Deep reinforcement learning for controlled Piecewise Deterministic Markov Process

Application to optimising medical treatment

Orlane Rossini [1], Alice Cleynen [1,2], Benoîte de Saporta [1],
Régis Sabbadin [3] and 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

28 Mai 2024

Medical context

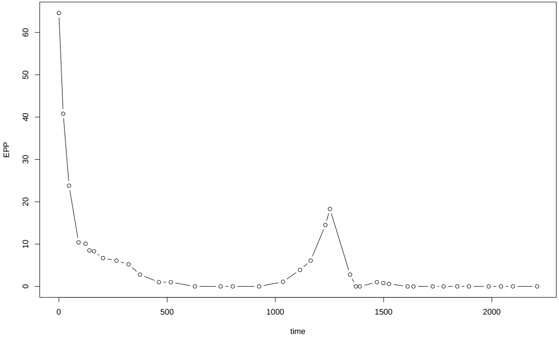


Figure: Example of patient data[a]

- Patients who have had cancer benefit from regular monitoring;
- the concentration of cancer cells is measured over time;
- The doctor has to make new decisions at each visit.

[a] IUCT Oncopole et CRCT, Toulouse, France

Medical context

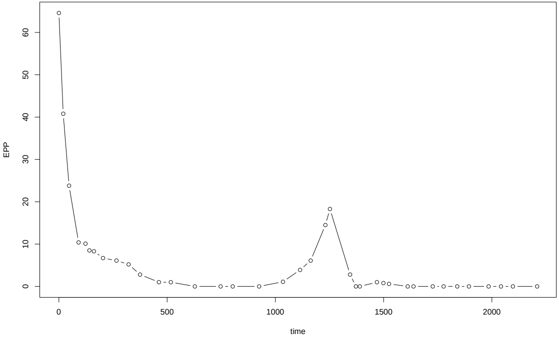


Figure: Example of patient data[a]

- Patients who have had cancer benefit from regular monitoring;
- the concentration of cancer cells is measured over time;
- The doctor has to make new decisions at each visit.

$\Longrightarrow$ **Optimising decision-making to ensure the patient's quality of life**

[a] IUCT Oncopole et CRCT, Toulouse, France

Methods

A real-life problem

$\downarrow$

Controlled PDMP[1]

[1] Piecewise Deterministic Markov Process

Methods

A real-life problem

$\downarrow$

Controlled PDMP[1]

$\downarrow$

Model-free framework

POMDP[2]

Modelling

[1] Piecewise Deterministic Markov Process

[2] Partially Observable Markov Decision Process

Methods

A real-life problem

Controlled PDMP[1]

Model-free framework
POMDP[2]

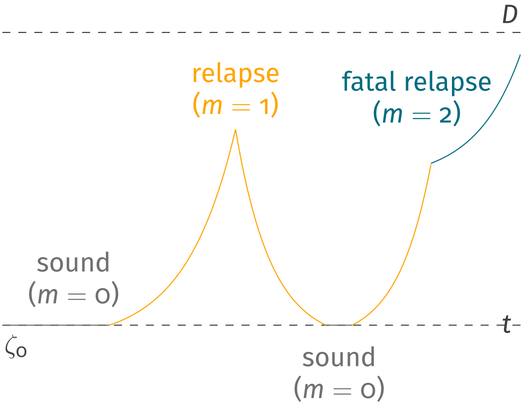
Modelling

Model-free algorithms

Solving

[1] Piecewise Deterministic Markov Process
[2] Partially Observable Markov Decision Process

Model-free framework[3]



Let the patient's state $s = (m, k, \zeta, u, t, \tau)$:

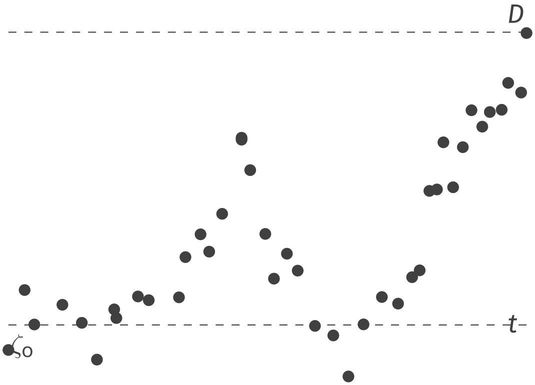
- m the overall state of the patient;
- k relapses number;
- ζ the biomarker;
- u time since last jump;
- t time since the beginning of follow-up;
- τ time since a treatment is applied.

