Deep Reinforcement Learning for Bayes-Adaptive Impulse Control of PDMPs

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

October 2025

Medical context

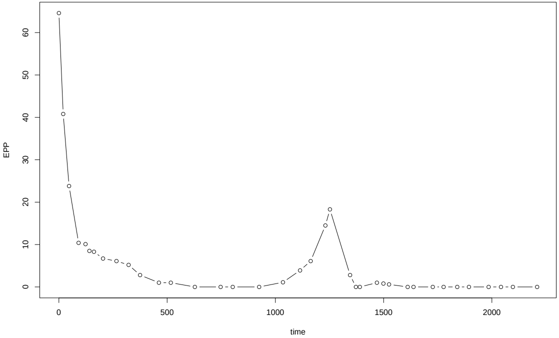


FIGURE: Example of patient data[a]

- 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

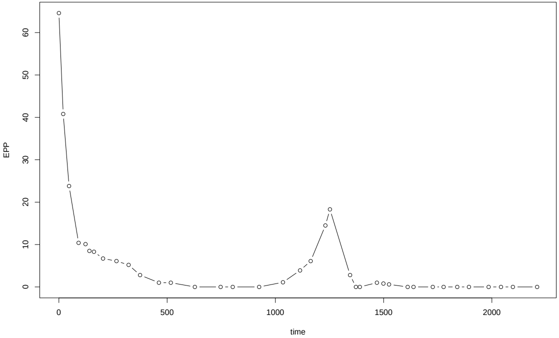


Figure: Example of patient data[a]

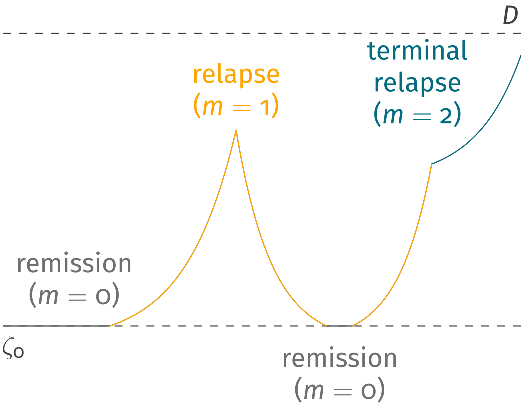
- Patients who have had cancer benefit from regular follow-up;
- The concentration of clonal immunoglobulin is measured over time;
- The doctor has to make new decisions at each visit.

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

[a] IUCT Oncopole and CRCT, Toulouse, France

Controlled PDMP[1]

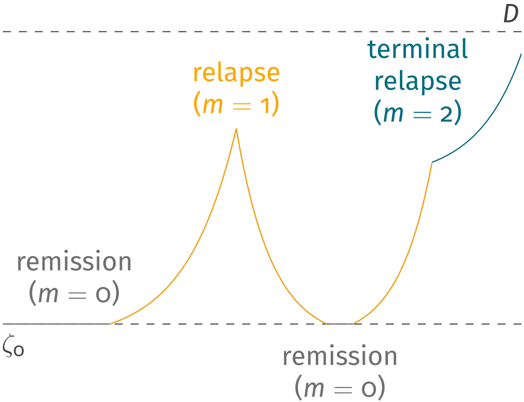
We switch randomly from one deterministic regime to another.



[1]Piecewise Deterministic Markov Processes

Controlled PDMP[1]

We switch randomly from one deterministic regime to another.



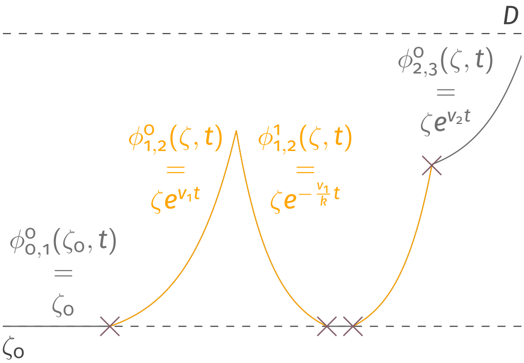
Let $x = (m, \ell, k, \zeta, u)$ the patient's condition:

