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

Application à l'optimisation d'un traitement médical

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

June 2024

Le contexte médical

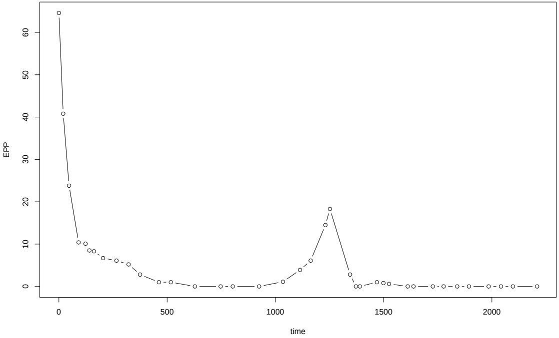


FIGURE: Exemple de donnée d'un patient[a]

- 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

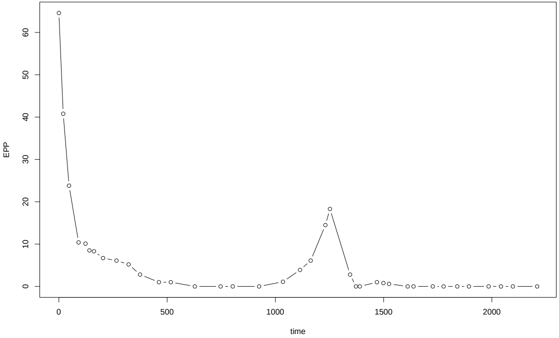


FIGURE: Exemple de donnée d'un patient[a]

- 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

Problème vie réelle simplifié

temps continu
espace d'état continu
partiellement observé
dynamique partiellement connue
simulable

PDMP[1] contrôlé

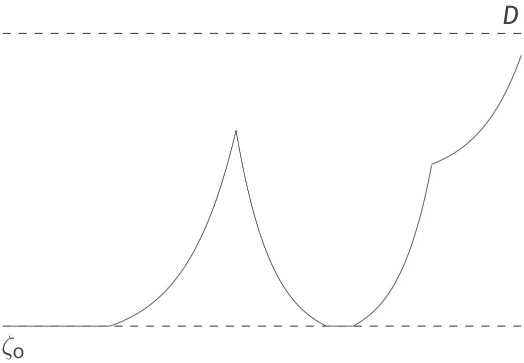
Modélisation

Résolution

[1] Processus Markovien Déterministe par Morceaux

Le modèle PDMP[2] contrôlé

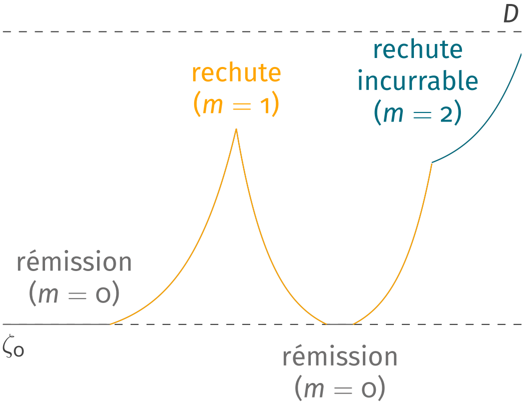
On passe aléatoirement d'un régime déterministe à un autre.



[2] Processus Markovien Déterministe par Morceaux

Le modèle PDMP[2] contrôlé

On passe aléatoirement d'un régime déterministe à un autre.



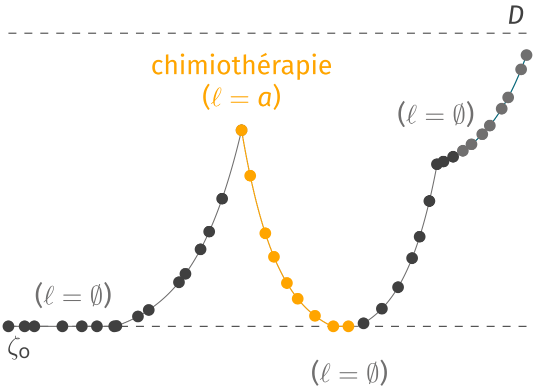
Soit l'état du patient $x = (m, k, \zeta, u)$:

- m l'état du patient;
- k le nombre de rechute;
- ζ le biomarqueur;
- u le temps depuis le dernier saut.

