Deep Reinforcement Learning for Impulse Control in PDMPs through BAPOMDP framework

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

July 2025

Medical context

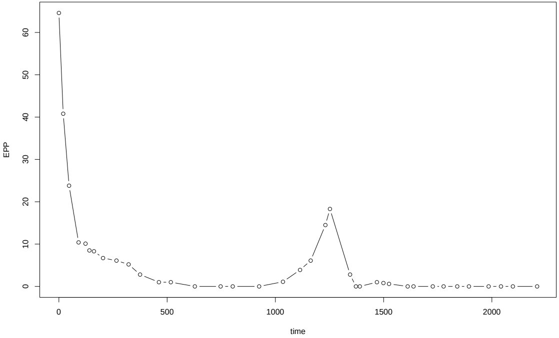


FIGURE: Example of patient data[a]

- Patients who have had cancer benefit from regular follow-up;
- The concentration of clonal immunoglobulin is measured over time;
- The doctor has to make new decisions at each visit.

[a] IUCT Oncopole and CRCT, Toulouse, France

Medical context

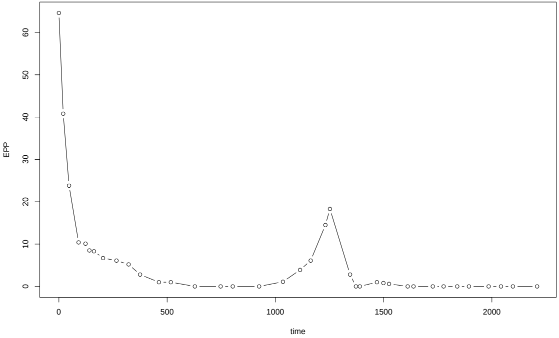


Figure: Example of patient data[a]

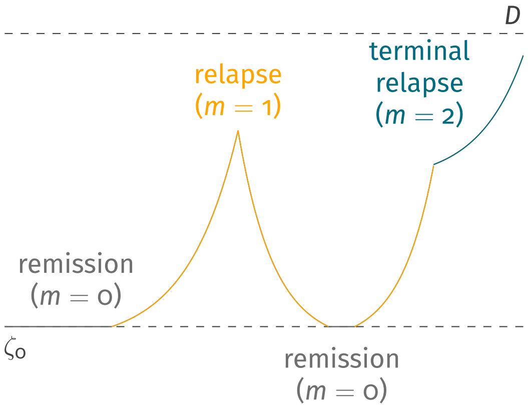
- Patients who have had cancer benefit from regular follow-up;
- The concentration of clonal immunoglobulin is measured over time;
- The doctor has to make new decisions at each visit.

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

[a] IUCT Oncopole and CRCT, Toulouse, France

Controlled PDMP[1]

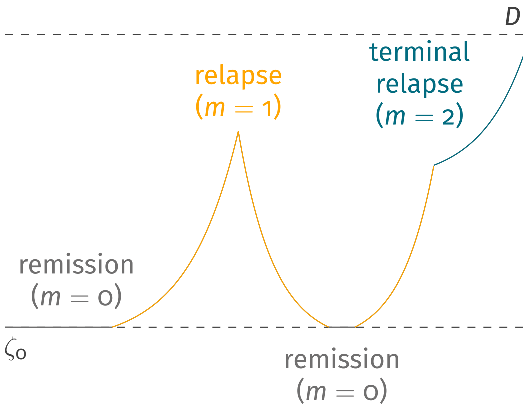
We switch randomly from one deterministic regime to another.



[1]Piecewise Deterministic Markov Processes

Controlled PDMP[1]

We switch randomly from one deterministic regime to another.



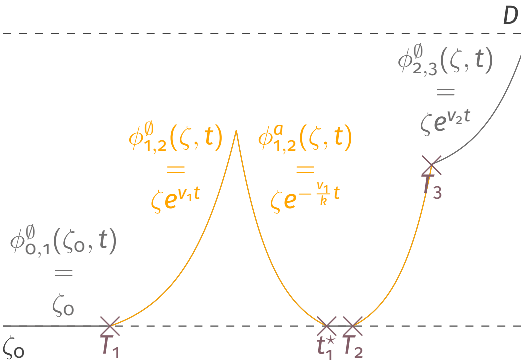
Let $x = (m, \ell, k, \zeta, u)$ the patient's condition:

- m the patient's condition;
- ℓ the current treatment;
- k the number of treatments;
- ζ the biomarker;
- u the time since the last jump.

