 2012, May 2012, Villers-lès-Nancy, France. 16 p. hal-00735602

HAL Id: hal-00735602

https://inria.hal.science/hal-00735602

Submitted on 26 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BRL Quasi-Optimal à l'aide de Transitions Locales Optimistes

Mauricio Araya-López[12], Vincent Thomas[12], Olivier Buffet[21]

Université de Lorraine[1] / INRIA[2]
LORIA – Campus scientifique, BP 239,
54506 Vandœuvre-lès-Nancy CEDEX
{mauricio.araya|vincent.thomas|olivier.buffet}@loria.fr
http://www.loria.fr/∼{maraya|vthomas|buffet}/

Résumé : L'apprentissage par renforcement bayésien basé modèle (BRL) permet une formalisation saine du problème consistant à agir optimalement face à un environnement inconnu, c'est-à-dire en évitant le dilemme exploration-exploitation. Toutefois, les algorithmes s'attaquant explicitement au BRL souffrent d'une telle explosion combinatoire qu'un grand nombre de travaux repose sur des algorithmes heuristiques. Cet article introduit BOLT, un algorithme heuristique simple et (presque) déterministe pour le BRL qui est optimiste vis à vis de la fonction de transition. Nous analysons la complexité d'échantillon de BOLT et montrons que, pour certains paramètres, l'algorithme est quasi-optimal au sens bayésien avec une grande probabilité. Puis, des résultats expérimentaux mettent en valeur les principales différences entre cette méthode et des travaux antérieurs.

1 Introduction

Agir en environnement inconnu requiert de faire un compromis entre *exploration* (agir de manière à acquérir de la connaissance) et *exploitation* (agir de manière à maximiser le retour espéré). Les algorithmes *d'apprentissage par renforcement bayésien à base de modèle* (BRL) accomplissent cela en maintenant et en utilisant une distribution de probabilité sur les modèles possibles (ce qui requiert une connaissance experte sous la forme d'une distribution a priori). Ces algorithmes tombent typiquement dans l'une des trois classes suivantes (Asmuth *et al.*, 2009).

Les approches **belief-lookahead** essayent de faire un compromis optimal entre exploration et exploitation en reformulant le BRL comme le problème de résoudre un POMDP dans lequel l'état est le couple $\boldsymbol{\omega} = (s, \boldsymbol{b})$, s étant l'état observé et $\boldsymbol{b}$ la distribution sur les modèles possibles (Duff, 2002) ; toutefois, ce problème est insoluble en pratique, ne permettant que des solutions approchées coûteuses en temps de calcul (Poupart *et al.*, 2006)(Dimitrakakis, 2008).

Les approches **optimistes** proposent des mécanismes d'exploration qui tentent explicitement de réduire l'incertitude sur le modèle (Brafman & Tennenholtz, 2003; Kolter & Ng, 2009; Sorg *et al.*, 2010; Asmuth *et al.*, 2009) en reposant sur le principe de "l'optimisme face à l'incertitude".

Les approches **non dirigées**, telles que les stratégies d'exploration ϵ-gloutonne ou boltzmannienne (Sutton & Barto, 1998), effectuent des actions exploratoires indépendamment de la connaissance actuelle de l'environnement.

Nous nous concentrons ici sur les approches optimistes et, comme la plupart des recherches dans le domaine et sans perte de généralité, nous considérons uniquement l'incertitude sur la fonction de transition, faisant l'hypothèse que la fonction de récompense est connue. Des travaux récents prouvent que certains algorithmes sont soit PAC-MDP – avec une grande probabilité ils agissent souvent comme une politique optimale le ferait (si le modèle MDP était connu) – ou PAC-BAMDP – avec une grande probabilité ils agissent souvent comme un algorithm belief-lookahead optimal le ferait.

Cet article présente d'abord le contexte sur le BRL basé modèle en section 2, et sur les analyses PAC-MDP et PAC-BAMDP en section 3. Ensuite, la section 4 introduit un nouvel algorithme, BOLT, qui (1), comme BOSS (Asmuth *et al.*, 2009), est optimiste à propos du modèle de transition – ce qui est intuitivement satisfaisant puisque l'incertitude est sur ce modèle – et (2), comme BEB (Kolter & Ng, 2009), est (presque) déterministe – ce qui conduit à un meilleur contrôle sur cette approche. Nous démontrons ensuite

en section 5 que BOLT est PAC-BAMDP pour des horizons infinis, en généralisant les résultats antérieurs connus pour BEB en horizon fini. Des expérimentations dans la section 6 éclairent ensuite le comportement de ces algorithmes en pratique, montrant en particulier que BOLT parait moins sensible au réglage des paramètres que BEB. Du fait de contraintes d'espace et de clareté, nous avons reporté les preuves des lemmes en annexe, et avons conservé les preuves principales des théorèmes dans le corps de l'article.

2 Contexte

2.1 Apprentissage par renforcement

Un *processus de décision markovien* (MDP) (Puterman, 1994) est défini par un uplet $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ où $\mathcal{S}$ est un ensemble fini d'*états*, $\mathcal{A}$ est un ensemble fini d'*actions*, la fonction de *transition* T donne la probabilité de passer de l'état s à l'état s' quand une action a est effectuée : $T(s, a, s') = Pr(s'|s, a)$, et $R(s, a, s')$ est la *récompense* scalaire instantanée obtenue pendant cette transition. L'apprentissage par renforcement (RL pour *Reinforcement Learning*) (Sutton & Barto, 1998) est le problème consistant à trouver une politique de décision optimale – une application $\pi : \mathcal{S} \mapsto \mathcal{A}$ – quand le modèle (T sans R dans notre cas) est inconnu mais en interagissant avec le système. Un critère de performance typique est l'espérance du retour γ-pondéré :

$$V^{\pi}_{\boldsymbol{\mu}}(s) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \mid s_0 = s\right],$$

où $\boldsymbol{\mu} \in \mathcal{M}$ est le modèle inconnu et $\gamma \in [0, 1]$ est un facteur d'atténuation. Dans cet article nous ne tenons pas compte de l'incertitude à propos de la fonction de récompense puisque celle-ci peut-être transformée en une incertitude sur la fonction de transition, signifiant que seule T est inconnue et que R est donnée. Sous une politique optimale, cette fonction de valeur d'état vérifie l'équation d'optimalité de Bellman (Bellman, 1954) (pour tout $s \in \mathcal{S}$) :

$$V^{*}_{\boldsymbol{\mu}}(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s, a, s') \left[R(s, a, s') + \gamma V^{*}_{\boldsymbol{\mu}}(s')\right],$$

et calculer cette fonction de valeur optimale permet de dériver une politique optimale en se comportant de manière gloutonne, c'est-à-dire en prenant les actions dans l'ensemble $\arg\max_{a \in \mathcal{A}} Q^{*}_{\boldsymbol{\mu}}(s, a)$, où la fonction de valeur état-action $Q^{*}_{\boldsymbol{\mu}}$ est définie par

$$Q^{*}_{\boldsymbol{\mu}}(s, a) = \sum_{s' \in \mathcal{S}} T(s, a, s') \left[R(s, a, s') + \gamma V^{*}_{\boldsymbol{\mu}}(s')\right].$$

Les algorithmes de RL typiques soit (i) estiment directement la fonction de valeur état-action $Q^{*}_{\boldsymbol{\mu}}$ (RL sans modèle), ou (ii) apprennent T pour calculer $V^{*}_{\boldsymbol{\mu}}$ ou $Q^{*}_{\boldsymbol{\mu}}$ (RL avec modèle). Toutefois, dans les deux cas, une difficulté majeure est de choisir des actions de manière à faire un compromis entre l'exploitation de la connaissance courante et l'exploration pour acquérir plus de connaissances.

2.2 RL bayésien basé modèle

Nous considérons ici l'*apprentissage par renforcement bayésien basé modèle* (Strens, 2000), c'est-à-dire le RL basé modèle où la connaissance sur le modèle est représentée à l'aide d'une distribution de probabilité $\boldsymbol{b}$ sur tous les modèles de transition possibles. Une distribution a priori initiale $\boldsymbol{b}_0 = Pr(\boldsymbol{\mu})$ doit être spécifiée, laquelle est ensuite mise à jour en utilisant la règle de Bayes. Au temps t la distribution a posteriori $\boldsymbol{b}_t$ dépend de la distribution initiale b_0 et de l'historique état-action jusqu'ici $h_t = s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t$. Cette mise à jour peut être appliquée séquentiellement grâce à l'hypothèse de Markov qui dit que, au temps $t + 1$, on n'a besoin que de $\boldsymbol{b}_t$ et du triplet (s_t, a_t, s_{t+1}) pour calculer la nouvelle distribution :

$$\boldsymbol{b}_{t+1} = Pr(\boldsymbol{\mu}|h_{t+1}, \boldsymbol{b}_0) = Pr(\boldsymbol{\mu}|s_t, a_t, s_{t+1}, \boldsymbol{b}_t). \tag{1}$$

Cette distribution $\boldsymbol{b}_t$ est connue comme la *croyance* (*belief*) sur le modèle, et résume l'information qui a été collectée à propos du modèle jusqu'au pas de temps courant.

Si on considère la croyance comme faisant partie de l'état, le belief-MDP résultant sur cet espace d'état infini multi-varié peut, en théorie, être résolu optimalement. Toutefois c'est en pratique insoluble du fait de la complexité croissante avec l'horizon de planification. De manière remarquable, modéliser des problèmes de RL comme des belief-MDP fournit une façon saine de traiter le dilemme exploration-exploitation, parce que ces deux objectifs sont naturellement inclus dans le même critère d'optimisation.