[2] Processus Markovien Déterministe par Morceaux

Le modèle PDMP[2] contrôlé

On passe aléatoirement d'un régime déterministe à un autre.



Soit l'état du patient $x = (m, k, \zeta, u)$:

- m l'état du patient;
- k le nombre de rechute;
- ζ le biomarqueur;
- u le temps depuis le dernier saut.

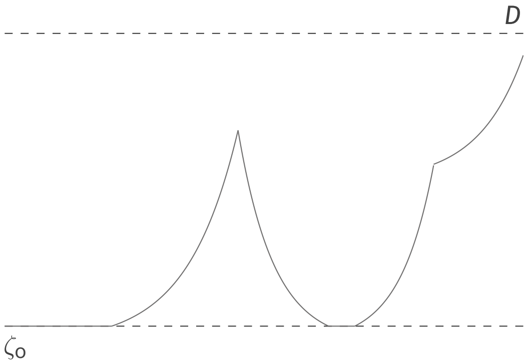
Soit d la décision telle que: $d = (\ell, r)$:

- ℓ le traitement;
- r le temps avant la prochaine visite.

[2] Processus Markovien Déterministe par Morceaux

Caractéristiques d'un PDMP[3]

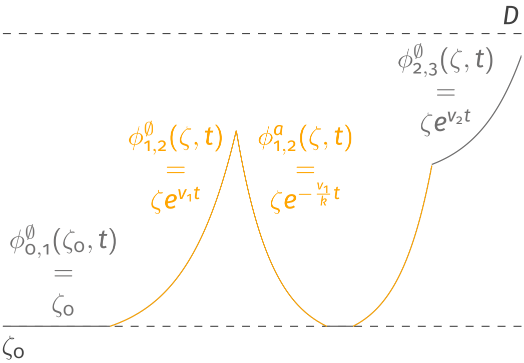
Un PDMP se définit par trois caractéristiques locales.



[3] Processus Markovien Déterministe par Morceaux

Caractéristiques d'un PDMP[3]

Un PDMP se définit par trois caractéristiques locales.



Le flot

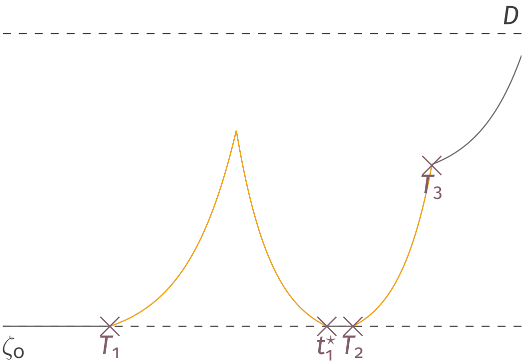
Description de la partie déterministe du processus.

$$\Phi^{\ell}(x,t) = (m, k, \phi^{\ell}_{m,k}(\zeta,t), u+t)$$

[3] Processus Markovien Déterministe par Morceaux

Caractéristiques d'un PDMP[3]

Un PDMP se définit par trois caractéristiques locales.



L'intensité de saut

Description des mécanismes de saut du processus.

