Echantillonnage adaptatif optimal dans les champs de Markov discrets

Mathieu Bonneau, Nathalie Dubois Peyrard Peyrard, Régis Sabbadin

▶ **To cite this version:**

Mathieu Bonneau, Nathalie Dubois Peyrard Peyrard, Régis Sabbadin. Echantillonnage adaptatif optimal dans les champs de Markov discrets. JFPDA 2011 - Sixièmes journées francophones de planification, décision et apprentissage pour la conduite de systèmes, Labo/service de l'auteur, Ville service, Pays service., Jun 2011, Rouen, France. ⟨hal-02744532⟩

HAL Id: hal-02744532

https://hal.inrae.fr/hal-02744532v1

Submitted on 3 Jun 2020

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.

HAL Authorization

Echantillonnage adaptatif optimal dans les champs de Markov discrets

Mathieu BONNEAU[1], Nathalie PEYRARD[1], and Régis SABBADIN[1]

INRA Toulouse - Unité de Biométrie et Intelligence Artificielle - UR 875

Résumé : La question de l'échantillonnage optimal d'information spatiale a été largement étudiée dans le cas d'observations à valeurs réelles. A l'inverse, peu de travaux traitent du cas d'observations discrètes, pourtant pertinant dans de nombreux problèmes de cartographie de systèmes biologiques. Nous proposons d'exploiter le cadre des champs de Markov, classiquement utilisés en analyse d'image, à la fois pour représenter la distribution spatiale du phéomène étudié, et également pour définir une notion d'utilisé d'un échantillon. Nous considérons le problème de la conception d'une stratégie adaptative d'éhantillonnage et l'échantillon optimal est défini comme celui maximisant un critère d'utilité en espérance quantifiant la qualité de la carte reconstruire et les coûts d'échantillonnage. Nous présentons ensuite un algorithme de résolution de type apprentissage par renforcement, reposant sur la modélisation du problème d'origine dans le cadre des processus décisionnels de Markov. En nous appuyant sur une validation empirique, nous illustrons les avantages de cette approche qui apporte un gain significatif en temps de calcul et fournit des politiques approchées satisfaisantes.

1 Introduction

Le problème de l'échantillonnage spatial peut être défini de la manière suivante : considérons un ensemble d'unités géographiques sur lesquelles une information sur un phénomène spatial peut-être mesurée, mais les ressources disponibles pour collecter ces informations sont limitées. La question est alors de décider sur quelles unités mesurer le phénomène afin d'obtenir les observations les plus utiles pour reconstruire le phénomène complet. Dans une approche basée sur la modélisation, cela pose la question de l'échantillonnage optimal dans un champ aléatoire spatial. Il s'agit d'un problème complexe qui mobilise plusieurs disciplines en statitiques spatiales (de Gruijter *et al.*, 2006; Müller, 2007) et en intelligence artificielle (Krause *et al.* 2008; Krause & Guestrin 2009; Peyrard *et al.* 2010), et soulève des défis méthodologiques en modélisation, inférence et algorithmique. Un champ de recherche très actif s'intéresse au cas où les observations sont à valeurs continues (par exemple pour la surveillance des températures ou d'indices de pollution). Des modèles et des algorithmes efficaces ont été proposés (M. Fuentes & Holland, 2007; Krause *et al.*, 2008), reposant principalement sur les outils de la géostatistique (champ aléatoire gaussien et krigeage). A l'inverse, peu de travaux concernent le cas des observations discrètes. Pourtant cette situation se retrouve dans de nombreuses études sur les systèmes biologiques (Wiles 2005). L'observation discrète peut être par exemple la classe d'abondance d'une espèce rare ou invasive, une classe de sévérité en épidémiologie, ou simplement une valeur binaire traduisant la présence ou l'absence.
Résoudre des problèmes d'échantillonnage optimal dans les champs aléatoires discrets reste donc une question ouverte. Les verrous méthodologiques portent principalement sur la modélisation et l'efficacité computationnelle de la résolution. Les modèles utilisés doivent être suffisamment réalistes pour représenter l'incertitude sur la carte à reconstruire ainsi que le coût de l'échantillonnage. Il s'agit ensuite de concevoir des méthodes de résolution approchée à la fois de complexité raisonnable et de qualité satisfaisante. Nous proposons de nous placer dans le cadre des champs de Markov discrets (CM, Geman & Geman 1984) classiquement utilisé en analyse d'image, et de définir l'utilité des observations à partir du critère Maximum Posterior Marginals (MPM) souvent utilisé pour la reconstruction d'image à partir d'observations bruitées. Avec ces éléments, nous définissons le problème d'échantillonnage adaptatif optimal dans un CM. En échantillonnage adaptatif, l'ensemble des sites échantillonnés est choisi de manière séquentielle et les observations

acquises lors des étapes précédentes sont prises en compte pour sélectionner les prochains sites à explorer. L'échantillonnage adaptatif est à préférer à l'option statique (Thompson & Seber, 1996) lorsqu'il est compatible avec les contraintes dues à la mise en œuvre de cette méthode sur le terrain. Cependant, le problème d'optimisation correspondant est plus complexe. Dans cet article nous proposons un algorithme d'optimisation basé sur la simulation. Pour cela nous modélisons le problème d'échantillonnage adaptatif optimal dans le cadre des Processus Décisionnels de Markov (PDM, Puterman 1994). Cela nous permet d'exploiter les principes de l'Apprentissage par Renforcement (AR, Sutton & Barto 1998), une technique efficace de résolution des PDM basée sur la simulation. La formalisation du problème d'échantillonnage adaptatif optimal dans un CM est introduite dans la section 2. Nous montrons ensuite comment le modéliser dans le cadre PDM (section 3). Enfin, nous décrivons un algorithme de résolution de type AR dans la section 4 et nous présentons une évaluation empirique de la méthode dans la section 5. Des perspectives méthodologiques et finalisées à ces travaux sont discutées dans la section 6.