L'état de croyance peut ainsi être écrit comme $\boldsymbol{\omega} = (s, \boldsymbol{b})$, ce qui définit un MDP *Bayes-Adaptive* (BAMDP) [1] (Duff, 2002), un type particulier de belief-MDP où l'état de croyance est factorisé en l'état (visible) du système et la croyance sur le modèle (caché). De plus, du fait de l'intégration sur tous les modèles possibles dans la fonction de valeur du BAMDP, la fonction de transition $T(\omega, a, \omega')$ est donnée par

$$Pr(\boldsymbol{\omega}'|\boldsymbol{\omega}, a) = Pr(\boldsymbol{b}'|\boldsymbol{b}, s, a, s')E[Pr(s'|s, a)|\boldsymbol{b}],$$

où la première probabilité vaut 1 si $\boldsymbol{b}'$ satisfait l'équation (1) et 0 sinon. La politique bayésienne optimale peut alors être obtenue en calculant la fonction de valeur bayésienne optimale (Duff, 2002; Poupart *et al.*, 2006) :

$$\begin{aligned}\mathbb{V}^*(s, \boldsymbol{b}) &= \max_a \left[\sum_{s'} E[Pr(s'|s, a)|\boldsymbol{b}](R(s, a, s') + \gamma \mathbb{V}^*(s', \boldsymbol{b}'))\right] \\ &= \max_a \left[\sum_{s'} T(s, a, s', \boldsymbol{b})(R(s, a, s') + \gamma \mathbb{V}^*(s', \boldsymbol{b}'))\right], \qquad (2)\end{aligned}$$

avec $\boldsymbol{b}'$ la distribution a posteriori après la mise à jour de Bayes avec (s, a, s'). Pour le cas de l'horizon fini nous pouvons utiliser le même raisonnement, de sorte que la valeur optimale peut être calculée en théorie pour un horizon fini ou infini en effectuant des mises à jour de Bayes et en calculant des espérances. Pourtant, en pratique, calculer cette fonction de valeur de façon exacte est insoluble du fait du grand facteur de branchement dans le développement de l'arbre.

A première vue, une piste naturelle à suivre consiste à considérer des algorithmes pour POMDP, parce qu'ils traitent aussi des belief-MDP. Malheureusement, on ne peut pas bénéficier directement des algorithmes pour POMDP classiques à cause du nombre infini de dimensions de l'espace d'états. D'autres méthodes approchées – hors-ligne ou en-ligne – ont été introduites pour cette raison, permettant dans un certain nombre de cas de prouver des propriétés théoriques.

Ici, nous sommes intéressés par des approches heuristiques suivant le principe de *l'optimisme face à l'incertain*, lequel consiste à supposer un retour plus important sur les transitions les plus incertaines. Certaines d'entre elles résolvent le MDP généré par le modèle espéré (à une certaine étape) avec une récompense d'exploration supplémentaire qui favorise les transitions dont les modèles sont moins connus, comme dans R-MAX (Brafman & Tennenholtz, 2003), BEB (Kolter & Ng, 2009), ou des récompenses basées sur la variance (Sorg *et al.*, 2010). Une autre approche, utilisée dans BOSS (Asmuth *et al.*, 2009), est de résoudre, quand le modèle a changé suffisamment, une estimation optimiste du vrai MDP (obtenue en fusionnant plusieurs modèles échantillonnés).

2.3 A priori plats et structurés

Le choix d'une distribution a priori appropriée est un problème important dans les algorithmes BRL, parce qu'il a un impact direct sur la qualité de la solution et le temps de calcul. Une approche naïve est de considérer une distribution de Dirichlet indépendante pour chaque couple état-action, ce qui est connu comme la distribution *Flat-Dirichlet-Multinomial* (FDM), et dont la densité est définie comme

$$\boldsymbol{b} = f(\boldsymbol{\mu}; \boldsymbol{\theta}) = \prod_{s,a} D(\boldsymbol{\mu}_{s,a}; \boldsymbol{\theta}_{s,a}),$$

où $D(\cdot;\cdot)$ sont des distributions de Dirichlet indépendantes. Les FDM peuvent être utilisées pour modéliser n'importe quelle distribution de MDP à états et actions discrètes, mais ne sont appropriées que sous l'hypothèse forte d'indépendance. Toutefois, cette distribution a priori a été largement utilisée du fait de sa simplicité pour calculer des mises à jour bayésienne et la valeur espérée. Considérons que les paramètres du

1. BAMDP signifie aussi *Belief-Augmented* MDP (Dimitrakakis, 2008).

vecteur $\boldsymbol{\theta}$ sont les compteurs des transitions observées, alors la valeur espérée de la probabilité de transition est $E[Pr(s'|s,a)|\boldsymbol{b}] = \frac{\boldsymbol{\theta}_{s,a}(s')}{\sum_{s''}\boldsymbol{\theta}_{s,a}(s'')}$, et la mise à jour de Bayes, après une transition (s,a,s'), est réduite à $\boldsymbol{\theta}'_{s,a}(s') = \boldsymbol{\theta}_{s,a}(s') + 1$.

Même si les FDM sont utiles pour analyser et évaluer les algorithmes, en pratique ils sont inefficaces parce qu'ils n'exploitent pas l'information sur la structure du problème. On peut par exemple encoder le fait que de multiples actions partagent le même modèle en factorisant plusieurs distributions de Dirichlet, ou permettre à l'algorithme d'identifier de telles structures en utilisant des distributions de Dirichlet combinées avec des processus de restaurant chinois ou des processus de buffet indien (Asmuth *et al.*, 2009). Aussi, des problèmes spécifiques peuvent mener à d'autres familles de distributions différentes de celles de Dirichlet, par exemple la distribution bayésienne sur des MDP déterministes présentée par Sorg *et al.* (2010).

3 Algorithmes PAC

L'apprentissace probablement approximativement correct (PAC) fournit une manière d'analyser la qualité des algorithmes d'apprentissage (Valiant, 1984). L'idée générale est que, avec une probabilité élevée $1-\delta$ (probablement), une machine avec une erreur d'apprentissage produit une erreur en généralisation faible bornée par ϵ (approximativement correct). Si le nombre de pas requis pour atteindre cette condition est majoré par une fonction polynomiale, alors l'algorithme est PAC-efficient.

3.1 Analyse PAC-MDP

En RL, la propriété PAC-MDP (Strehl *et al.*, 2009) garantit qu'un algorithme génère une politique ϵ-proche avec probabilité $1-\delta$ sauf pour un nombre polynomial de pas de temps. La différence principale avec l'apprentissage PAC est qu'il n'y a pas de garantie concernant les instants où les pas non-ϵ-proches vont avoir lieu, mais le nombre de ces pas est majoré par un polynôme. Pour vérifier la propriété PAC-MDP, trois conditions doivent usuellement être satisfaites. D'abord, l'algorithme doit utiliser au moins des valeurs proches d'être *optimistes* avec une grande probabilité. L'algorithme doit aussi garantir avec une grande probabilité qu'il est *précis*, c'est-à-dire que, pour des parties connues du modèle, son évaluation réelle sera ϵ-proche de la fonction de valeur optimale. Finalement, le nombre de pas non-ϵ-proches (aussi appelé *complexité d'échantillon*) doit être majoré par une fonction *polynomiale*. Ces trois exigences sont des conditions nécessaires pour le théorème général PAC-MDP (théorème 10 dans (Strehl *et al.*, 2009)).

En termes mathématiques, les algorithmes PAC-MDP sont ceux pour lesquels, avec probabilité $1-\delta$, l'évaluation de la politique $\boldsymbol{A}_t$, générée par l'algorithme $\boldsymbol{A}$ au temps t sur le modèle sous-jacent réel $\boldsymbol{\mu}_0$, est ϵ-proche de la politique optimale sur le même modèle sauf pour un nombre polynomial de pas de temps :

$$V^{\boldsymbol{A}_t}_{\boldsymbol{\mu}_0}(s) \geq V^*_{\boldsymbol{\mu}_0}(s) - \epsilon. \tag{3}$$

Plusieurs algorithmes de RL vérifient la propriété PAC-MDP, différent les uns des autres principalement par la finesse du majorant de la complexité en échantillon. Par exemple, R-MAX (Brafman & Tennenholtz, 2003), MBIE-EB (Strehl & Littman, 2005) et Delayed Q-Learning (Strehl *et al.*, 2009) sont quelques algorithmes de RL classiques pour lesquels cette propriété a été prouvée, alors que BOSS (Asmuth *et al.*, 2009) est un algorithme de RL bayésien qui est aussi PAC-MDP.

Dans l'analyse PAC-MDP, la politique produite par un algorithme devrait être proche de la politique optimale dérivée du modèle MDP réel sous-jacent. Toutefois, cette politique *utopique* (Poupart *et al.*, 2006) ne peut être calculée, parce qu'il est impossible d'apprendre le modèle exact avec un nombre fini d'échantillons.

Dans le RL basé modèle, le dilemme exploration-exploitation apparaît comme le compromis entre produire des estimations précises du modèle réel (exploration) et agir optimalement par rapport aux estimations courantes (exploitation). Ainsi, la correction d'un algorithme dépendra de ces deux critères différents, lesquels ne peuvent être combinés sans biaiser les résultats vers certains modèles.

3.2 Analyse PAC-BAMDP

Une alternative à l'approche PAC-MDP est d'être PAC par rapport à la politique *bayésienne* optimale, plutôt que d'utiliser la politique *utopique* optimale. Nous appellerons cela *l'analyse PAC-BAMDP*, parce qu'elle vise à garantir la proximité à la solution optimale du MDP Bayes-Adaptive. Ce type d'analyse

a été d'abord introduit dans (Kolter & Ng, 2009), sous le nom de propriété *quasi-bayésienne*, où il est montré qu'une version *modifiée* de BEB est PAC-BAMDP dans le cas d'un horizon fini et sans facteur d'atténuation [2].

Définissons maintenant la façon d'évaluer une politique au sens bayésien :

Définition 3.1
L'évaluation bayésienne $\mathbb{V}$ d'une politique π est la valeur espérée étant donnée une distribution sur les modèles $\boldsymbol{b}$:

$$\mathbb{V}^{\pi}(s, \boldsymbol{b}) = E_{\boldsymbol{\mu}}[V^{\pi}_{\boldsymbol{\mu}}(s)|\boldsymbol{b}] = \int_{\mathcal{M}} V^{\pi}_{\boldsymbol{\mu}}(s) Pr(\boldsymbol{\mu}|\boldsymbol{b}) d\boldsymbol{\mu}.$$

Cette définition a déjà été présentée implicitement dans (Duff, 2002) et explicitement dans (Dimitrakakis, 2008), mais il est très important de souligner la différence entre une évaluation MDP normale sur un MDP connu, et l'évaluation bayésienne [3]. Cette définition est cohérente avec l'équation 2, où

$$\begin{aligned}\mathbb{V}^{*}(s, \boldsymbol{b}) &= \max_{\pi} \int_{\mathcal{M}} V^{\pi}_{\boldsymbol{\mu}}(s) Pr(\boldsymbol{\mu}|\boldsymbol{b}) d\boldsymbol{\mu} \\ &= \max_{a} \left[\sum_{s'} E[Pr(s'|s,a)|\boldsymbol{b}](R(s,a,s') + \gamma \mathbb{V}^{*}(s', \boldsymbol{b}')) \right].\end{aligned}$$

Définissons la propriété PAC-BAMDP :

Définition 3.2
On dit qu'un algorithme est PAC-BAMDP si, avec probabilité $1-\delta$, l'évaluation bayésienne –paramétrée par la croyance $\boldsymbol{b}$– d'une politique $\boldsymbol{A}_t$ générée par l'algorithme $\boldsymbol{A}$ au temps t est ϵ-proche de la politique bayésienne optimale sauf pour un nombre polynomial de pas de temps :

$$\mathbb{V}^{\boldsymbol{A}_t}(s, \boldsymbol{b}) \geq \mathbb{V}^{*}(s, \boldsymbol{b}) - \epsilon, \tag{4}$$

avec $\delta \in [0, 1)$ et $\epsilon > 0$.