- Saut à la frontière (déterministe)

$$t^\star(x) = t_m^{\ell\star}(\zeta) = \inf\{t > 0 : \phi_{m,k}^{\ell}(\zeta, t) \in \{\zeta_0, D\}\}$$

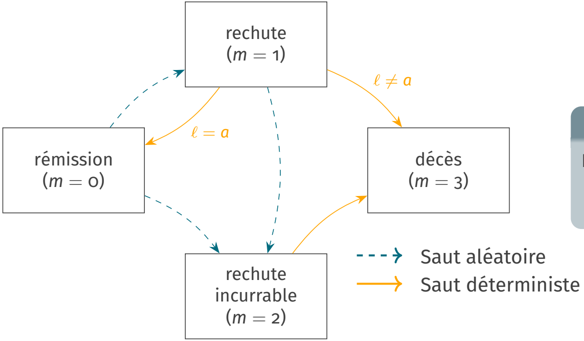
- Saut aléatoire

$$\mathbb{P}(T > t) = e^{-\int_0^t \lambda_m^{\ell}(\Phi(x,s))\,\mathrm{ds}}$$

[3] Processus Markovien Déterministe par Morceaux

Caractéristiques d'un PDMP[3]

Un PDMP se définit par trois caractéristiques locales.



Le noyau

Description de l'état du processus après chaque saut.

$$\mathbb{P}(X' \in A | X = x) = \int_A Q_m^d(\Phi^\ell(x, T), \mathrm{d}x')$$

[3]Processus Markovien Déterministe par Morceaux

Méthodes

Problème vie réelle simplifié

PDMP[4] contrôlé

POMDP[5]

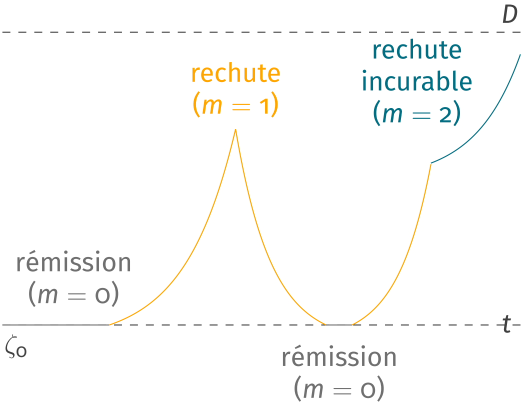
Modélisation

dates d'observations discrètes
espace d'état continu
partiellement observé
dynamique partiellement connue
simulable

Résolution

[4] Processus Markovien Déterministe par Morceaux
[5] Processus de Décision Markovien Partiellement Observé

Le modèle POMDP[6]



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

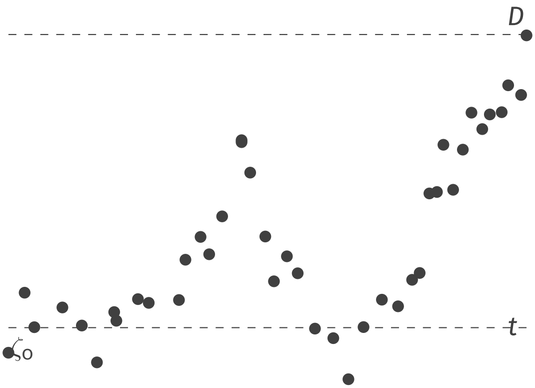
- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

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

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

[6] Processus de Décision Markovien Partiellement Observé

Le modèle POMDP[6]



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

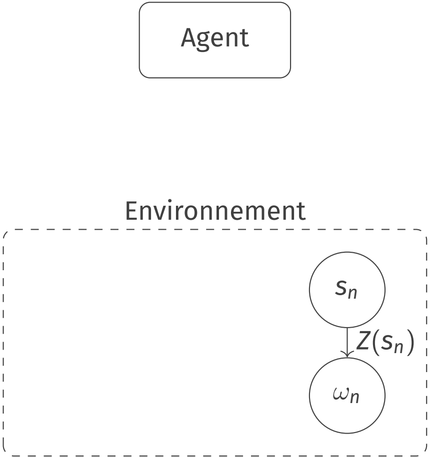
- m état général du patient;
- k nombre de rechute;
- ζ biomarqueur;
- u temps depuis le dernier saut;
- t temps écoulé depuis le début du suivi;
- τ temps depuis l'application d'un traitement.