2 Echantillonnage adaptatif optimal dans les champs de Markov

Soit $X = (X_1, \ldots, X_n)$ un ensemble de variables aléatoires à valeur dans $\Omega^n = \{1, \ldots, K\}^n$. L'ensemble $V = \{1, \ldots, n\}$ est l'ensemble des indices du vecteur X et un élément $i \in V$ est appelé un site. Nous noterons $x = (x_1, \ldots, x_n)$ une réalisation de X et $x_B = \{x_i\}_{i \in B}$, $\forall B \subseteq V$ une réalisation d'un sous-ensemble de X. La distribution de X, notée $\mathbb{P}(. \mid \theta)$ (θ est un vecteur de paramètres [1]) est celle d'un Champ de Markov (CM) associé à un graphe de voisinage $G = (V, E)$ où E est l'ensemble des arêtes : la distribution s'écrit donc de manière factorisée $\mathbb{P}(X = x \mid \theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \Psi_c(x_c, \theta)$, avec $\mathcal{C}$ l'ensemble des clique de G et $\Psi_c, c \in \mathcal{C}$ des fonctions potentielles (positives). Enfin, $Z(\theta)$ est la constante de normalisation de la distribution.
Nous considérons que l'objectif du problème d'échantillonnage est de reconstruire le vecteur X sur un sous-ensemble $R \subseteq V$ indiçant les variables dites d'intérêt. Nous supposerons que les observations peuvent être acquises uniquement sur un sous-ensemble $O \subseteq V$ de sites indiçant les variables observables, et tel que $R \cup O = V$. L'intersection entre O et R peut être non vide. Le problème est alors celui du choix d'un ensemble de sites $A \subseteq O$, appelé *plan d'échantillonnage*, où X sera observé, dans le but de reconstruire X_R, tout en assurant un compromis entre qualité de la reconstruction et coût d'échantillonnage.
Une observation x_A définit un nouveau CM, $\mathbb{P}(. \mid x_A, \theta)$, qui peut alors être utilisé pour reconstruire X_R. Nous nous basons pour cela sur le critère Maximum Posterior Marginals(MPM). L'estimateur MPM x_R^* de l'état des variables X_R est défini comme la configuration qui maximise l'espérance du nombre de variables bien restaurées sous l'hypothèse que la vraie carte suit la distribution $\mathbb{P}(. \mid x_A, \theta)$ [2] :

$$x_R^* = \Big\{ x_i^*, i \in R, \quad x_i^* = \operatorname*{argmax}_{x_i \in \Omega} \mathbb{P}(x_i \mid x_A, \theta) \Big\}.$$

Notons ici que la probabilité marginale de x_i^* ne dépend pas de l'ordre dans lequel les observations x_A ont été obtenues. Elle dépend uniquement de leur valeur et leur position.
Afin de quantifier la qualité de l'*échantillon* (A, x_A) en terme de reconstruction de X_R, nous utilisons les probabilités marginales de l'estimateur MPM :

$$MPM(A, x_A) \quad = \quad \sum_{i \in R} \left[\max_{x_i \in \Omega} \Big\{ \mathbb{P}(x_i \mid x_A, \theta) \Big\} \right]. \tag{1}$$

Par ailleurs, si les observations sont nécessaires pour une bonne reconstruction, leur acquisition consomme des ressources (temps, argent). Nous définissons donc une fonction de coût c qui associe

1. Nous considèrerons que θ est connu.

2. Un autre choix aurait pu être le critère de Maximum a Posteriori. Nous ne l'avons pas retenu car le mode d'une distribution sur un espace d'état de grande dimension peut ne pas être très marqué. D'autre part, une quantification de la performance de l'échantillonnage en terme d'espérance du nombre de sites bien restaurés est plus facilement interprétable qu'une probabilité de bonne restauration de la carte dans sa globalité.

à chaque plan d'échantillonnage A une valeur réelle positive qui quantifie les ressources nécessaires pour observer l'état des variables X_A. La modélisation du coût d'échantillonnage est une question à part entière. Ici nous illustrons cette notion sur une définition très simple où les coûts d'échantillonnage sont additifs

$$c(A) \quad = \quad \sum_{i \in A} c_i, \ c_i \in \mathbb{R}^+. \tag{2}$$

Cette définition particulière du coût, là encore, ne dépend pas de l'ordre dans lequel les sites ont été visités. Cela ne serait certainement plus vérifié pour des coûts plus réalistes, basés par exemple sur la distance parcourue lors d'une étape d'échantillonnage.
Finalement, l'utilité U d'un échantillon (A, x_A) est définie comme un compromis entre la qualité de la reconstruction définie par l'équation (1) et le coût de A défini par l'équation (2) :

$$U\big(A, x_A\big) = MPM(A, x_A) - \alpha c(A),$$

où la constante α permet d'homogénéiser les échelles de la qualité et du coût.
Afin d'améliorer la qualité de la reconstruction, le plan d'échantillonnage peut être choisi de manière séquentielle. On parle alors d'échantillonnage adaptatif. L'échantillonnage est alors découpé en H plans d'échantillonnage successifs $(A^1, \ldots, A^H)$, choisis en fonction des résultats des plans précédents. Une politique d'échantillonnage adaptative $\delta = (\delta^1, \ldots, \delta^H)$ se définie alors de manière naturelle : $\delta^1 = A^1$ est fixée pour une politique donnée δ et à chaque étape $t \in \{2, \ldots, H\}$: $A^t = \delta^t((A^1, x_{A^1}), \ldots, (A^{t-1}, x_{A^{t-1}}))$.
Nous appellerons *historique* une trajectoire $(A^1, x_{A^1}), \ldots, (A^H, x_{A^H})$ suivie en appliquant la politique δ. L'ensemble des historiques atteignables pour une politique δ est τ_δ. Dans la suite, nous considèrerons uniquement l'ensemble Δ_L des politiques adaptatives telles qu'à chaque étape la taille de l'échantillon est au plus égal à L $(1 \leq L \leq |O|)$. Par ailleurs, nous supposerons que les observations sont fiables, ce qui implique qu'il nous suffit de considérer les politiques qui visitent chaque site au plus une fois. Dans ce cas, l'information nécessaire contenue dans un historique peut se résumer par la paire (A, x_A), où $A = \cup_h A^h$ et $x_A = \cup_h x_{A^h}$.
La qualité d'une politique est définie comme l'espérance de l'utilité U du résumé d'un historique atteignable :

$$V(\delta) = \sum_{(A, x_A) \in \tau_\delta} \mathbb{P}\big(x_A \mid \theta\big) U\big(A, x_A\big).$$

Enfin, le problème d'échantillonnage adaptatif optimal dans un CM consiste à calculer la politique maximisant V :

$$\delta^* = \arg\max_{\delta \in \Delta_L} V(\delta). \tag{3}$$

L'optimisation exacte n'est pas accessible. En effet, le seul calcul de $\mathbb{P}\big(x_i \mid x_A, \theta\big)$ est connu pour être un problème $\#P$-difficile et l'approcher à $2^{n^{1-\varepsilon}}$ près pour un $\varepsilon > 0$ quelconque est NP-difficile (Roth, 1996). Il est donc indispensable de disposer de méthodes de résolution approchées. Dans la section suivante, nous décrivons comment modéliser le problème d'échantillonnage adaptatif optimal comme un Processus Décisionnel de Markov (PDM) (Puterman, 1994) de taille exponentielle. Cela nous permet ensuite de le résoudre par Apprentissage par Renforcement (AR, Sutton & Barto 1998).