- 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

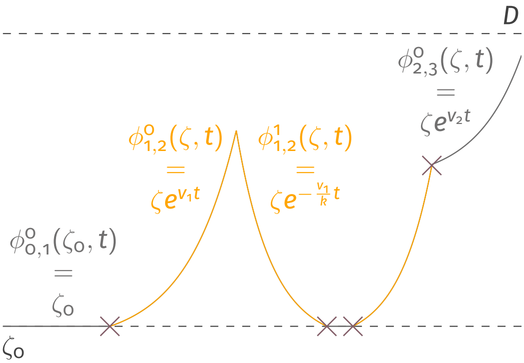
Description of the deterministic part of the process.

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

[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)

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

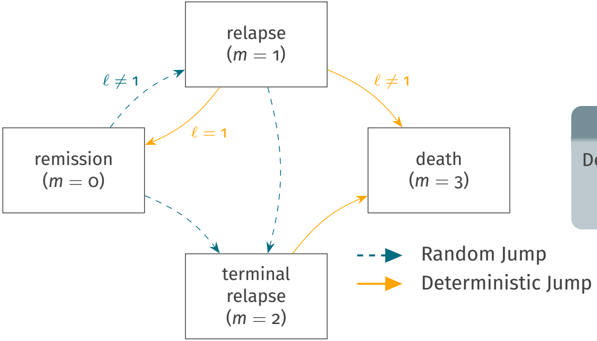
- Random jump

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

[2] Piecewise Deterministic Markov Processes

Local Characteristics of a PDMP[2]

A PDMP is defined by three local characteristics.



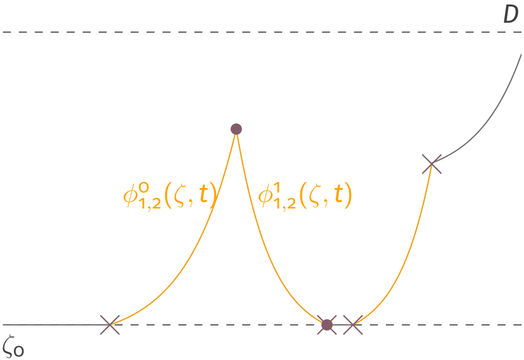
Markov kernel

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

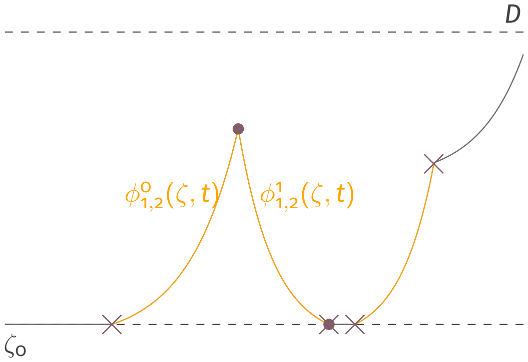
Impulse control for PDMP



Choosing a new starting point :

- date for the next impulse;
- point from which to restart the process.

Impulse control for PDMP



Choosing a new starting point :

- date for the next impulse;
- point from which to restart the process.

Restrictions:

- the delay between consecutive impulses is in a finite set;
- only change the current treatment ℓ.

Solving impulse control for PDMP

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

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

Solving impulse control for PDMP

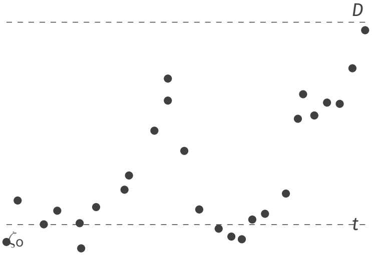
Identify an ϵ-optimal strategy $\mathcal{S} = (\tau_n, \chi_n)_{n \geq 1}$

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

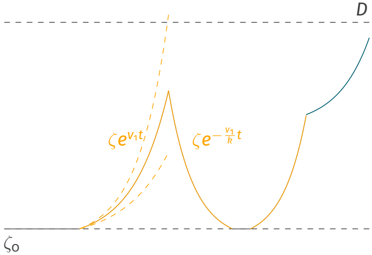
$$\mathcal{V}^\star(x) = \inf_{\mathcal{S} \in S} \mathcal{V}(\mathcal{S}, x)$$