[1] Piecewise Deterministic Markov Processes

Local Characteristics of a PDMP[2]

A PDMP is defined by three local characteristics.



FLOW

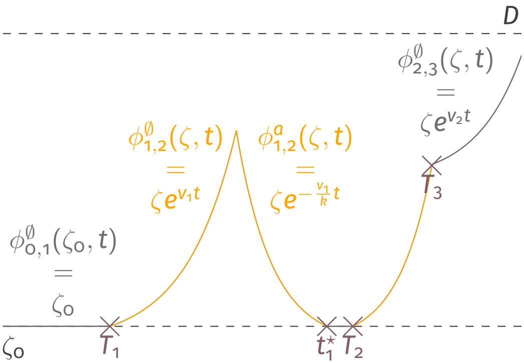
Description of the deterministic part of the process.

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

[2] Piecewise Deterministic Markov Processes

Local Characteristics of a PDMP[2]

A PDMP is defined by three local characteristics.



Jump intensity

Description of the process jump mechanisms.

- Boundary jump (deterministic)

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

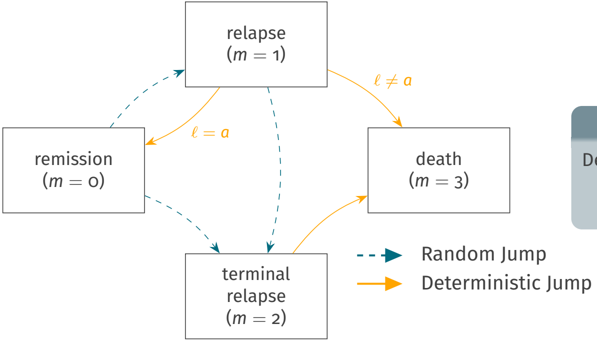
- Random jump

$$\mathbb{P}(T > t) = e^{-\int_0^t \lambda^{\ell}_{m,k}(\Phi(x,s))\,\mathrm{d}s}$$

[2]Piecewise Deterministic Markov Processes

Local Characteristics of a PDMP[2]

A PDMP is defined by three local characteristics.



MARKOV KERNEL

Description of the state of the process after each jump.

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

[2] Piecewise Deterministic Markov Processes

Solving impulse control for PDMP[3]

Identify an ϵ-optimal strategy $\mathcal{S} = (\tau_n, \chi_n)_{n\geq 1}$

$$\underbrace{\mathcal{V}(\mathcal{S}, x)}_{\text{Expected cost of strategy}\mathcal{S}} = \mathbb{E}_x^{\mathcal{S}}\left[\int_0^{+\infty} e^{-\gamma t} \underbrace{c_R(X_t)}_{\text{current trajectory cost}} dt + \sum_{n=1}^{\infty} \underbrace{c_I}_{\text{impulse cost}} (X_{\tau_n}, X_{\tau_n^+})\right],$$

[3]Piecewise Deterministic Markov Processes

Solving impulse control for PDMP[3]

Identify an ϵ-optimal strategy $\mathcal{S} = (\tau_n, \chi_n)_{n\geq 1}$

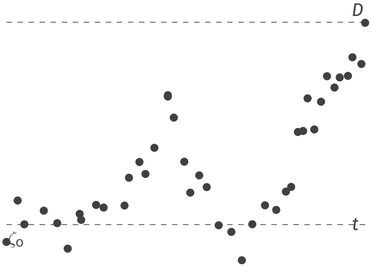
$$\underbrace{\mathcal{V}(\mathcal{S}, x)}_{\text{Expected cost of strategy}\mathcal{S}} = \mathbb{E}_x^{\mathcal{S}}\left[\int_0^{+\infty} e^{-\gamma t} \underbrace{c_R(X_t)}_{\text{current trajectory cost}} dt + \sum_{n=1}^{\infty} \underbrace{c_I}_{\text{impulse cost}} (X_{\tau_n}, X_{\tau_n^+})\right],$$