3 Modélisation en processus décisionnel de Markov

Les PDM offrent un cadre mathématique et des algorithmes d'optimisation efficaces pour la résolution de problèmes de décision séquentielle dans l'incertain. Le problème d'échantillonnage Adaptatif Optimal dans un CM (EAOCM) peut être modélisé comme un PDM à horizon fini dont les espaces d'état et de décisions sont de taille exponentielle en la taille du problème d'origine. Cette représentation exponentielle ne peut être évitée puisqu'il est connu qu'un PDM peut être résolut en un temps polynomial en sa représentation, alors que les problèmes d'échantillonnage optimal dans un CM sont $\#P$-difficiles (voir section précédente).
Nous rappelons d'abord rapidement la définition d'un PDM avant de présenter la modélisation du

problème EAOCM. Un PDM à horizon fini est défini par un 5-tuple $< S, D, T, p, r >$, où S est un ensemble fini de N *états* possibles du système, D est un ensemble fini de M *décisions* possibles, $T = \{1, \ldots, H\}$ est un ensemble fini d'étapes de décision appelé *horizon*. Un *état terminal* $s^{H+1} \in S$ est atteint après que la dernière décision ait été appliquée à l'étape H. La *récompense* $r^t(s^t, d^t) \in \mathbb{R}$ est obtenue lorsque le système est dans l'état s^t au temps t et la décision d^t est appliquée. Lorsque l'état s^{H+1} est rencontré au temps $H+1$, une récompense finale $r^{H+1}(s^{H+1})$ est attribuée.
Une *politique* $\pi = \{\pi^1, \ldots, \pi^H\}$ est un ensemble de fonctions de décision $\pi^t : S \to D$. La *fonction de valeur* $V^\pi : S \times T \to \mathbb{R}$ d'une politique π est définie comme l'espérance de la somme des récompenses futures qui peuvent être obtenues à partir de l'état et du temps courant lorsque le système suit la chaîne de Markov définie par π. Résoudre un PDM consiste alors à trouver la politique optimale π^* dont la valeur est maximale pour tout état à toute étape de décision : $V^{\pi^*}(s,t) \geq V^\pi(s,t), \forall \pi, s, t$.

Nous modélisons le problème EAOCM comme un PDM de la manière suivante :
Espace d'état. L'état s^t $(t = 1, \ldots, H+1)$ du PDM va résumer l'information disponible au temps t sur les variables indicées dans O. Cette information au temps $t \geq 2$ est obtenue à partir des échantillons passés : $\big((A^1, x_{A^1}), \ldots, (A^{t-1}, x_{A^{t-1}})\big)$. Comme nous l'avons déjà précisé, il n'est pas nécessaire de connaître l'ordre dans lequel les observations ont été obtenues pour évaluer la qualité d'une politique d'échantillonnage. Ainsi nous pouvons définir l'état au temps t comme une paire (A, x_A) :

$$s^t = (A, x_A) = \left(\bigcup_{h=1}^{t-1} A^h, \bigcup_{h=1}^{t-1} x_{A^h} \right), \forall t = 2, \ldots, H+1.$$

Le PDM du problème EAOCM a un unique état initial $s^1 = (\emptyset, \emptyset)$. Avec cet encodage, le nombre d'états possibles du système est exponentiel en la taille de représentation du problème EAOCM. Il est majoré par $(2K)^{|O|}$.
Espace de décision. La décision d^t représente le plan d'échantillonnage choisi à l'étape t. Il peut être modélisé par exemple par un vecteur binaire de longueur $|O|$ avec au plus L entrées à 1 (les autres étant à 0). Dans ce cas, le nombre total de décisions est au plus $|\{A \subseteq O,\ |A| \leq L\}| = \sum_{l=0}^{L} \binom{|O|}{l}$.
Horizon. Les étapes de décision du PDM correspondent aux étapes de décision dans le problème EAOCM. Ainsi, $T = \{1, \ldots, H\}$. Après que la décision d^H ait été appliquée, dans l'état s^H et à l'étape H, un état final est atteint s^{H+1} qui encode l'ensemble des sites de O visités ainsi que les observations correspondantes récoltées.
Fonctions de transition. Si $s^t = (A, x_A)$ et $d^t = A^t$ sont donnés, l'état s^{t+1} est défini de manière unique par les observations x_{A^t} : $s^{t+1} = (A \cup A^t, x_{A \cup A^t})$. De manière évidente, la fonction de transition du PDM peut être obtenue à partir de la distribution du CM du problème d'origine $\mathbb{P}$: $\forall t \in \{1, \ldots, H\}$,

$$p^t\big(s^{t+1} \mid s^t, d^t\big) = \mathbb{P}\big(x_{A^t} \mid x_A, \theta\big).$$

Ces probabilités peuvent être calculées en ligne (pour s^t et d^t fixés), mais pour un coût en temps élevé, puisque ce calcul est équivalent au calcul de probabilités marginales dans un CM. Par contre, ces transitions peuvent être simulées de manière efficace, en utilisant par exemple l'échantillonneur de Gibbs (Geman & Geman, 1984).
Fonctions de récompense. Pour tout $t \in \{1, \ldots, H\}$, la récompense représente le coût d'échantillonnage. Ainsi, dans l'état s^t

$$r^t(s^t, d^t) = r^t(d^t) = -\alpha c(A^t).$$

Après que la décision d^H ait été appliquée à l'étape H et que l'état $s^{H+1} = (A, x_A)$ a été atteint, la récompense finale est attribuée, définie comme la qualité de la reconstruction MPM :

$$r^{H+1}(s^{H+1}) = \sum_{i \in R} \left[\max_{x_i \in \Omega} \Big\{ \mathbb{P}(x_i \mid x_A, \theta) \Big\} \right].$$

La figure 1 décrit de manière schématique cette modélisation PDM du problème EAOCM.
La politique optimale de ce PDM est un ensemble de fonctions associant des plans d'échantillonnage à des unions d'observations sur les étapes passées. Elle a donc la même structure que celle d'une

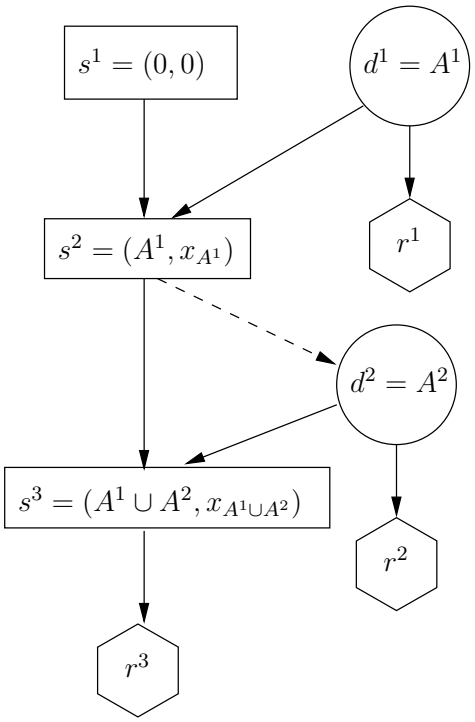

FIGURE 1 – Modélisation PDM du problème EAOCM, pour un horizon $H = 2$.

politique du problème EAOCM, puisqu'encore une fois l'ordre d'acquisition des observations de compte pas (ceci sera formellement démontré dans la preuve de la proposition suivante). Ainsi, nous pouvons établir que :

Proposition 1
Une politique optimale pour le modèle PDM du problème EAOCM fournit une politique optimale pour le problème EAOCM défini par l'équation (3).