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[3]

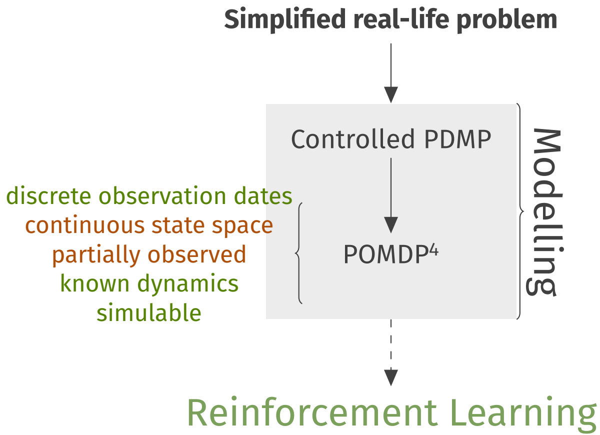
Modelling

Deep Reinforcement Learning

Data consuming

[3] Piecewise Deterministic Markov Processes

Previous work: Partial observation and known dynamics[5]



[4] Partially Observed Markov Decision Process

[5] de Saporta B, Thierry d'Argenlieu A, Sabbadin R, Cleynen A (2024) A Monte-Carlo planning strategy for medical follow-up optimization: Illustration on multiple myeloma data. PLOS ONE 19(12): e0315661

Methods

Simplified real-life problem

↓

Controlled PDMP

↓

BAPOMDP[6]

Modelling

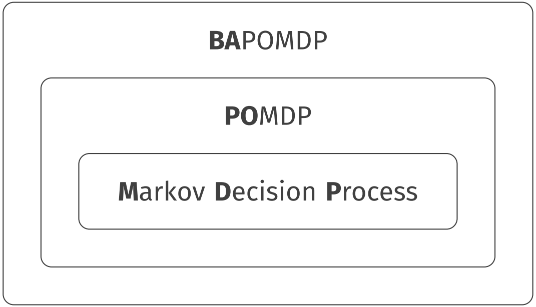
discrete observation dates
continuous state space
partially observed
known dynamics
simulable

↓

Deep Reinforcement Learning

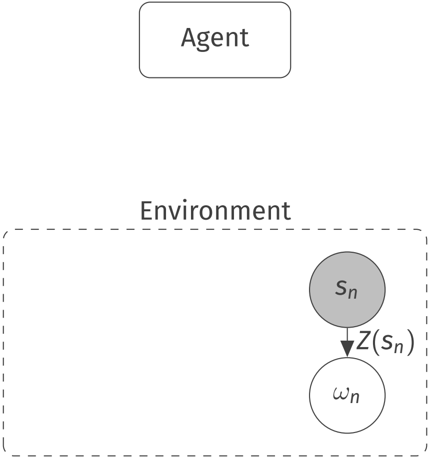
[6] Bayes-Adaptive Partially Observed Markov Decision Process

Characteristics of a MDP[7]



[7] Markov Decision Process

Characteristics of a POMDP[8]



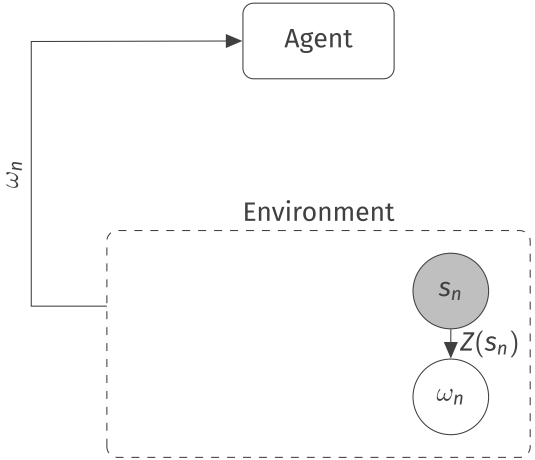
POMDP Definition

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]



