Apprentissage par renforcement pour le contrôle de processus de Markov déterministe par morceaux

Application à l'optimisation d'un traitement médical

Orlane Rossini [1], Alice Cleynen [1,2], Benoîte de Saporta [1],
Régis Sabbadin [3] et Meritxell Vinyals [3]

[1]IMAG, Univ Montpellier, CNRS, Montpellier, France
[2]John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia
[3]Univ Toulouse, INRAE-MIAT, Toulouse, France

Octobre 2024

Le contexte médical

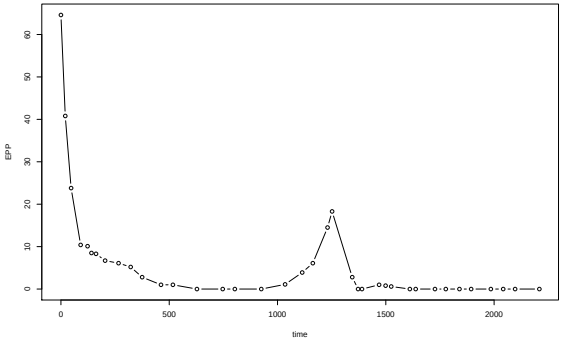


FIGURE: Exemple de donnée d'un patient[a]

- Des patients ayant eu un cancer bénéficient d'un suivi régulier;
- La concentration d'immunoglobuline clonale est mesurée dans le temps;
- Le médecin doit prendre de nouvelles décisions à chaque visite.

[a] IUCT Oncopole et CRCT, Toulouse, France

Le contexte médical

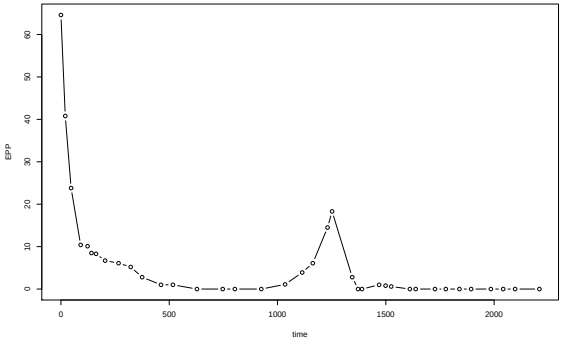


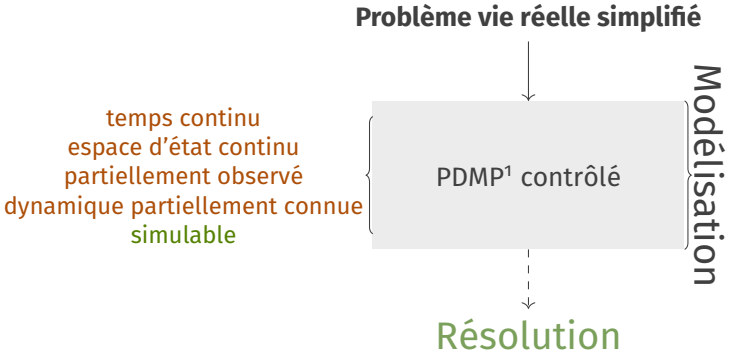
FIGURE: Exemple de donnée d'un patient[a]

- Des patients ayant eu un cancer bénéficient d'un suivi régulier;
- La concentration d'immunoglobuline clonale est mesurée dans le temps;
- Le médecin doit prendre de nouvelles décisions à chaque visite.

$\Longrightarrow$ **Optimiser la prise de décision pour assurer la qualité de vie du patient**

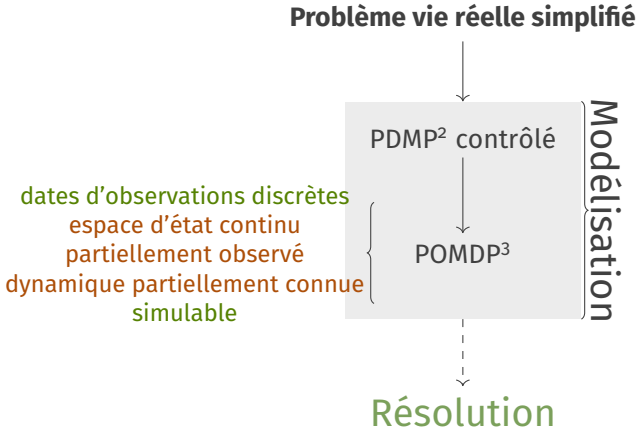
[a] IUCT Oncopole et CRCT, Toulouse, France