Une différence conceptuelle majeure est que, dans l'analyse PAC-BAMDP, l'objectif est de garantir la correction approchée parce que la politique bayésienne optimale est difficile à calculer, alors que, dans l'analyse PAC-MDP, la garantie d'une correction approchée est requise parce que la politique utopique optimale est impossible à trouver en un nombre fini de pas.

4 Algorithmes BRL Optimistes

La section 2.2 a montré comment calculer en théorie la fonction de valeur bayésienne optimale. Ce calcul étant trop coûteux, il est courant d'utiliser des algorithmes sous-optimaux – mais efficaces. Une technique populaire est de maintenir une distribution sur le modèle, de choisir un MDP représentatif d'après cette distribution, et d'agir suivant sa fonction de valeur. L'algorithme de base dans cette famille est appelé EXPLOIT (Poupart *et al.*, 2006), et consiste à sélectionner à chaque pas de temps le modèle moyen de $\boldsymbol{b}$. Ainsi, à chaque pas de temps t, l'algorithme doit résoudre un MDP différent d'horizon H – un paramètre de l'algorithme, pas l'horizon du problème – comme on peut le voir sur la figure 1. Nous considérerons pour l'analyse que H est le nombre d'itérations i que *value iteration* effectue à chaque pas de temps t, mais en pratique la convergence peut être atteinte bien avant le H dérivé de la théorie pour le cas à horizon infini.

BEB (Kolter & Ng, 2009) suit la même idée qu'EXPLOIT, mais ajoute un bonus d'exploration à la fonction de récompense. A l'inverse, BOSS (Asmuth *et al.*, 2009) n'utilise pas l'approche EXPLOIT, mais échantillonne différents modèles de la distribution et les utilise pour construire un MDP optimiste. BEB a l'avantage d'être un algorithme quasi-déterministe [4] et ne repose pas sur l'échantillonnage comme BOSS. D'autre part, BOSS est optimiste vis à vis des transitions, là où repose l'incertitude, alors que BEB est optimiste par rapport à la fonction de récompense, alors que cette fonction est connue.

2. Toutefois, quelques erreurs – rectifiables – ont été localisées dans la preuve que BEB est quasi-bayésien dans (Kolter & Ng, 2009), comme discuté avec les auteurs.

3. Nous utilisons une notation différente pour l'évaluation bayésienne, $\mathbb{V}$, pour la distinguer d'une évaluation MDP normale V.

4. Dans le cas de valeurs égales, les actions sont échantillonnées uniformément.

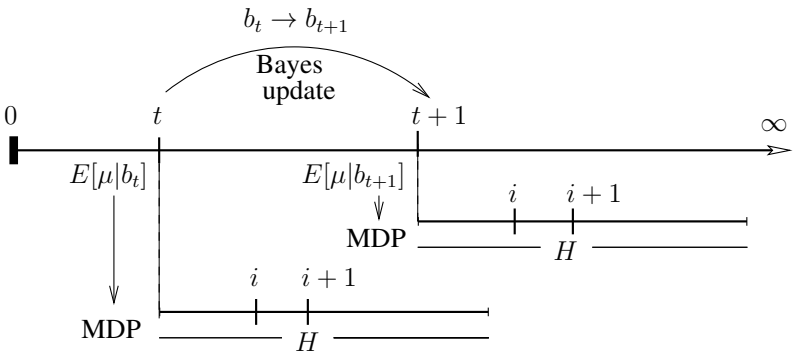

FIGURE 1 – Algorithme de type EXPLOIT. A chaque pas de temps t, l'algorithme effectue une mise à jour bayésienne de la distribution, et résout le MDP dérivé du modèle moyen de la croyance.

4.1 Transitions Locales Optimistes Bayésiennes

Dans cette section, nous introduisons un nouvel algorithme appelé BOLT (*Bayesian Optimistic Local Transitions*), qui repose sur le fait d'agir, à chaque pas de temps t, en suivant une politique optimale pour une variante optimiste du modèle moyen courant. Cette variante est obtenue, pour chaque couple état-action, en renforçant les mises à jour bayésiennes avant de calculer le modèle de transition local moyen. Ceci est accompli en utilisant un nouveau MDP avec un espace d'action augmenté $\mathfrak{A} = \mathcal{A} \times \mathcal{S}$,, où le modèle de transition pour l'action $\alpha = (a, \sigma)$ dans l'état s est le modèle moyen local dérivé de $\boldsymbol{b}_t$ mis à jour avec une évidence artificielle de transitions $\lambda^{\eta}_{s,a,\sigma} = \{(s, a, \sigma), \ldots, (s, a, \sigma)\}$ de taille η (un paramètre de l'algorithme). Dit autrement, nous prenons à la fois une action a et un prochain état σ que nous souhaiterions voir arriver avec une plus grande probabilité. Ce MDP peut être résolu comme suit :

$$\begin{aligned} V_i^{\text{BOLT}}(s, \boldsymbol{b}_t) \quad &= \max_{\alpha} \sum_{s'} \hat{T}(s, \alpha, s', \boldsymbol{b}_t) \left[R(s, a, s') + \gamma V_{i-1}^{\text{BOLT}}(s', \boldsymbol{b}_t)\right] \\ &\text{avec} \qquad \hat{T}(s, \alpha, s') = E[Pr(s'|s, a)|\boldsymbol{b}_t, \lambda^{\eta}_{s,a,\sigma}]. \end{aligned}$$

Le *value iteration* de BOLT néglige l'évolution de $\boldsymbol{b}_t$, mais la fonction de transition modifiée fonctionne comme une approximation optimiste de cette évolution bayésienne négligée.

Modifier la fonction de transition semble être une approche plus naturelle que modifier la fonction de récompense comme dans BEB puisque l'incertitude que nous considérons dans ces problèmes concerne justement la fonction de transition, et non la fonction de récompense.

D'un point de vue computationnel, chaque mise à jour de BOLT requiert $|\mathcal{S}|$ fois plus de calculs que chaque mises à jour de BEB. Cela implique des temps de calcul multipliés par $|\mathcal{S}|$ quand on résout des problèmes à horizon fini par programmation dynamique, et probablement une augmentation similaire pour *value iteration*. Toutefois, avec des distributions structurées, tous les prochains états σ n'ont pas à être explorés, mais seulement ceux qui correspondent à des transitions possibles.

Ici, l'optimisme est contrôlé par le paramètre positif η – un paramètre à valeur entière ou réelle selon la famille de distributions – et le comportement avec différentes valeurs de paramètres dépendra de la famille de distributions utilisée. Toutefois, pour des distributions classiques comme FDM, il peut être prouvé que BOLT est toujours optimiste par rapport à la fonction de valeur bayésienne optimale.

Lemme 4.1 (Optimisme de BOLT)
Soit $(s_t, \boldsymbol{b}_t)$ l'état de croyance courant depuis lequel on applique le value iteration *de* BOLT *avec un horizon H et $\eta = H$. Soit aussi $\boldsymbol{b}_t$ une distribution dans la famille FDM, et soit $\mathbb{V}_H(s_t, \boldsymbol{b}_t)$ la fonction de valeur bayésienne optimale. Alors, on a*

$$V_H^{\text{BOLT}}(s_t, \boldsymbol{b}_t) \geq \mathbb{V}_H(s_t, \boldsymbol{b}_t).$$

[*Preuve dans App. A.1*]

5 Analyse de BOLT

Dans cette section nous prouvons que BOLT est PAC-BAMDP dans le cas γ-pondéré avec un horizon infini [5] quand on utilise une distribution FDM. L'autre algorithme dont il est prouvé qu'il est PAC-BAMDP est BEB, mais l'analyse fournie dans (Kolter & Ng, 2009) ne vaut que pour les domaines à horizon fini avec une condition d'arrêt imposée de la mise à jour de Bayes. Pour cette raison, nous incluons dans App. B une analyse de BEB utilisant les résultats de cette section de manière à être capable de comparer ensuite ces algorithmes d'un point de vue théorique.

D'après la définition 3.2, nous devons analyser la politique $\boldsymbol{A}_t$ générée par BOLT au temps t, c'est-à-dire $\boldsymbol{A}_t = \operatorname{argmax}_\pi V_H^{\text{BOLT},\pi}(s_t)$, et montrer que, avec une grande probabilité et sauf pour un nombre polynomial de pas de temps, cette politique est ϵ-proche de la politique optimale bayésienne.

Théorème 5.1 (BOLT est PAC-BAMDP)
Dénotons $\boldsymbol{A}_t$ la politique suivie par BOLT *au temps t avec $\eta = H$. Soit aussi s_t et $\boldsymbol{b}_t$ l'état et la croyance correspondants à cet instant. Alors, avec probabilité au moins $1-\delta$,* BOLT *est ϵ-proche de la politique bayésienne optimale*

$$\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) \geq \mathbb{V}^*(s_t, \boldsymbol{b}_t) - \epsilon$$

sauf pour $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\eta^2}{\epsilon^2(1-\gamma)^2}\right) = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^2}{\epsilon^2(1-\gamma)^2}\right)$pas de temps.
[*Preuve en section 5.2*]

Dans la preuve nous verrons que H dépend de ϵ et γ. Ainsi, le majorant de la complexité d'échantillon et le paramètre optimiste η dépendront seulement de la correction désirée (ϵ,δ) et des caractéristiques du problème (γ,$|\mathcal{S}|$,$|\mathcal{A}|$).