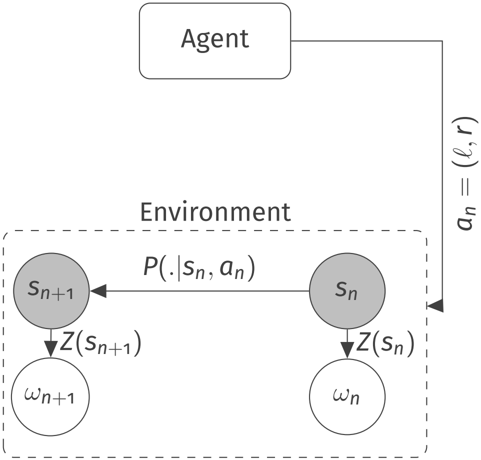
POMDP Definition

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

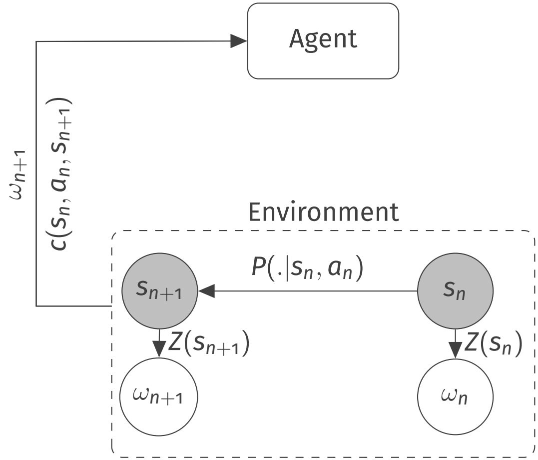
A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- 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 $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

The transition function $P(s'|s, a)$ is a combination of PDMP local characteristics.

[8] Partially Observed Markov Decision Process

Characteristics of a POMDP[8]



POMDP Definition

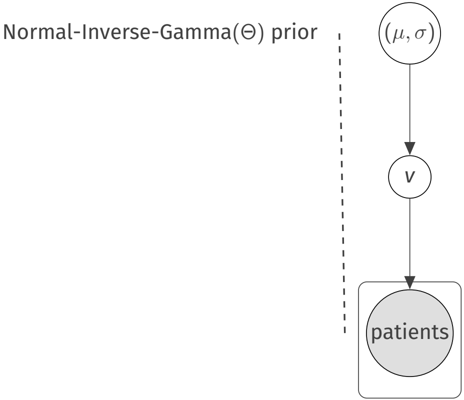
A POMDP is defined by a tuple $(\mathbb{S}, \mathbb{A}, P, \Omega, Z, c)$.

- 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 $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

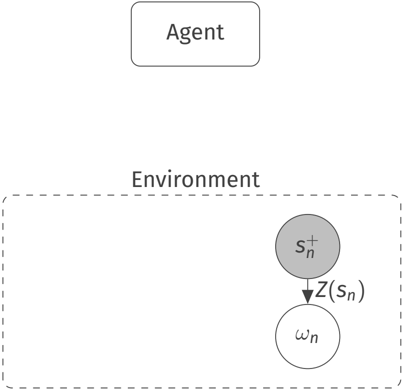
The transition function $P(s'|s, a)$ is a combination of PDMP local characteristics.

[8] Partially Observed Markov Decision Process

Handle uncertainty with Bayesian framework



Characteristics of a BAPOMDP[9]



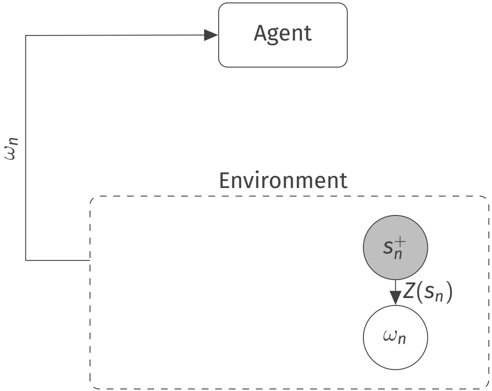
BAPOMDP Definition

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]