$$\mathcal{V}^\star(x) = \inf_{\mathcal{S}\in S} \mathcal{V}(\mathcal{S}, x)$$

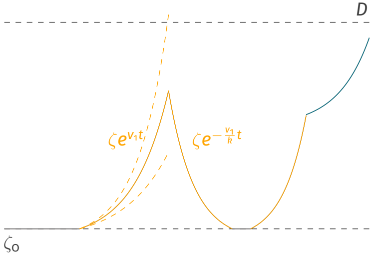
[3]Piecewise Deterministic Markov Processes

Difficulties

Partial observation



Partially known dynamics



Hypothesis: $v_1 \sim$ Log-Normal (μ, σ^{-2}), with μ and σ unknown.

Methods

Simplified real-life problem

continuous time
continuous state space
partially observed
partially known dynamics
simulable

Controlled PDMP[4]

Modelling

Deep Reinforcement Learning

Data consuming

[4] Piecewise Deterministic Markov Processes

Methods

Simplified real-life problem

↓

Controlled PDMP[5]

↓

BAPOMDP[6]

Modelling

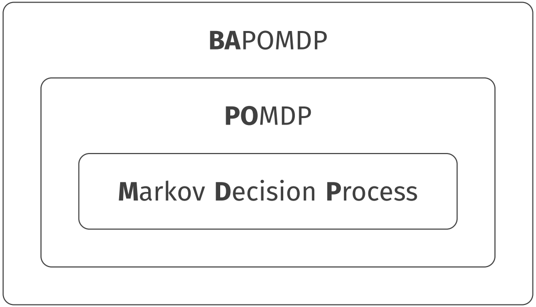
discrete observation dates
continuous state space
partially observed
known dynamics
simulable

↓

Deep Reinforcement Learning

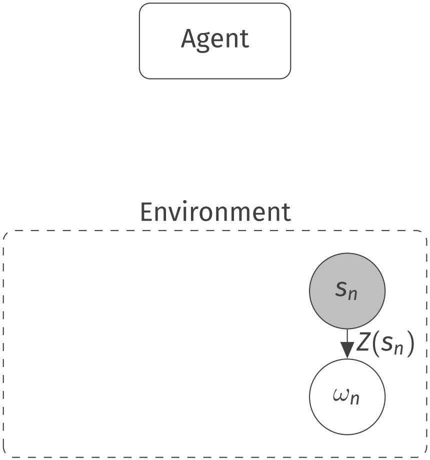
[5] Piecewise Deterministic Markov Processes
[6] Bayes-Adaptive Partially Observed Markov Decision Process

Characteristics of a MDP[7]



[7] Markov Decision Process

Characteristics of a POMDP[8]



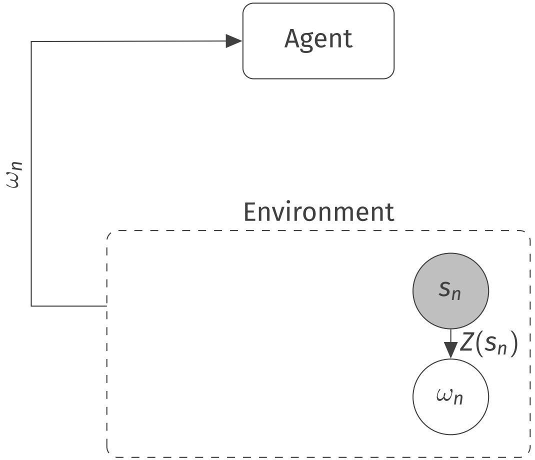
POMDP Definition

A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- Patient condition $s = (m, k, \zeta, u) \in \mathbb{S}$;
- Actions $a = (\ell, r) \in \mathbb{A}$;
- Transition function $P(s'|s, a)$;
- **Observation** $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- **Observation function** $Z(\omega|s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[8] Partially Observed Markov Decision Process

Characteristics of a POMDP[8]