Let d be the decision such that: $d = (\ell, r)$:

- ℓ the treatment (*nothing, chemotherapy, palliatif cares*);
- r time before the next visit (*15, 30, 60 days*).

[3] Partially Observable Markov Decision Process (POMDP)

Model-free framework[3]



Let the patient's state $s = (m, k, \zeta, u, t, \tau)$:

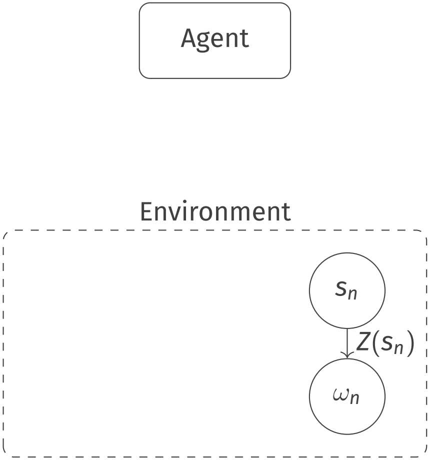
- m the overall state of the patient;
- k relapses number;
- ζ the biomarker;
- u time since last jump;
- t time since the beginning of follow-up;
- τ time since a treatment is applied.

Let d be the decision such that: $d = (\ell, r)$:

- ℓ the treatment (*nothing, chemotherapy, palliatif cares*);
- r time before the next visit (*15, 30, 60 days*).

[3] Partially Observable Markov Decision Process (POMDP)

POMDP[4] Characteristics



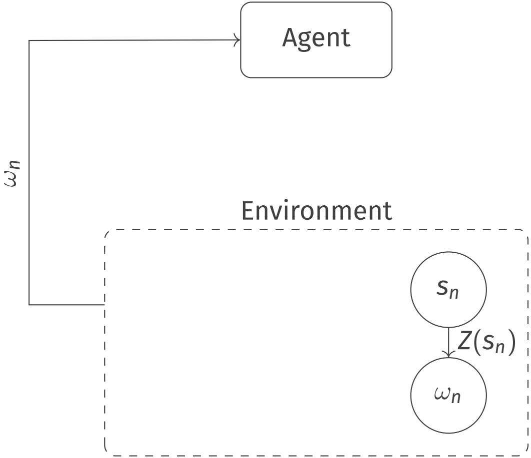
POMDP Definition

A POMDP is defined by a tuple $(\mathcal{S}, \mathcal{D}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Patient's state $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Decisions $d = (\ell, r) \in \mathcal{D}$;
- Transition probabilities $\mathcal{P}(s, d)(s')$;
- Observations $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- The observations function $\mathcal{Z}(s)(\omega)$;
- Cost function $C(s, d, s')$.

[4] Partially Observable Markov Decision Process

POMDP[4] Characteristics



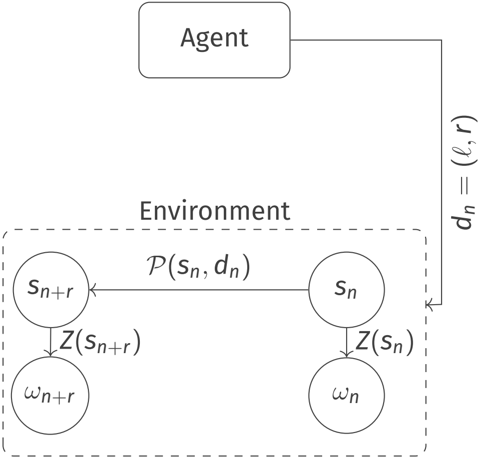
POMDP Definition

A POMDP is defined by a tuple $(\mathcal{S}, \mathcal{D}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Patient's state $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Decisions $d = (\ell, r) \in \mathcal{D}$;
- Transition probabilities $\mathcal{P}(s, d)(s')$;
- Observations $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- The observations function $\mathcal{Z}(s)(\omega)$;
- Cost function $C(s, d, s')$.