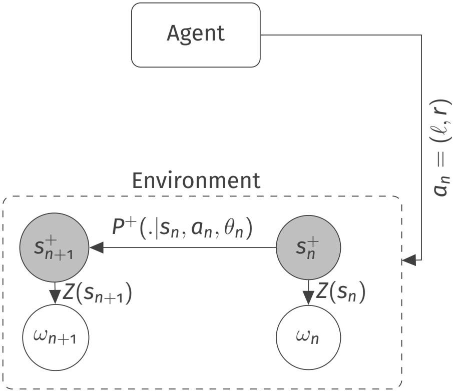
BAPOMDP Definition

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 $(\mathbb{S}^+, \mathbb{A}, P^+, \Omega, Z, c)$.

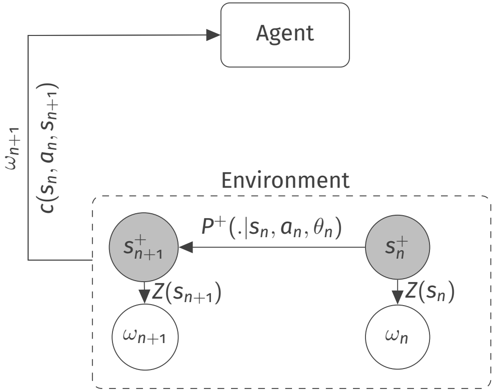
- **Space of hyperstate** $\mathbb{S}^+ = \mathbb{S} \times \Theta$**;**
- Actions $a = (\ell, r) \in \mathbb{A}$;
- **Transition function** $P^+(s', \theta' | s, a, \theta)$**;**
- Observation $\omega = (k, F(\zeta, \epsilon), \mathbb{1}_{m=3}) \in \Omega$;
- Observation function $Z(\omega|s)$;
- Cost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

$$P^+\big((s', \theta') \in B_E \times B_\Theta \mid (s, \theta), a\big)$$

$$= \int_{B_E} 1_{B_\Theta} \mathcal{U}(\theta, s, a, s') \times P(ds' \mid s, a, \theta).$$

[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 $c : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to \mathbb{R}$.

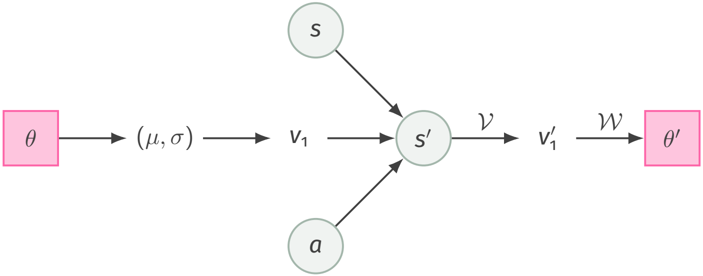
$$P^+\left((s', \theta') \in B_E \times B_\Theta \mid (s, \theta), a\right)$$

$$= \int_{B_E} 1_{B_\Theta} \mathcal{U}(\theta, s, a, s') \times P(ds' \mid s, a, \theta).$$

[9] Bayes Adaptive Partially observed Markov decision process

Generate transition from prior

$$\mathcal{U}(\theta, s, a, s') = \mathcal{W}(\theta, \mathcal{V}(s, a, s')),$$



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 $\pi^\star$

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

[10] Bayes Adaptative Partially Observable Markov Decision Process

Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

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

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

[10]Bayes Adaptative Partially Observable Markov Decision Process

Solving a BAPOMDP[10]

Identify an optimal policy $\pi^\star$

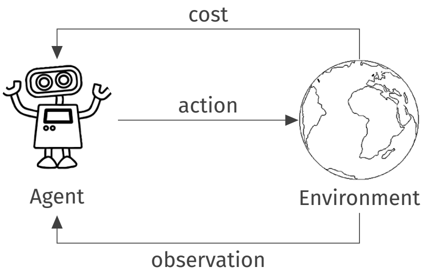
In reality, we do not observe state space!