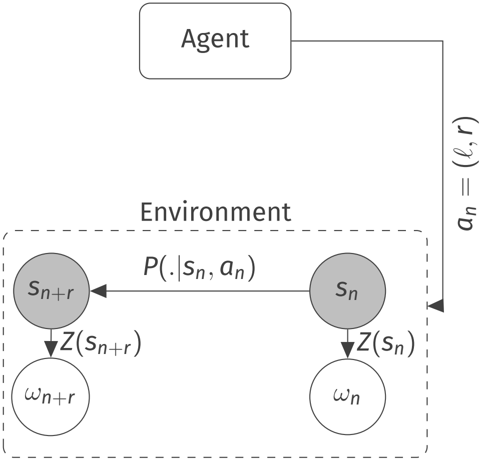
POMDP Definition

A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- Patient condition $s = (m, k, \zeta, u) \in \mathbb{S}$;
- Actions $a = (\ell, r) \in \mathbb{A}$;
- Transition function $P(s'|s, a)$;
- **Observation** $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- **Observation function** $Z(\omega|s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[8] Partially Observed Markov Decision Process

Characteristics of a POMDP[8]



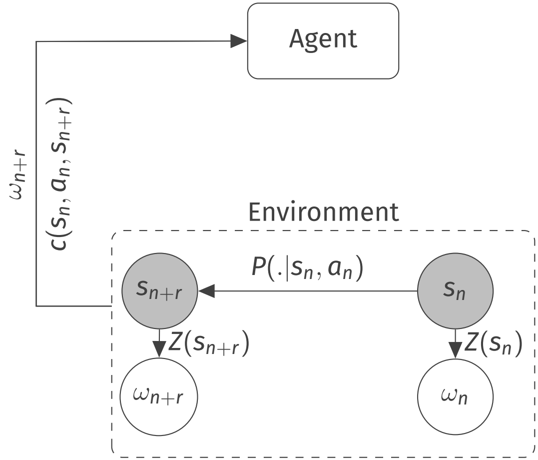
POMDP Definition

A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- Patient condition $s = (m, k, \zeta, u) \in \mathbb{S}$;
- Actions $a = (\ell, r) \in \mathbb{A}$;
- Transition function $P(s'|s, a)$;
- **Observation** $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$**;**
- **Observation function** $Z(\omega|s)$**;**
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[8] Partially Observed Markov Decision Process

Characteristics of a POMDP[8]



POMDP Definition

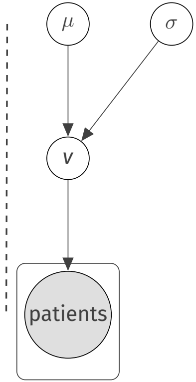
A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- Patient condition $s = (m, k, \zeta, u) \in \mathbb{S}$;
- Actions $a = (\ell, r) \in \mathbb{A}$;
- Transition function $P(s'|s, a)$;
- **Observation** $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- **Observation function** $Z(\omega|s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

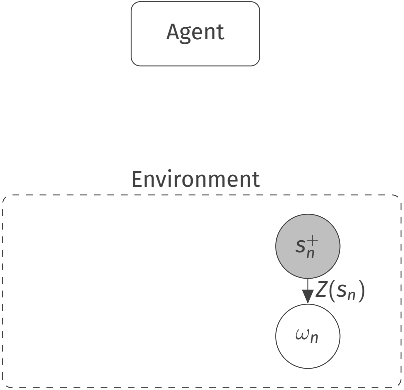
[8] Partially Observed Markov Decision Process

Handle uncertainty with Bayesian framework

Normal-Inverse-Gamma(Θ) prior



Characteristics of a BAPOMDP[9]



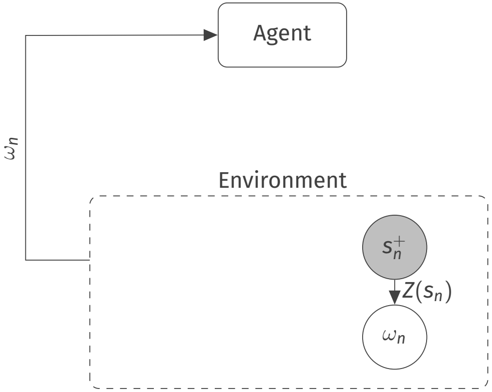
BAPOMDP Definition

Un BAPOMDP se définit par un tuple $(\mathbb{S}^+, \mathbb{A}, P^+, \Omega, Z, c)$.