5.1 Fonction de valeur mixte

Pour prouver que BOLT est PAC-BAMDP nous introduisons quelques concepts et résultats préliminaires. D'abord, supposons pour l'analyse que nous maintenons un vecteur de compteurs de transitions θ, même si les distributions peuvent être autres que FDM pour les lemmes spécifiques présentés dans cette section. Comme la croyance est contrôlée, à chaque pas de temps nous pouvons définir un ensemble $K = \{(s,a) | \|\boldsymbol{\theta}_{s,a}\| \geq m\}$ de *couples état-action connus (known)* (Kearns & Singh, 1998), c'est-à-dire des couples état-action avec "assez" d'évidences. Aussi, pour analyser un algorithme de type EXPLOIT $\boldsymbol{A}$ en général (comme) EXPLOIT, BOLT ou BEB), nous introduisons une fonction de valeur *mixte* $\tilde{\mathbb{V}}$ obtenue en effectuant une mise à jour bayésienne exacte quand un couple état-action est dans K et la mise à jour de $\boldsymbol{A}$ hors de K. En utilisant ces concepts, nous pouvons revisiter le lemme 5 de (Kolter & Ng, 2009) pour le cas γ-pondéré.

Lemme 5.2 (Inégalité induite revisitée)
Soit $\mathbb{V}_H^\pi(s_t, \boldsymbol{b}_t)$ l'évaluation bayésienne d'une politique π, et $a = \pi(s, \boldsymbol{b})$ une action de la politique. Nous définissons

$$\tilde{\mathbb{V}}_i^\pi(s, \boldsymbol{b}) = \begin{cases} \sum_{s'} T(s,a,s',\boldsymbol{b})(R(s,a,s') + \gamma \tilde{\mathbb{V}}_{i-1}^\pi(s', \boldsymbol{b}')) & \text{if } (s,a) \in K \\ \sum_{s'} \tilde{T}(s,a,s')(\tilde{R}(s,a,s') + \gamma \tilde{\mathbb{V}}_{i-1}^\pi(s', \boldsymbol{b}')) & \text{if } (s,a) \notin K \end{cases} \tag{5}$$

la fonction de valeur mixte, *où $\tilde{T}$ et $\tilde{R}$ peuvent être différents respectivement de T et R. Ici, $\boldsymbol{b}'$ est le vecteur de paramètre postérieur, après la mise à jour de Bayes avec (s,a,s'). Soit aussi A_K l'événement qu'un couple $(s,a) \notin K$ est généré pour la première fois en partant d'un état s_t et en suivant la politique π pour H pas de temps. En supposant des récompenses normalisées pour R et une récompense maximale $\tilde{R}_{max}$ pour $\tilde{R}$, alors*

$$\mathbb{V}_H^\pi(s_t, \boldsymbol{b}_t) \geq \tilde{\mathbb{V}}_H^\pi(s_t, \boldsymbol{b}_t) - \frac{(1-\gamma^H)}{(1-\gamma)} \tilde{R}_{max} Pr(A_K), \tag{6}$$

où $Pr(A_K)$ est la probabilité de l'événement A_K.
[*Preuve dans App. A.2*]

5. Pour suivre les preuves, gardez, s'il vous plaît, en tête que H n'est pas l'horizon du problème (lequel est infini dans notre analyse), mais *l'horizon de calcul des MDP* résolus à chaque étape.

5.2 BOLT est PAC-BAMDP

Soit $\tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t)$ l'évaluation de la politique de BOLT $\boldsymbol{A}_t$ utilisant une fonction de valeur *mixte* où $\tilde{R}(s,a,s') = R(s,a,s')$ est la fonction de récompense, et $\tilde{T}(s,a,s') = \hat{T}(s,\alpha,s',\boldsymbol{b}_t) = E[Pr(s'|s,a)|\boldsymbol{b}_t, \lambda^{\eta}_{s,a,\sigma}]$ est le modèle de transition de BOLT, où a et σ sont obtenus de la politique $\boldsymbol{A}_t$. Notons que, même si nous appliquons la mise à jour de BOLT, nous contrôlons toujours la croyance à chaque pas de temps comme présenté dans l'équation 5. Toutefois, pour $\hat{T}$ nous considérons la croyance au temps t, et non la croyance contrôlée $\boldsymbol{b}$ comme dans la mise à jour bayésienne.

Lemme 5.3 (Majorant mixte de BOLT)
La différence entre la valeur optimiste obtenue par BOLT *et la valeur bayésienne obtenue par la fonction de valeur* mixte *sous la politique $\boldsymbol{A}_t$ générée par* BOLT *avec $\eta = H$ est majorée par*

$$V_H^{\text{BOLT}}(s_t, \boldsymbol{b}_t) - \tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) \leq \frac{(1-\gamma^H)\eta^2}{(1-\gamma)m}. \tag{7}$$

[*Preuve dans App. A.3*]

Preuve [Preuve du théorème 5.1] Nous commençons par l'inégalité induite (lemme 5.2) avec $\boldsymbol{A}_t$ la politique générée par BOLT au temps t, et $\tilde{\mathbb{V}}$ une fonction de valeur *mixte* utilisant la mise à jour de BOLT quand $(s,a) \notin K$. Comme $\tilde{R}_{max} = 1$, la chaîne d'inégalités est

$$\begin{aligned}
\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) &\geq \mathbb{V}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) \\
&\geq \tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) - \frac{1-\gamma^H}{1-\gamma} Pr(A_K) \\
&\geq V_H^{\text{BOLT}}(s_t, \boldsymbol{b}_t) - \frac{\eta^2(1-\gamma^H)}{m(1-\gamma)} - \frac{1-\gamma^H}{1-\gamma} Pr(A_K) \\
&\geq \mathbb{V}_H^{*}(s_t, \boldsymbol{b}_t) - \frac{\eta^2(1-\gamma^H)}{m(1-\gamma)} - \frac{1-\gamma^H}{1-\gamma} Pr(A_K) \\
&\geq \mathbb{V}^{*}(s_t, \boldsymbol{b}_t) - \frac{\eta^2(1-\gamma^H)}{m(1-\gamma)} - \frac{1-\gamma^H}{1-\gamma} Pr(A_K) - \frac{\gamma^H}{(1-\gamma)}
\end{aligned}$$

où la 3ème étape est due au lemme 5.3 (précision) et la 4ème étape au lemme 4.1 (optimisme). Pour simplifier l'analyse, supposons que $\frac{\gamma^H}{(1-\gamma)} = \frac{\epsilon}{2}$ et fixons $m = \frac{4\eta^2}{\epsilon(1-\gamma)}$.

Si $Pr(A_K) > \frac{\eta^2}{m} = \frac{\epsilon(1-\gamma)}{4}$, d'après les inégalités de Hoeffding et de Boole, nous savons que A_K ne se produit pas dans plus de

$$O\left(\frac{|\mathcal{S}||\mathcal{A}|m}{Pr(A_K)} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right) = O\left(\frac{|\mathcal{S}||\mathcal{A}|\eta^2}{\epsilon^2(1-\gamma)^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)$$

pas de temps avec une probabilité $1-\delta$. En négligeant les logarithmes nous avons le théorème désiré. Ce majorant est dérivé du fait que, si A_K se produit plus de $|\mathcal{S}||\mathcal{A}|m$ fois, alors tous les couples état-action sont connus [6] et nous ne nous échapperons plus jamais de K.

Pour $Pr(A_K) \leq \frac{\eta^2}{m}$, nous avons que

$$\begin{aligned}
\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) &\geq \mathbb{V}^{*}(s_t, \boldsymbol{b}_t) - \frac{\epsilon(1-\gamma^H)}{4} - \frac{\epsilon(1-\gamma^H)}{4} - \frac{\epsilon}{2} \\
&\geq \mathbb{V}^{*}(s_t, \boldsymbol{b}_t) - \frac{\epsilon}{4} - \frac{\epsilon}{4} - \frac{\epsilon}{2} \\
&= \mathbb{V}^{*}(s_t, \boldsymbol{b}_t) - \epsilon
\end{aligned}$$

ce qui vérifie le théorème propose. □

En suivant les résultats de Kolter & Ng (2009) pour BEB, l'optimisme peut être assuré avec $\beta \geq 2H^2$, pour $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^4}{\epsilon^2(1-\gamma)^2}\right)$ pas de temps non ϵ-proches (voir App. B.2), ce qui est un plus mauvais résultat que BOLT. Néanmoins les majorants utilisés dans les preuves sont assez lâches pour espérer que la propriété d'optimisme soit en pratique valide pour de bien plus petites valeurs de β et η.

6. Ici nous ne considérons pas l'information déjà encodée dans la distribution a priori, de sorte que ce majorant sera bien plus fin en pratique.

6 Expérimentations

Pour illustrer les caractéristiques de BOLT, nous présentons ici des résultats expérimentaux sur un certain nombre de domaines. Pour tous les domaines nous avons essayé différents paramètres pour BOLT et BEB, mais avons aussi utilisé une variante ε-gloutonne d'EXPLOIT, dans laquelle il y a une probabilité ε de choisir une action au hasard plutôt que de suivre une politique optimale pour le modèle moyen. Toutefois, pour tous problèmes présentés, EXPLOIT standard ($\varepsilon = 0.0$) surpasse la variante ε-gloutonne.

Rappelons que les valeurs théoriques des paramètres β et η –qui garantissent l'optimisme– dépendent de l'horizon H des MDP résolus à chaque pas de temps. Dans ces expérimentations, au lieu d'utiliser l'horizon employé pour *value iteration* asynchrone, l'arrêt a lieu quand $\|V_{i+1} - V_i\|_\infty < \epsilon$. De plus, nous ré-utilisons la fonction de valeur finale au temps t comme fonction de valeur initiale au temps $t + 1$ pour accélérer les calculs. Pour résoudre ces MDP à horizon infini nous avons utilisé $\gamma = 0,95$ et $\epsilon = 0,01$, mais il faut faire attention à ce que le critère de performance employé ici soit la moyenne de la récompense totale *non γ-pondérée*, comme dans (Poupart *et al.*, 2006; Asmuth *et al.*, 2009).