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



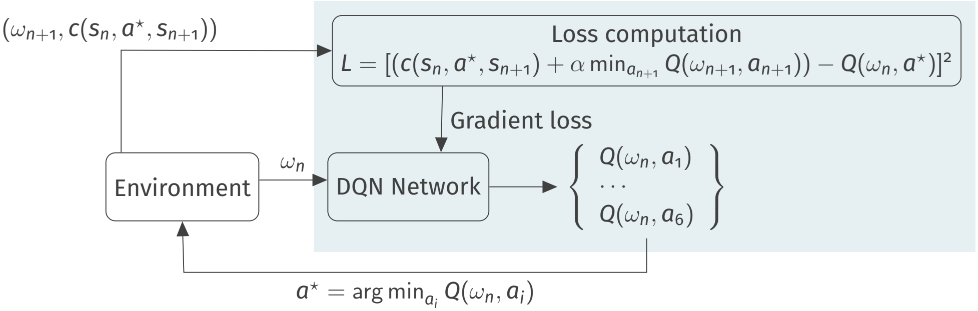
The optimal policy is obtained from the experiments $<\omega, a, \omega', c>$, generated from P^+ transition function

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

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

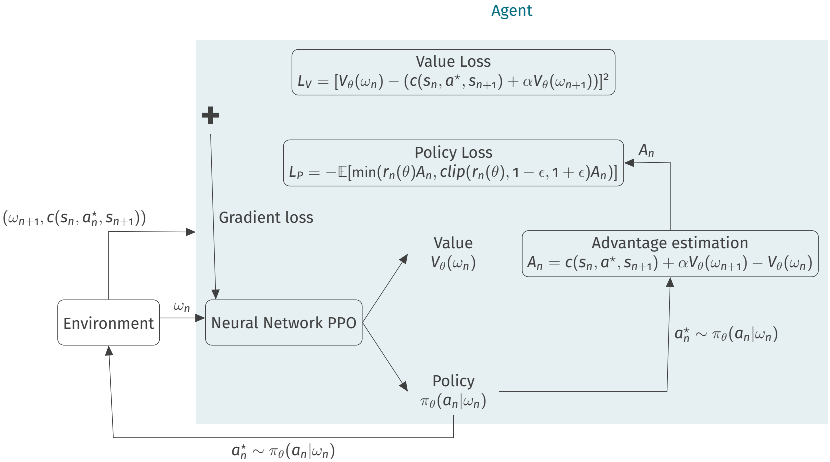
$$\underbrace{A(s,a)}_{\text{Advantage function}} = \underbrace{Q(s,a) - V(s)}_{\text{Extra cost obtained by the agent by taking the action}}$$

Algorithm example: DQN[11]



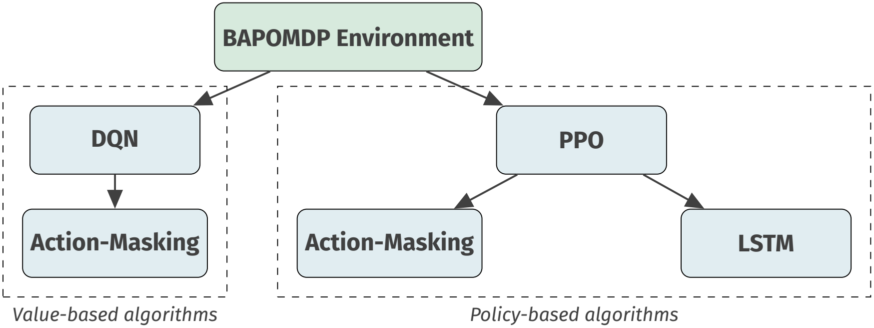
[11] Deep Q-Network

Algorithm example: PPO[12]

