Estimation Monte Carlo sans modèle de politiques de décision

Raphael Fonteneau* — **Susan A. Murphy****
Louis Wehenkel* — **Damien Ernst***

** Département d'Electricité, Electronique et Informatique - Université de Liège*
Grande Traverse, 10 - 4000 Liège - Belgium
{raphael.fonteneau,l.wehenkel,dernst}@ulg.ac.be

*** Département de Statistiques - Université du Michigan*
439 West Hall, 1085 S. Univ. - University of Michigan - Ann Arbor, MI 48109-1107
samurphy@umich.edu

RÉSUMÉ. *Cet article propose un estimateur de l'espérance du retour de politiques de décision déterministes en boucle fermée à partir d'un échantillon de transitions d'un système dynamique. Cet estimateur, appelé en anglais "Model-free Monte Carlo (MFMC) estimator", calcule une moyenne des retours d'un ensemble de "trajectoires artificielles" construites à partir de de la politique à évaluer ainsi que de transitions du système disponibles dans un échantillon fixé dont l'acquisition s'est faite indépendamment de la politique à évaluer. Sous certaines hypothèses de continuité Lipschitzienne de la dynamique du système, de la fonction de récompense et de la politique de décision à évaluer, on montre que le biais et la variance de l'estimateur proposé sont bornés par des termes qui dépendent des constantes de Lipschitz, du nombre de trajectoires artificielles, de la parcimonie de l'échantillon de transitions ainsi que de la variance "naturelle" du retour de la politique.*

ABSTRACT. *We propose an algorithm for estimating the finite-horizon expected return of a closed loop control policy from an a priori given (off-policy) sample of one-step transitions. This algorithm, named Model-free Monte Carlo (MFMC) estimator, averages cumulated rewards along a set of "broken trajectories" made of one-step transitions selected from the sample on the basis of the control policy. Under some Lipschitz continuity assumptions on the system dynamics, reward function and control policy, we provide bounds on the bias and variance of the estimator that depend only on the Lipschitz constants, on the number of broken trajectories used in the estimator, and on the sparsity of the sample of one-step transitions.*

MOTS-CLÉS : *Apprentissage par renforcement, Evaluation de politiques de décision, Estimation par méthode de Monte Carlo*

KEYWORDS: *Reinforcement Learning, Policy Evaluation, Monte Carlo Estimation*

RSTI - RIA – 25/2011. Apprentissage artificiel, pages 321 à 343

1. Introduction

Les problèmes de contrôle optimaux sous incertitudes se rencontrent dans de nombreux domaines, notamment en finance, en médecine, en sciences de l'ingénieur ou en intelligence artificielle. Parmi les techniques couramment utilisées pour résoudre de tels problèmes, une vaste majorité utilisent un critère d'évaluation des politiques de décision dans le but de s'orienter efficacement dans l'espace des politiques admissibles, et de converger vers une politique de décision (quasi-)optimale.

Dans le cas où le système à contrôler peut être simulé à moindre coût, une technique fréquemment utilisée consiste à estimer le retour des politiques de décision par méthode de Monte Carlo (MC). Concrètement, cette méthode génère de nombreuses trajectoires indépendantes du système lorsque celui-ci est contrôlé par la politique de décision qu'on souhaite évaluer, et les retours des trajectoires ainsi obtenues sont moyennés. Ceci permet d'obtenir une estimation non biaisée de l'espérance du retour de la politique de décision. Cependant, lorsque générer de telles trajectoires coûte trop cher, cette approche n'est pas réaliste.

Dans cet article, on décrit un critère d'évaluation des politiques de décision récemment introduit (Fonteneau *et al.*, 2010), fonctionnant dans les cas où l'on ne peut pas simuler la politique à évaluer. Dans ce contexte, les seules informations disponibles sur le système se résument à un échantillon de transitions du système, collectées au préalable au moyen de protocoles expérimentaux, et ce de manière indépendante de la politique à évaluer. Ce critère d'évaluation est un estimateur de l'espérance du retour de la politique de décision dont le calcul s'inspire de l'estimateur MC. De manière analogue aux méthodes MC, l'espérance du retour de la politique de décision est estimée en moyennant les retours obtenus par des trajectoires indépendantes, c'est-à-dire les sommes des récompenses observées tout au long de chaque trajectoire. Ce qui différencie l'estimateur MC classique de ce nouvel estimateur − nommé MFMC de l'anglais *Model-free Monte Carlo* −, c'est que les trajectoires utilisées ne sont pas des trajectoires "réelles" (c'est-à-dire des trajectoires qui pourraient être obtenues par simulation), mais des trajectoires "artificielles" reconstruites à partir de la politique de décision à évaluer et des transitions disponibles dans l'échantillon. Sous certaines hypothèses de continuité Lipschitzienne de la dynamique du système, de la fonction de récompense ainsi que de la politique de décision, il est possible de borner le biais et la variance de l'estimateur MFMC. On montre alors que l'estimateur MFMC tend à se comporter comme l'estimateur MC lorsque la parcimonie de l'échantillon de transitions converge vers zéro.

La suite de cet article est organisée de la manière suivante. La section 2 propose une brève présentation de travaux connexes. La section 3 formalise le problème soulevé dans cet article, tandis que la section 4 détaille l'estimateur MFMC ainsi que ses principales propriétés théoriques. Des résultats de simulation sont présentés en section 5. Les démonstrations des résultats théoriques sont données en annexe.

2. Travaux connexes

Le problème consistant à évaluer l'espérance du retour d'une politique de décision a déjà fait l'objet de nombreuses recherches, notamment dans le cadre de l'apprentissage par renforcement. En apprentissage par renforcement, l'espérance du retour d'une politique de décision est souvent évaluée au travers d'une *fonction de valeur* qui, à un état initial, associe l'espérance du retour de la politique de décision partant de cet état. Les méthodes basées sur les différences temporelles (*Temporal Difference methods*, (Sutton, 1988; Watkins *et al.*, 1992; Rummery *et al.*, 1994; Bradtke *et al.*, 1996)) font partie des techniques qui permettent d'estimer une fonction de valeur à partir de la connaissance de transitions du système. Ces techniques ont l'avantage d'avoir été étudiées en profondeur d'un point de vue théorique (Dayan, 1992; Tsitsiklis, 1994). Dans le cas où l'espace d'état est de grande taille, infini et/ou continu, ces méthodes doivent être combinées avec des approximateurs de fonction (Sutton *et al.*, 2009; Busoniu *et al.*, 2010).

Depuis quelques années, une nouvelle classe d'algorithmes d'apprentissage par renforcement permet d'estimer des fonctions de valeur en l'absence de modèle du système, cela en utilisant des approximateurs de fonction ajustés à partir d'une collection de transitions du système fixée au préalable (algorithmes en mode *batch*). De bons résultats expérimentaux ont été obtenus (Ormoneit *et al.*, 2002; Ernst *et al.*, 2005; Riedmiller, 2005), et de nombreux papiers se sont focalisés sur l'analyse des propriétés théoriques de ces algorithmes (Antos *et al.*, 2008; Munos *et al.*, 2008).

Le principal point négatif de ces techniques est la forte dépendance que la qualité de leur solution entretient avec le choix des approximateurs de fonction, choix loin d'être trivial (Busoniu *et al.*, 2010). A la différence de ces techniques, l'estimateur MFMC décrit dans cet article n'utilise pas d'approximateurs de fonction. Il s'agit d'une extension de l'estimateur de Monte Carlo traditionnel à un contexte où l'on ne connaît pas de modèle du système. Dès lors, ce travail est étroitement lié avec les recherches visant à construire des estimateurs de Monte Carlo performants en présence d'un modèle (Dimitrakakis *et al.*, 2008).