Méthodes



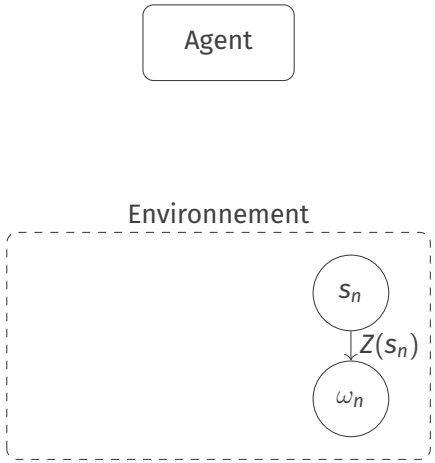
[1]Piecewise Deterministic Markov Process

Méthodes



[2] Piecewise-deterministic Markov process
[3] Partially Observable Markov Decision Process

Caractéristiques d'un POMDP

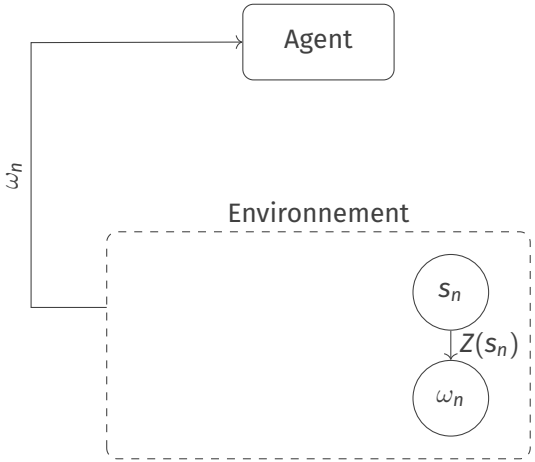


Définition d'un POMDP

Un POMDP se définit par un tuple $(\mathcal{S}, \mathcal{D}, \mathcal{K}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Etat du patient $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Décisions $d = (\ell, r) \in \mathcal{D}$;
- $\mathcal{K}(s) \subseteq D$ l'espace des décisions admissibles dans l'état s;
- Probabilité de transition $\mathcal{P}(s, d)(s')$;
- Observation $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Fonction d'observation $\mathcal{Z}(s)(\omega)$;
- Fonction coût $C(s, d, s')$.

Caractéristiques d'un POMDP

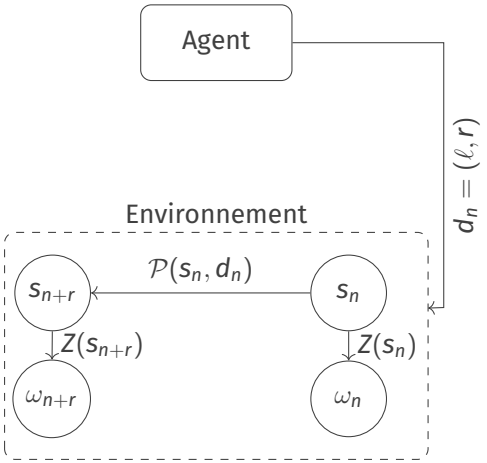


Définition d'un POMDP

Un POMDP se définit par un tuple $(\mathcal{S}, \mathcal{D}, \mathcal{K}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Etat du patient $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Décisions $d = (\ell, r) \in \mathcal{D}$;
- $\mathcal{K}(s) \subseteq D$ l'espace des décisions admissibles dans l'état s;
- Probabilité de transition $\mathcal{P}(s, d)(s')$;
- Observation $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Fonction d'observation $\mathcal{Z}(s)(\omega)$;
- Fonction coût $C(s, d, s')$.

Caractéristiques d'un POMDP

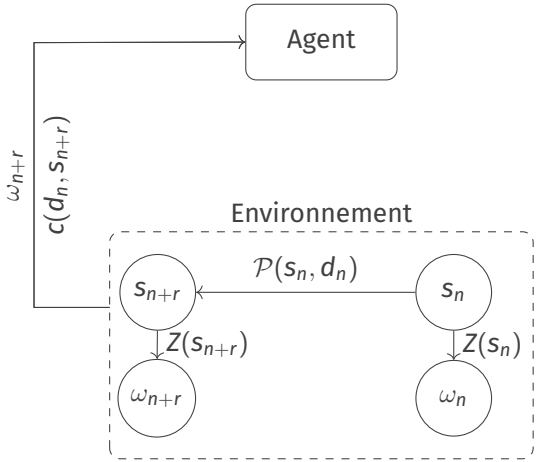


Définition d'un POMDP

Un POMDP se définit par un tuple $(\mathcal{S}, \mathcal{D}, \mathcal{K}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Etat du patient $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Décisions $d = (\ell, r) \in \mathcal{D}$;
- $\mathcal{K}(s) \subseteq D$ l'espace des décisions admissibles dans l'état s;
- Probabilité de transition $\mathcal{P}(s, d)(s')$;
- Observation $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Fonction d'observation $\mathcal{Z}(s)(\omega)$;
- Fonction coût $C(s, d, s')$.