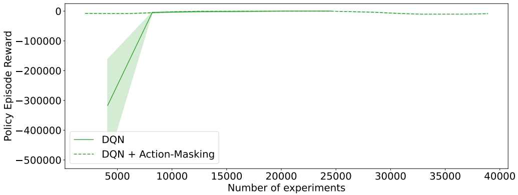


[12] Proximal policy optimization

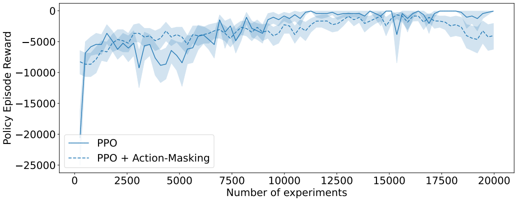
Algorithms Benchmark



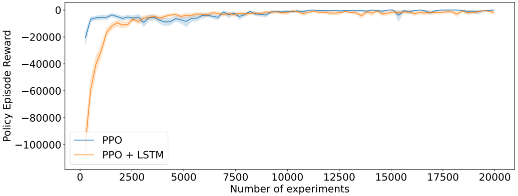
Results



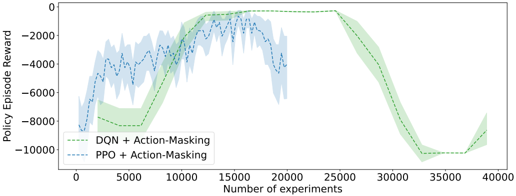
DQN
DQN + Action-Masking
Policy Episode Reward
Number of experiments



PPO
PPO + Action-Masking
Policy Episode Reward
Number of experiments

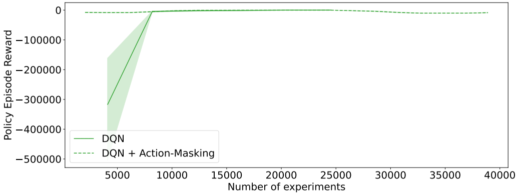


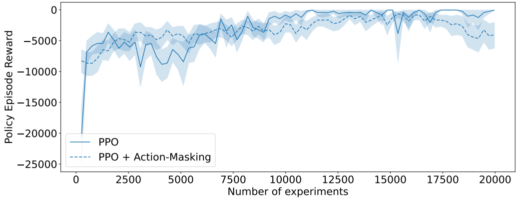
PPO
PPO + LSTM
Policy Episode Reward
Number of experiments

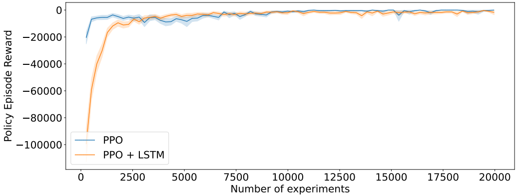


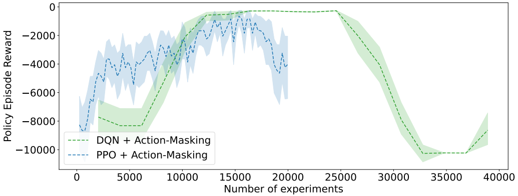
DQN + Action-Masking
PPO + Action-Masking
Policy Episode Reward
Number of experiments

Results



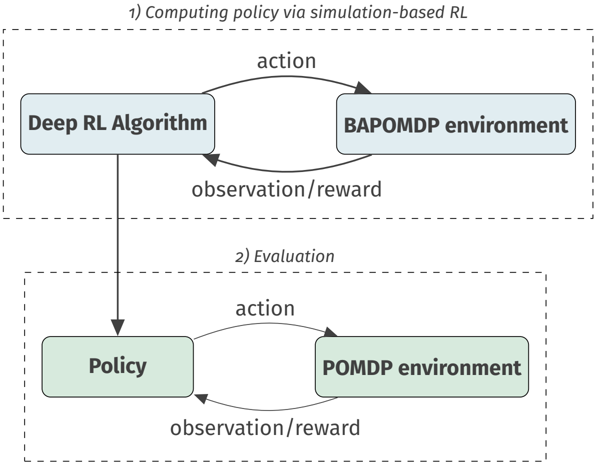






$\Longrightarrow$ DQN with Action-Masking outperformed all baseline algorithms.

