Un exemple d'optimisation de traitement médical avec modèle incertain

Alice Cleynen [1] & Benoîte de Saporta [2] & Orlane Le Quellennec [3] & Régis Sabbadin [4]

[1] *IMAG, Univ Montpellier, CNRS, Montpellier, France orlane.le-quellennec@umontpellier.fr*
[2] *IMAG, Univ Montpellier, CNRS, Montpellier, France benoite.de-saporta@umontpellier.fr*
[3] *IMAG, Univ Montpellier, CNRS, Montpellier, France alice.cleynen@umontpellier.fr*
[4] *Univ Toulouse, INRAE-MIAT, Toulouse, France regis.sabbadin@inrae.fr*

Résumé. Les maladies humaines telles que le cancer impliquent un suivi à long terme. Un·e patient·e alterne des phases de rémission et de rechutes. Un biomarqueur est monitoré tout au long du suivi. Sa dynamique est modélisée par un processus de Markov déterministe par morceaux (PDMP) contrôlé. Le PDMP évolue en temps et en espace continus, le processus est partiellement observé à travers un bruit et certains de ses paramètres sont inconnus, ce qui rend le problème du contrôle particulièrement difficile. À notre connaissance, il n'existe pas de méthode pour contrôler un tel PDMP, c'est-à-dire pour maximiser la vie du·de la patient·e tout en minimisant le coût du traitement et les effets secondaires. Dans un premier temps, nous considérons des dates discrètes uniquement pour les décisions, transformant ainsi le PDMP contrôlé en un processus de décision markovien partiellement observé (POMDP). Ensuite, par le biais de simulations, nous comparons les méthodes d'apprentissage par renforcement bayésiennes et non-bayésiennes pour résoudre ce POMDP.

Mots-clés. Processus de décision markovien, contrôle stochastique, apprentissage par renforcement, bayésien, optimisation de traitement

Abstract. Human diseases such as cancer involve long-term follow-up. A patient alternates phases of remission with relapses. A biomarker is monitored throughout the follow-up. Its dynamic is modeled by a controlled piecewise deterministic Markov process (PDMP). The PDMP evolves in continuous time and space, the process is partially observed through noise and some of its parameters are unknown, making the control problem especially difficult. To our knowledge, there is no method to control such a PDMP, i.e. to maximize the life of the patient while minimizing the treatment cost and side effects. First, we consider discrete dates only for the decisions, thus turning the controlled PDMP into a partially observed Markov decision process (POMDP). Then, through simulations, we compare Bayesian and non-Bayesian reinforcement learning methods to solve this POMDP.

Keywords. Markov decision process, stochastic control, reinforcement learning, bayesian, treatment optimisation

1 Introduction

Le suivi et le traitement d'un cancer correspondent à un problème de décisions séquentielles sous incertitude. En effet, le·la médecin a pour objectif d'identifier et d'adapter le traitement pour préserver au mieux la qualité et l'espérance de vie du·de la patient·e tout en minimisant le coût du traitement et les effets secondaires au cours du temps. La décision du·de la médecin s'appuie, par exemple, sur des prélèvements sanguins qui représentent l'évolution du cancer.

Dans notre exemple d'application, un·e patient·e intègre un essai-clinique au début d'une phase de rémission. Pendant les phases de rémission, le biomarqueur reste au seuil nominal ζ_0. Lors d'une rechute et en l'absence de traitement, le niveau du biomarqueur augmente exponentiellement et lorsqu'il atteint une valeur critique D, le·la patient·e décède. Sous traitement, le niveau redescend, mais la probabilité que le·la patient·e développe une résistance au traitement augmente avec le nombre de rechutes.

Ce problème peut être modélisé par un processus de Markov déterministe par morceaux (PDMP) contrôlé (Davis, 1984). Plusieurs difficultés apparaissent lorsque l'on veut contrôler le processus : (1) l'espace d'état est continu, (2) une partie du processus est cachée (les dates de rechutes et le type de rechute ne sont pas observés) et (3) le modèle est partiellement connu (la vitesse de croissance du cancer est inconnue). Les problématiques (1) et (2) ont déjà été étudiées dans de précédent travaux (Cleynen and de Saporta, 2018) et (Cleynen and de Saporta, 2021). À notre connaissance, il n'existe pas de travaux sur l'association de ces trois complexités.

