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

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

July 1st 2024

Le contexte médical

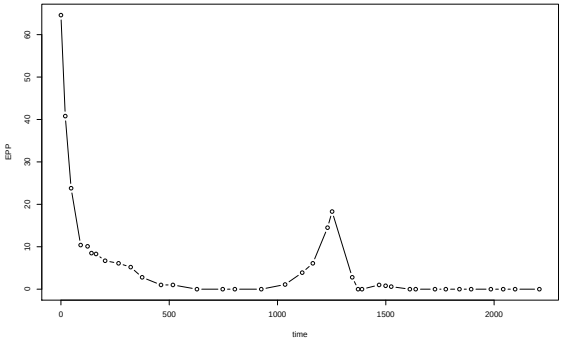


FIGURE: Exemple de donnée d'un patient[a]

- 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

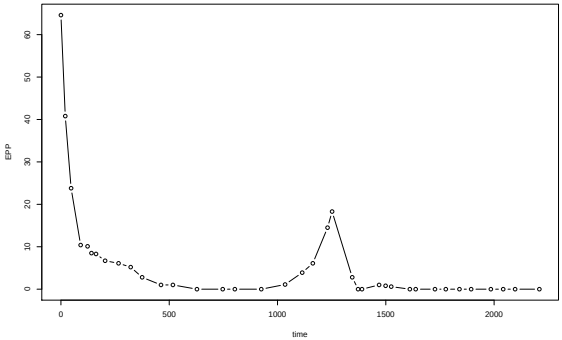


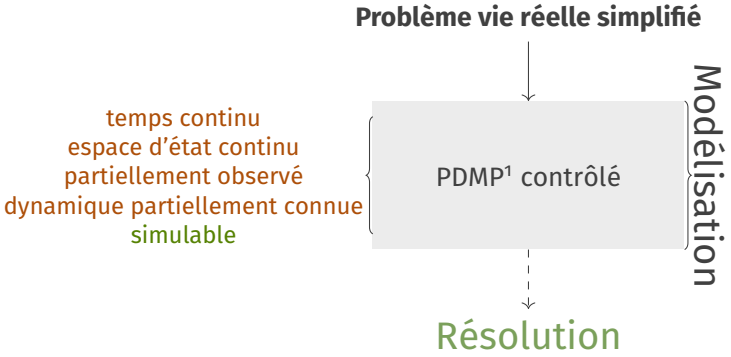
FIGURE: Exemple de donnée d'un patient[a]

- 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**

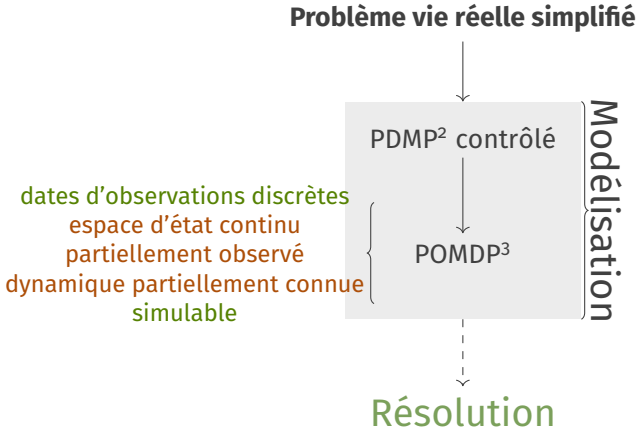
[a] IUCT Oncopole et CRCT, Toulouse, France

Méthodes



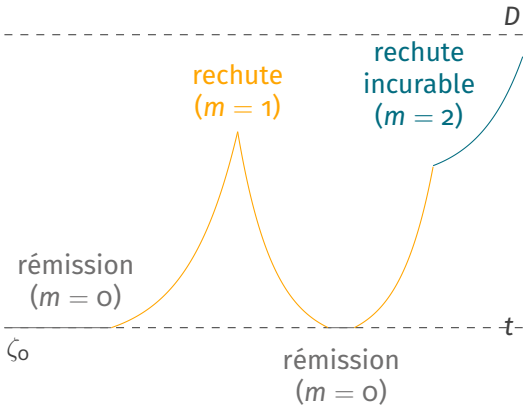
[1]Processus Markovien Déterministe par Morceaux