3. Formulation du problème

Tout au long de cet article, on considère un système à temps discret, stationnaire, dont la dynamique est modélisée par l'équation

$$x_{t+1} = f(x_t, u_t, w_t) \quad t = 0, 1, \ldots, T-1 \; .$$

$T \in \mathbb{N}^*$ est l'horizon d'optimisation du problème, x_t appartient à un espace d'état normé $\mathcal{X}$, u_t appartient à un espace de décisions normé $\mathcal{U}$. Une récompense instantanée

$$r_t = \rho(x_t, u_t, w_t) \in \mathbb{R}$$

est associée à chaque transition du système entre les instants t et $t+1$. Les incertitudes du problème sont modélisées par un processus aléatoire non observable

$$w_t \in \mathcal{W}, \quad t = 0 \dots T-1$$

que l'on suppose indépendamment et identiquement distribué (i.i.d.) selon une loi de probabilité $p_{\mathcal{W}}(.), \forall t = 0, \dots, T-1$. Dans la suite de cet article, on note $w_t \sim p_{\mathcal{W}}(.)$, et comme cela est sous-entendu par la notation, on suppose que $p_{\mathcal{W}}(.)$ ne dépend ni du couple état-décision $(x_t, u_t) \in \mathcal{X} \times \mathcal{U}$, ni de $t \in [\![0, T-1]\!]$ (on utilise la notation $[\![0, T-1]\!] = \{0, \dots, T-1\}$).

Soit h une politique de décision déterministe, en boucle fermée, non stationnaire :

$$h : [\![0, T-1]\!] \times \mathcal{X} \to \mathcal{U} \ .$$

A un instant $t \in [\![0, T-1]\!]$ et un état du système $x_t \in \mathcal{X}$, la politique de décision h associe une décision $u_t = h(t, x_t) \in \mathcal{U}$. Soit $J^h(x_0)$ l'espérance du retour de la politique de décision h partant de l'état initial $x_0 \in \mathcal{X}$:

$$J^h(x_0) = \mathop{\mathbb{E}}_{w_0, \dots, w_{T-1} \sim p_{\mathcal{W}}(.)} \left[R^h(x_0) \right] \ ,$$

avec

$$\begin{aligned} R^h(x_0) &= \sum_{t=0}^{T-1} \rho\left(x_t, h(t, x_t), w_t\right) \ , \\ x_{t+1} &= f\left(x_t, h(t, x_t), w_t\right), \forall t \in [\![0, T-1]\!] \ . \end{aligned}$$

Une réalisation de la variable aléatoire $R^h(x_0)$ correspond à une somme de récompenses obtenues lorsqu'on simule une trajectoire de la politique de décision h partant de l'état initial $x_0 \in \mathcal{X}$, perturbée par le processus aléatoire $w_t \sim p_{\mathcal{W}}(.), t = 0 \dots T-1$. On suppose également que $R^h(x_0)$ admet une variance finie $\sigma^2_{R^h}(x_0)$,

$$\sigma^2_{R^h}(x_0) = \mathop{Var}_{w_0, \dots, w_{T-1} \sim p_{\mathcal{W}}(.)} \left[R^h(x_0) \right] \ .$$

Dans cet article, les fonctions f et ρ, ainsi que la distribution de probabilités $p_{\mathcal{W}}(.)$, sont fixées et *inconnues*, donc non simulables. On dispose simplement d'un échantillon de $n \in \mathbb{N}^*$ transitions du système $\mathcal{F}_n = \left[\left(x^l, u^l, r^l, y^l \right) \right]_{l=1}^n$. Pour chaque transition $\left(x^l, u^l, r^l, y^l \right) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R} \times \mathcal{X}$, les deux premiers éléments $\left(x^l, u^l \right)$ sont choisis arbitrairement, tandis que le couple (r^l, y^l) est déterminé par le couple $\left(\rho(x^l, u^l, .), f(x^l, u^l, .) \right)$ tiré selon la distribution $p_{\mathcal{W}}(.)$. A partir d'un tel échantillon $\mathcal{F}_n$, l'objectif est d'estimer $J^h(x_0)$.

4. Un estimateur de $J^h(x_0)$ en l'absence de modèle du système

On rappelle préalablement en section 4.1 la définition de l'estimateur MC, ainsi que son biais et sa variance. L'estimateur MFMC, qui procède par imitation de l'estimateur MC dans un contexte où aucun modèle du système n'est connu, est décrit en

section 4.2 . Le biais et la variance de ce nouvel estimateur sont analysés théoriquement en section 4.3.

4.1. *L'estimateur de Monte Carlo*

L'estimateur MC est utilisé dans un contexte où l'on dispose d'un modèle du système, c'est-à-dire quand les fonctions f, ρ et $p_{\mathcal{W}}(.)$ sont connues. Une estimation de $J^h(x_0)$ est alors obtenue en moyennant les retours de $p \in \mathbb{N}^*$ trajectoires du système lorsque celui ci est contrôlé par la politique de décision h partant de l'état initial $x_0 \in \mathcal{X}$. Formellement, l'estimateur MC s'écrit

$$\mathbb{M}_p^h(x_0) = \frac{1}{p} \sum_{i=1}^{p} \sum_{t=0}^{T-1} \rho\left(x_t^i, h(t, x_t^i), w_t^i\right)$$

où,

$$\begin{aligned} \forall t \in [\![0, T-1]\!], \forall i \in [\![1, p]\!], \qquad & w_t^i \sim p_{\mathcal{W}}(.) \ , \\ & x_0^i = x_0 \ , \\ & x_{t+1}^i = f\left(x_t^i, h(t, x_t^i), w_t^i\right) \ . \end{aligned}$$

Le biais et la variance de l'estimateur MC se calculent de façon immédiate :

$$\begin{aligned} \underset{w_t^i \sim p_{\mathcal{W}}(.), i=1\ldots p, t=0\ldots T-1}{\mathbb{E}} \left[\mathbb{M}_p^h(x_0) - J^h(x_0)\right] &= 0 \ , \\ \underset{w_t^i \sim p_{\mathcal{W}}(.), i=1\ldots p, t=0\ldots T-1}{Var} \left[\mathbb{M}_p^h(x_0)\right] &= \frac{\sigma^2_{R^h}(x_0)}{p} \ . \end{aligned}$$

4.2. *L'estimateur MFMC*

L'estimateur MFMC, de l'anglais "Model-Free Monte Carlo" (Fonteneau *et al.*, 2010), fonctionne en construisant, à partir d'un échantillon $\mathcal{F}_n$ de transitions du système, $p \in \mathbb{N}^*$ séquences de transitions appelées "trajectoires artificielles". Ces trajectoires artificielles servent d'approximation des p trajectoires "réelles" qui seraient obtenues si la politique de décision h pouvait être simulée sur le système. L'estimateur MFMC estime $J^h(x_0)$ en moyennant les retours de ces trajectoires artificielles, c'est-à-dire, pour chaque séquence de transitions, les sommes des récompenses portées par les transitions qui la composent. Les transitions sont choisies de manière à minimiser les cassures au sein des trajectoires artificielles ainsi que leur "éloignement" des trajectoires réelles qu'on obtiendrait en simulant la politique de décision h.

Les p trajectoires artificielles sont créées de manière séquentielle, et chaque transition de l'échantillon $\mathcal{F}_n$ est utilisée au plus une fois. Il est donc nécessaire que

$p \times T \leq n$. Etant donnée une trajectoire artificielle en cours de construction, la $t-$ème transition ajoutée est choisie parmi les transitions non encore utilisées de telle sorte que ses deux premiers éléments minimisent la distance - cela en utilisant une mesure de distance Δ dans $\mathcal{X} \times \mathcal{U}$ - avec le couple formé par le dernier élément de la $(t-1)-$ème transition et de la décision suggérée par la politique h au niveau du dernier élément de la transition $(t-1)$.

Estimation MFMC *(arguments : $\mathcal{F}_n, h(.,.), x_0, \Delta(.,.), T, p$)*

On note $\mathcal{G}$ l'échantillon courant de transitions issues de $\mathcal{F}_n$ non encore utilisées ;

Initialement, $\mathcal{G} \leftarrow \mathcal{F}_n$;

Pour $i = 1$ **à** p, construire une trajectoire artificielle :

$t \leftarrow 0$;

$x_t^i \leftarrow x_0$;

Tant que $t < T$,

$u_t^i \leftarrow h(t, x_t^i)$;

$\mathcal{H} \leftarrow \underset{(x,u,r,y)\in\mathcal{G}}{\arg\min} \left(\Delta((x,u),(x_t^i,u_t^i))\right)$;

Soit l_t^i le plus petit indice dans $\mathcal{F}_n$ des transitions de $\mathcal{H}$;

$t \leftarrow t+1$,

$x_t^i \leftarrow y^{l_t^i}$;

$\mathcal{G} \leftarrow \mathcal{G} \setminus \{(x^{l_t^i}, u^{l_t^i}, r^{l_t^i}, y^{l_t^i})\}$;

fin **Tant que**

fin **Pour**

Retourner l'ensemble des indices $\left\{l_t^i\right\}_{i=1,t=0}^{i=p,t=T-1}$.