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

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

[6] Processus de Décision Markovien Partiellement Observé

Caractérisitiques d'un POMDP[7]



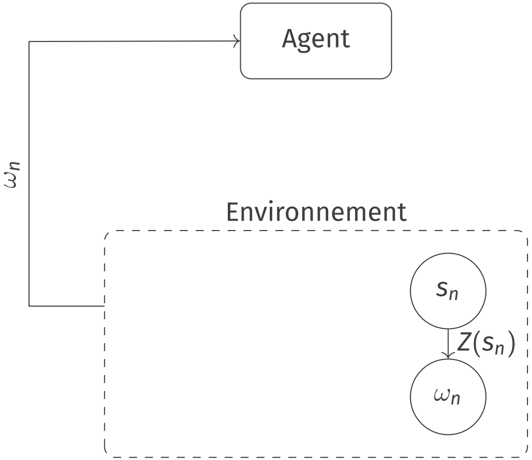
POMDP Definition

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

[7] Processus de Décision Markovien Partiellement Observé

Caractérisitiques d'un POMDP[7]



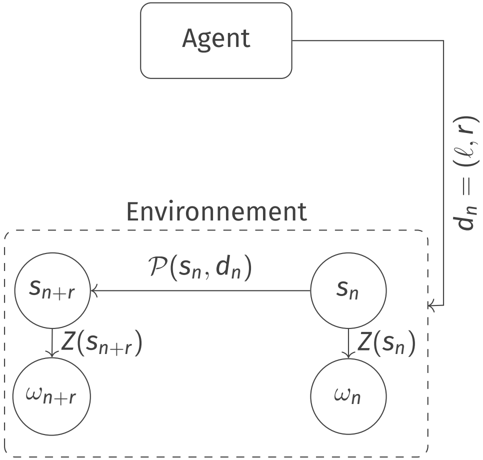
POMDP Definition

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

[7] Processus de Décision Markovien Partiellement Observé

Caractérisitiques d'un POMDP[7]



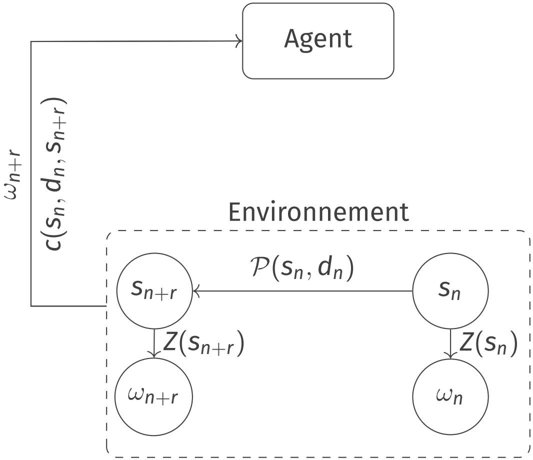
POMDP Definition

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

[7] Processus de Décision Markovien Partiellement Observé

Caractérisitiques d'un POMDP[7]



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

[7] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[8]

Identifier une politique optimale!

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

[8] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[8]

Identifier une politique optimale!

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

[8] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[8]

Identifier une politique optimale!

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

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

[8] Processus de Décision Markovien Partiellement Observé

Résoudre un POMDP[8]

Identifier une politique optimale!

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

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

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

[8]Processus de Décision Markovien Partiellement Observé

Etat de l'art

Quand la dynamique est connue

Problème vie réelle simplifié

PDMP[9] contrôlé

POMDP[10]

Discrétisation

Modélisation

Programmation dynamique[11]

Monte-Carlo Planning[12]

---

[9] Processus Markovien Déterministe par Morceaux

[10] Processus de Décision Markovien Partiellement Observé

[11] A. Cleynen and B. de Saporta. Numerical method to solve impulse control problems for partially observed piecewise deterministic Markov processes. 2023

[12] A. Cleynen, B. de Saporta, A. Thierry D'Argenlieu, and R. Sabbadin. Medical follow-up optimization : A Monte-Carlo planning strategy. 2024

Méthodes

Meritxell Vinyals [15]