- **Space of hyperstate** $\mathbb{S}^+ = \mathbb{S} \times \Theta$**;**
- Actions $a = (\ell, r) \in \mathbb{A}$;
- **Transition function** $P^+(s', \theta' | s, a, \theta)$**;**
- Observation $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Observation function $Z(\omega | s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[9] Bayes Adaptive Partially observed Markov decision process

Characteristics of a BAPOMDP[9]



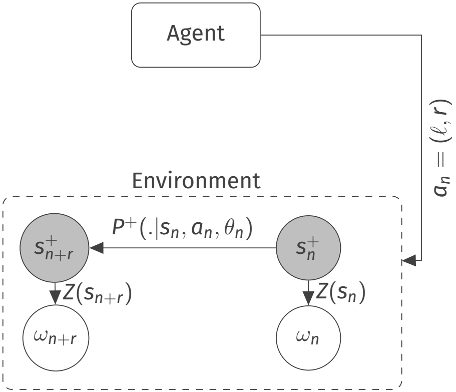
BAPOMDP Definition

Un BAPOMDP se définit par un tuple $(\mathbb{S}^+, \mathbb{A}, P^+, \Omega, Z, c)$.

- **Space of hyperstate** $\mathbb{S}^+ = \mathbb{S} \times \Theta$**;**
- Actions $a = (\ell, r) \in \mathbb{A}$;
- **Transition function** $P^+(s', \theta' | s, a, \theta)$**;**
- Observation $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Observation function $Z(\omega | s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[9] Bayes Adaptive Partially observed Markov decision process

Characteristics of a BAPOMDP[9]



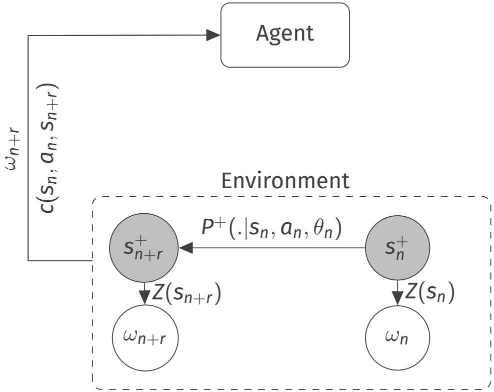
BAPOMDP Definition

Un BAPOMDP se définit par un tuple $(\mathbb{S}^+, \mathbb{A}, P^+, \Omega, Z, c)$.

- **Space of hyperstate** $\mathbb{S}^+ = \mathbb{S} \times \Theta$**;**
- Actions $a = (\ell, r) \in \mathbb{A}$;
- **Transition function** $P^+(s', \theta' | s, a, \theta)$**;**
- Observation $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Observation function $Z(\omega | s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[9] Bayes Adaptive Partially observed Markov decision process

Characteristics of a BAPOMDP[9]