6.1 Le problème de la chaîne

Dans le problème de la chaîne à 5 états (Strens, 2000; Poupart *et al.*, 2006), tout état est connecté à l'état s_1 en prenant l'action b et tout état s_i est connecté à l'état suivant s_{i+1} avec l'action a, sauf s_5 qui est connecté à lui-même. A chaque pas de temps, l'agent peut "glisser" avec la probabilité p, effectuant l'autre action que celle souhaitée. Rester dans s_5 donne une récompense de $1,0$ alors que revenir en s_1 donne une récompense de $0,2$. Toutes les autres récompenses sont 0. Les variantes diffèrent selon les distributions utilisées : **Full** (FDM), **Tied**, dans laquelle la probabilité p est factorisée pour toutes les transitions, et **Semi**, où chaque action a une probabilité factorisée indépendante.

Algorithme	Tied	Semi	Full
EXPLOIT	366,1	354,9	230,2
BEB ($\beta = 1$)	365,9	362,5	**343,0**
BEB ($\beta = 150$)	366,5	297,5	165,2
BOLT ($\eta = 7$)	**367,9**	**367,0**	289,6
BOLT ($\eta = 150$)	366,6	358,3	278,7
BEETLE *	365,0	364,8	175,4
BOSS *	365,7	365,1	300,3

TABLE 1 – **Résultat du problème de la chaîne pour différentes distributions**. Récompense moyenne sur 500 essais pour un horizon de 1000 avec $p = 0,2$. Les résultats avec une * proviennent de publications antérieures.

Les résultats de la table 1 montrent que BEB surpasse les autres algorithmes avec une valeur de β réglée pour la distribution FDM, comme déjà montré par Kolter & Ng (2009). Toutefois, pour de grandes valeurs de β, cette performance décroît dramatiquement. BOLT par contre produit des résultats comparables avec BOSS pour un paramètre réglé, mais ne décroît pas beaucoup pour de grandes valeurs de η. En effet, cette valeur correspond au majorant théorique qui garantit l'optimisme, $\eta = H \approx \log(\epsilon(1-\gamma))/\log(\gamma) \approx 150$. Sans surprise, les résultats de BEB et BOLT avec des distributions informatives ne sont pas très différents des autres techniques, parce que le problème dégénère en un problème très facile. Néanmoins, BOLT obtient les meilleurs résultats pour ces distributions en utilisant le paramètre optimisé, et d'assez bons résultats avec un grand η, au contraire de BEB qui échoue à fournir un résultat compétitif pour la distribution Semi avec un grand β.

Cette variabilité dans les résultats soulève la question de la sensibilité au réglage de paramètre. Dans un domaine de RL, on ne peut généralement pas régler les paramètres de l'algorithme pour chaque problème, parce que le modèle complet du problème est inconnu. Un bon algorithme de RL doit donc avoir de bonnes performances pour différents problèmes sans modifier ses paramètres.

La figure 2 montre comment BEB et BOLT se comportent pour différents paramètres. Dans l'analyse basse résolution, la performance de BEB décroît très vite, alors que BOLT tend aussi à décroître, mais maintient de bons résultats. Nous avons aussi conduit des expérimentations pour d'autres valeurs de la probabilité de

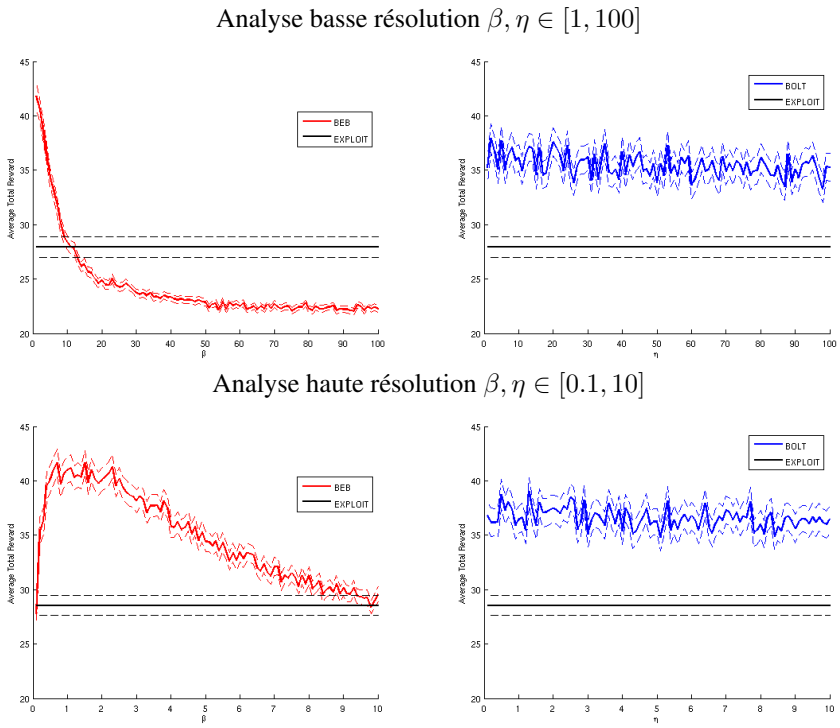

FIGURE 2 – **Problème de la chaîne**. Moyenne de la récompense totale sur 300 essais pour un horizon de 150, et pour des paramètres β et η entre 1 et 100, et entre 0,1 et 10. Pour référence, la valeur obtenue par EXPLOIT est aussi tracée. Tous les résultats sont montrés avec un intervalle de confiance de 95%.

glisser p, le même schéma (i) étant amplifié quand p est proche de 0, c'est-à-dire en donnant une décroissance pire pour BEB et des résultats presque constants pour BOLT, et (ii) obtenant un comportement presque identique quand p est proche de $0,5$. Dans les résultats haute résolution, BEB a un pic de performance autour de 1, alors que BOLT maintient un comportement similaire à l'expérimentation basse résolution.

6.2 Autres problèmes structurés

Un autre exemple illustratif est le problème *Paint/Polish* dans lequel l'objectif est de livrer plusieurs objets polis et peints sans éraflure, en utilisant plusieurs actions stochastiques aux probabilités inconnues. La description complète de ce problème peut être trouvée dans (Walsh *et al.*, 2009). Ici, les conséquences possibles de chaque action sont données à l'agent, mais les probabilités de chaque conséquence ne le sont pas. Nous avons utilisé une distribution structurée qui encode cette information, les résultats obtenus étant résumés dans la figure 3 en utilisant à la fois des analyses basse et haute résolution. Nous avons aussi conduit cette expérimentation avec une distribution FDM, obtenant des résultats similaires que pour le problème de la chaîne. Sans surprise, utiliser une distribution structurée fournit de meilleurs résultat qu'en utilisant FDM. Toutefois, l'impact élevé du fait d'être trop optimiste, montré dans la figure 3, n'a pas lieu avec FDM, principalement parce que la phase d'apprentissage est beaucoup plus courte en utilisant une distribution structurée. Encore une fois, la baisse de performance de BEB est bien plus forte que celle de BOLT, mais, contrairement au problème de la chaîne, le meilleur paramètre de BOLT bât le meilleur paramètre de BEB.

Le dernier exemple est le problème du *Marble Maze* [7] (Asmuth *et al.*, 2009) où nous avons encodé explicitement les 16 clusters possibles dans la distribution, ce qui conduit à de faibles besoins d'exploration. Comme attendu, EXPLOIT fournit une très bonne solution pour ce problème, et BOLT fournit des résultats similaires avec plusieurs paramétrages différents. Au contraire, pour toutes les valeurs de β testées, BEB se comporte bien plus mal qu'EXPLOIT. Par exemple, pour le meilleur η $(= 2,0)$, BOLT fait un score de $-0,455$, alors que pour le meilleur β $(= 0,9)$, BEB fait un score de $-2,127$, et EXPLOIT de $-0,590$.

7. Moyenné sur 100 essais avec $H = 100$.

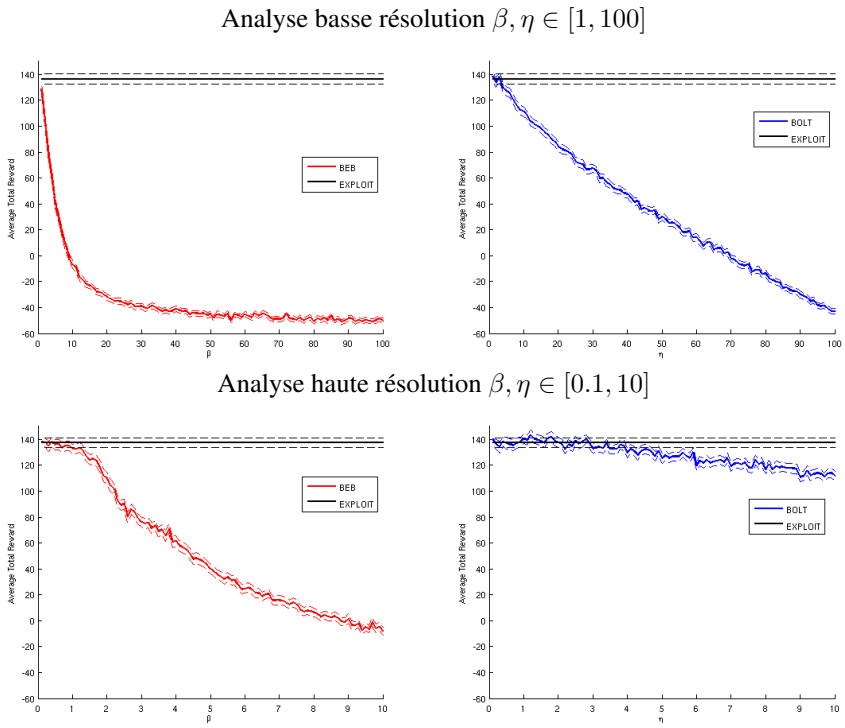

FIGURE 3 – **Problème Paint/Polish**. Moyenne de la récompense totale sur 300 essais pour un horizon de 150, pour plusieurs valeurs de β et η en utilisant une distribution structuée. Pour référence, la valeur obtenue par EXPLOIT est aussi tracée. Tous les résultats sont montrés avec un intervalle de confiance de 95%.