Figure 1. *Version tabulaire de l'algorithme MFMC : construction d'un ensemble de p trajectoires artificielles de T transitions à partir d'un échantillon de n transitions*

Une version tabulaire de l'algorithme MFMC permettant la construction des trajectoires artificielles est donnée à la figure 1. A partir de la politique de décision h, de l'état initial x_0, de la mesure de distance Δ et du nombre de trajectoires p, l'algorithme MFMC renvoie un ensemble d'indices de transitions $\{l_t^i\}_{i=1,t=0}^{i=p,t=T-1}$ (indices correspondant au positionnement des transitions dans $\mathcal{F}_n$). En utilisant cet ensemble d'indices, on définit formellement l'estimateur MFMC de l'espérance du retour de la politique de décision h partant de l'état initial x_0, noté $\mathfrak{M}_p^h(\mathcal{F}_n, x_0)$:

$$\mathfrak{M}_p^h(\mathcal{F}_n, x_0) = \frac{1}{p}\sum_{i=1}^{p}\sum_{t=0}^{T-1} r^{l_t^i}.$$

Une illustration de l'estimateur MFMC est donnée en figure 2. Cette figure représente les transitions sélectionnées par l'algorithme MFMC, ainsi que les trajectoires réelles

qui seraient obtenues si elles étaient soumises aux mêmes perturbations que celles qui ont affecté la génération des transitions sélectionnées.

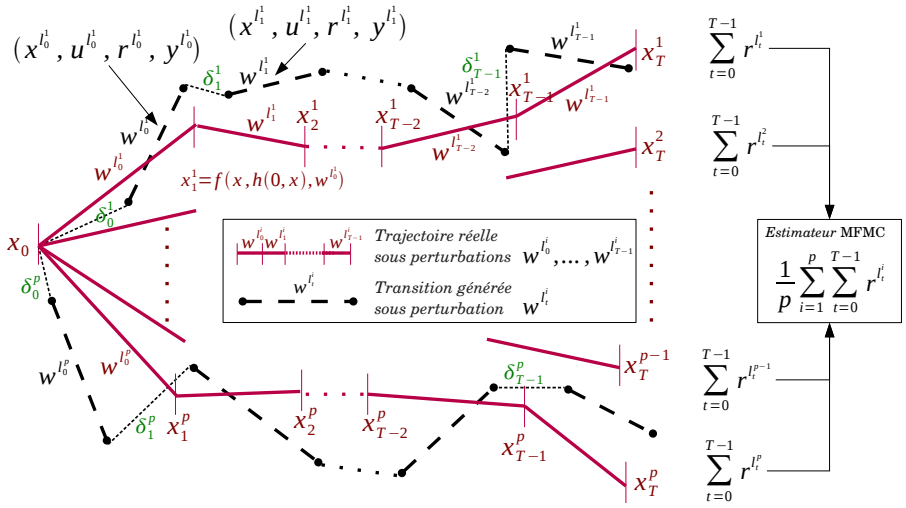

Figure 2. *L'estimateur MFMC reconstruit p trajectoires artificielles à partir de transitions du système*

La complexité de l'algorithme MFMC est linéaire en fonction de la cardinalité n de l'ensemble de transitions $\mathcal{F}_n$, de la longueur T des trajectoires artificielles et du nombre de trajectoires artificielles p.

4.3. *Etude théorique de l'estimateur MFMC*

Cette section est dédiée à l'analyse des principales propriétés théoriques de l'estimateur MFMC, à savoir son biais et sa variance.

1) Tout d'abord, on définit un ensemble d'échantillons de transitions compatibles avec l'échantillon $\mathcal{F}_n$ de la manière suivante : de l'échantillon $\mathcal{F}_n = \left[\left(x^l, u^l, r^l, y^l\right)\right]_{l=1}^{n}$, on extrait l'échantillon de couples état - décision

$$\mathcal{P}_n = \left[\left(x^l, u^l\right)\right]_{l=1}^{n} \in (\mathcal{X} \times \mathcal{U})^n,$$

et on considère tous les échantillons de n transitions qui pourraient être générés en complétant chaque paire (x^l, u^l) de $\mathcal{P}_n$ par un couple $(\rho(x^l, u^l, w^l), f(x^l, u^l, w^l)) \in \mathbb{R} \times \mathcal{U}$, w^l étant tirée selon la distribution de probabilité $p_{\mathcal{W}}(.)$. On désigne par $\tilde{\mathcal{F}}_n$ un tel échantillon "aléatoire" de transitions, dont le caractère aléatoire est induit par les n perturbations $w^l \quad l = 1 \ldots n$. L'échantillon $\mathcal{F}_n$ peut ainsi être considéré comme une réalisation de la variable aléatoire $\tilde{\mathcal{F}}_n$;

2) Ensuite, on étudie la distribution de l'estimateur $\mathfrak{M}_p^h(\tilde{\mathcal{F}}_n, x_0)$, considéré comme une fonction de $\tilde{\mathcal{F}}_n$; la distribution de $\mathfrak{M}_p^h(\tilde{\mathcal{F}}_n, x_0)$ est caractérisée par l'intermédiaire de son biais et de sa variance, que l'on exprime en fonction de la $k-$parcimonie de l'échantillon de couples état-décision $\mathcal{P}_n$. La $k-$parcimonie peut être interprétée comme le plus petit rayon γ tel que toutes les boules de rayon γ - selon la mesure de distance Δ - contiennent au moins k couples de $\mathcal{P}_n$. L'utilisation de la notion de $k-$parcimonie implique naturellement que l'espace $\mathcal{X} \times \mathcal{U}$ est borné selon la mesure de distance Δ.

Les caractérisations du biais et de la variance de l'estimateur MFMC nécessitent quelques hypothèses et définitions supplémentaires détaillées ci-dessous, et les théorèmes exprimant ces caractérisations sont donnés à la suite. Les démonstrations des théorèmes sont reportées en annexes.

Hypothèse 4.1 (Continuité Lipschitzienne des fonctions f, ρ et h) *La dynamique du système f, la fonction de récompense ρ ainsi que la politique de décision h sont supposées Lipschitziennes, c'est-à-dire qu'il existe trois constantes L_f, L_ρ et $L_h \in \mathbb{R}^+$ telles que* $\forall\, (x, x', u, u', w) \in \mathcal{X}^2 \times \mathcal{U}^2 \times \mathcal{W}$,

$$\begin{aligned}
&\|f(x,u,w) - f(x',u',w)\|_{\mathcal{X}} \leq L_f \left(\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}}\right), \\
&|\rho(x,u,w) - \rho(x',u',w)| \leq L_\rho \left(\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}}\right), \\
&\forall t \in [\![0, T-1]\!], \|h(t,x) - h(t,x')\|_{\mathcal{U}} \leq L_h \|x - x'\|_{\mathcal{X}} \ ,
\end{aligned}$$

où $\|.\|_{\mathcal{X}}$ et $\|.\|_{\mathcal{U}}$ désignent les normes sur les espaces $\mathcal{X}$ et $\mathcal{U}$, respectivement.

La mesure de distance utilisée dans cet article est définie de la manière suivante :

Définition 4.2 (Mesure de distance Δ)

$$\forall (x, x', u, u') \in \mathcal{X}^2 \times \mathcal{U}^2, \Delta\left((x,u),(x',u')\right) = \left(\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}}\right).$$

On suppose que l'espace $\mathcal{X} \times \mathcal{U}$ est borné selon la mesure de distance Δ, et, étant donné un entier $k \in \mathbb{N}^*$ tel que $k \leq n$, on définit la *$k-$parcimonie*, $\alpha_k(\mathcal{P}_n)$ de l'échantillon de couples état-décision $\mathcal{P}_n$ de la manière suivante :

Définition 4.3 ($k-$parcimonie)

$$\alpha_k\left(\mathcal{P}_n\right) = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} \left\{ \Delta_k^{\mathcal{P}_n}(x,u) \right\} \ ,$$

où $\Delta_k^{\mathcal{P}_n}(x,u)$ représente la distance entre $(x,u) \in \mathcal{X} \times \mathcal{U}$ et le $k-$ème couple le plus près (en utilisant la mesure de distance Δ) dans $\mathcal{P}_n$.