Dans ce papier, on s'intéresse uniquement à la problématique de modèle partiellement connu. Le suivi et le traitement du·de la patient·e sont modélisés par un processus de décision markovien (MDP) à espace d'états fini. Pour contrôler le MDP, on compare deux types d'algorithmes d'apprentissage par renforcement. Le Q-learning (Watkins and Dayan, 1992; Vivek and Bhatnagar, 2022) est une méthode dite *model-free* et le BAMCP (Guez et al., 2012) une méthode dite *model-based*.

2 Le problème d'optimisation

Un processus de décision markovien (MDP) se définit par le tuple $(\mathbb{S}, \mathbb{A}, P, c)$, où $\mathbb{S}$ est l'ensemble fini des états et $\mathbb{A}$ l'ensemble des actions possibles, P est la matrice de transition entre les différents états et c la fonction de coût. Le processus est à horizon fini H.

L'espace d'états $\mathbb{S}$. Un état $s \in \mathbb{S}$ se définit par un tuple : $s = (m, \zeta, v)$, où :

- $m \in \mathcal{M} = \{0, 1, 2\}$ est le mode, noté m, qui correspond à l'état général du·de la patient·e ($m = 0$: rémission, $m = 1$: rechute, $m = 2$: décès) ;
- $\zeta \in \mathcal{Z} = \{0, 1, 2, 3, 4\}$ est le marqueur sanguin mesuré ;
- $v \in \mathcal{V} = \{0, 1, 2\}$ est l'intensité de la rechute ($v = 0$: phase de rémission ou décès, $v = 1$ ou $v = 2$: phase de rechute).

TABLE 1 – **Matrice de transition selon l'action choisie**, où $p^a_{s'}$ est une probabilité de transition de l'état $s = (0, 0, 0)$ vers l'état s' sous l'action a.

$s \backslash s'$	$(0,0,0)$	$(1,0,1)$	$(1,0,2)$	$(1,1,1)$	$(1,1,2)$	$(1,2,1)$	$(1,2,2)$	$(1,3,1)$	$(1,3,2)$	$(2,4,0)$
$(0,0,0)$	$p^{\emptyset}_{(0,0,0)}$	$p^{\emptyset}_{(1,0,1)}$	$p^{\emptyset}_{(1,0,2)}$	0	0	0	0	0	0	0
$(1,0,1)$	0	0	0	1	0	0	0	0	0	0
$(1,0,2)$	0	0	0	0	0	0	1	0	0	0
$(1,1,1)$	0	0	0	0	0	1	0	0	0	0
$(1,1,2)$	0	0	0	0	0	0	0	0	1	0
$(1,2,1)$	0	0	0	0	0	0	0	1	0	0
$(1,2,2)$	0	0	0	0	0	0	0	0	0	1
$(1,3,1)$	0	0	0	0	0	0	0	0	0	1
$(1,3,2)$	0	0	0	0	0	0	0	0	0	1
$(2,4,0)$	0	0	0	0	0	0	0	0	0	1

(a) Patient·e n'est pas sous traitement $a = \emptyset$

$s \backslash s'$	$(0,0,0)$	$(1,0,1)$	$(1,0,2)$	$(1,1,1)$	$(1,1,2)$	$(1,2,1)$	$(1,2,2)$	$(1,3,1)$	$(1,3,2)$	$(2,4,0)$
$(0,0,0)$	$p^{\rho}_{(0,0,0)}$	$p^{\rho}_{(1,0,1)}$	$p^{\rho}_{(1,0,2)}$	0	0	0	0	0	0	0
$(1,0,1)$	1	0	0	0	0	0	0	0	0	0
$(1,0,2)$	1	0	0	0	0	0	0	0	0	0
$(1,1,1)$	1	0	0	0	0	0	0	0	0	0
$(1,1,2)$	1	0	0	0	0	0	0	0	0	0
$(1,2,1)$	0	0	0	1	0	0	0	0	0	0
$(1,2,2)$	0	0	0	0	1	0	0	0	0	0
$(1,3,1)$	0	0	0	0	0	1	0	0	0	0
$(1,3,2)$	0	0	0	0	0	1	0	0	0	0
$(2,4,0)$	0	0	0	0	0	0	0	0	0	1