En résumé, il est difficile de savoir a priori quel algorithme aura les meilleures performances pour un problème donné avec une distribution donnée et étant donnés certains paramètres. Toutefois, BOLT généralise bien (en théorie et en pratique) pour un ensemble large de valeurs de paramètres, principalement parce que l'optimisme est majoré par les lois de probabilité et non par un paramètre libre comme dans BEB.

7 Conclusion

Nous avons présenté BOLT, un algorithme nouveau et simple qui utilise une augmentation optimiste de la mise à jour de Bayes, laquelle est ainsi optimiste *à propos* de l'incertitude plutôt que juste *face à* l'incertitude. Nous avons montré que BOLT est strictement optimiste pour certains paramètres η, et avons utilisé ce résultat pour montrer qu'il est aussi PAC-BAMDP. Les majorants de complexité d'échantillon pour BOLT sont plus fins que pour BEB. Des expérimentations montrent que BOLT est plus efficace que BEB quand on utilise les paramètres dérivés de manière théorique dans le problème de la chaîne, et qu'il semble, en général, plus robuste au réglage de son paramètre. Les travaux futurs incluent l'utilisation d'un bonus η dynamique pour BOLT, ce qui devrait être particulièrement approprié pour des horizons finis, et l'exploration de preuves générales pour garantir la propriété PAC-BAMDP pour une famille plus large de distributions que FDM.

Références

ASMUTH J., LI L., LITTMAN M., NOURI A. & WINGATE D. (2009). A Bayesian sampling approach to exploration in reinforcement learning. In *Proc. of UAI*.

BELLMAN R. (1954). The theory of dynamic programming. *Bull. Amer. Math. Soc.*, **60**, 503–516.

BRAFMAN R. & TENNENHOLTZ M. (2003). R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, **3**, 213–231.

DIMITRAKAKIS C. (2008). Tree exploration for Bayesian RL exploration. In *CIMCA/IAWTIC/ISE*.

DUFF M. (2002). *Optimal learning : Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst.
KEARNS M. & SINGH S. (1998). Near-optimal reinforcement learning in polynomial time. In *Machine Learning*, p. 260–268.
KOLTER J. & NG A. (2009). Near-Bayesian exploration in polynomial time. In *Proc. of ICML*.
POUPART P., VLASSIS N., HOEY J. & REGAN K. (2006). An analytic solution to discrete Bayesian reinforcement learning. In *Proc. of ICML*.
PUTERMAN M. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley-Interscience.
SORG J., SINGH S. & LEWIS R. (2010). Variance-based rewards for approximate Bayesian reinforcement learning. In *Proc. of UAI*.
STREHL A., LI L. & LITTMAN M. (2009). Reinforcement learning in finite MDPs : PAC analysis. *JMLR*, **10**, 2413–2444.
STREHL A. & LITTMAN M. (2005). A theoretical analysis of model-based interval estimation. In *Proc. of ICML*.
STRENS M. J. A. (2000). A Bayesian framework for reinforcement learning. In *Proc. of ICML*.
SUTTON R. & BARTO A. (1998). *Reinforcement Learning : An Introduction*. MIT Press.
VALIANT L. G. (1984). A theory of the learnable. In *Proc. of STOC* : ACM.
WALSH T., SZITA I., DIUK C. & LITTMAN M. (2009). Exploring compact reinforcement-learning representations with linear regression. In *Proc. of UAI*.

A Preuves techniques

A.1 Preuve du lemme 4.1

Preuve Nous prouvons ce lemme en trouvant un majorant pour la fonction de valeur optimale bayésienne à l'itération i. Ensuite, par induction, nous prouvons que la différence accumulée entre la fonction de valeur de BOLT et ce majorant est minorée par une quantité positive.

Soit $\mathbb{Q}^*_H(s_t, \boldsymbol{b}_t, a)$ la fonction de valeur état-action bayésienne optimale. On peut majorer cette fonction de valeur état-action bayésienne optimale à l'itération i par

$$\begin{aligned}\mathbb{Q}^*_i(s, \boldsymbol{b}, a) &= \sum_{s'} \frac{\boldsymbol{\theta}_{s,a}(s')}{\|\boldsymbol{\theta}_{s,a}\|}(R(s,a,s') + \gamma \mathbb{V}^*_{i-1}(s', \boldsymbol{b}')) \\ &\leq \max_{\phi} \sum_{s'} \frac{\boldsymbol{\theta}^t_{s,a}(s') + \phi(s')}{\|\boldsymbol{\theta}^t_{s,a}\| + \|\phi\|}(R(s,a,s') + \gamma \mathbb{V}^*_{i-1}(s', \boldsymbol{b}')),\end{aligned}$$

où $\boldsymbol{b}$ est un *descendant* de $\boldsymbol{b}_t$, dans le sens que $\boldsymbol{b}$ est une croyance obtenue en appliquant $H - i$ mises à jour bayésienne à partir de $\boldsymbol{b}_t$, et où $\|\phi\| = \|\boldsymbol{\theta}_{s,a}\| - \|\boldsymbol{\theta}^t_{s,a}\|$. Maintenant définissons

$$\begin{aligned} f(\phi) &= \sum_{s'} \frac{\boldsymbol{\theta}^t_{s,a}(s') + \phi(s')}{\|\boldsymbol{\theta}^t_{s,a}\| + \|\phi\|} g(s', s, a, \boldsymbol{b}), \text{ avec} \\ g(s', s, a, \boldsymbol{b}) &= R(s,a,s') + \gamma \mathbb{V}^*_{i-1}(s', \boldsymbol{b}'). \end{aligned}$$

Veuillez noter que $f(\phi)$ est une fonction linéaire parce que $\|\phi\| = C_{s,a}$ est constante pour un couple état-action donné. Ainsi, maximiser $f(\cdot)$ peut être transformé en maximiser $g(\cdot, s, a, \boldsymbol{b})$ comme suit :

$$\begin{aligned}\mathbb{Q}^*_i(s, \boldsymbol{b}, a) &\leq \max_{\phi} f(\phi) \\ &= \frac{1}{\|\boldsymbol{\theta}^t_{s,a}\| + C_{s,a}} \left(\left[\sum_{s'} \boldsymbol{\theta}^t_{s,a}(s') g(s', s, a, \boldsymbol{b}) \right] + \max_{\phi} \left[\sum_{s'} \phi(s') g(s', s, a, \boldsymbol{b}) \right] \right) \\ &\leq \frac{1}{\|\boldsymbol{\theta}^t_{s,a}\| + C_{s,a}} \left(\left[\sum_{s'} \boldsymbol{\theta}^t_{s,a}(s') g(s', s, a, \boldsymbol{b}) \right] + C_{s,a} \max_{\sigma}(g(\sigma, s, a, \boldsymbol{b})) \right). \end{aligned} \tag{8}$$

Analysons maintenant la différence entre la fonction de valeur état-action de BOLT, $Q^{\text{BOLT}}_H(s_t, \boldsymbol{b}_t, a) = \max_\sigma Q^{\text{BOLT}}_H(s_t, \boldsymbol{b}_t, (a, \sigma))$, et la fonction de valeur bayésienne optimale $\mathbb{V}^*_H(s_t, \boldsymbol{b}_t)$. Pour prouver le pas

d'induction $i+1$, faisons l'hypothèse que la différence entre les fonctions de valeur à l'itération i est minorée par une quantité positive : $0 \leq \Delta_i \leq V_i^{\text{BOLT}}(s, \boldsymbol{b}_t) - \mathbb{V}_i^*(s, \boldsymbol{b})$. Alors, on peut minorer la valeur de BOLT à l'itération $i+1$ avec $\eta = H$:

$$\begin{aligned}
Q_{i+1}^{\text{BOLT}}(s, \boldsymbol{b}_t, a) &= \max_{\sigma} \sum_{s'} \frac{\boldsymbol{\theta}_{s,a}^t(s') + H\delta(\sigma, s')}{\|\boldsymbol{\theta}_{s,a}^t\| + H} (R(s,a,s') + \gamma V_i^{\text{BOLT}}(s', \boldsymbol{b}_t)) \\
&= \max_{\sigma} \sum_{s'} \frac{\boldsymbol{\theta}_{s,a}^t(s') + H\delta(\sigma, s')}{\|\boldsymbol{\theta}_{s,a}^t\| + H} (R(s,a,s') + \gamma(\Delta_i + \mathbb{V}_i^*(s, \boldsymbol{b}))) \\
&\geq \gamma \Delta_i + \frac{H \max_\sigma(g(\sigma, s, a, \boldsymbol{b})) + \sum_{s'} \boldsymbol{\theta}_{s,a}^t(s') g(s', s, a, \boldsymbol{b})}{\|\boldsymbol{\theta}_{s,a}^t\| + H}. \qquad (9)
\end{aligned}$$

Les équations 8 et 9 permettent de montrer que la différence à l'itération $i+1$ peut être majorée par

$$\begin{aligned}
Q_{i+1}^{\text{BOLT}}(s, \boldsymbol{b}_t, a) - \mathbb{Q}_{i+1}(s, \boldsymbol{b}, a) &\geq \gamma \Delta_i + \frac{H \max_\sigma(g(\sigma, s, a, \boldsymbol{b})) + \sum_{s'} \boldsymbol{\theta}_{s,a}^t(s') g(s', s, a, \boldsymbol{b})}{\|\boldsymbol{\theta}_{s,a}^t\| + H} \\
&\quad - \frac{C_{s,a} \max_\sigma(g(\sigma, s, a, \boldsymbol{b})) + \sum_{s'} \boldsymbol{\theta}_{s,a}^t(s') g(s', s, a, \boldsymbol{b})}{\|\boldsymbol{\theta}_{s,a}^t\| + C_{s,a}} \\
&\geq \gamma \Delta_i + \frac{(H - C_{s,a}) \sum_{s'} \boldsymbol{\theta}_{s,a}^t (\max_\sigma(g(\sigma, s, a, \boldsymbol{b})) - g(s', s, a, \boldsymbol{b}))}{\|\boldsymbol{\theta}_{s,a}^t\| + H} \\
&\geq \gamma \Delta_i,
\end{aligned}$$

où la dernière étape est due à ce que tous les éléments dans la fraction ont des valeurs positives ($H > C_{s,a}$). Cela implique que $V_{i+1}^{\text{BOLT}}(s, \boldsymbol{b}_t) - \mathbb{V}_{i+1}^*(s, \boldsymbol{b}) \geq \gamma \Delta_i$. Ainsi, en notant que l'étape de départ est $\Delta_0 = 0$ –parce que $\mathbb{V}_0(s, \boldsymbol{b}) = V^{\text{BOLT}}(s, \boldsymbol{b}_t) = 0$– et en appliquant cette dernière équation de manière répétée, nous obtenons le résultat désiré par induction. □