Problème vie réelle simplifié

PDMP[13] contrôlé

dates d'observations discrètes
espace d'état continu
partiellement observé
dynamique partiellement connue
simulable

POMDP[14]

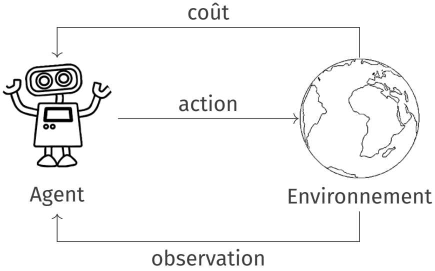
Modélisation

Apprentissage par renforcement profond

---

[13] Processus Markovien Déterministe par Morceaux
[14] Processus de Décision Markovien Partiellement Observé
[15] Univ Toulouse, INRAE-MIAT, Toulouse, France

Apprentissage par renforcement



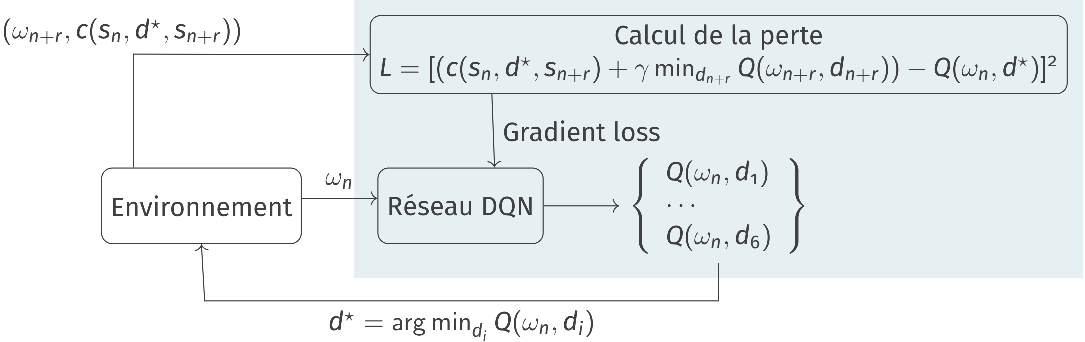
La politique optimale est obtenue à partir des expériences $< \omega, d, \omega', c >$

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

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

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

Algorithme DQN[16]



[16] 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	8.47	[7.61, 9.33]

TABLE: Policy evaluation performance on 10^5 simulations

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	8.47	[7.61, 9.33]	96.95%	0.65

Table: Policy evaluation performance on 10^5 simulations

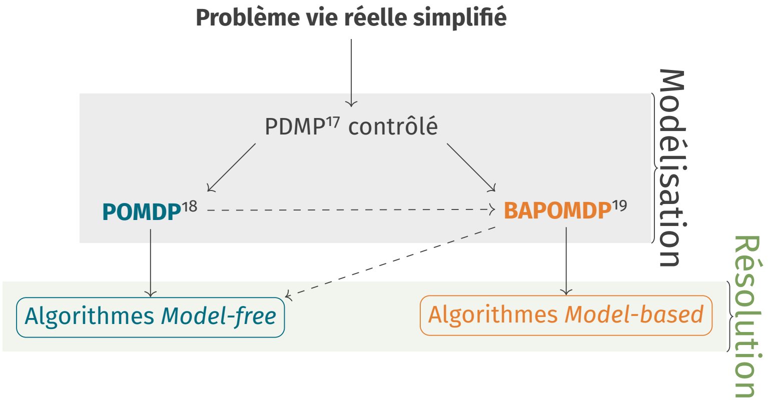
Résultats

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	8.47	[7.61, 9.33]	96.95%	0.65

TABLE: Policy evaluation performance on 10^5 simulations

Nécessite beaucoup de données pour apprendre la politique optimale !