Méthodes



[2] Processus Markovien Déterministe par Morceaux
[3] Processus de Décision Markovien Partiellement Observé

Le modèle POMDP[4]



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

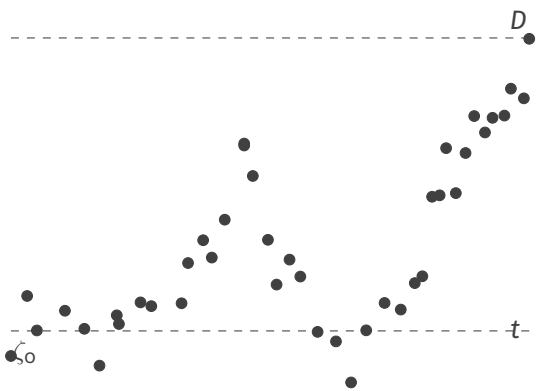
- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

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

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

[4]Processus de Décision Markovien Partiellement Observé

Le modèle POMDP[4]



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

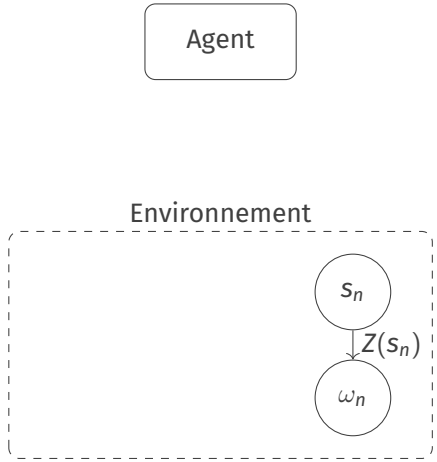
- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

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

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

[4]Processus de Décision Markovien Partiellement Observé

Caractéristiques d'un POMDP[5]



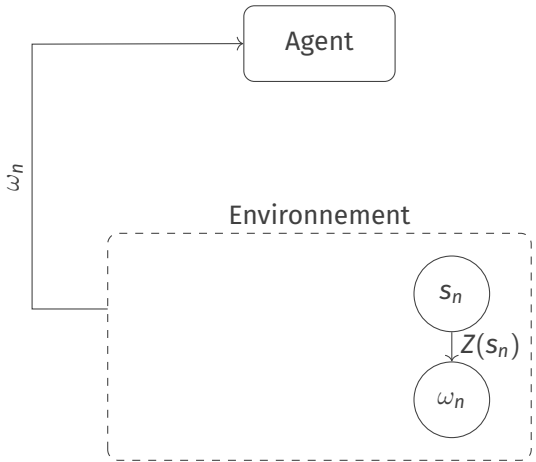
POMDP Definition

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 $C(s, d, s')$.

[5] Processus de Décision Markovien Partiellement Observé

Caractéristiques d'un POMDP[5]



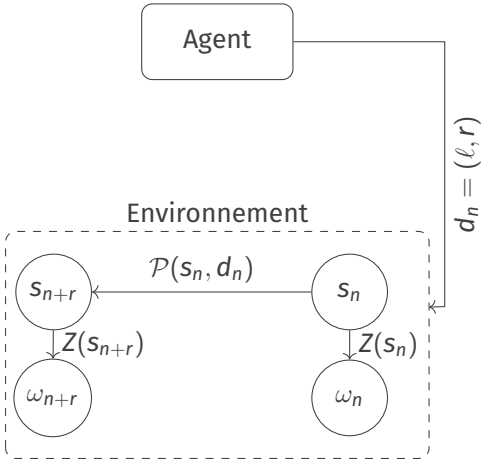
POMDP Definition

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 $C(s, d, s')$.