Caractéristiques d'un POMDP



Définition d'un POMDP

Un POMDP se définit par un tuple $(\mathcal{S}, \mathcal{D}, \mathcal{K}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Etat du patient $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Décisions $d = (\ell, r) \in \mathcal{D}$;
- $\mathcal{K}(s) \subseteq D$ l'espace des décisions admissibles dans l'état s;
- Probabilité de transition $\mathcal{P}(s, d)(s')$;
- Observation $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Fonction d'observation $\mathcal{Z}(s)(\omega)$;
- Fonction coût $C(s, d, s')$.

Résoudre un POMDP

Identifier une politique $\pi : h \rightarrow d$ optimale déterministe!

On n'observe pas l'espace d'état ! Soit l'historique $h = (\omega_0, d_1, \omega_1, d_2, \ldots, \omega_n)$

Résoudre un POMDP

Identifier une politique $\pi : h \to d$ optimale déterministe!

On n'observe pas l'espace d'état ! Soit l'historique $h = (\omega_0, d_1, \omega_1, d_2, \ldots, \omega_n)$

$$\underbrace{V(\pi, h)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}_h^{\pi}[\sum_{n=1}^{H-1} c(D_n, S_n)]}_{\text{Coût attendu à long terme suite à la politique menée } \pi}$$

Résoudre un POMDP

Identifier une politique $\pi : h \to d$ optimale déterministe!

On n'observe pas l'espace d'état ! Soit l'historique $h = (\omega_0, d_1, \omega_1, d_2, \ldots, \omega_n)$

$$\underbrace{V(\pi, h)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}_h^{\pi}[\sum_{n=1}^{H-1} c(D_n, S_n)]}_{\text{Coût attendu à long terme suite à la politique menée } \pi}$$

Résoudre un POMDP

Identifier une politique $\pi : h \to d$ optimale déterministe!

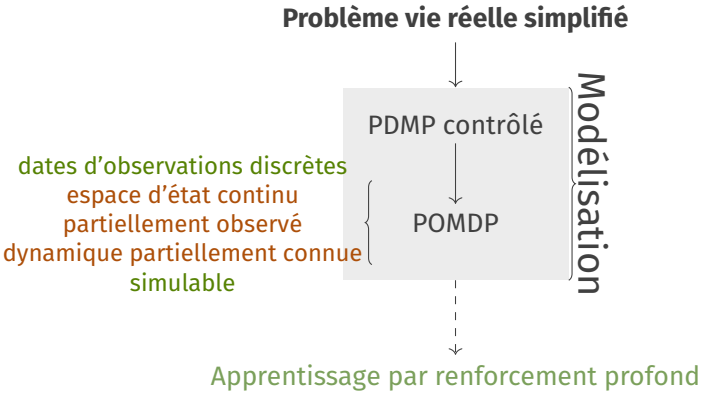
On n'observe pas l'espace d'état ! Soit l'historique $h = (\omega_0, d_1, \omega_1, d_2, \ldots, \omega_n)$

$$\underbrace{V(\pi, h)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}_h^{\pi}[\sum_{n=1}^{H-1} c(D_n, S_n)]}_{\text{Coût attendu à long terme suite à la politique menée } \pi}$$

$$\underbrace{V^{\star}(h)}_{\text{Fonction valeur}} = \underbrace{\min_{\pi \in \Pi} V(\pi, h)}_{\text{Minimisation sur l'ensemble des politiques } \Pi.}$$

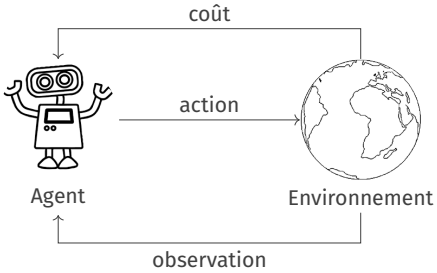
Méthodes

Meritxell Vinyals [4]