On définit également l'espérance de l'estimateur MFMC, notée $E_{p,\mathcal{P}_n}^h(x_0)$:

Définition 4.4 (Espérance de l'estimateur MFMC)

$$E^h_{p,\mathcal{P}_n}(x_0) = \underset{w^1,\dots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\mathfrak{M}^h_p(\tilde{\mathcal{F}}_n, x_0)\right].$$

Dès lors, on peut calculer des bornes supérieures sur le biais et la variance de l'estimateur MFMC, données dans les théorèmes suivants.

Théorème 4.5 (Biais de l'estimateur MFMC)

$$\left|J^h(x_0) - E^h_{p,\mathcal{P}_n}(x_0)\right| \leq C\alpha_{pT}(\mathcal{P}_n)$$
$$\textit{avec } C = L_\rho \sum_{t=0}^{T-1}\sum_{i=0}^{T-t-1} \left[L_f(1+L_h)\right]^i .$$

La démonstration de ce théorème est donnée en annexe A. L'expression donnée ci-dessus montre que le biais de l'estimateur converge vers 0 lorsque la $pT-$parcimonie tend vers 0. Il est à noter que la $pT-$parcimonie ne dépend que de l'échantillon de couples état-décision $\mathcal{P}_n$ et du paramètre p, le biais augmentant avec le nombre de trajectoires artificielles. On définit également la variance $V^h_{p,\mathcal{P}_n}(x_0)$ de l'estimateur MFMC :

Définition 4.6 (Variance de l'estimateur MFMC)

$$\begin{aligned} V^h_{p,\mathcal{P}_n}(x_0) &= \underset{w^1,\dots,w^n\sim p_{\mathcal{W}}(.)}{Var}\left[\mathfrak{M}^h_p(\tilde{\mathcal{F}}_n, x_0)\right] \\ &= \underset{w^1,\dots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\left(\mathfrak{M}^h_p(\tilde{\mathcal{F}}_n, x_0) - E^h_{p,\mathcal{P}_n}(x_0)\right)^2\right] . \end{aligned}$$

On a alors le théorème suivant.

Théorème 4.7 (Variance de l'estimateur MFMC)

$$V^h_{p,\mathcal{P}_n}(x_0) \leq \left(\frac{\sigma_{R^h}(x_0)}{\sqrt{p}} + 2C\alpha_{pT}\left(\mathcal{P}_n\right)\right)^2$$
$$\textit{avec } C = L_\rho \sum_{t=0}^{T-1}\sum_{i=0}^{T-t-1} \left[L_f(1+L_h)\right]^i .$$

La preuve de ce théorème est donnée en annexe B. On observe que la variance de l'estimateur MFMC converge vers la variance de l'estimateur MC lorsque la parcimonie de l'échantillon de transitions converge vers 0.

En ce qui concerne les deux théorèmes présentés dans cette section, il est à noter que les bornes calculées peuvent s'avérer particulièrement larges, étant donné le caractère assez général des hypothèses faites sur le système.

5. Illustration

Cette section propose une étude expérimentale du comportement de l'estimateur MFMC au travers d'un exemple illustratif.

5.1. *Description du problème*

La dynamique du système et la fonction de récompense sont données par :

$$\forall (x_t, u_t, w_t) \in \mathcal{X} \times \mathcal{U} \times \mathcal{W}, x_{t+1} = \sin\left(\frac{\pi}{2}(x_t + u_t + w_t)\right)$$

et

$$\forall (x_t, u_t, w_t) \in \mathcal{X} \times \mathcal{U} \times \mathcal{W}, \rho(x_t, u_t, w_t) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_t^2 + u_t^2)} + w_t \ ,$$

où l'espace d'état $\mathcal{X}$ est égal à $[-1, 1]$ et l'espace de décisions $\mathcal{U}$ est égal à $[-\frac{1}{2}, \frac{1}{2}]$. Chaque perturbation w_t appartient à l'intervalle $\mathcal{W} = [-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$ avec $\epsilon = 0.1$. $p_{\mathcal{W}}(.)$ est une distribution de probabilité uniforme sur l'intervalle $\mathcal{W}$. L'horizon d'optimisation T est égal à 15. La politique de décision h dont on souhaite évaluer l'espérance du retour est définie de la manière suivante :

$$\forall x \in \mathcal{X}, \forall t \in [\![0, T-1]\!], h(t, x) = -\frac{x}{2} \ .$$

L'état initial du sytème est fixé à $x_0 = -0.5$. L'échantillon de transitions $\mathcal{F}_n$, utilisé en remplacement du modèle supposé inconnu, a été généré selon le mécanisme décrit en section 4.3.

5.2. *Résultats*

Influence de la parcimonie. Dans un premier jeu d'expériences, on travaille à valeur de paramètre p fixée, $p = 10$. L'estimateur MFMC va ainsi construire $p = 10$ trajectoires artificielles afin d'estimer $J^h(-0.5)$. Pour différentes cardinalités

$$n_j = (10j)^2 \quad j = 1 \ldots 10,$$

50 échantillons de transitions $\mathcal{F}_{n_j}^1, \ldots, \mathcal{F}_{n_j}^{50}$ sont générés. L'estimateur MFMC est calculé pour chacun de ces échantillons. Etant donné une valeur de cardinalité $n_j = m_j^2$, les 50 différents échantillons $\mathcal{F}_{n_j}^1, \ldots, \mathcal{F}_{n_j}^{50}$ sont générés en considérant les mêmes couples état-décision $(x^l, u^l) \quad l = 1 \ldots n_j$ (c'est-à-dire l'échantillon $\mathcal{P}_{n_j}$). Cet échantillon de couples couvre l'espace $\mathcal{X} \times \mathcal{U}$ de manière "uniforme", c'est à dire

$$x^l = -1 + \frac{2j_1}{m_j} \text{ et } u^l = -1 + \frac{2j_2}{m_j} \text{ avec } j_1, j_2 \in [\![0, m_j - 1]\!].$$

Les résultats de ce premier jeu d'expériences sont rassemblés sur la figure 3. Pour chaque valeur de cardinalité n_j, les 50 calculs de l'estimateur MFMC sont représentés par une boîte à moustache. Chaque boîte s'étend du premier au troisième quartile,

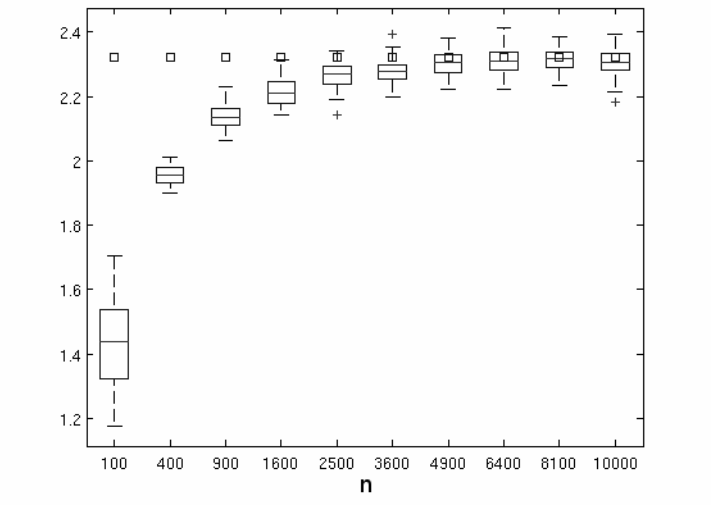

Figure 3. *Calculs de l'estimateur MFMC pour différentes cardinalités de l'échantillon de transitions, avec* $p = 10$*. Les carrés représentent* $J^h(x_0)$