Preuve Nous présentons ici les grandes lignes de la démonstration, qui est disponible en annexe. Elle s'articule autour des trois étapes suivantes :

(i) tout d'abord nous définissons une fonction ϕ qui transforme toute politique PDM en une politique EAOCM valide $\delta = \phi(\pi)$, qui définit les décisions indépendamment de l'ordre dans lequel les observations précédentes ont été acquises. Nous montrons alors que $V(\phi(\pi)) = V^{\pi}((\emptyset, \emptyset), 1)$.

(ii) ensuite nous montrons que pour toute trajectoire partielle (les observations passées), la valeur d'une politique EAOCM optimale ayant pour état initial ces observations ne dépend pas de l'ordre dans lequel elles ont été acquises. Donc nous pouvons limiter la recherche de la politique optimale du problème EAOCM aux politiques ayant cette propriété.

(iii) enfin, nous montrons qu'une telle politique δ peut être transformée en une politique PDM, via une transformation μ, et que $V(\delta) = V^{\mu(\delta)}((\emptyset, \emptyset), 1)$.

Nous en déduisons que si π^* est une politique optimale pour la représentation PDM du problème EAOCM, alors $\phi(\pi^*)$ est optimale pour ce problème. □

4 Algorithme de résolution par apprentissage par renforcement

Puisque la résolution exacte du problème EAOCM modélisé en PDM n'est possible en pratique que pour des problèmes de très petite taille, nous présentons maintenant une méthode approchée

de résolution utilisant l'apprentissage par renforcement. Pour cela nous rappelons la définition de la fonction de valeur état-action $Q^\pi(s^t, d^t, t)$, également appelée Q-fonction, et qui est égale à l'espérance des récompenses lorsque la décision d^t est appliquée dans l'état s^t au temps t et qu'ensuite la politique π est suivie :

$$Q^\pi(s^t, d^t, t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{H+1} r^{t'} \mid s^t, d^t \right].$$

L'intérêt d'utiliser Q^π plutôt que la fonction de valeur V^π réside dans le fait que si la fonction de valeur état-action optimale Q^{π^*} est connu, alors la politique optimale peut être calculée sans calculer et stocker les fonctions de transition. En effet, $\forall t \in T,\ \forall s \in S$:

$$\pi^{t*}(s^t) \stackrel{def}{=} \arg\max_{d^t} Q^{\pi^*}(s^t, d^t, t) \tag{4}$$

Le principe de l'AR est de simuler des *épisodes* du PDM afin d'apprendre la valeur de la politique optimale à partir de ces épisodes. Un épisode est une séquence de transitions $\left\{(s^t, d^t, s^{t+1}, r)_{1 \leq t \leq H}\right\}$ et une récompense finale $r^{H+1}(s^{H+1})$. La transition de l'état s^t à l'état s^{t+1} pour une décision d^t est simulée selon la fonction de transition $p^t(. \mid s^t, d^t)$ et la récompense r^t est observée. Plusieurs algorithmes utilisant une représentation tabulaire de la Q-fonction existent dans la littérature de l'AR. Nous allons illustrer la résolution du problème d'échantillonnage avec l'algorithme Watkins's $Q(\lambda)$ naïf, connu pour offrir de meilleures performances que l'algorithme basique de Q-learning, même s'il reste lui-même améliorable. L'idée principale de cet algorithme est de combiner la notion de *traces d'éligibilité* et l'algorithme du Q-learning afin de mettre à jour de manière incrémentale l'estimation de la Q-fonction après chaque épisode. Nous renvoyons le lecteur vers (Sutton & Barto, 1998) pour une description détaillé de l'algorithme Watkins's $Q(\lambda)$ naïf.
Considérons maintenant la question de la simulation des épisodes pour la modélisation PDM du problème EAOCM. Si l'on suppose que la fonction de coût est connue alors la simulation d'un épisode consiste à simuler une trajectoire d'états cohérente avec une trajectoire de décisions données. Chaque transition peut alors être simulée en utilisant l'échantillonneur de Gibbs. Cependant en pratique cette approche est très coûteuse en temps (voir section suivante) puisque la simulation d'un seul épisode demande de lancer H échantillonneurs de Gibbs. Nous montrons maintenant qu'il est possible de mettre en œuvre l'algorithme Watkins's $Q(\lambda)$ naïf en simulant des trajectoires d'états "augmentées", où dans un premier temps une configuration x_V est simulée selon $\mathbb{P}$ puis l'unique trajectoire d'états $(s^1, s^2, \ldots, s^{H+1})$ cohérente avec la trajectoire d'actions est extraite itérativement à partir de x_V. Plus précisément, si un état s^t est représenté par un vecteur de taille n, avec $s^t(i) = 0$ si le site i n'a pas encore été visité, et $s^t(i)$ égal à l'état observé x_i sinon, nous pouvons établir le lemme suivant :

Lemme 1
Soit $(d^1, \ldots, d^H)$ une trajectoire de décisions donnée, une trajectoire d'états $(s^1, \ldots, s^{H+1})$ simulée selon les deux étapes suivantes a la même distribution de probabilité jointe qu'une trajectoire d'états simulée selon le modèle PDM du problème EAOCM :

1. *Simulation d'une configuration x_V selon la distribution jointe $\mathbb{P}(. \mid \theta)$.*
2. *Déduction itérative des valeurs $(s^1, \ldots, s^{H+1})$ selon $s^1(i) = 0\ \forall i \in V$ et $\forall t \in \{1, \ldots, H\},\ s^{t+1}(i) = s^t(i) + d^t(i)x_i$.*

(Nous rappelons qu'un site est visité au plus une fois lors d'une trajectoire.)

Preuve La démonstration repose sur le fait que (i) l'ensemble des trajectoires d'états atteignables par le schéma classique de simulation des transitions et par le schéma proposé sont les mêmes, (ii) la probabilité d'observer une trajectoire d'états particulière est la même dans les deux cas, elle est égale à la probabilité de la configuration x_A correspondante selon $\mathbb{P}(. \mid \theta)$, $A = \cup_{t=1}^{H} d^t$ étant l'ensemble des sites visités. Une version détaillée de cette preuve est disponible en annexe. □
A partir de ce lemme, nous établissons facilement la proposition suivante :

Proposition 2
Si les trajectoires d'états sont simulées selon le schéma augmenté du Lemme 1, l'algorithme Watkins's $Q(\lambda)$ naïf correspondant converge vers la Q-fonction optimale pour le modèle PDM du problème EAOCM.