[4]Univ Toulouse, INRAE-MIAT, Toulouse, France

Apprentissage par renforcement



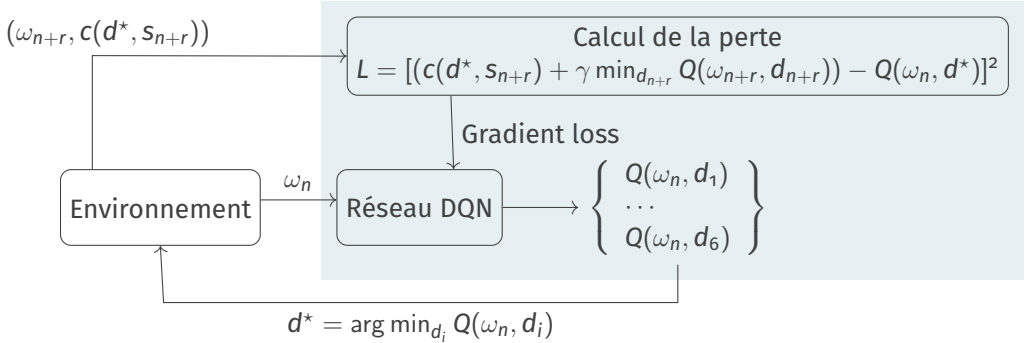
La politique optimale est obtenue à partir des expériences $< \omega, d, \omega', c >$

$$\underbrace{Q^{\pi}(h,d)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}[\sum_{n=0}^{H-1} c(D_n, S_n)|h,d,\pi]}_{\text{Valeur d'une action dans un état suivant la politique } \pi}$$

$$\underbrace{Q^{\star}(h,d)}_{\text{Q fonction}} = \min_{\pi \in \Pi} Q^{\pi}(h,d)$$

$$\underbrace{\pi^{\star}}_{\text{Politique Optimale}} = \arg\min_{d \in \mathcal{D}} Q^{\star}(h,d)$$

Algorithme DQN[5]



[5]Deep Q-Network

Résultats

Politique	Coût moyen (log)	IC	Taux de survie	Rechutes
OH	5.05	[3.58, 5.62]	91.88%	2.12
Inactive	6.39	[5.98, 6.67]	0.08%	1.00
Threshold	16.03	[14.83, 16.56]	74.04%	1.01
DQN observé	6.40	[5.99, 6.69]	0.84%	1.01
DQN	5.64	[4.49, 6.16]	76.76%	0.99
R2D2[6]	6.49	[6.19, 6.71]	0.08%	1.00

TABLE: Policy evaluation performance on 10^5 simulations

[6] R2D2 $\approx$ DQN + LSTM

Conclusion

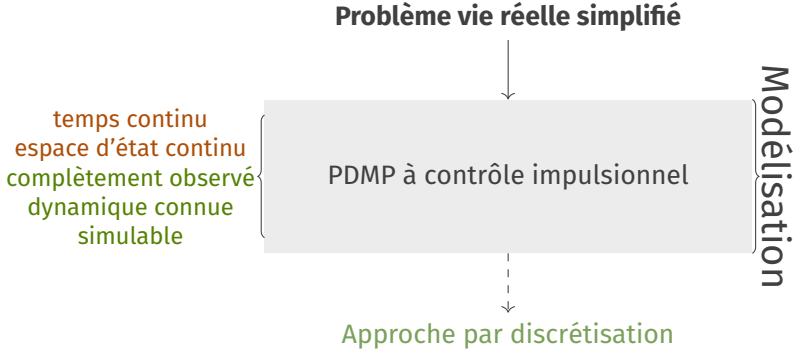
Politique	Coût moyen (log)	IC	Taux de survie	Rechutes
OH	5.05	[3.58, 5.62]	91.88%	2.12
Inactive	6.39	[5.98, 6.67]	0.08%	1.00
Threshold	16.03	[14.83, 16.56]	74.04%	1.01
DQN observé	6.40	[5.99, 6.69]	0.84%	1.01
DQN	5.64	[4.49, 6.16]	76.76%	0.99
R2D2	6.49	[6.19, 6.71]	0.08%	1.00

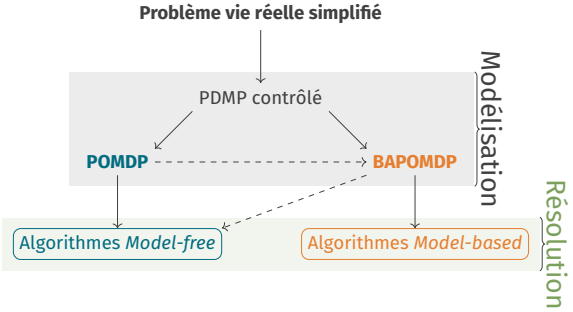
- Politiques peu conforme à la réalité
- Paramétrisation de la fonction de coût
- Work in Progress

Nécessite beaucoup de données pour apprendre la politique optimale !