[4] Partially Observable Markov Decision Process

POMDP[4] Characteristics



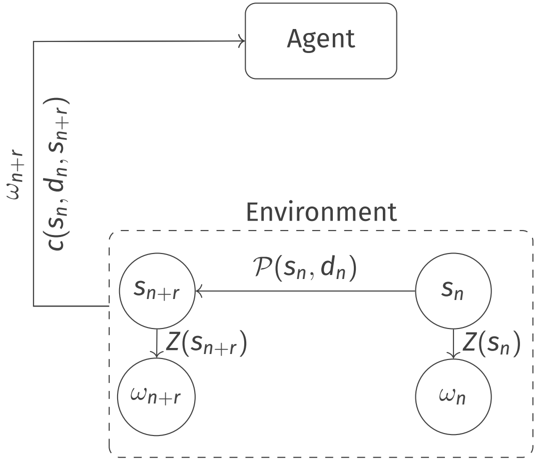
POMDP Definition

A POMDP is defined by a tuple $(\mathcal{S}, \mathcal{D}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Patient's state $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Decisions $d = (\ell, r) \in \mathcal{D}$;
- Transition probabilities $\mathcal{P}(s, d)(s')$;
- Observations $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- The observations function $\mathcal{Z}(s)(\omega)$;
- Cost function $C(s, d, s')$.

[4] Partially Observable Markov Decision Process

POMDP[4] Characteristics



POMDP Definition

A POMDP is defined by a tuple $(\mathcal{S}, \mathcal{D}, \mathcal{P}, \Omega, \mathcal{Z}, C)$.

- Patient's state $s = (m, k, \zeta, u, t, \tau) \in \mathcal{S}$;
- Decisions $d = (\ell, r) \in \mathcal{D}$;
- Transition probabilities $\mathcal{P}(s, d)(s')$;
- Observations $\omega = (\tau, t, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- The observations function $\mathcal{Z}(s)(\omega)$;
- Cost function $C(s, d, s')$.

POMDP Characteristics

- Discrete observation dates
- Continuous state space
- Partially observable
- Partially known model
- Simulator

[4] Partially Observable Markov Decision Process

What does it mean *to solve* ?

Identifying the optimal policy!

$$\underbrace{C(s,d,s')}_{\text{Cost function}} = \underbrace{C_V}_{\text{Visit cost}}$$

$$+ \underbrace{C_D(H-t') \times \left(1 - \frac{1}{2}\mathbb{1}_{m=2 \text{ and } \ell=b}\right)\mathbb{1}_{m'=3}}_{\text{Death cost}}$$

$$+ \underbrace{\beta|\zeta - \zeta_{TH}| \times \left(1 - \frac{1}{2}\mathbb{1}_{\ell=a \text{ or } \ell=b}\right)\mathbb{1}_{\zeta>\zeta_{TH}}}_{\text{Disease cost}}$$

$$+ \underbrace{\kappa_C \times r(1 - \frac{1}{2}\mathbb{1}_{m=1 \text{ and } \ell=a})\mathbb{1}_{\ell=a}}_{\text{Chemotherapy cost}} + \underbrace{\kappa_P \times r(1 - \frac{1}{2}\mathbb{1}_{m=2 \text{ and } \ell=b})\mathbb{1}_{\ell=b}}_{\text{Palliatif care cost}}$$

What does it mean *to solve* ?

Identifying the optimal policy!

$$\underbrace{V(\pi, s)}_{\text{Criterion to optimise}} = \underbrace{\mathbb{E}_s^\pi[\sum_{n=0}^{H-1} c(S_{n-1}, D_n, S_n)]}_{\text{Long-term expected cost following the policy } \pi}$$

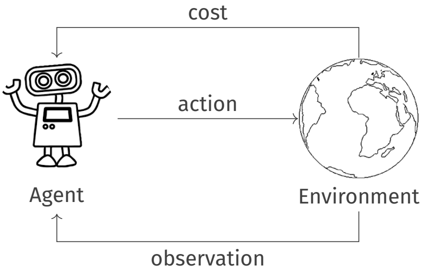
What does it mean *to solve* ?