Preuve Puisque, d'après le Lemme 1, les distributions jointes des trajectoires d'états sont les mêmes pour le schéma classique et augmenté de simulation, les probabilités de transition de s^t vers s^{t+1} sous la décision d^t sont également identiques. Donc l'algorithme Watkins's $Q(\lambda)$ naïf mis en œuvre en utilisant le schéma de simulation augmenté a les même espace d'état, espace d'action, fonctions de transition et récompenses que l'algorithme original. Les deux versions de l'algorithme convergent donc vers la même limite : la Q-fonction optimale pour le modèle PDM du problème EAOCM. □

Le fait de se baser sur la Proposition 2, pour simuler une trajectoire d'états est bien plus efficace car un seul échantillonneur de Gibbs est nécessaire, au lieu de H. Notons par ailleurs que cette proposition repose uniquement sur des considérations en terme de simulation, elle reste donc valable pour n'importe quel algorithme AR basé sur la simulation d'épisodes.

5 Evaluation expérimentale

La mise en œuvre de l'algorithme Watkins's $Q(\lambda)$ naïf demande de faire certains choix de paramétrisation. Ainsi, nous avons utilisé cet algorithme avec la méthode ϵ-*greedy*, pour $\epsilon = \frac{1}{10}$. Le taux d'apprentissage a été fixé à $\frac{1}{n^k(s^t,d^t)}$, où $n^k(s^t, d^t)$ est le numéro de l'épisode durant lequel (s^t, d^t) a été rencontré pour la $k^{\text{ième}}$ fois. La Q-fonction peut être initialisée arbitrairement, cependant l'initialisation peut affecter grandement le temps de convergence. Dans les expériences présentées, lorsqu'un triplet (s, d, t) est rencontré pour la première fois dans la formule de mise à jour, nous avons initialisé de la manière suivante :

$$Q(s,d,t) \leftarrow \sum_{r\in R} \big\{ \max_{x_r\in\Omega} \mathbb{P}^{BP}(x_r \mid x_A, \theta)\big\},$$

où A est l'ensemble des sites visités dans s et $\mathbb{P}^{BP}(x_r \mid x_A, \theta)$ est une approximation de la probabilité marginale calculée par l'algorithme de *belief propagation* (BP, Pearl 1988). L'algorithme BP a également été utilisé pour calculer une approximation de la récompense finale r^{H+1} (qui demande une évaluation du MPM).

L'algorithme Watkins's $Q(\lambda)$ naïf fournit une estimation de la fonction de valeur état-action. À partir de cette estimation, une solution approchée du PDM est obtenue d'après (4). Elle est notée π^*_W et désignée comme la *politique Watkins* dans la suite. À partir de π^*_W nous construisons une solution approchée du problème EAOCM δ^*_W, de la manière suivante : $\delta^{*^1}_W = \pi^{*1}_W((\emptyset,\emptyset))$ et pour tout $t \in \{2,\dots,H\}$ et tout historique atteignable,

$$\delta^{*^t}_W\big((A^1, x_A), \dots, (A^{t-1}, x_{A^{t-1}})\big) = \pi^{*t}_W\big(A, x_A\big),$$

où $A = \cup_{h=1}^{t-1} A^h$.

Afin de mesurer la qualité de δ^*_W, nous avons considéré des problèmes EAOCM sur des grilles régulières avec $R = O = V$. Le modèle CM est un modèle de Potts binaire ($\Omega = \{1, 2\}$) :

$$\mathbb{P}\big(x_V \mid \alpha, \beta\big) = \frac{1}{Z} exp\Big(\sum_{i\in V} \alpha(x_i - 1) + \sum_{(i,j)\in E} \beta \mathbb{1}_{\{x_i = x_j\}}\Big)$$

Enfin, tous les coûts c_i sont nuls.

Les performances en terme de temps de calcul et de qualité de la politique approchée de notre résolution par AR ont d'abord été évaluées sur une grille 4x4, avec $H = 6$, et $L = 1$, et pour un modèle de Potts de paramètres $(\alpha, \beta) = (0, 0.5)$. L'algorithme Watkins's $Q(\lambda)$ naïf a été lancé sur un temps d'exécution de 24h avec une évaluation exacte de la valeur courante de la politique tous les 5000 épisodes. L'utilisation du schéma de simulation augmenté (Lemme 1) apporte un gain en temps significatif puisque cela permet de simuler 5000 trajectoires en 13min environ au lieu de 1h15min environ avec le schéma classique (H échantillonneurs de Gibbs par trajectoire). Nous avons comparé la valeur exacte de la politique π^*_W à celle de trois autres politiques : la vraie politique optimale (définie par l'équation (3)), une politique gloutonne qui choisit à chaque étape le site qui apporte la plus grande augmentation de l'information mutuelle (Krause, 2008), et une politique purement aléatoire. Le graphe de la figure 2 montre l'évolution de la valeur de la

politique Watkins lorsque le nombre d'épisodes simulés augmente, ainsi que les valeurs exactes de la politique optimale (évaluée de manière exhaustive en programmation dynamique en 240h !) et des deux autres politiques de référence.

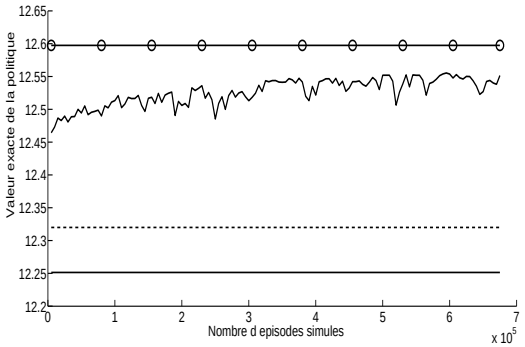

FIGURE 2 – Valeur exacte de la politique Watkins en fonction du nombre d'épisodes simulés. Comparaison avec la valeur de la politique optimale (ligne cerclée), la politique basée sur le critère d'information mutuelle (pointillés) et une politique aléatoire (ligne pleine).