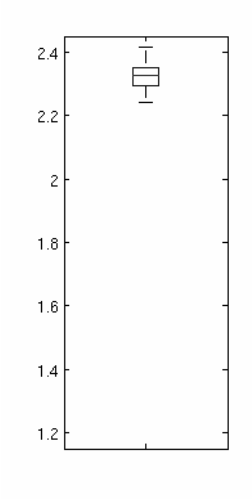

Figure 4. *Calculs de l'estimateur MC avec* $p = 10$

coupée par la médiane. Les valeurs extrêmes sont représentées par les moustaches qui s'étendent depuis les extrémités de la boîte jusqu'aux valeurs adjacentes des données, cela dans une limite de 1.5 fois la largeur de l'intervalle interquartile partant

de l'extrémité de la boîte. Les *outliers*, dont les valeurs sont au-dessus de la limite des moustaches, sont représentés par des croix. Les carrés représentent une estimation précise de $J^h(-0.5)$ obtenue par méthode de Monte Carlo pour un très grand nombre de simulations. On peut observer sur ce graphique que lorsque la cardinalité de l'échantillon de transitions augmente, ce qui correspond ici à une diminution de la $pT-$parcimonie $\alpha_{pT}\left(\mathcal{P}_n\right)$, la précision de l'estimation de $J^h(-0.5)$ par l'estimateur MFMC augmente. Comme cela a été expliqué précédemment, il existe des similitudes entre l'estimateur MFMC et l'estimateur MC. Cela s'observe expérimentalement en mettant la figure 3 en parallèle avec la figure 4, sur laquelle on a représenté les résultats obtenus par 50 calculs indépendants de l'estimateur MC pour $p = 10$ trajectoires réelles. On observe ainsi que l'estimateur MFMC tend à se comporter de manière similaire à l'estimateur MC quand la parcimonie de l'échantillon de transitions tend vers zéro.

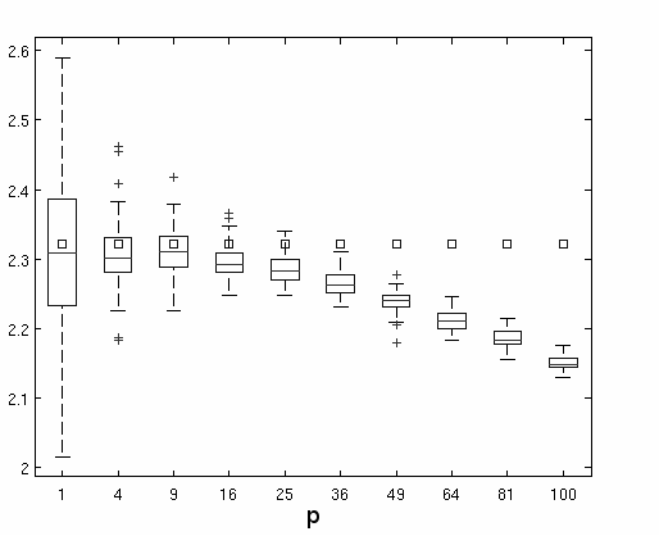

Figure 5. *Calculs de l'estimateur MFMC pour différentes valeurs du nombre de trajectoires artificielles p. Les carrés représentent* $J^h(x_0)$

Influence du nombre de trajectoires artificielles. Dans un deuxième jeu d'expériences, on souhaite observer l'influence sur la qualité de l'estimation du nombre de trajectoires artificielles p à partir desquelles l'estimateur MFMC calcule ses prédictions. Pour chaque valeur du nombre de trajectoires

$$p_j = j^2 \quad j = 1 \ldots 10,$$

50 échantillons $\mathcal{F}^1_{10,000}, \ldots, \mathcal{F}^{50}_{10,000}$ contenant chacun 10000 transitions sont générés. L'estimateur MFMC est calculé pour chaque échantillon de transitions, et les résultats sont rassemblés sur la figure 5. Cette figure montre que le biais de l'estimateur MFMC est assez faible pour les premières valeurs du nombre de trajectoires artificielles p, puis

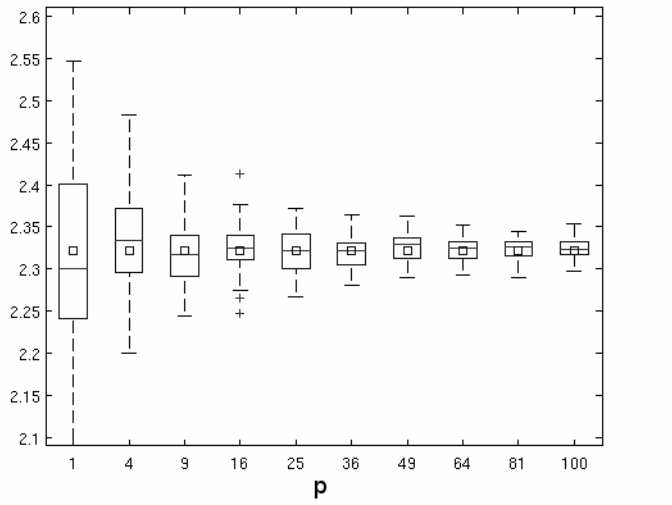

Figure 6. *Calculs de l'estimateur MC pour diférentes valeurs du nombre de trajectoires p. Les carrés représentent $J^h(x_0)$.*

augmente avec p. Ceci concorde avec le théorème 4.5 qui borne le biais de l'estimateur MFMC par un terme croissant en p.

La figure 6 représente l'évolution de l'estimateur MC lorsque le nombre de trajectoires réelles augmente. On constate que, pour les premières valeurs du paramètre p, le comportement de l'estimateur MC est peu différent de l'estimateur MFMC. En revanche, la précision de l'estimateur MC croît avec le paramètre p, ce qui n'est pas le cas de l'estimateur MFMC dont la précision se détériore fortement au-delà d'un certain seuil. Il est à noter que la valeur de ce seuil peut être augmentée en faisant croître la cardinalité de l'échantillon de transitions de manière à diminuer la $pT-$parcimonie.

6. Conclusions et perspectives d'améliorations

Cet article propose un estimateur de l'espérance du retour des politiques de décision dans un contexte où aucun modèle du système n'est disponible. Ce nouvel estimateur, nommé MFMC, fonctionne en reconstruisant des trajectoires artificielles à partir de transitions du système, préalablement collectées dans un échantillon, puis en moyennant les retours de ces trajectoires artificielles. A cet égard, l'estimateur MFMC peut être considéré comme une extension de l'estimateur de Monte Carlo classique au contexte où les simulations de la politique de décision sont impossibles. Cet article propose également des bornes théoriques sur le biais et la variance de l'estimateur MFMC ; ces bornes dépendent de la parcimonie de l'échantillon de trajectoires, ainsi

que des constantes de Lipschitz des fonctions de dynamique, de récompense et de la politique de décision à évaluer. Quand la parcimonie de l'échantillon tend vers 0, le biais et la variance de l'estimateur MFMC convergent vers les biais et variance de l'estimateur de Monte Carlo classique.

Différentes perspectives d'améliorations des travaux présentés dans cet article sont envisageables. Tout d'abord, il serait certainement plus réaliste de considérer un processus de perturbations qui dépendrait des couples état-décision, et il serait intéressant d'analyser comment les bornes calculées dans ce papier évolueraient sous ces nouvelles hypothèses. Ces bornes pourraient certainement permettre d'optimiser certaines valeurs des paramètres, comme par exemple le nombre de trajectoires artificielles p, où bien le choix de la mesure de distance Δ utilisée pour construire les trajectoires artificielles. Il est à noter que le calcul explicite des bornes nécessite la connaissance des constantes de Lipschitz (ou de bornes supérieures sur ces constantes), ce qui implique la mise au point d'un mécanisme permettant d'estimer de telles constantes à partir d'un échantillon de transitions.

D'autre part, la borne sur la variance de l'estimateur MFMC dépend explicitement de la variance "naturelle" du retour de la politique de décision. Utiliser cette borne dans l'optique de déterminer des valeurs optimales de p et Δ implique également la mise au point d'une technique de calcul de borne sur la variance naturelle à partir de l'échantillon de transitions.

Finalement, l'estimateur MFMC s'ajoute au catalogue des techniques capables d'estimer l'espérance du retour d'une politique de décision. Les avantages ainsi que les inconvénients de l'estimateur MFMC par rapport aux autres techniques d'estimations proposées dans la littérature méritent certainement d'être investigués dans de futurs travaux.