A.2 Preuve du lemme 5.2

Preuve Cette preuve suit le même raisonnement que le lemme 5 de (Kolter & Ng, 2009), mais est généralisée pour être appliquée à une fonction de valeur *mixte*, et aussi pour le cas à récompense γ-pondérée. Soit p_i un chemin partiel généré par la politique π, $p_i = \langle s_0, a_0, s_1, \ldots, a_{i-1}, s_i \rangle$, soit $Pr(p_i)$ la probabilité d'obtenir ce chemin étant donnée la croyance a priori $\boldsymbol{b}$, et $r(p_i)$ la récompense pour la dernière étape du chemin partiel p_i, c'est-à-dire (s_{i-1}, a_i, s_i). En outre, soient $\tilde{Pr}(p_i)$ et $\tilde{r}(p_i)$ respectivement le produit des probabilités de transition le long de p_i et la récompense pour la dernière étape de la fonction de valeur mixte $\tilde{\mathbb{V}}$ pour un chemin donné p_i. Alors,

$$\begin{aligned}
\tilde{\mathbb{V}}_H^\pi(s_t, \boldsymbol{b}_t) - \mathbb{V}_H^\pi(s_t, \boldsymbol{b}_t) &= \sum_{i=1}^{H-1} \gamma^i \sum_{p_i} \tilde{Pr}(p_i)\tilde{r}(p_i) - \sum_{i=1}^{H-1} \gamma^i \sum_{p_i} Pr(p_i) r(p_i) \\
&= \sum_{i=1}^{H-1} \gamma^i \left[\sum_{p_i \in K} \left(\tilde{Pr}(p_i)\tilde{r}(p_i) - Pr(p_i) r(p_i) \right) \right. \\
&\qquad \left. + \sum_{p_i \notin K} \left(\tilde{Pr}(p_i)\tilde{r}(p_i) - Pr(p_i) r(p_i) \right) \right] \\
&= \sum_{i=1}^{H-1} \gamma^i \sum_{p_i \notin K} \left(\tilde{Pr}(p_i)\tilde{r}(p_i) - Pr(p_t) r(p_i) \right) \\
&\leq \sum_{i=1}^{H-1} \gamma^i \sum_{p_i \notin K} \tilde{Pr}(p_i)\tilde{r}(p_i) \leq \frac{(1-\gamma^H)}{(1-\gamma)} \tilde{R}_{max} Pr(A_K),
\end{aligned}$$

où on divise la somme des chemins entre ceux qui sont toujours dans K et ceux qui s'en échappent. Si un chemin est dans K, alors les récompenses et les probabilités sont égales, de sorte qu'ils peuvent être retirés (troisième étape), et dans la dernière étape on majore la différence en prenant la valeur de récompense maximale possible, et la définition de $Pr(A_K)$. □

A.3 Preuve du lemme 5.3

Preuve Nous prouvons ce lemme par induction, où l'étape d'induction $i+1$ repose sur l'hypothèse que la différence entre les deux évaluations à l'itération i est majorée par une quantité maximale Δ_i :

$$V_i^{\text{BOLT}}(s, \boldsymbol{b}_t) - \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}) \leq \Delta_i. \tag{10}$$

Nous calculons maintenant la différence delta pour $i+1$ en distinguant deux cas. Le premier cas est quand $(s,a) \notin K$ (avec $\alpha = \boldsymbol{A}_t(s, \boldsymbol{b})$), ce qui veut dire que les probabilités et récompenses pour chaque terme sont les mêmes. Formellement,

$$\begin{aligned}
\Delta_{i+1}^{(\notin K)} &= V_{i+1}^{\text{BOLT}}(s, \boldsymbol{b}_t) - \tilde{\mathbb{V}}_{i+1}^{\boldsymbol{A}_t}(s, \boldsymbol{b}) \\
&= \sum_{s'} \hat{T}(s, \alpha, s', \boldsymbol{b}_t) \left(R(s,a,s') + \gamma V_i^{\text{BOLT}}(s, \boldsymbol{b}_t) - R(s,a,s') - \gamma \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}') \right) \\
&\leq \gamma \Delta_i \sum_{s'} \hat{T}(s, \alpha, s', \boldsymbol{b}_t) \\
&= \gamma \Delta_i.
\end{aligned}$$

Le second cas est quand $(s,a) \in K$, les probabilités différant alors,

$$\begin{aligned}
&\Delta_{i+1}^{(\in K)} = V_{i+1}^{\text{BOLT}}(s, \boldsymbol{b}_t) - \tilde{\mathbb{V}}_{i+1}^{\boldsymbol{A}_t}(s, \boldsymbol{b}) \\
&= \sum_{s'} \hat{T}(s, \alpha, s', \boldsymbol{b}_t)(R(s,a,s') + \gamma V_i^{\text{BOLT}}(s, \boldsymbol{b}_t)) - \sum_{s'} T(s,a,s',\boldsymbol{b})(R(s,a,s') + \gamma \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}')) \\
&\leq \gamma \Delta_i + \sum_{s'} \left(\hat{T}(s, \alpha, s', \boldsymbol{b}_t) - T(s,a,s',\boldsymbol{b}) \right) \times \left(R(s,a,s') + \gamma \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}') \right) \\
&= \gamma \Delta_i + \sum_{s'} \left(\frac{\boldsymbol{\theta}_{s,a}^t(s') + \eta\delta(\sigma, s')}{\|\boldsymbol{\theta}_{s,a}^t\| + \eta} - \frac{\boldsymbol{\theta}_{s,a}(s')}{\|\boldsymbol{\theta}_{s,a}\|} \right) \times \left(R(s,a,s') + \gamma \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}') \right) \\
&\leq \gamma \Delta_i + \sum_{s'} \frac{\eta\delta(\sigma, s')}{\|\boldsymbol{\theta}_{s,a}^t\| + \eta} (R(s,a,s') + \gamma \tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}')) \\
&\leq \gamma \Delta_i + \frac{\eta^2}{m}.
\end{aligned}$$

Ici, la 3ème étape est due à la définition de Δ_i. La 5ème inégalité est vraie parce que $\boldsymbol{\theta}_{s,a}^t(s') \leq \boldsymbol{\theta}_{s,a}(s')$ et $\|\boldsymbol{\theta}_{s,a}^t\| + H \geq \|\boldsymbol{\theta}_{s,a}\|$. La dernière étape est obtenue parce que la valeur maximale de $\tilde{\mathbb{V}}$ est toujours plus basse que $H = \eta$, et $\|\boldsymbol{\theta}_{s,a}^t\| + H \geq m$ parce que nous sommes en train d'analyser les couples état-action qui sont dans K.

Nous pouvons maintenant définir notre différence maximale à l'itération $i+1$ comme le pire cas : $\Delta_{i+1} = \max(\Delta_{i+1}^{(\notin K)}, \Delta_{i+1}^{(\in K)}) = \frac{\eta^2}{m} + \gamma \Delta_i$. En appliquant cette équation de manière répétée avec $\Delta_0 = 0$, parce que $V_0^{\text{BOLT}}(s, \boldsymbol{b}_t) = \tilde{\mathbb{V}}_0^{\boldsymbol{A}_t}(s, \boldsymbol{b}) = 0$, nous obtenons le résultat souhaité. □

B Description et analyse de BEB

B.1 Bonus d'Exploration Bayésien

L'algorithme Bonus d'Exploration Bayésien (BEB) consiste à résoudre à chaque pas de temps t un MDP généré par le modèle moyen courant avec une fonction de récompense modifiée dépendant de la croyance courante. Soit $R(s,a,s')$ la fonction de récompense donnée et $\boldsymbol{b}_t$ la croyance au temps t, alors BEB effectue *value iteration* comme suit :

$$V_i^{\text{BEB}}(s, \boldsymbol{b}_t) = \max_a \sum_{s'} T(s,a,s',\boldsymbol{b}_t) \left[\hat{R}(s,a,s',\boldsymbol{b}_t) + \gamma V_{i-1}^{\text{BEB}}(s', \boldsymbol{b}_t) \right]$$

$$\text{avec} \qquad \hat{R}(s,a,s',\boldsymbol{b}_t) = R(s,a,s') + \frac{\beta}{1 + \|\theta_{s,a}\|},$$

où $T = E_{\mathcal{M}}[\mu \mid \boldsymbol{b}_t]$ est le modèle moyen de $\boldsymbol{b}_t$, c'est-à-dire $T(s,a,s',\boldsymbol{b}_t) = E[Pr(s'|s,a)|\boldsymbol{b}_t] = \frac{\boldsymbol{\theta}_{s,a}(s')}{\|\boldsymbol{\theta}_{s,a}\|}$, et β un paramètre qui contrôle l'optimisme de l'algorithme. Veuillez noter que le *value iteration* de BEB néglige l'évolution de $\boldsymbol{b}_t$, de sorte qu'à chaque pas de temps t l'algorithme résout un MDP avec des fonctions de transition et de récompense fixes. Si $\beta \geq 2H^2$, cet algorithme est toujours optimiste comparé à la fonction de valeur bayésienne optimale dans le cas de la récompense non γ-pondérée, comme dit dans le lemme 4 de (Kolter & Ng, 2009). De plus, nous pouvons aisément étendre ce résultat au cas de la récompense γ-pondérée, l'optimisme étant maintenu pour $\beta \geq 2H(1-\gamma^H)/(1-\gamma)$. Toutefois, par simplicité nous choisirons pour l'analyse $\beta = 2H^2 \geq 2H(1-\gamma^H)/(1-\gamma)$.

B.2 Analyse de BEB

Théorème B.1 (BEB est PAC-BAMDP)
Notons $\boldsymbol{A}_t$ la politique suivie par BEB *au temps t avec $\beta = 2H^2$. Soient aussi s_t et $\boldsymbol{b}_t$ les état et croyance correspondant à cet instant. Alors, avec une probabilité au moins $1-\delta$,* BEB *est ϵ-proche de la politique bayésienne optimale*

$$\mathbb{V}^{\boldsymbol{A}_t}(s_t,\boldsymbol{b}_t) \geq \mathbb{V}^*(s_t,\boldsymbol{b}_t) - \epsilon$$

sauf dans $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\beta^2}{\epsilon^2(1-\gamma)^2}\right) = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^4}{\epsilon^2(1-\gamma)^2}\right)$ pas de temps.