[5]Processus de Décision Markovien Partiellement Observé

Caractéristiques d'un POMDP[5]



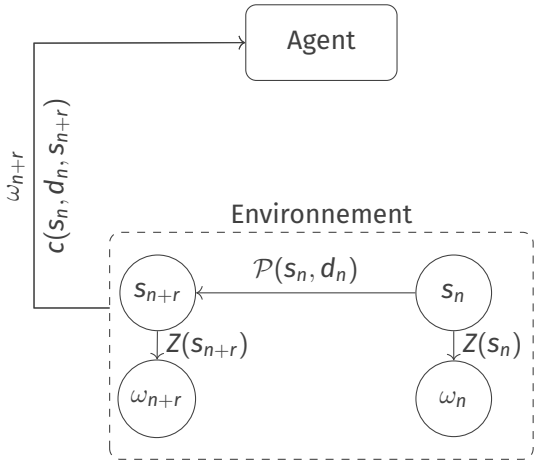
POMDP Definition

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 $C(s, d, s')$.

[5]Processus de Décision Markovien Partiellement Observé

Caractéristiques d'un POMDP[5]



POMDP Definition

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 $C(s, d, s')$.

[5] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[6]

Identifier une politique optimale!

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

[6] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[6]

Identifier une politique optimale!

$$\underbrace{V(\pi, s)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}_s^{\pi}[\sum_{n=0}^{H-1} c(S_{n-1}, D_n, S_n)]}_{\text{Coût attendu à long terme suite à la politique menée } \pi}$$

[6] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[6]

Identifier une politique optimale!

$$\underbrace{V(\pi, s)}_{\text{Critère à optimiser}} = \underbrace{\mathbb{E}_s^{\pi}[\sum_{n=0}^{H-1} c(S_{n-1}, D_n, S_n)]}_{\text{Coût attendu à long terme suite à la politique menée } \pi}$$

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

[6] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[6]

Identifier une politique optimale!

En réalité on observe pas l'espace d'état !

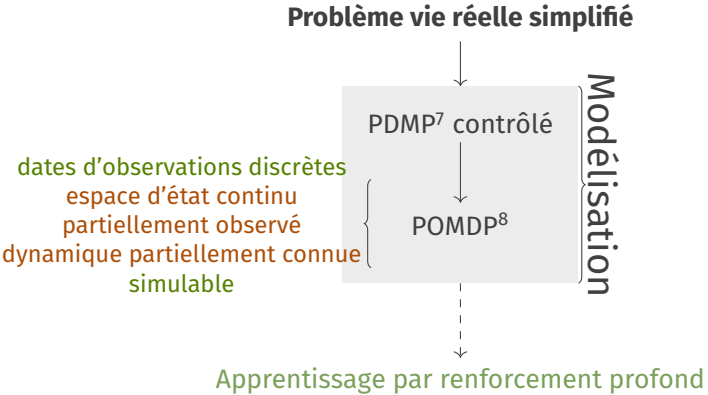
Soit l'historique $h = (\omega_0, d_0, \omega_1, d_1, \ldots, \omega_n)$

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

[6] Processus de Décision Markovien Partiellement Observé

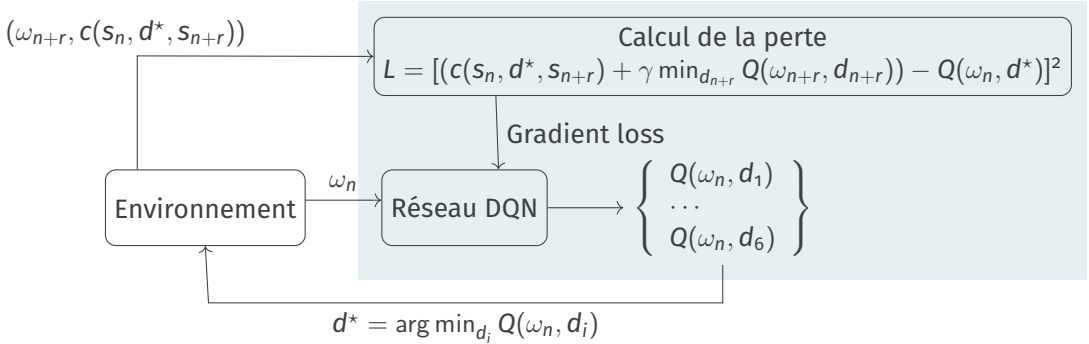
Méthodes

Meritxell Vinyals [9]