Sur cet exemple nous constatons les bonnes performances de la politique aléatoire. Ceci s'explique par le fait que sur des cartes simulées selon les valeurs de paramètres du modèle de Potts choisies la structure spatiale est très homogène et les états 1 et 2 sont présents en proportions à peu près égales. En conséquence une exploration aléatoire de l'espace permet de récolter suffisamment d'information pour une restauration MPM. Cependant, il faut noter que la valeur de la politique aléatoire est de 2.1% inférieure à celle de la politique optimale, alors que la valeur de la politique Watkins est inférieure d'environ 0.8 %. La politique basée sur l'information mutuelle se situe entre Watkins et l'aléatoire. Enfin, nous avons observé sur l'étude sur grille 4x4 que la valeur de la politique Watkins atteint rapidement un plateau et nous avons estimé qu'il suffit d'allouer un budget temps de 30 min à l'algorithme pour l'atteindre.
Nous avons ensuite testé l'algorithme Watkins's $Q(\lambda)$ naïf sur une grille 6x6 et pour des CM plus structurés. Pour cela, la grille est divisée en quatre sous-grilles 3x3. Un paramètre α_A est attaché aux deux sous-grilles sur la première diagonale et un paramètre $\alpha_B \neq \alpha_A$ aux deux sous-grilles sur l'autre diagonale. Les valeurs utilisées sont $((\alpha_A, \alpha_B), \beta) = ((0, -2), 0.8)$, ce qui conduit à des configurations où les sites dans l'état 2 sont plus nombreux dans la zone A que dans la zone B (une estimation par MCMC des proportions de sites dans l'état 2 est de 10% dans la zone A et 1% dans la zone B). La résolution exacte du PDM est inaccessible pour cette taille de problème. L'algorithme glouton qui détermine la politique basée sur l'information mutuelle est également très coûteux en temps car il implique des calculs exacts sur un CM. Nous n'avons donc pas poursuivi plus loin la comparaison avec ce critère car cela demanderait de développer un algorithme de résolution approché. Nous avons donc comparé la politique Watkins obtenue après 30min de temps de calcul (les résultats sont similaires après 3h) à la politique aléatoire. Les valeurs de ces deux politiques ont été estimées par MCMC (pour les vraies valeurs des paramètres du modèle), le calcul exact étant impossible. Cette valeur est une estimation de l'espérance du pourcentage de variables correctement restaurées, notées CR. Le tableau 1 reporte ces valeurs ainsi que les estimations CR_1 et CR_2 de l'espérance du pourcentage de variables dans l'état 1 et 2 correctement restaurées. Pour

	politique Watkins	politique aléatoire
CR	98.3%	94%
CR_1	98%	94.4%
CR_2	87.6%	66.5%

TABLE 1 – Estimation de l'espérance du pourcentage de variables correctement restaurées (CR), ainsi que de variables dans l'état 1 (resp. 2) correctement restaurées (CR_1 et CR_2) pour les politiques Watkins et aléatoires.

ce problème qui est plus structuré spatialement, nous observons que la politique Watkins conduit à une bien meilleure restauration des sites dans l'état 2. Cela est dû au fait que la politique Watkins est adaptative et qu'elle peut donc cibler la position des prochains sites à échantillonner vers les zones où il reste le plus d'incertitude (pour un problème plus homogène l'incertitude resterait semblable sur toute la grille). Une interprétation classique du modèle de Potts à deux états est celle d'un modèle spatial de carte d'occurrence : les sites dans l'état 1 sont vides, ceux dans l'état 2 sont occupés. Il est donc important de disposer d'une méthode qui soit capable de retrouver les sites occupés avant que la propagation ne soit trop importante, par exemple dans le cas du contrôle d'espèces invasives.

6 Conclusion

Nous avons proposé un cadre pour la modélisation de problèmes d'échantillonnage spatial adaptatif sous contrainte de ressources limitées, lorsque l'objectif est la cartographie d'une information à valeur discrète. Ce cadre explore à la fois des outils de statistiques spatiales (champs de Markov) et de la planification dans l'incertain et la théorie de la décision (processus décisionnels de Markov). La combinaison d'outils de ces différents domaines à conduit à un cadre de modélisation riche et à un algorithme de résolution approchée efficace de type apprentissage par renforcement. Le cadre peut par ailleurs être facilement adapté aux modèles orientés, les réseaux Bayésiens. La validation empirique réalisée nous a permis d'observer des résultats prometteurs, en terme de complexité temporelle et de qualité de la politique approchée retournée. En particulier, si nous avons constaté que pour des problèmes trop homogènes spatialement et où la proportion de sites "occupés" est trop importante, il est inutile de rechercher des politiques d'échantillonnage plus complexe que l'aléatoire, pour des situations plus hétérogènes et avec incidence plus faible, la politique adaptative Watkins apporte une réelle amélioration. L'approche proposée pour l'échantillonnage spatial est actuellement mise en œuvre sur un problème de cartographie d'abondances des adventices dans une parcelle cultivée. Plus largement, l'approche est pertinente pour de nombreux problèmes de surveillance ou suivi en agriculture (contrôle d'épidémies) et en écologie (évaluation de la biodiversité) mais aussi pour des problèmes de surveillance robotique ou de placements de capteurs en analyse d'image. Il est probable que la plupart de ces problèmes soient de nature hétérogènes, en particulier lorsque le graphe sous-jacent est non régulier.

Le modèle peut être amélioré suivant plusieurs directions. Puisque le budget disponible est en général limité, il serait souhaitable de pouvoir représenter des contraintes sur l'espace d'action et/ou de modéliser de manière réaliste des coûts d'actions. Dans notre étude, la seule contrainte représentée limite la taille de l'échantillon à chaque étape. En pratique les contraintes sont souvent plus complexes : le coût d'un échantillon peut être lié au temps passé à explorer un site ou au temps nécessaire pour se déplacer d'un site échantillonné au suivant. L'introduction de coûts réalistes dans le problème d'optimisation et l'évaluation de l'impact sur la politique d'échantillonnage restent des questions ouvertes qui sont néanmoins cruciales pour aborder des problèmes de gestion en agronomie/écologie/environnement. Le modèle proposé pourrait également être étendu par la prise en compte d'un bruit sur les observations (observations non fiables). Enfin, notre objectif est d'ancrer notre cadre de modélisation dans un cadre plus riche, celui des *champs de Markov logiques* (Domingos & Lowd, 2009). Ce cadre bénéficie du pouvoir d'expression de la logique (propositionnelle ou du premier ordre) et d'algorithmes d'apprentissage pour représenter et raisonner sur des connaissances exprimées par un champ de Markov. Incorporer notre cadre pour l'échantillonnage dans le cadre des champs de Markov logiques permettrait d'aborder d'autres problèmes d'intelligence artificielle (élicitation de préférences, diagnostic de systèmes, ...) classiquement exprimés dans le cadre de la logique.

Notre approche pourrait également être améliorée d'un point de vue computationel. Le modèle de PDM qui représente le problème EAOCM a une très grande taille d'espace d'état et d'espace d'action. Même si des améliorations de l'algorithme de Watkins ont été proposées, nous pensons que le gain serait faible à rester dans cette famille d'algorithmes par apprentissage par renforcement. L'alternative classique en apprentissage par renforcement est d'utiliser des approximations

paramétrisées de la fonction de valeur (Sutton & Barto, 1998) ou de la politique solution (Sutton *et al.*, 2000) comme par exemple la méthode Least-Squares Policy Iteration algorithm (Lagoudakis & Parr, 2003). Ces approches sont adaptées pour la résolution du problème d'échantillonnage adaptatif et restent à explorer. Enfin, des méthodes d'approximation basées sur un découpage du graphe du champ de Markov pourraient être utilisées afin de réduire des problèmes trop grands en sous-problèmes traités indépendamment. Cette solution pourrait notamment exploiter les algorithmes de décomposition parallèle pour les processus décisionnels de Markov factorisés (Guestrin & Gordon, 2002).