Remerciements

Ces travaux ont été financés par le FRIA (Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture), le FRS-FNRS (Fonds de la Recherche Scientifique) ainsi que le NIH (financements P50 DA10075 et R01 MH080015). Les résultats décrits dans cet article ont été obtenus grâce aux Pôles d'Attraction Interuniversitaire BIOMAGNET et DYSCO ainsi qu'au réseau européen d'excellence PASCAL2.

7. Bibliographie

Antos A., Munos R., Szepesvari C., « Fitted Q-iteration in continuous action-space MDPs », *in* J. Platt, D. Koller, Y. Singer, S. Roweis (eds), *Advances in Neural Information Processing Systems 20*, MIT Press, Cambridge, MA, p. 9-16, 2008.

Bradtke S., Barto A., « Linear least-squares algorithms for temporal difference learning », *Machine Learning*, vol. 22, p. 33-57, 1996.

Busoniu L., Babuska R., De Schutter B., Ernst D., *Reinforcement Learning and Dynamic Programming using Function Approximators*, Taylor & Francis CRC Press, 2010.

Dayan P., « The Convergence of TD(λ) for general λ », *Machine Learning*, vol. 8, p. 341-162, 1992.

Dimitrakakis C., Lagoudakis M. G., « Rollout sampling approximate policy iteration », *Machine Learning*, vol. 72, p. 157-171, 2008.

Ernst D., Geurts P., Wehenkel L., « Tree-based batch mode reinforcement learning », *Journal of Machine Learning Research*, vol. 6, p. 503-556, 2005.

Fonteneau R., Murphy S., Wehenkel L., Ernst D., « Model-free Monte Carlo–like policy evaluation », *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, JMLR : W&CP 9*, Chia Laguna, Sardinia, Italy, p. 217-224, 2010.

Munos R., Szepesvári C., « Finite-time bounds for fitted value iteration », *Journal of Machine Learning Research*, vol. 9, p. 815-857, 2008.

Ormoneit D., Sen S., « Kernel-based reinforcement learning », *Machine Learning*, vol. 49, n° 2-3, p. 161-178, 2002.

Riedmiller M., « Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method », *Proceedings of the Sixteenth European Conference on Machine Learning (ECML 2005)*, Porto, Portugal, p. 317-328, 2005.

Rummery G., Niranjan M., On-line Q-learning using connectionist systems, Technical Report n° 166, Cambridge University Engineering Department, 1994.

Sutton R., « Learning to predict by the methods of temporal difference », *Machine Learning*, vol. 3, p. 9-44, 1988.

Sutton R. S., Maei H. R., Precup D., Bhatnagar S., Silver D., Szepesvári C., Wiewiora E., « Fast gradient-descent methods for temporal-difference learning with linear function approximation », *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, ACM, New York, NY, USA, p. 993-1000, 2009.

Tsitsiklis J., « Asynchronous stochastic approximation and Q-learning », *Machine Learning*, vol. 16, p. 185-202, 1994.

Watkins C., Dayan P., « Q-Learning », *Machine Learning*, vol. 8, n° 3-4, p. 179-192, 1992.

Annexes

A. Démonstration du théorème 4.5

Avant de démontrer le théorème 4.5, on donne trois résultats préliminaires. Etant donné un vecteur de perturbations $\Omega = [\Omega(0), \ldots, \Omega(T-1)] \in \mathcal{W}^T$, on définit la fonction de valeur état-décision $Q^{h,\Omega}_{T-t}(x,u)$ pour $t \in [\![0, T-1]\!]$ perturbée par Ω de la manière suivante :

Définition A.1 (Fonction de valeur perturbée)

$$Q^{h,\Omega}_{T-t}(x,u) = \rho\left(x,u,\Omega(t)\right) + \sum_{t'=t+1}^{T-1} \rho(x_{t'}, h(t', x_{t'}), \Omega(t'))$$

avec

$$\begin{aligned} x_{t+1} &= f(x,u,\Omega(t))\,, \\ x_{t'+1} &= f(x_{t'}, h(t', x_{t'}), \Omega(t')), \forall t' \in [\![t+1, T-1]\!]\,. \end{aligned}$$

On définit ensuite l'espérance du retour de la politique h conditionnée à un vecteur de perturbations Ω :

Définition A.2 (Espérance du retour conditionnée à un vecteur de perturbations)

$$\mathbb{E}\left[R^h(x_0)|\Omega\right] = \mathop{\mathbb{E}}_{w_0,\ldots,w_{T-1}\sim p_{\mathcal{W}}(.)} \left[R^h(x_0)|w_0 = \Omega(0), \ldots, w_{T-1} = \Omega(T-1)\right].$$

A partir des définitions données ci-dessus, on obtient immédiatement les deux lemmes suivants :

Lemme A.3

$$\forall(\Omega, x_0) \in \mathcal{W}^T \times \mathcal{X}, \mathbb{E}\left[R^h(x_0)|\Omega\right] = Q^{h,\Omega}_T(x_0, h(0, x_0))$$

Lemme A.4
$\forall(x,u) \in \mathcal{X} \times \mathcal{U}$, $\forall \Omega \in \mathcal{W}^T$,

$$Q^{h,\Omega}_{T-t+1}(x,u) = \rho\left(x,u,\Omega(t-1)\right) + Q^{h,\Omega}_{T-t}\left(f\big(x,u,\Omega(t-1)\big), h\big(t, f(x,u,\Omega(t-1))\big)\right)$$

Ensuite, on démontre la continuité Lipschitzienne des fonctions de valeur état-décision perturbées :

Lemme A.5 (Continuité Lipschitzienne de $Q_{T-t}^{h,\Omega}$)
$\forall t \in [\![0, T-1]\!], \forall (x, x', u, u') \in \mathcal{X}^2 \times \mathcal{U}^2,$

$$\left|Q_{T-t}^{h,\Omega}(x,u) - Q_{T-t}^{h,\Omega}(x',u')\right| \leq L_{Q_{T-t}} \Delta((x,u),(x',u'))$$

avec $L_{Q_{T-t}} = L_\rho \sum_{i=0}^{T-t-1} \left[L_f(1+L_h)\right]^i$.

Démonstration du lemme A.5. La démonstration se fait par récurrence. On considère la propriété $\mathcal{H}(T-t)$ correspondant à la continuité Lipschitzienne de la fonction de valeur $Q_{T-t}^{h,\Omega}$:

$$\begin{aligned}
\mathcal{H}(T-t) \quad : \quad & \forall (x, x', u, u') \in \mathcal{X}^2 \times \mathcal{U}^2, \\
& \left|Q_{T-t}^{h,\Omega}(x,u) - Q_{T-t}^{h,\Omega}(x',u')\right| \leq L_{Q_{T-t}} \Delta((x,u),(x',u')) \\
& \text{avec } L_{Q_{T-t}} = L_\rho \sum_{i=0}^{T-t-1} \left[L_f(1+L_h)\right]^i,
\end{aligned}$$

et on montre par récurrence que $\mathcal{H}(T-t)$ est vraie, $\forall t \in [\![0, T-1]\!]$. Par souci de concision, on utilise dans cette démonstration la notation :

$$\Delta_{T-t}^{Q} = \left|Q_{T-t}^{h,\Omega}(x,u) - Q_{T-t}^{h,\Omega}(x',u')\right| .$$

– **Initialisation :** $t = T-1$**.**

On a

$$\Delta_1^Q = \left|\rho(x,u,\Omega(T-1)) - \rho(x',u',\Omega(T-1)\right| ,$$

et la continuité Lipschitzienne de ρ permet d'écrire :

$$\Delta_1^Q \leq L_\rho\big(\|x-x'\|_{\mathcal{X}} + \|u-u'\|_{\mathcal{U}}\big) \ = L_\rho \Delta((x,u),(x',u')) ,$$

ce qui démontre que $\mathcal{H}(1)$ est vraie.