Identifying the optimal policy!

$$\underbrace{V(\pi, s)}_{\text{Criterion to optimise}} = \underbrace{\mathbb{E}_s^\pi[\sum_{n=0}^{H-1} c(S_{n-1}, D_n, S_n)]}_{\text{Long-term expected cost following the policy } \pi}$$

$$\underbrace{V^\star(s)}_{\text{Value function}} = \underbrace{\min_{\pi \in \Pi} V(\pi, s)}_{\text{Minimising over the all set of policy } \Pi}$$

Reinforcement learning



Learning optimal policy from all experiences

$$\underbrace{Q^\pi(s,d)}_{\text{Criterion to optimise}} = \underbrace{\mathbb{E}^\pi[\sum_{n=0}^{H-1} c(S_{n-1}, D_n, S_n)|s, d = (\ell, r)]}_{\text{Value of an action in a state, following } \pi}$$

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

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

Model-free algorithms

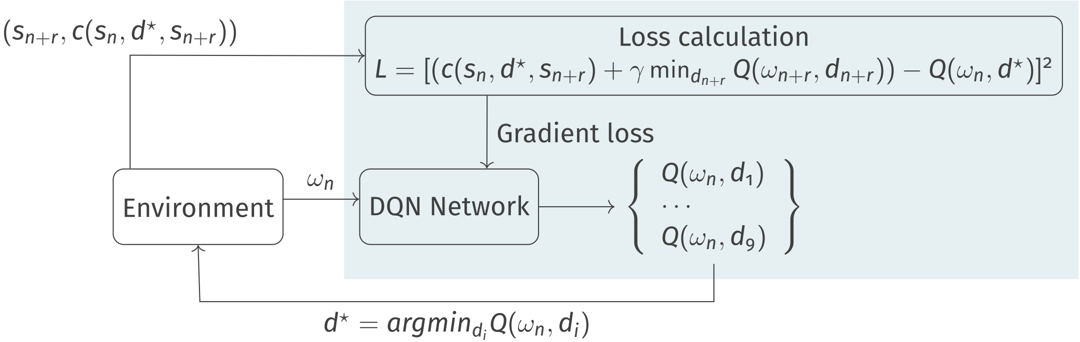


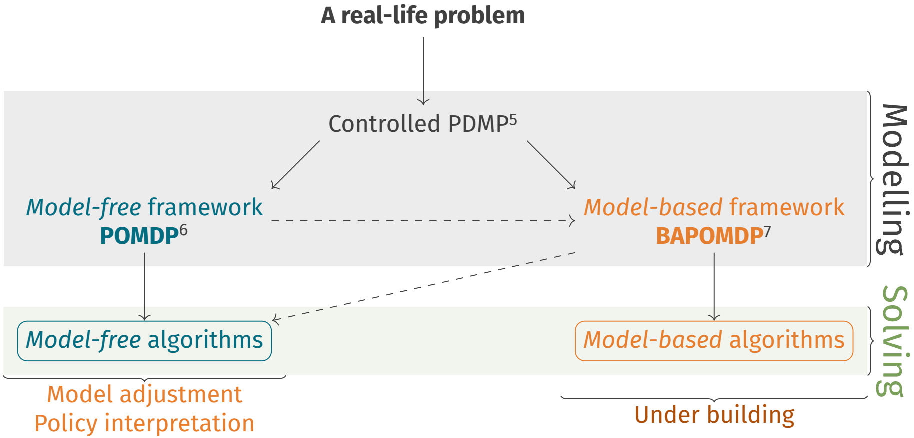
FIGURE: Deep Q-Network

Results

Policy	Mean Cost	Confidence interval	Survival rate	Relapse rate
OH	119.16	[119.15, 119.17]	95.71%	2.57
Random	158.09	[158.08, 158.10]	18.32%	1.27
Inactive	164.09	[164.08, 164.10]	0.14%	1.00
Threshold	157.27	[157.26, 157.28]	80.52%	1.01
DQN	121.52	[121.51, 121.53]	99.83%	0.59

Table: Policy evaluation performance on 10^5 simulations

Conclusion and Future Works



[5] Piecewise Deterministic Markov Process
[6] Partially Observable Markov Decision Process
[7] Bayes Adaptive Partially Observable Markov Decision Process