[7] Processus Markovien Déterministe par Morceaux
[8] Processus de Décision Markovien Partiellement Observé
[9] Univ Toulouse, INRAE-MIAT, Toulouse, France

Algorithme DQN[10]



[10] Deep Q-Network

Résultats

Politique	Coût moyen (log)	Interval de confiance
OH	8.79	[7.89, 9.69]
Random	11.82	[10.80, 12.84]
Inactive	12.49	[11.54, 13.45]
Threshold	9.89	[8.94, 10.83]
DQN	12.49	[11.54, 13.45]
R2D2[11]	8.47	[7.61, 9.33]

TABLE: Policy evaluation performance on 10^5 simulations

[11] R2D2 $\approx$ DQN + LSTM

Résultats

Politique	Coût moyen (log)	IC	Taux de survie	Rechutes
OH	8.79	[7.89, 9.69]	93.45%	2.14
Random	11.82	[10.80, 12.84]	27.45%	1.01
Inactive	12.49	[11.54, 13.45]	0.01%	1.00
Threshold	9.89	[8.94, 10.83]	78.95%	1.01
DQN	12.49	[11.54, 13.45]	0.02%	1
R2D2[11]	8.47	[7.61, 9.33]	96.95%	0.65

TABLE: Policy evaluation performance on 10^5 simulations

[11] R2D2 $\approx$ DQN + LSTM

Conclusion

Politique	Coût moyen (log)	IC	Taux de survie	Rechutes
OH	8.79	[7.89, 9.69]	93.45%	2.14
Random	11.82	[10.80, 12.84]	27.45%	1.01
Inactive	12.49	[11.54, 13.45]	0.01%	1.00
Threshold	9.89	[8.94, 10.83]	78.95%	1.01
DQN	12.49	[11.54, 13.45]	0.02%	1
R2D2	8.47	[7.61, 9.33]	96.95%	0.65

- Politiques peu conforme à la réalité
- Algorithmes sensibles à la fonction de coût et sa paramétrisation
- Historique important dans l'apprentissage

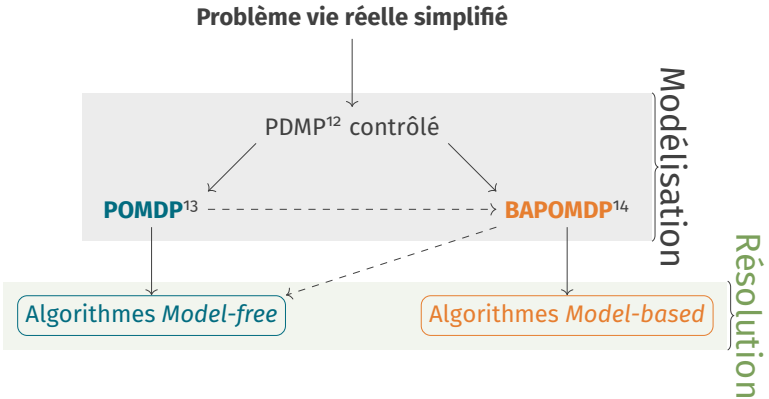
Conclusion

Politique	Coût moyen (log)	IC	Taux de survie	Rechutes
OH	8.79	[7.89, 9.69]	93.45%	2.14
Random	11.82	[10.80, 12.84]	27.45%	1.01
Inactive	12.49	[11.54, 13.45]	0.01%	1.00
Threshold	9.89	[8.94, 10.83]	78.95%	1.01
DQN	12.49	[11.54, 13.45]	0.02%	1
R2D2	8.47	[7.61, 9.33]	96.95%	0.65

- Politiques peu conforme à la réalité
- Algorithmes sensibles à la fonction de coût et sa paramétrisation
- Historique important dans l'apprentissage

Nécessite beaucoup de données pour apprendre la politique optimale !

Perspectives



[12] Processus Markovien Déterministe par Morceaux
[13] Processus de Décision Markovien Partiellement Observé
[14] Processus de Décision Markovien Partiellement Observé Bayes Adaptif

Résumé

- Politiques difficilement interprêtables
- Très sensibles à la fonction de coût
- L'historique améliore les performances
- Nécessite beaucoup de données pour apprendre la politique optimale !