(b) Patient·e est sous traitement $a = \rho$

L'espace des actions $\mathbb{A}$. On note $a \in \mathbb{A} = \{\emptyset, \rho\}$, où $\emptyset$ correspond à la décision de ne pas traiter et ρ à celle d'administrer un traitement au·à la patient·e. Les deux actions sont toujours possibles.

La matrice de transition P. La matrice de transition $\{P(s'|s, a), s', s \in \mathbb{S}, a \in \mathbb{A}\}$, où $P(s'|s, a)$ est la probabilité de passer dans l'état $s' \in \mathbb{S}$ en partant de l'état $s \in \mathbb{S}$ et en appliquant l'action $a \in \mathbb{A}$, voir Table 1. Les transitions sont déterministes, excepté à partir de l'état courant $s = (0, 0, 0)$.

Fonction de coût $c : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$. Les valeurs des coûts associés aux couples (s, a) sont données dans la Table 2.

Lorsque la matrice de transitions est connue, un MDP à espace d'états fini (pas trop grand) pourrait être résolu de manière exacte par programmation dynamique (Bellman, 2010). La matrice de transition, du MDP, détaillée ci-dessus, est partiellement connue et la résolution exacte ne s'applique pas.

TABLE 2 – **Détails des coûts selon l'état courant et l'action choisie.**

$s \backslash a$	$\emptyset$	ρ
$(0,0,0)$	0	300
$(1,0,1)$	200	500
$(1,0,2)$	300	600
$(1,1,1)$	200	500
$(1,1,2)$	300	600
$(1,2,1)$	200	500
$(1,2,2)$	300	600
$(1,3,1)$	200	500
$(1,3,2)$	300	600
$(2,4,0)$	1000	1900

3 Estimations des probabilités de transitions

Le processus de décision markovien (MDP), détaillé dans la section 2, peut être résolu avec des méthodes d'apprentissage par renforcement dites *model-free*. Cette famille d'algorithmes simule de nombreuses trajectoires, alternant états et actions, pour converger vers la politique optimale. Les méthodes dites *model-based* apprennent également une représentation des probabilités de transitions pour converger plus rapidement vers la solution (Atkeson and Santamaria, 1997). Lorsque la loi de transition a une forme paramétrique, l'apprentissage par renforcement bayésien permet d'inférer les paramètres inconnus tout en apprenant la politique *bayes* optimale, à partir d'interactions simulées avec le MDP contrôlé.

Les méthodes d'apprentissage par renforcement bayésien transforment le MDP à modèle inconnu en un MDP à modèle connu sur un espace *d'hyper-états*, appelé *Bayes Adaptative* MDP (BAMDP) (Duff, 2002). Un hyper-état regroupe à la fois l'état du MDP et les valeurs courantes des paramètres de la loi de transitions. Le BAMDP pourrait être résolu par des méthodes classiques de résolution de MDP, à ceci près que la taille de son espace d'état nécessite des méthodes de résolution adaptées. De plus, les méthodes classiques ne permettent pas d'inférer les paramètres du modèle et d'apprendre la politique optimale en même temps.

Un BAMDP se définit par le tuple $(\mathbb{S}', \mathbb{A}, P', c)$, où $\mathbb{S}'$ est l'espace des hyper-états et P' est la matrice de transition entre les hyper-états. L'espace des actions $\mathbb{A}$ et la fonction de coût c restent inchangés.