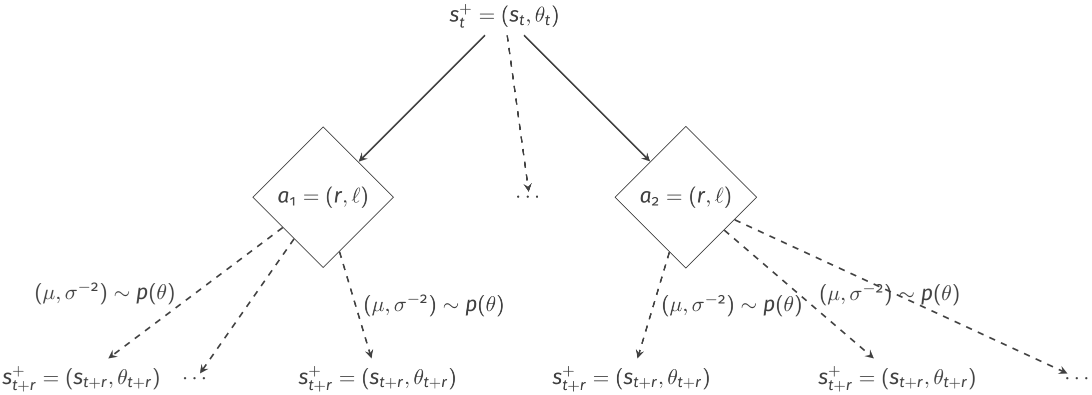
BAPOMDP Definition

Un BAPOMDP se définit par un tuple $(\mathbb{S}^+, \mathbb{A}, P^+, \Omega, Z, c)$.

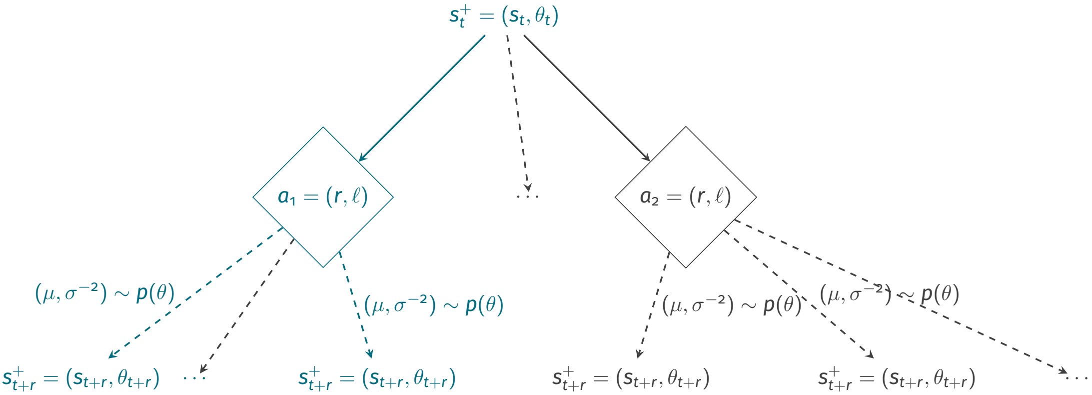
- **Space of hyperstate** $\mathbb{S}^+ = \mathbb{S} \times \Theta$**;**
- Actions $a = (\ell, r) \in \mathbb{A}$;
- **Transition function** $P^+(s', \theta' | s, a, \theta)$**;**
- Observation $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Observation function $Z(\omega|s)$;
- Cost function $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