Evaluate BAPOMDP framework



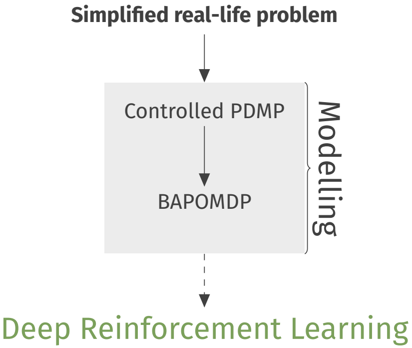
Prior	$(\mu_0, \kappa_0, \alpha_0, \beta_0)$
θ_{weak}	$(1,\ 0.001,\ 1.01,\ 1)$
θ_{medium}	$(-6.785,\ 5.001,\ 3.51,\ 1)$
θ_{high}	$(-6.23,\ 10,\ 6.01,\ 1)$

Evaluate BAPOMDP framework

What would happen if we modeled the unknown parameter v_1 as a fixed Normal–Gamma random variable in a POMDP?

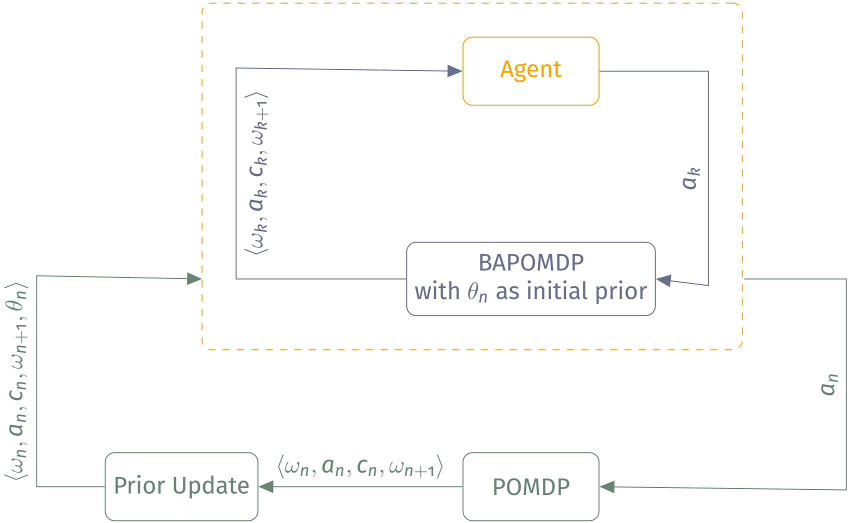
Prior	BAPOMDP	Non-adaptive POMDP
θ_{weak}	-5.74 ± 0.00	-5.70 ± 0.00
θ_{medium}	-5.66 ± 0.00	-5.97 ± 0.00
θ_{high}	-5.76 ± 0.00	-5.45 ± 0.00

Conclusion



- Bayes-adaptive method to address the PDMP control problem
- Comparable test-time performance to non-adaptive models
- No estimates of unknown parameters

Future work



Agent
$\langle \omega_k, a_k, c_k, \omega_{k+1} \rangle$
a_k
BAPOMDP
with θ_n as initial prior
$\langle \omega_n, a_n, c_n, \omega_{n+1}, \theta_n \rangle$
a_n
Prior Update
$\langle \omega_n, a_n, c_n, \omega_{n+1} \rangle$
POMDP

Policy behavior indicators

Table: **Summary of policy behavior indicators based on** 5 000 **Monte-Carlo simulations.**

Indicator	PPO with AM	DQN with AM
Survival rates	99.80% ± 0.00	99.70% ± 0.00
Average number of treatment	19.99 ± 0.00	19.99 ± 0.01
Average time spend under treatment	1199.63 ± 00.04	1199.56 ± 0.05
Average number of visit	58.99 ± 0.01	38.99 ± 0.01
Average delay between two visits	40.00 ± 0.00	60.00 ± 0.00
Rate of visits occurring within 15 days	0.01 ± 0.00	0.00 ± 0.00
Rate of visits occurring within 30 days	66.66 ± 0.17	0.00 ± 0.00
Rate of visits occurring within 60 days	33.33 ± 0.17	100 ± 0.00