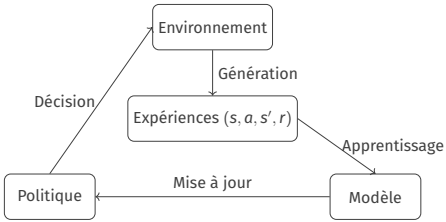
Perspectives

Evaluer les politiques obtenues avec les différents algorithmes



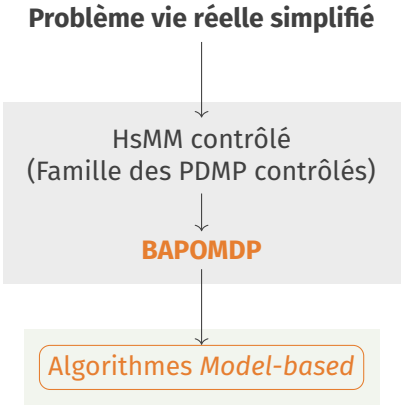


Apprentissage par renforcement
model-based



Perspectives

Extension de la méthode aux HsMM[7] contrôlés



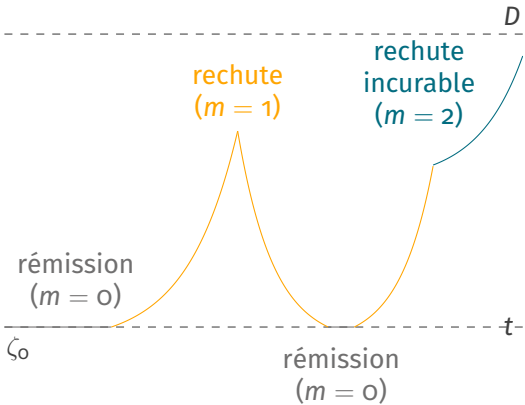
- Politique d'**échantillonnage adaptative** pour **reconstruire un processus stochastique**
- Application à la migration des oiseaux

[7] Hidden semi-Markov Model

Résumé

Programme 3ème Année

- Travaux sur les familles de politiques ϵ-optimales (publication contrôle)
- Transformation et résolution du PDMP contrôlé en BAPOMDP (acte de conférence ML)
- Extension de la méthode aux HsMM contrôlés - Application sur la migration des oiseaux (publication d'écologie)
- Rédaction et soutenance de thèse

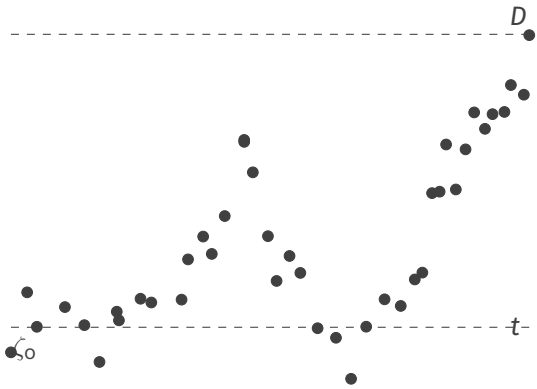


Soit $s = (m, k, \zeta, u, t, \tau)$ l'état du patient:

- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

Soit d la décision telle que: $d = (\ell, r)$:

- ℓ traitement (*rien, chimiothérapie*);
- r temps avant la prochaine visite (*15, 30, 60 jours*).



Soit $s = (m, k, \zeta, u, t, \tau)$ l'état du patient:

- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

Soit d la décision telle que: $d = (\ell, r)$:

- ℓ traitement (*rien, chimiothérapie*);
- r temps avant la prochaine visite (*15, 30, 60 jours*).

$$\underbrace{C(d, s')}_{\text{Fonction de coût}} = \underbrace{C_V}_{\text{coût de la visite}} + \underbrace{C_D(H - t') \times \mathbb{1}_{m'=3}}_{\text{coût de la mort}} + \underbrace{\kappa_C \times r \times \mathbb{1}_{\ell=a}}_{\text{coût de la chimiothérapie}}$$

avec $C_V = 1$, $C_D = 0.08$ et $\kappa_C = 0.4$