L'espace des hyper-états $\mathbb{S}'$. La première ligne de la matrice de transition (Table 1) a une distribution multinomiale : $P(.|s = (0,0,0), a) \sim \mathcal{M}(p^a_{(0,0,0)}, p^a_{(1,0,1)}, p^a_{(1,0,2)})$. Une loi naturelle conjuguée est une loi de Dirichlet de paramètre $(\theta^a_{(0,0,0)}, \theta^a_{(1,0,1)}, \theta^a_{(1,0,2)})$ avec $\theta^a_{s'} \in \Theta$. On note w notre croyance *a priori* avec $w(p^a_{(0,0,0)}, p^a_{(1,0,1)}, p^a_{(1,0,2)}) \sim \mathcal{D}(\theta^a_{(0,0,0)}, \theta^a_{(1,0,1)}, \theta^a_{(1,0,2)})$. Un état $s' \in \mathbb{S} \times \Theta$ se défini par un tuple : $s' = (s, \begin{bmatrix} \theta^\emptyset_1 & \theta^\emptyset_2 & \theta^\emptyset_3 \\ \theta^\rho_1 & \theta^\rho_2 & \theta^\rho_3 \end{bmatrix})$, où :

- s est la chaîne de Markov du MDP détaillé dans la section 2
- $(\theta^\emptyset_{(0,0,0)}, \theta^\emptyset_{(1,0,1)}, \theta^\emptyset_{(1,0,2)})$ sont les paramètres de la loi a priori $w(p^\emptyset_{(0,0,0)}, p^\emptyset_{(1,0,1)}, p^\emptyset_{(1,0,2)}) \sim$

$\mathcal{D}(\theta^{\emptyset}_{(0,0,0)}, \theta^{\emptyset}_{(1,0,1)}, \theta^{\emptyset}_{(1,0,2)})$;

- $(\theta^{\rho}_{(0,0,0)}, \theta^{\rho}_{(1,0,1)}, \theta^{\rho}_{(1,0,2)})$ sont les paramètres de la loi a priori $w(p^{\rho}_{(0,0,0)}, p^{\rho}_{(1,0,1)}, p^{\rho}_{(1,0,2)}) \sim \mathcal{D}(\theta^{\rho}_{(0,0,0)}, \theta^{\rho}_{(1,0,1)}, \theta^{\rho}_{(1,0,2)})$.

La matrice de transition P'. A chaque nouvelle observation, on met à jour la loi (de Dirichlet) a posteriori sur les paramètres de la matrice de transition : $\Theta' = \Theta + \Delta^{a}_{s'}$, où $\Delta^{a}_{s'}$ est une matrice de taille $A \times N$ avec 1 associé à la position de l'observation $(s = (0,0,0), a, s')$ et 0 sur les autres positions. Les transitions sont déterministes pour tous les états, excepté lorsque la chaîne de Markov du MDP vaut $s = (0,0,0)$. D'après la formule des probabilités composées, on a : $P'(s', \theta'|s, a, \theta) = \mathbb{P}(s'|s, a, \theta)\mathbb{P}(\theta'|s, a, s', \theta)$.

En théorie, la politique *bayes* optimale peut être obtenue de façon exacte par la programmation dynamique mais l'espace des hyper-états est généralement trop important pour que ce soit réalisable. En effet, initialement $(t = 0)$, il y a autant d'hyper-états que d'états dans le MDP. Par la suite, la taille de l'espace d'état correspond à tous les états atteignables à partir de $s'_0 = (s_0, \Theta_0)$. Donc la taille de l'espace des hyper-états d'un BAMDP est exponentiellement plus grande que celle du MDP $(\mathcal{O}(|S|^t))$ P d'origine (voir Table **??**). Il faut donc appliquer des méthodes de résolution adaptées, comme les algorithmes de type *Monte-Carlo planning*.

4 Comparaison d'algorithmes bayésiens et non-bayésiens

On compare deux méthodes d'apprentissage par renforcement. Le Q-learning (Watkins and Dayan, 1992; Vivek and Bhatnagar, 2022) est une méthode sans modèle et le BAMCP (Guez et al., 2012) une méthode avec modèle. Dans les deux algorithmes, la mise à jour de la politique se fait à chaque nouvelle observation et la connaissance acquise est réutilisée sur les nouvelles trajectoires.