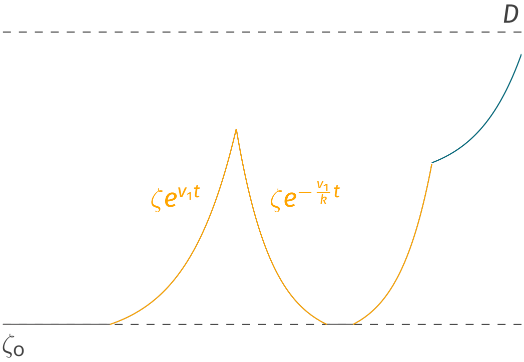
Conclusion et futurs travaux



[17] Processus Markovien Déterministe par Morceaux
[18] Processus de Décision Markovien Partiellement Observé
[19] Processus de Décision Markovien Partiellement Observé Bayes Adaptif

Une dynamique partiellement connue

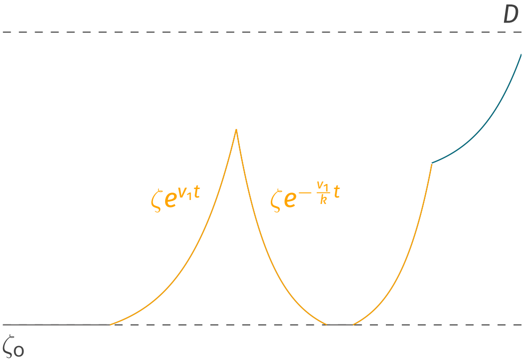
On ne connait pas le paramètre de pente v_1 de la maladie.



<u>Hypothèse:</u>
$v_1 \sim$ log-normale (μ, σ^{-2}).

Une dynamique partiellement connue

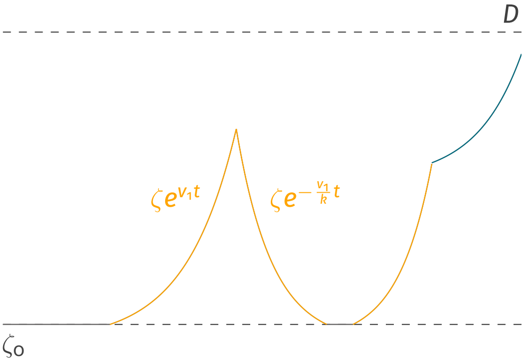
On ne connait pas le paramètre de pente v_1 de la maladie.



Hypothèse:
$v_1 \sim$ log-normale (μ, σ^{-2}).
Inférence bayésienne:
$(\mu, \sigma^{-2}) \sim$ gamma-log-normale$(\alpha, \beta, \kappa, \nu)$.

Une dynamique partiellement connue

On ne connait pas le paramètre de pente v_1 de la maladie.

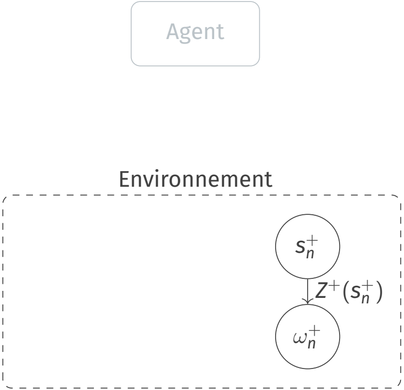


Hypothèse:
$v_1 \sim$ log-normale (μ, σ^{-2}).
Inférence bayésienne:
$(\mu, \sigma^{-2}) \sim$ gamma-log-normale$(\alpha, \beta, \kappa, \nu)$.

Mise à jour des hyperparamètres

- $\alpha_{n+1} = \frac{\beta_n \alpha_n + \log(\hat{v_1})}{\beta_n + 1}$
- $\beta_{n+1} = \beta_n + 1$
- $\kappa_{n+1} = \kappa_n + \frac{1}{2}$
- $\nu_{n+1} = \nu_n + \frac{\beta_n (\log(\hat{v_1} - \alpha_n)^2)}{2(\beta_n + 1)}$