7 Annexe

Démonstration de la Proposition 1

Soit $h^1 = ((\emptyset,\emptyset))$ et $\forall t = 2,...,H$, $h^t = ((A^1, x_{A^1}),\ldots,(A^{t-1}, x_{A^{t-1}}))$. Pour tout historique h^t, un unique état du PDM $s^t(h^t)$ peut être défini comme $s^1(h^1) = (\emptyset,\emptyset)$ et $\forall t = 2,...,H$, $s^t(h^t) = \left(\cup_{k=1}^t A^k, x_{\cup_{k=1}^t A^k}\right)$. Nous définissons une transformation ϕ de l'ensemble des politiques du PDM vers l'ensemble des politiques du problème EAOCM. Soit π, une politique du PDM, $\delta = \phi(\pi)$ est définie ainsi : pour tout $t = 1 \ldots H$ et toute trajectoire atteignable h^t, $\delta^t(h^t) = \pi^t(s^t(h^t))$.

(i) Montrons d'abord que $V^\pi((\emptyset,\emptyset),1) = V(\phi(\pi))$. En effet, rappelons que

$$\begin{aligned} V^\pi((\emptyset,\emptyset),1) &= \mathbb{E}_\pi\Big[\sum_{t=1}^{H+1} r^t \mid s^1\Big] \\ &= \sum_{s^2,\ldots,s^{H+1}} \mathbf{P}(s^2,\ldots,s^{H+1} \mid \pi, s^1 = (\emptyset,\emptyset))\Big[\sum_{t=1}^{H} r^t(\pi^t(s^t)) + r^{H+1}(s^{H+1})\Big], \end{aligned}$$

où $\mathbf{P}(s^2,\ldots,s^{H+1} \mid \pi, s^1)$ est la probabilité de la trajectoire d'état $(s^2,\ldots,s^{H+1})$, sachant l'état initial s^1 et pour la politique π. Pour toute trajectoire d'états possible du PDM $(s^1,\ldots,s^{H+1})$ nous pouvons déduire un unique historique $h^{H+1} = ((A^1, x_{A^1}),\ldots,(A^H, x_{A^H}))$, où A^t est l'ensemble des sites impliqués dans s^{t+1} mais pas dans s^t. Alors,

$$\mathbf{P}(s^2,\ldots,s^{H+1} \mid \pi, s^1) = \begin{cases} 0 \text{ si trajectoire non atteignable,} \\ \mathbb{P}(x_A \mid \theta) \text{ sinon.} \end{cases}$$

avec $A = \cup_{t=1}^H A^t$,. De plus, nous avons $r^t(\pi(s^t)) = -\alpha \sum_{a\in A^t} c_a$ et $r^{H+1}(s^{H+1}) = \sum_{r\in R} \max_{x_r\in\Omega} \left\{\mathbb{P}(x_r \mid x_A,\theta)\right\}$. Finalement

$$\begin{aligned} V^\pi((\emptyset,\emptyset),1) &= \sum_{h^{H+1}\in\tau_{\phi(\pi)}} \mathbb{P}(x_A \mid \theta)\Big[-\alpha\sum_{t=1}^{H}\sum_{a\in A^t} c_a + \sum_{r\in R}\max_{x_r\in\Omega}\left\{\mathbb{P}(x_r \mid x_A,\theta)\right\}\Big] \\ &= \sum_{h^{H+1}\in\tau_{\phi(\pi)}} \mathbb{P}(x_A \mid \theta) U(A, x_A) \\ &= V(\phi(\pi)). \end{aligned}$$

(ii) Nous établissons ensuite par récurrence arrière qu'il est possible de définir une politique optimale δ^* pour le problème EAOCM, telle qu'elle sélectionne les plans d'échantillonnage successifs indépendamment de l'ordre d'acquisition des observations passées. Nous désignerons par $\mathcal{D}_{ioa}$ l'ensemble des politiques partageant cette propriété. Considérons tout d'abord $\delta^{*,H}$:

$$\delta^{*,H}((A^1, x_{A^1}),\ldots,(A^{H-1}, x_{A^{H-1}})) = \arg\max_{A^H} \sum_{x_{A^H}} \mathbb{P}(x_{A^H} | x_{A^1},\ldots,x_{A^{H-1}},\theta) U(A, x_A),$$

où $A = A^1 \cup \ldots \cup A^H$. Cependant $\mathbb{P}(x_{A^H} | x_{A^1},\ldots,x_{A^{H-1}},\theta)$ et $U(A, x_A)$ ne dépendent pas de l'ordre d'acquisition des observations $x_{A^1},\ldots,x_{A^{H-1}}$. Donc, $\delta^{*,H}$ ne dépend pas de l'ordre de ces

arguments.
Regardons maintenant l'étape $h = H - 1$:

$$\delta^{*,H-1}((A^1, x_{A^1}), \ldots, (A^{H-2}, x_{A^{H-2}})) = \arg\max_{A^{H-1}} \sum_{x_{A^{H-1}}}$$

$$\sum_{x_{A^H}} \mathbb{P}(x_{A^{H-1}}, x_{A^H} | x_{A^1}, \ldots, x_{A^{H-2}}, \delta^{*,H}(\ldots), \theta) U(A, x_A).$$

Puisque $\delta^{*,H}$ ne dépend pas de l'ordre de ses arguments, il en est de même $\delta^{*,H-1}$.
En suivant le même raisonnement pour $h = H - 2, \ldots, 1$, nous montrons qu'une politique δ^* peut être calculée, qui appartient à $\mathcal{D}_{ioa}$. Ce résultat implique que la recherche de la politique optimale du problème EAOCM peut être limitée à une recherche parmi l'ensemble $\mathcal{D}_{ioa}$.

(*iii*) Considérons maintenant une politique δ dans $\mathcal{D}_{ioa}$. Il est possible de construire à partir de δ une politique π du modèle PDM. La construction se fait également par induction : $\pi^1((\emptyset, \emptyset)) = \delta^1$, et pour $t = 2$ à H et un état atteignable s^t nous définissons un historique $((A^1, x_{A^1}), \ldots, (A^{t-1}, x_{A^{t-1}}))$ de taille $t - 1$, pour lequel l'ordre dans lequel les observations sont obtenues est choisi arbitrairement, et nous posons $\pi(s^t) = \delta^t((A^1, x_{A^1}), \ldots, (A^{t-1}, x_{A^{t-1}}))$. Avec cette procédure, π est définie uniquement sur les états s^t atteignables en suivant δ. Pour les autres états, la politique est fixée à une valeur arbitraire (la valeur de π ne dépend pas de ce choix puisque l'état correspondant ne sera jamais atteint). Désignons par μ cette transformation d'une politique du problème EAOCM vers une politique PDM. En suivant le même raisonnement que dans le point (*i*) il est facile de voir que $V^{\mu(\delta)}((\emptyset, \emptyset), 1) = V(\delta)$.