Soit $\tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t,\boldsymbol{b}_t)$ l'évaluation de la politique BEB $\boldsymbol{A}_t$ au temps t en utilisant une fonction de valeur *mixte*. Dans ce cas, $\tilde{R}(s,a,s') = \hat{R}(s,a,s',\boldsymbol{b}_t)$ est la récompense avec bonus, et $\tilde{T}(s,a,s') = T(s,a,s',\boldsymbol{b}_t)$ est le modèle de transition moyen de $\boldsymbol{b}_t$.

Lemme B.2 (Majorant Mixte de BEB)
La différence entre la valeur obtenue par BEB *et la valeur obtenue par la fonction de valeur* mixte *sous la politique générée par* BEB, *$\boldsymbol{A}_t$, avec $\beta = 2H^2$ est majorée par*

$$V_H^{\text{BEB}}(s_t,\boldsymbol{b}_t) - \tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t,\boldsymbol{b}_t) \leq \frac{2\beta(1-\gamma^H)}{m(1-\gamma)}. \tag{11}$$

Preuve Suivant la même technique par induction que pour la preuve du lemme 5.3, supposons que la différence entre deux évaluations à l'itération i est majorée par une quantité maximale Δ_i. A nouveau, nous pouvons diviser la différence Δ_{i+1} en deux cas. Le premier cas est quand $(s,a) \notin K$, et, pour la même raison que dans la preuve du lemme B.2, nous avons :

$$\Delta_{i+1}^{(\notin K)} \leq \gamma\Delta_i.$$

Le second cas est quand $(s,a) \in K$, où les probabilités et récompenses diffèrent,

$$\begin{aligned}
\Delta_{i+1}^{(\in K)} &= V_{i+1}^{\text{BEB}}(s,\boldsymbol{b}_t) - \tilde{\mathbb{V}}_{i+1}^{\boldsymbol{A}_t}(s,\boldsymbol{b}) \\
&= \sum_{s'} T(s,a,s',\boldsymbol{b}_t)(\hat{R}(s,a,s',\boldsymbol{b}_t) + \gamma V_i^{\text{BEB}}(s,\boldsymbol{b}_t)) - \sum_{s'} T(s,a,s',\boldsymbol{b})(R(s,a,s') + \gamma\tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s,\boldsymbol{b}')) \\
&\leq \sum_{s'} T(s,a,s',\boldsymbol{b}_t)\ \left(\hat{R}(s,a,s',\boldsymbol{b}_t) + \gamma V_i^{\text{BEB}}(s,\boldsymbol{b}_t) - R(s,a,s') - \gamma\tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s,\boldsymbol{b}')\right) \\
&\quad + \sum_{s'} |T(s,a,s',\boldsymbol{b}_t) - T(s,a,s',\boldsymbol{b})| \times (R(s,a,s') + \gamma\tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s,\boldsymbol{b}')) \\
&\leq \sum_{s'} \left[T(s,a,s',\boldsymbol{b}_t)\left(\frac{\beta}{1+\|\boldsymbol{\theta}_{s,a}^t\|} + \gamma\Delta_i\right) + \frac{(1-\gamma^i)}{(1-\gamma)}\sum_{s'}|T(s,a,s',\boldsymbol{b}_t) - T(s,a,s',\boldsymbol{b})|\right] \\
&\leq \frac{\beta}{(1+\|\boldsymbol{\theta}_{s,a}^t\|)} + \gamma\Delta_i + \frac{2H(1-\gamma^i)}{(1-\gamma)(1+\|\boldsymbol{\theta}_{s,a}^t\|)} \\
&\leq \frac{2\beta}{m} + \gamma\Delta_i.
\end{aligned}$$

Ici, la 3ème étape est due à la propriété

$$\sum_x p(x)f(x) - \sum q(x)g(x) \leq \sum_x p(x)(f(x) - g(x)) + \sum_x |p(x) - q(x)|g(x),$$

qui est vérifiée si toutes les fonctions sont positives. La prochaine étape utilise les définitions de $\hat{R}$ et Δ_i dans le terme de gauche, et en considérant la valeur maximale possible de $R(s, a, s') + \gamma\tilde{\mathbb{V}}_i^{\boldsymbol{A}_t}(s, \boldsymbol{b}')$. La 5ème étape est obtenue par le lemme 3 de (Kolter & Ng, 2009), où la somme des différences absolues est $\sum_s' |T(s, a, s', \boldsymbol{b}_t) - T(s, a, s', \boldsymbol{b})| \leq \frac{2H}{(1+\|\boldsymbol{\theta}_{s,a}^t\|)}$. La dernière étape est due aux faits que $(1-\gamma^i)/(1-\gamma) \leq H$, et que $1 + \|\boldsymbol{\theta}_{s,a}^t\| \geq m$ parce que nous analysons des couples état-action qui sont dans K.

Nous pouvons maintenant définir notre différence maximale à l'itération $i+1$ comme le pire cas : $\Delta_{i+1} = \max(\Delta_{i+1}^{(\notin K)}, \Delta_{i+1}^{(\in K)}) = \frac{2\beta}{m} + \gamma\Delta_i$. En appliquant cette équation de manière répétée avec le pas de départ $\Delta_0 = 0$, parce que $V_0^{\text{BEB}}(s, \boldsymbol{b}_t) = \tilde{\mathbb{V}}_0^{\boldsymbol{A}_t}(s, \boldsymbol{b}) = 0$, nous obtenons le résultat souhaité. □

Avec ces lemmes, nous sommes maintenant prêts à prouver le théorème B.1, lequel montre que BEB est PAC-BAMDP dans le cas γ-pondéré à horizon infini (sans modifier l'algorithme pour arrêter de contrôler les croyances comme dans (Kolter & Ng, 2009)).

Preuve [Preuve du théorème B.1] Considérons l'inégalité induite (lemme 5.2) avec $\boldsymbol{A}_t$ la politique générée par BEB au temps t, et $\tilde{\mathbb{V}}$ une fonction de valeur *mixte* utilisant la mise à jour de BEB quand $(s, a) \notin K$. Comme le bonus de BEB décroît toujours, on peut définir $\tilde{R}_{max} = 2\beta$ ($> 1 + \beta$). Le lemme B.2 est défini pour un horizon de calcul H, de sorte que, pour utiliser ce résultat, nous avons d'abord besoin d'observer qu'une somme tronquée est toujours plus petite que la somme infinie. En partant de là, et en supposant des récompenses normalisées, nous avons

$$\begin{aligned}
\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) &\geq \mathbb{V}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) \\
&\geq \tilde{\mathbb{V}}_H^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) - \frac{2\beta(1-\gamma^H)}{(1-\gamma)} Pr(A_K) \\
&\geq V_H^{\text{BEB}}(s_t, \boldsymbol{b}_t) - \frac{2\beta(1-\gamma^H)}{m(1-\gamma)} - \frac{2\beta(1-\gamma^H)}{(1-\gamma)} Pr(A_K) \\
&\geq \mathbb{V}_H^*(s_t, \boldsymbol{b}_t) - \frac{2\beta(1-\gamma^H)}{m(1-\gamma)} - \frac{2\beta(1-\gamma^H)}{(1-\gamma)} Pr(A_K) \\
&\geq \mathbb{V}^*(s_t, \boldsymbol{b}_t) \ - \frac{2\beta(1-\gamma^H)}{m(1-\gamma)} - \frac{2\beta(1-\gamma^H)}{(1-\gamma)} Pr(A_K) - \frac{\gamma^H}{(1-\gamma)}
\end{aligned}$$

où la 3ème étape est due au lemme B.2 (précision), la 4ème étape au lemme 4 de (Kolter & Ng, 2009) (optimisme), et la dernière étape au lemme 2 de (Kearns & Singh, 1998) (majorant de l'erreur entre les calculs à horizon fini et infini) [8].

Ici, l'erreur entre $\mathbb{V}^*(s_t, \boldsymbol{b}_t)$ et $\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t)$ dépend de H, β, m et γ, en plus de $Pr(A_K)$. Pour simplifier l'analyse, supposons que $\frac{\gamma^H}{(1-\gamma)} = \frac{\epsilon}{2}$ et fixons $m = \frac{8\beta}{\epsilon(1-\gamma)}$.

Maintenant considérons deux cas pour $Pr(A_K)$. Supposons d'abord que $Pr(A_K) > \frac{1}{m} = \frac{\epsilon(1-\gamma)}{8\beta}$; alors, par les inégalités de Hoeffding et de Boole (Valiant, 1984), cela n'a lieu dans pas plus de

$$O\left(\frac{|\mathcal{S}||\mathcal{A}|m}{Pr(A_K)} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right) = O\left(\frac{|\mathcal{S}||\mathcal{A}|\beta^2}{\epsilon^2(1-\gamma)^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)$$

pas de temps avec la probabilité $1-\delta$. En négligeant les logarithmes nous avons la complexité d'échantillon du théorème. Ce majorant est dérivé du fait que, si l'événement A_K survient plus de $|\mathcal{S}||\mathcal{A}|m$ fois, alors tous les couples état-action sont connus et on ne s'échappera plus jamais de K.

Le second cas est quand $Pr(A_K) \leq \frac{1}{m}$, où

$$\begin{aligned}
\mathbb{V}^{\boldsymbol{A}_t}(s_t, \boldsymbol{b}_t) &\geq \mathbb{V}^*(s_t, \boldsymbol{b}_t) - \frac{2\epsilon(1-\gamma^H)}{8} - \frac{2\epsilon(1-\gamma^H)}{8} - \frac{\epsilon}{2} \\
&\geq \mathbb{V}^*(s_t, \boldsymbol{b}_t) - \frac{\epsilon}{4} - \frac{\epsilon}{4} - \frac{\epsilon}{2} = \mathbb{V}^*(s_t, \boldsymbol{b}_t) - \epsilon
\end{aligned}$$

qui vérifie le théorème proposé. □

8. Le lemme 2 de (Kearns & Singh, 1998) est présenté pour des MDP normaux, maus sont applicabilité aux BAMDP est évidente avec les mêmes arguments.