Un BAPOMDP[20] partiellement observé



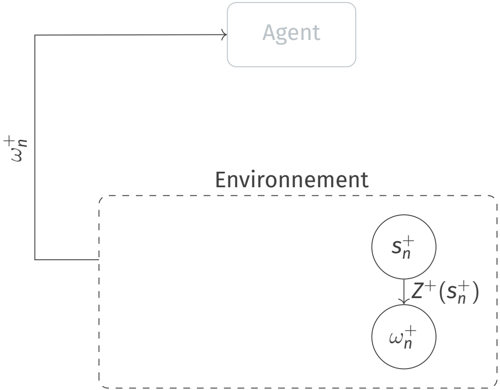
BAMDP PO

Un BAMDP-PO se définit par un tuple $(\mathcal{S}^+, \mathcal{D}, \mathcal{K}, \mathcal{P}^+, \Omega^+, \mathcal{Z}^+, C)$.

- L'hyper-état du patient $s^+ = (m, k, \zeta, u, \tau, t, \alpha, \beta, \kappa, \nu)$;
- Les décisions restent inchangées;
- $\mathcal{K}(\omega) \subseteq D$ l'espace des décisions admissibles selon l'observation ω;
- La probabilité de transition $\mathcal{P}(s^+, d)(s')$;
- Les observations $\omega^+ = (z, F(\zeta), \tau, t, \tilde{\alpha}, \tilde{\beta}, \tilde{\kappa}, \tilde{\nu})$;
- La fonction d'observations $\mathcal{Z}(s^+)(\omega^+)$;
- La fonction de coût $C : \mathcal{D} \times \mathcal{S} \to \mathbb{R}$.

[20] Processus de décision Markovien Partiellement Observé Bayes adaptatif

Un BAPOMDP[20] partiellement observé



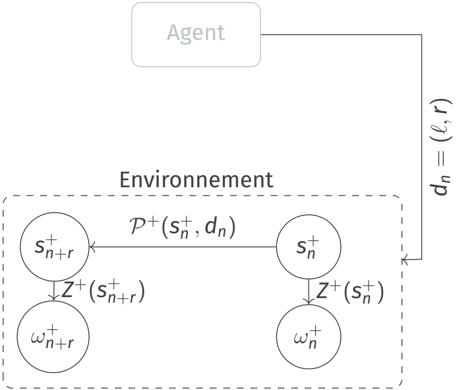
BAMDP PO

Un BAMDP-PO se définit par un tuple $(\mathcal{S}^+, \mathcal{D}, \mathcal{K}, \mathcal{P}^+, \Omega^+, \mathcal{Z}^+, C)$.

- L'hyper-état du patient $s^+ = (m, k, \zeta, u, \tau, t, \alpha, \beta, \kappa, \nu)$;
- Les décisions restent inchangées;
- $\mathcal{K}(\omega) \subseteq D$ l'espace des décisions admissibles selon l'observation ω;
- La probabilité de transition $\mathcal{P}(s^+, d)(s')$;
- Les observations $\omega^+ = (z, F(\zeta), \tau, t, \tilde{\alpha}, \tilde{\beta}, \tilde{\kappa}, \tilde{\nu})$;
- La fonction d'observations $\mathcal{Z}(s^+)(\omega^+)$;
- La fonction de coût $C : \mathcal{D} \times \mathcal{S} \to \mathbb{R}$.

[20] Processus de décision Markovien Partiellement Observé Bayes adaptatif

Un BAPOMDP[20] partiellement observé



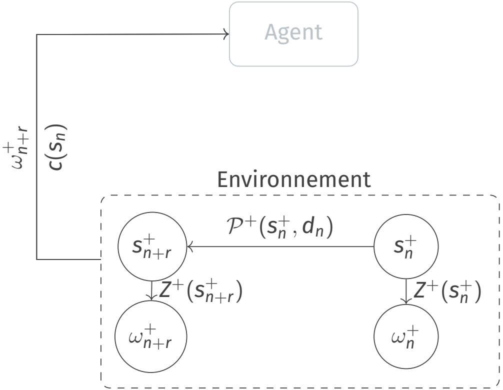
BAMDP PO

Un BAMDP-PO se définit par un tuple $(\mathcal{S}^+, \mathcal{D}, \mathcal{K}, \mathcal{P}^+, \Omega^+, \mathcal{Z}^+, C)$.