– **Récurrence :** $1 \leq t \leq T-1$**.**

Supposons que $\mathcal{H}(T-t)$ soit vraie, pour $1 \leq t \leq T-1$. En utilisant le lemme A.4, on peut écrire :

$$\begin{aligned}
\Delta_{T-t+1}^{Q} &= \left|Q_{T-t+1}^{h,\Omega}(x,u) - Q_{T-t+1}^{h,\Omega}(x',u')\right| \\
&= \Big|\rho(x,u,\Omega(t-1)) - \rho(x',u',\Omega(t-1)) \\
&+ \ Q_{T-t}^{h,\Omega}(f(x,u,\Omega(t-1)), h(t,f(x,u,\Omega(t-1)))) \\
&- \ Q_{T-t}^{h,\Omega}(f(x',u',\Omega(t-1)), h(t,f(x',u',\Omega(t-1))))\Big| .
\end{aligned}$$

De là,

$$\begin{aligned}\Delta^{Q}_{T-t+1} &\leq \big|\rho(x,u,\Omega(t-1)) - \rho(x',u',\Omega(t-1))\big| \\ &+ \big|Q^{h,\Omega}_{T-t}(f(x,u,\Omega(t-1)), h(t,f(x,u,\Omega(t-1)))) \\ &- Q^{h,\Omega}_{T-t}(f(x',u',\Omega(t-1)), h(t,f(x',u',\Omega(t-1))))\big|.\end{aligned}$$

$\mathcal{H}(T-t)$ et la continuité Lipschitzienne de ρ permettent d'écrire :

$$\begin{aligned}\Delta^{Q}_{T-t+1} &\leq L_\rho \Delta((x,u),(x',u')) \\ &+ L_{Q_{T-t}} \Delta\Big((f(x,u,\Omega(t-1)), h(t,f(x,u,\Omega(t-1)))), \\ &\quad (f(x',u',\Omega(t-1)), h(t,f(x',u',\Omega(t-1))))\Big).\end{aligned}$$

Les continuités Lipschitziennes des fonctions f et h permettent d'écrire :

$$\begin{aligned}\Delta^{Q}_{T-t+1} &\leq L_\rho \Delta((x,u),(x',u')) \\ &+ L_{Q_{T-t}}\big(L_f \Delta((x,u),(x',u')) + L_h L_f \Delta((x,u),(x',u'))\big),\end{aligned}$$

puis,

$$\Delta^{Q}_{T-t+1} \leq L_{Q_{T-t+1}} \Delta((x,u),(x',u'))$$

étant donné que

$$L_{Q_{T-t+1}} \doteq L_\rho + L_{Q_{T-t}} L_f (1+L_h).$$

Cela démontre $\mathcal{H}\ (T-t+1)$ et clôt la démonstration. ■

Etant donnée une trajectoire artificielle $\tau^i = \left[\left(x^{l^i_t}, u^{l^i_t}, r^{l^i_t}, y^{l^i_t}\right)\right]_{t=0}^{T-1}$, on désigne par Ω^{τ^i} le vecteur de perturbations associé $\Omega^{\tau^i} = \left[w^{l^i_0}, \ldots, w^{l^i_{T-1}}\right]$, c'est à dire le vecteur constitué des T perturbations (inconnues) qui ont affecté la génération des transitions $\left(x^{l^i_t}, u^{l^i_t}, r^{l^i_t}, y^{l^i_t}\right)$ (cf. premier point de la section 4.3). On démontre alors le lemme suivant.

Lemme A.6 (Bornes sur l'espérance du retour conditionnée à Ω)

$$\forall i \in [\![1,p]\!],\ b^h(\tau^i, x_0) \leq \mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right] \leq a^h(\tau^i, x_0)\ ,$$

avec

$$b^h(\tau^i, x_0) = \sum_{t=0}^{T-1} \left[r^{l_t^i} - L_{Q_{T-t}} \delta_t^i \right] ,$$

$$a^h(\tau^i, x_0) = \sum_{t=0}^{T-1} \left[r^{l_t^i} + L_{Q_{T-t}} \delta_t^i \right] ,$$

$$\delta_t^i = \Delta \left((x^{l_t^i}, u^{l_t^i}), (y^{l_{t-1}^i}, h(t, y^{l_{t-1}^i})) \right) , \forall t \in [\![0, T-1]\!] ,$$

$$y^{l_{-1}^i} = x_0, \forall i \in [\![1, p]\!].$$

Démonstration du lemme A.6. On donne la démonstration pour le calcul de la borne inférieure. La borne supérieure se calcule suivant le même principe. Avec $u_0 = h(0, x_0)$, la continuité Lipschitzienne de $Q_T^{h,\Omega^{\tau^i}}$ implique :

$$\left| Q_T^{h,\Omega^{\tau^i}}(x_0, u_0) - Q_T^{h,\Omega^{\tau^i}}(x^{l_0^i}, u^{l_0^i}) \right| \leq L_{Q_T} \Delta \left((x_0, u_0), (x^{l_0^i}, u^{l_0^i}) \right) .$$

Le lemme (A.3) donne :

$$Q_T^{h,\Omega^{\tau^i}}(x_0, u_0) = \mathbb{E} \left[R^h(x_0) | \Omega^{\tau^i} \right] .$$

On a donc :

$$\begin{aligned} \left| \mathbb{E} \left[R^h(x_0) | \Omega^{\tau^i} \right] - Q_T^{h,\Omega^{\tau^i}}(x^{l_0^i}, u^{l_0^i}) \right| &= \left| Q_T^{h,\Omega^{\tau^i}}(x_0, h(0, x_0)) - Q_T^{h,\Omega^{\tau^i}}(x^{l_0^i}, u^{l_0^i}) \right| \\ &\leq L_{Q_T} \Delta \left((x_0, h(0, x_0)), (x^{l_0^i}, u^{l_0^i}) \right) . \end{aligned} \quad [1]$$

Il s'ensuit que :

$$Q_T^{h,\Omega^{\tau^i}}(x^{l_0^i}, u^{l_0^i}) - L_{Q_T} \delta_0^i \leq \mathbb{E} \left[R^h(x_0) | \Omega^{\tau^i} \right] .$$

En utilisant le lemme (A.4), on a :

$$Q_T^{h,\Omega^{\tau^i}}(x^{l_0^i}, u^{l_0^i}) = \rho \left(x^{l_0^i}, u^{l_0^i}, w^{l_0^i} \right) + Q_{T-1}^{h,\Omega^{\tau^i}} \left(f(x^{l_0^i}, u^{l_0^i}, w^{l_0^i}), h(1, f(x^{l_0^i}, u^{l_0^i}, w^{l_0^i})) \right) .$$

Par définition de Ω^{τ^i}, on a :

$$\rho \left(x^{l_0^i}, u^{l_0^i}, w^{l_0^i} \right) = r^{l_0^i} \text{ et } f \left(x^{l_0^i}, u^{l_0^i}, w^{l_0^i} \right) = y^{l_0^i} .$$

Il s'ensuit que :

$$Q_T^{h,\Omega^{\tau^i}} \left(x^{l_0^i}, u^{l_0^i} \right) = r^{l_0^i} + Q_{T-1}^{h,\Omega^{\tau^i}} \left(y^{l_0^i}, h(1, y^{l_0^i}) \right) ,$$

et

$$Q_{T-1}^{h,\Omega^{\tau^i}}\left(y^{l_0^i},h(1,y^{l_0^i})\right)+r^{l_0^i}-L_{Q_T}\delta_0^i\leq\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\ .$$

La continuité Lipschitzienne de $Q_{T-1}^{h,\Omega^{\tau^i}}$ donne :

$$\begin{aligned}\left|Q_{T-1}^{h,\Omega^{\tau^i}}(y^{l_0^i},h(1,y^{l_0^i}))-Q_{T-1}^{h,\Omega^{\tau^i}}(x^{l_1^i},u^{l_1^i})\right|&\leq L_{Q_{T-1}}\Delta\left((y^{l_0^i},h(1,y^{l_0^i})),(x^{l_1^i},u^{l_1^i})\right)\\&=L_{Q_{T-1}}\delta_1^i,\end{aligned}$$

ce qui implique :

$$Q_{T-1}^{h,\Omega^{\tau^i}}\left(x^{l_1^i},u^{l_1^i}\right)-L_{Q_{T-1}}\delta_1^i\leq Q_{T-1}^{h,\Omega^{\tau^i}}\left(y^{l_0^i},h(1,y^{l_0^i})\right)\ .$$

Finalement, on obtient :

$$Q_{T-1}^{h,\Omega^{\tau^i}}\left(x^{l_1^i},u^{l_1^i}\right)+r^{l_0^i}-L_{Q_T}\delta_0^i-L_{Q_{T-1}}\delta_1^i\leq\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right].$$

La calcul de la borne inférieure s'obtient en développant l'itération ci-dessus. ■

On démontre un lemme supplémentaire.