La première, de type *model-free*, consiste à résoudre le MDP simulé, décrit en section 2 avec l'algorithme de Q-learning. L'expérience est réalisée une première fois sur 10^2 patients simulés, puis 10^3 et enfin 10^6.

Pour la seconde méthode de résolution, dite *model-based*, l'algorithme BAMCP, est appliqué pour résoudre le BAMDP décrit dans la section 3. Cet algorithme est de type *Monte-Carlo planning*. De nouveau, l'expérience est réalisée une première fois sur 10^2 patients simulés, puis 10^3 et enfin 10^6.

Les probabilités de transitions inconnues sont fixées pour les simulations : $p^{\emptyset}_{(0,0,0)} = \frac{3}{6}$, $p^{\emptyset}_{(1,0,1)} = \frac{1}{6}$, $p^{\emptyset}_{(1,0,2)} = \frac{2}{6}$, $p^{\rho}_{(0,0,0)} = \frac{3}{6}$, $p^{\rho}_{(1,0,1)} = \frac{1}{6}$ et $p^{\rho}_{(1,0,2)} = \frac{2}{6}$. Afin de comparer les performances des deux algorithmes, la politique optimale, de cet exemple, est identifiée par programmation dynamique. Son coût exact est de 888.89. L'algorithme de Q-learning converge vers une politique aux performances comparables à la politique optimale après avoir vu 10^6 trajectoires (voir Table 3).

L'algorithme de BAMCP (Guez et al., 2012) devrait permettre d'obtenir une politique

TABLE 3 – **Performances de l'algorithme de Q-learning estimé par simulations de Monte-Carlo à 10^6 répétitions.**

Nombre de patients simulés	Coût estimé de la politique	Intervalle de confiance à 95%
10^2	1374.72	$[1373.56, 1375.89]$
10^3	921.16	$[920.48, 921.83]$
10^6	892.2	$[891.54, 892.86]$

bayes optimale plus performante sur un faible nombre de trajectoires. Les résultats de cet algorithme seront présentés dans la version définitive de ce papier.

5 Conclusion et perspectives

Le problème de suivi et de traitement d'un·e patient·e atteint·e d'un cancer a été simplifié et modélisé par un processus de décision markovien (MDP). Ce MDP est complètement observé et à espace d'états fini. La matrice de transition est partiellement connue. Nous avons comparé deux méthodes de résolutions, respectivement sans modèle (Q-learning) et bayésienne, à base de modèle (BAMCP).

Pour avoir une modélisation plus proche de la réalité, il faut considérer un espace d'états continu et des observations cachées. Le formalisme de processus de Markov déterministe par morceaux (PDMP) contrôlé est mieux adapté que le MDP. Nous envisageons de considérer une forme paramétrique de PDMP appropriée aux problèmes de contrôle de cancer (*Bayesian-Adaptive controlled PDMP*), et de développer un algorithme de résolution de type Monte-Carlo.

Bibliographie

Atkeson, C. G. and Santamaria, J. C. (1997). A comparison of direct and model-based reinforcement learning. In *Proceedings of International Conference on Robotics and Automation*, volume 4, pages 3557–3564vol.4. IEEE.

Bellman, R. E. (2010). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA.

Cleynen, A. and de Saporta, B. (2018). Change-point detection for piecewise deterministic Markov processes. *Automatica*, 97 :234–247.

Cleynen, A. and de Saporta, B. (2021). Sequential decision making for a class of hidden Markov processes, application to medical treatment optimisation. *arXiv :2112.09408*.

Davis, M. H. A. (1984). Piecewise-deterministic markov processes : A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society Series B (Methodological)*, 46 :353–376.

Duff, M. O. (2002). *Optimal learning : Computational procedures for Bayes -adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst.

Guez, A., Silver, D., and Dayan, P. (2012). Efficient bayes-adaptive reinforcement learning using sample-based search. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Vivek, V. and Bhatnagar, Dr. S. (2022). Finite Horizon Q-learning : Stability, Convergence, Simulations and an application on Smart Grids. *arXiv :2110.15093v3*.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Mach. Learn.*, 8(3) :279–292.