[9] Bayes Adaptive Partially observed Markov decision process

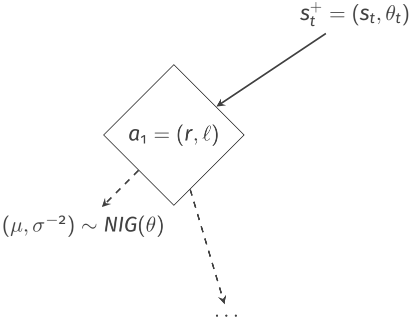
Generate transition from prior



Generate transition from prior



Generate transition from prior



Generate transition from prior

$s_t^+ = (s_t, \theta_t)$

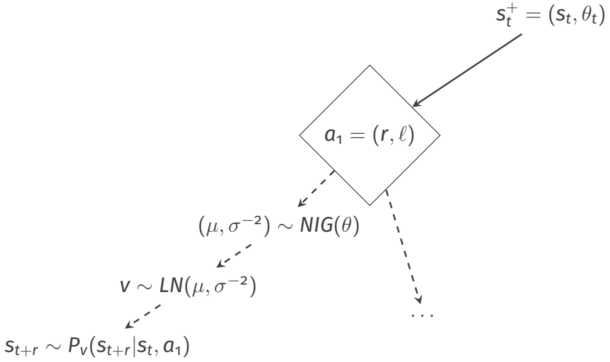
$a_1 = (r, \ell)$

$(\mu, \sigma^{-2}) \sim NIG(\theta)$

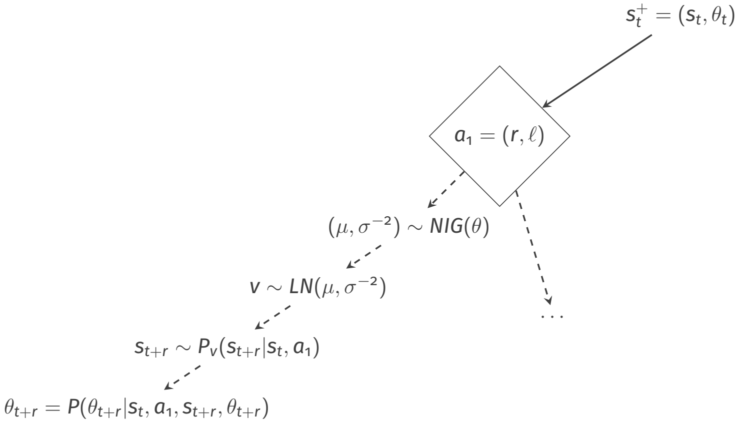
$v \sim LN(\mu, \sigma^{-2})$

$\ldots$

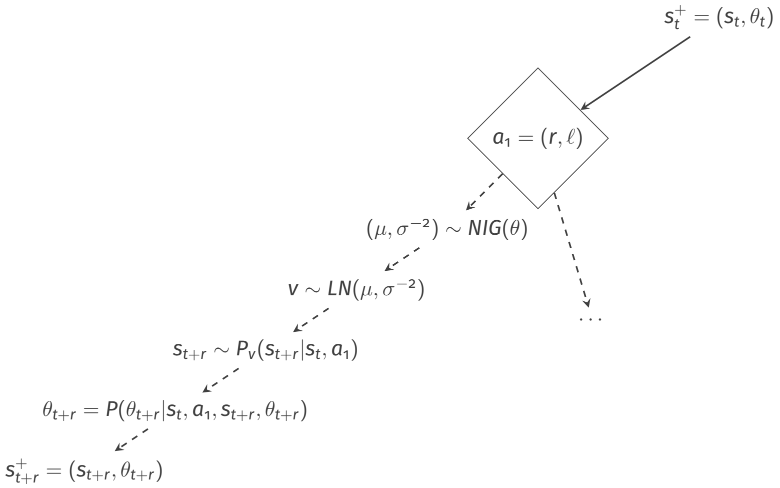
Generate transition from prior



Generate transition from prior



Generate transition from prior



Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

$$\underbrace{c(s,a,s')}_{\text{Cost function}} = \underbrace{C_V}_{\text{visit cost}} + \underbrace{C_D(H-t') \times \mathbb{1}_{m'=3}}_{\text{death cost}} + \underbrace{\kappa_C \times r \times \mathbb{1}_{\ell=a}}_{\text{treatment cost}}$$

[10] Bayes Adaptative Partially Observable Markov Decision Process

Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

$$\underbrace{V(\pi, s)}_{\text{Optimization criterion}} = \underbrace{\mathbb{E}_s^\pi\Big[\sum_{n=0}^{H-1} c(S_{n-1}, A_n, S_n)\Big]}_{\text{Expected long-term cost as a result of the policy } \pi}$$

[10] Bayes Adaptative Partially Observable Markov Decision Process

Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

$$\underbrace{V(\pi, s)}_{\text{Optimization criterion}} = \underbrace{\mathbb{E}_s^\pi\Big[\sum_{n=0}^{H-1} c(S_{n-1}, A_n, S_n)\Big]}_{\text{Expected long-term cost as a result of the policy } \pi}$$

$$\underbrace{V^\star(s)}_{\text{Value function}} = \underbrace{\min_{\pi \in \Pi} V(\pi, s)}_{\text{Minimisation across policy space}}$$

[10] Bayes Adaptative Partially Observable Markov Decision Process

Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

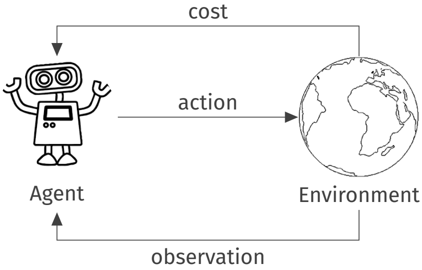
In reality, we do not observe state space!