(*iv*) Finalement, soit π^* la politique optimale du modèle PDM du problème EAOCM :

$$V^{\pi^*}(s, t) \geq V^{\pi}(s, t) \qquad \forall \pi, s, t$$

Donc la politique $\phi(\pi^*)$ est optimale pour le problème EAOCM. En effet, si δ est une politique du problème EAOCM (dans $\mathcal{D}_{ioa}$) et $\mu(\delta)$ est la politique correspondante pour le modèle PDM, alors

$$V^{\pi^*}((\emptyset, \emptyset), 1) \geq V^{\mu(\delta)}((\emptyset, \emptyset), 1),$$

et puisque $V^{\pi^*}((\emptyset, \emptyset), 1) = V(\phi(\pi^*))$ et $V^{\mu(\delta)}((\emptyset, \emptyset), 1) = V(\delta)$, nous obtenons $V(\phi(\pi^*)) \geq V(\delta)$. Ceci établit la Proposition 1.

Démonstration du Lemme 1 :
Soit une trajectoire d'actions $(d^1, \ldots, d^H)$, considérons une trajectoire d'états $(s^1, \ldots, s^{H+1})$ simulée suivant le schéma suivant à deux étapes

1. Simulation d'une configuration x_V selon la distribution jointe $\mathbb{P}(. \mid \theta)$.
2. Déduction itérative des valeurs $(s^1, \ldots, s^{H+1})$ selon $s^1(i) = 0 \ \forall i \in V$ et
$$\forall t \in \{1, \ldots, H\}, \ s^{t+1}(i) = s^t(i) + d^t(i) x_i. \tag{5}$$

Nous avons

$$\mathbf{P}(s^1, \ldots, s^{H+1} \mid d^1, \ldots, d^H) \quad = \quad \sum_{x_V \in \Omega^n} \mathbb{P}(x_V \mid \theta) \mathbf{P}(s^1, \ldots, s^{H+1} \mid x_V, d^1, \ldots, d^H).$$

La probabilité $\mathbf{P}(s^1, \ldots, s^{H+1} \mid x_V, d^1, \ldots, d^H)$ est égale soit à zéro soit à 1, puisque seulement une trajectoire d'états peut être atteinte sachant x_V et $(d^1, \ldots, d^H)$ d'après (5). De plus, sachant $(d^1, \ldots, d^H)$, la trajectoire d'états $(s^1, \ldots, s^{H+1})$ peut être atteinte depuis toute configuration x_V qui est en accord avec les observations de cette trajectoire sur le sous-ensemble A des sites visités par $(d^1, \ldots, d^H)$. D'où, si x'_A est l'ensemble des observations collectées sur A au cours de la trajectoire d'état $(s^1, \ldots, s^{H+1})$

$$\mathbf{P}(s^1, \ldots, s^{H+1} \mid d^1, \ldots, d^H) = \sum_{x_V \in \Omega^n} \mathbb{P}(x_V \mid \theta) \mathbb{1}_{\{x_A = x'_A\}},$$

qui par définition est égal à $\mathbb{P}(x'_A \mid \theta)$. Evaluons maintenant la probabilité d'observer la même trajectoire d'états $(s^1, \ldots, s^{H+1})$, sachant $(d^1, \ldots, d^H)$, lorsque l'on utilise les probabilités de transition définissant le PDM correspondant au problème EAOCM :

$$\mathbf{P}(s^1, \ldots, s^{H+1} \mid d^1, \ldots, d^H) = \mathbb{P}(x'_{d^1} \mid \theta) \prod_{t=2}^{H} \mathbb{P}(x'_{d^t} \mid x'_{d^{t-1}}, \ldots, x'_{d^1}, \theta).$$

Par application de la règle de Bayes, nous obtenons que $\mathbb{P}(x'_{d^1} \mid \theta) \prod_{t=2}^{H} \mathbb{P}(x'_{d^t} \mid x'_{d^{t-1}}, \ldots x'_{d^1}, \theta)$ est exactement $\mathbb{P}(x'_{d^1}, \ldots, x'_{d^H} \mid \theta)$, qui est égal à $\mathbb{P}(x'_A \mid \theta)$.
Donc, avec les deux schémas de simulation, pour une trajectoire d'actions donnée $(d^1, \ldots, d^H)$ les mêmes trajectoires d'états peuvent être atteintes (celles pour lesquelles les sites visités sont cohérents avec les actions) et chacune a la même probabilité dans les deux schémas. Ceci établit le Lemme 1.

Références

de Gruijter J., Brus D., Bierkens M. & Knotters K. (2006). *Sampling for Natural Resource Monitoring*. Springer.

Domingos P. & Lowd D. (2009). *Markov Logic : An Interface Layer for Artificial Intelligence*. Synthesis Lectures on AI and ML. San Rafael, CA : Morgan and Claypool.

Geman S. & Geman D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Guestrin C. & Gordon G. (2002). Distributed planning in hierarchical factored MDPs. In *Proc UAI'02*.

Krause A. (2008). *Optimizing Sensing Theory and Applications*. PhD thesis, School of computer Science, Carnegie Mellon University Pittsburg, PA 15213, 55-56.

Krause A. & Guestrin C. (2009). Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, **35**, 557–591.

Krause A., Singh A. & Guestrin C. (2008). Near-optimal sensor placements in Gaussian processes : theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, **9**, 235–284.

Lagoudakis M. & Parr R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*.

M. Fuentes A. C. & Holland D. (2007). Bayesian entropy for spatial sampling design of environmental data. *Environmental and Ecological Statistics*, **14**, 323–340.

Müller W. (2007). *Collecting spatial Data*. Springer Verlag : Heidelberg. 3rd ed.

Pearl J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA : Morgan Kaufmann.

Peyrard N., Sabbadin R. & Niaz U. F. (2010). Decision-theoretic optimal sampling with hidden Markov random fields. In *Proc ECAI'10*.

Puterman M. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Roth D. (1996). On the hardness of approximate reasoning. *Artificial Intelligence*, **82**, 273–302.

Sutton R. S. & Barto A. (1998). *Reinforcement Learning : An Introduction*. MIT Press.

Sutton R. S., McAllester D., Singh S. & Mansour Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In M. Press, Ed., *Advances in Neural Information Processing Systems*, volume 12, p. 1057–1063.

Thompson S. & Seber G. (1996). *Adaptive sampling*. Series in Probability and Statistics. New York : Wiley.

Wiles L. (2005). Sampling to make maps for site specific weed management. *Weed science*.