Lemme A.7 (Largeur des bornes)

$$\forall i\in[\![1,p]\!],a^h\left(\tau^i,x_0\right)-b^h\left(\tau^i,x_0\right)\leq 2C\alpha_{pT}\left(\mathcal{P}_n\right)\qquad[2]$$

avec

$$C=\sum_{t=0}^{T-1}L_{Q_{T-t}}\ .$$

Démonstration du lemme A.7. Par construction des bornes, on a la relation

$$a^h\left(\tau^i,x_0\right)-b^h\left(\tau^i,x_0\right)=\sum_{t=0}^{T-1}2L_{Q_{T-t}}\delta_t^i.$$

L'algorithme MFMC sélectionne $p\times T$ transitions distinctes en minimisant les distances $\Delta\left((y^{l_{t-1}^i},h(t,y^{l_{t-1}^i})),(x^{l_t^i},u^{l_t^i})\right)$. Par conséquent, étant donnée la définition de la $k-$parcimonie de l'échantillon $\mathcal{P}_n$ avec $k=pT$, on a

$$\begin{aligned}\delta_t^i=\Delta\left((y^{l_{t-1}^i},h(t,y^{l_{t-1}^i})),(x^{l_t^i},u^{l_t^i})\right)&\leq\Delta_{pT}^{\mathcal{P}_n}\left(y^{l_{t-1}^i},h(t,y^{l_{t-1}^i})\right)\\&\leq\alpha_{pT}(\mathcal{P}_n)\ ,\end{aligned}$$

ce qui termine la démonstration. ■

A partir des lemmes démontrés précédemment, on peut désormais calculer une borne supérieure sur le biais de l'estimateur MFMC.

Démonstration du théorème 4.5. Par définition des bornes $a^h\left(\tau^i, x_0\right)$ et $b^h\left(\tau^i, x_0\right)$, on a

$$\forall i \in [\![1, p]\!], \frac{b^h\left(\tau^i, x_0\right) + a^h\left(\tau^i, x_0\right)}{2} = \sum_{t=0}^{T-1} r^{l_t^i} .$$

Dès lors, étant donnés les lemmes A.6 et A.7, on a $\forall i \in [\![1, p]\!]$,

$$\begin{aligned}
&\left| \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\mathbb{E}\left[R^h(x_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right| \\
&\leq \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\left| \mathbb{E}\left[R^h(x_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right| \right] \leq C \alpha_{pT}(\mathcal{P}_n) .
\end{aligned}$$

Ensuite,

$$\begin{aligned}
&\left| \frac{1}{p} \sum_{i=1}^{p} \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\mathbb{E}\left[R^h(x_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right| \\
&\leq \frac{1}{p} \sum_{i=1}^{p} \left| \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\mathbb{E}\left[R^h(x_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right| \leq C \alpha_{pT}(\mathcal{P}_n) ,
\end{aligned}$$

ce qui peut être reformulé de la manière suivante :

$$\left| \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\frac{1}{p} \sum_{i=1}^{p} \mathbb{E}\left[R^h(x_0) | \Omega^{\tau^i} \right] \right] - E^h_{p,\mathcal{P}_n}(x_0) \right| \leq C \alpha_{pT}(\mathcal{P}_n) ,$$

puisque

$$\frac{1}{p} \sum_{i=1}^{p} \sum_{t=0}^{T-1} r^{l_t^i} = \mathfrak{M}^h_p(\tilde{\mathcal{F}}_n, x_0) .$$

Comme l'algorithme MFMC sélectionne $p \times T$ transitions distinctes, les perturbations $\left\{ w^{l_t^i} \right\}_{i=1,t=0}^{i=p,t=T-1}$ sont i.i.d. selon $p_{\mathcal{W}}(.)$. Quel que soit $i \in [\![1, p]\!]$, le Théorème de l'espérance totale implique que :

$$\begin{aligned}
&\underset{w^{l_0^i},\ldots,w^{l_{T-1}^i} \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[\underset{w^{l_0^i},\ldots,w^{l_{T-1}^i} \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[R^h(x_0) | \Omega^{\tau^i} \right] \right] \\
&= \underset{w_0,\ldots,w_{T-1} \sim p_{\mathcal{W}}(.)}{\mathbb{E}} \left[R^h(x_0) \right] = J^h(x_0) ,
\end{aligned}$$

ce qui conclut la démonstration. ■

B. Démonstration du théorème 4.7

Pour commencer, on démontre le lemme suivant.

Lemme B.1 (Variance d'une somme de variables aléatoires)
Soit $X_0, \ldots, X_{T-1}$ T variables aléatoires de variances finies $\sigma_0^2, \ldots, \sigma_{T-1}^2$, respectivement. On a alors :

$$Var\left[\sum_{t=0}^{T-1} X_t\right] \leq \left(\sum_{t=0}^{T-1} \sigma_t\right)^2 .$$

Démonstration du lemme B.1. La démonstration s'obtient par récurrence sur le nombre de variables aléatoires, cela en utilisant la formule

$$Cov(X_i, X_j) \leq \sigma_i \sigma_j \ , \forall i, j \in [\![0, T-1]\!],$$

qui découle directement de l'inégalité de Cauchy-Schwarz. ■

Démonstration du théorème 4.7. On désigne par $\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n, x_0)$ la variable aléatoire

$$\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n, x_0) = \mathfrak{M}_p^h(\tilde{\mathcal{F}}_n, x_0) - \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right].$$

En utilisant le lemme B.1, on peut écrire

$$\begin{aligned} V_{p,\mathcal{P}_n}^h(x_0) \quad \leq \quad & \left(\sqrt{\underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{Var}\left[\frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right]}\right. \\ + \quad & \left.\sqrt{\underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{Var}\left[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n, x_0)\right]}\right)^2 \end{aligned} \qquad [3]$$

Comme les perturbations $\left\{w^{l_t^i}\right\}_{i=1,t=0}^{i=p,t=T-1}$ sont i.i.d. suivant la distribution $p_{\mathcal{W}}(.)$ (cf démonstration du théorème 4.5), le Théorème de l'espérance totale permet d'écrire

$$\underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{Var}\left[\frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right] = \frac{\sigma_{R^h}^2(x_0)}{p} . \qquad [4]$$

Concentrons nous sur $\underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{Var}\left[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n, x_0)\right]$. Par définition, on a

$$\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n, x_0) = \frac{1}{p}\sum_{i=1}^{p}\left[\sum_{t=0}^{T-1} r^{l_t^i} - \mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right] .$$

Dès lors, en utilisant le lemme B.1, on a

$$\begin{aligned}
&\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\left[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0)\right]\\
&\leq \frac{1}{p^2}\left(\sum_{i=1}^{p}\sqrt{\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\left[\sum_{t=0}^{T-1}r^{l_t^i}-\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right]}\quad\right)^2 \qquad [5]
\end{aligned}$$

Ensuite,

$$\begin{aligned}
&\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\left[\sum_{t=0}^{T-1}r^{l_t^i}-\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right]\\
&\leq \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\left(\sum_{t=0}^{T-1}r^{l_t^i}-\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]\right)^2\right]\\
&\leq \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\left(a^h(\tau^i,x_0)-b^h(\tau^i,x_0)\right)^2\right]=\left(a^h(\tau^i,x_0)-b^h(\tau^i,x_0)\right)^2\\
&\qquad\qquad\leq 4C^2(\alpha_{pT}(\mathcal{P}_n))^2\,, \qquad [6]
\end{aligned}$$

puisque $\sum_{t=0}^{T-1}r^{l_t^i}$ et $\mathbb{E}\left[R^h(x_0)|\Omega^{\tau^i}\right]$ appartiennent à l'intervalle $\left[b^h(\tau^i,x_0),a^h(\tau^i,x_0)\right]$ dont la largeur est bornée par $2C\alpha_{pT}(\mathcal{P}_n)$, d'après le lemme A.7.

En rassemblant les equations (3), (4), (5) et (6), on obtient :

$$V_{p,\mathcal{P}_n}^h(x_0)\leq\left(\frac{\sigma_{R^h}(x_0)}{\sqrt{p}}+2C\alpha_{pT}(\mathcal{P}_n)\right)^2,$$

ce qui termine la démonstration. ■