Let $h_n = (\omega_0, a_0, \omega_1, a_1, \ldots, \omega_n)$ be the history

$$\underbrace{V^\star(h)}_{\text{Value function}} = \underbrace{\min_{\pi \in \Pi} V(\pi, h)}_{\text{Minimisation across policy space.}}$$

[10] Bayes Adaptative Partially Observable Markov Decision Process

Reinforcement Learning



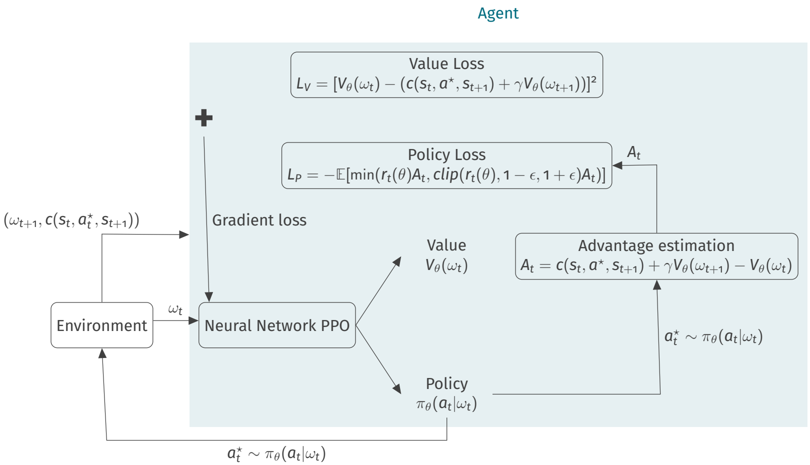
$$\underbrace{Q^{\pi}(s,a)}_{\text{Q value}} = \underbrace{\mathbb{E}^{\pi}[\sum_{n=0}^{H-1} c(S_{n-1}, A_n, S_n) | s, a = (\ell, r)]}_{\text{Value of an action in a state according to the policy } \pi}$$

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

The optimal policy is obtained from the experiments $< \omega, a, \omega', c >$, generate from P^{+} transition function

$$\underbrace{A(s,a)}_{\text{Advantage function}} = \underbrace{Q(s,a) - V(s)}_{\text{Extra cost obtained by the agent by taking the action}}$$

Algorithm example: PPO[11]



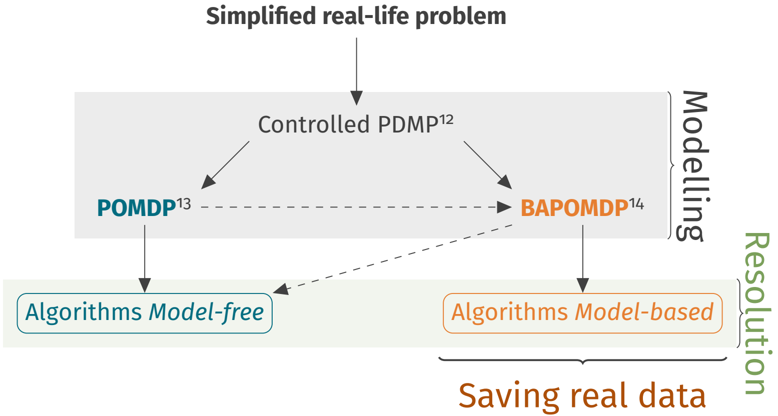
[11] Proximal policy optimization

Preliminary results

Policy	Mean cost (log)	CI	Death rate
OH	5.76	$[5.49, 6.03]$	58.69%
Real model	7.40	$[7.08, 7.72]$	99.66%
BAPOMDP model	7.46	$[7.14, 7.78]$	99.65%

Table: Policy evaluation performance on 10^5 simulations

Conclusion and future work



[12] Piecewise Deterministic Markov Processes
[13] Partially Observed Markov Decision Process
[14] Bayes Adaptative Partially Observed Markov Decision Process