- L'hyper-état du patient $s^+ = (m, k, \zeta, u, \tau, t, \alpha, \beta, \kappa, \nu)$;
- Les décisions restent inchangées;
- $\mathcal{K}(\omega) \subseteq D$ l'espace des décisions admissibles selon l'observation ω;
- La probabilité de transition $\mathcal{P}(s^+, d)(s')$;
- Les observations $\omega^+ = (z, F(\zeta), \tau, t, \tilde{\alpha}, \tilde{\beta}, \tilde{\kappa}, \tilde{\nu})$;
- La fonction d'observations $\mathcal{Z}(s^+)(\omega^+)$;
- La fonction de coût $C : \mathcal{D} \times \mathcal{S} \to \mathbb{R}$.

[20] Processus de décision Markovien Partiellement Observé Bayes adaptatif

Un BAPOMDP[20] partiellement observé



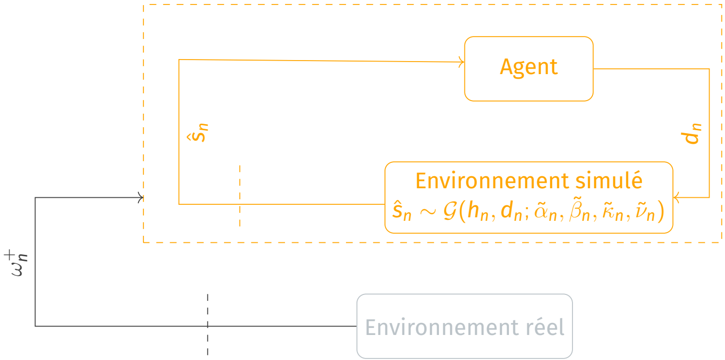
BAMDP PO

Un BAMDP-PO se définit par un tuple $(\mathcal{S}^+, \mathcal{D}, \mathcal{K}, \mathcal{P}^+, \Omega^+, \mathcal{Z}^+, C)$.

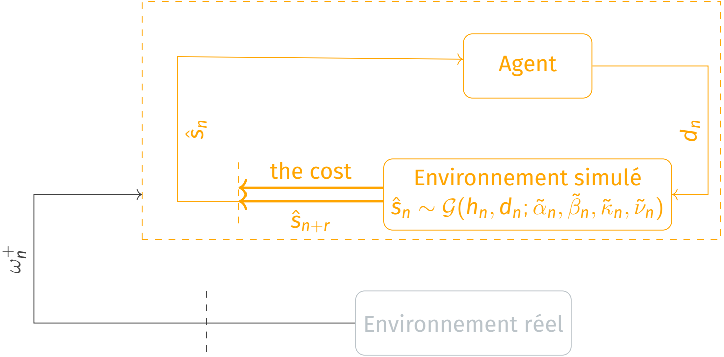
- L'hyper-état du patient $s^+ = (m, k, \zeta, u, \tau, t, \alpha, \beta, \kappa, \nu)$;
- Les décisions restent inchangées;
- $\mathcal{K}(\omega) \subseteq D$ l'espace des décisions admissibles selon l'observation ω;
- La probabilité de transition $\mathcal{P}(s^+, d)(s')$;
- Les observations $\omega^+ = (z, F(\zeta), \tau, t, \tilde{\alpha}, \tilde{\beta}, \tilde{\kappa}, \tilde{\nu})$;
- La fonction d'observations $\mathcal{Z}(s^+)(\omega^+)$;
- La fonction de coût $C : \mathcal{D} \times \mathcal{S} \to \mathbb{R}$.

[20] Processus de décision Markovien Partiellement Observé Bayes adaptatif

Une suggestion de résolution



Une suggestion de résolution



Une suggestion de résolution