An example of medical treatment optimization under model uncertainty

Orlane Rossini [1], Aymar Thierry d'Argenlieu [1], Alice Cleynen [1,2], Benoîte de Saporta [1] and 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

September 7, 2023

Contents

A medical context

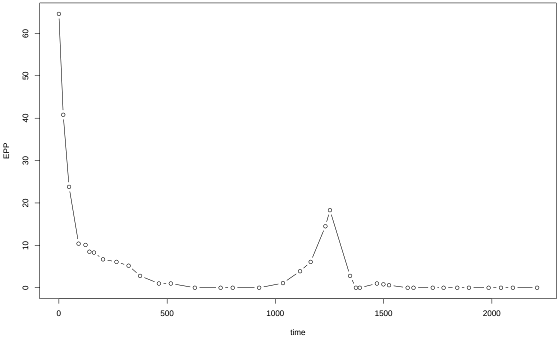
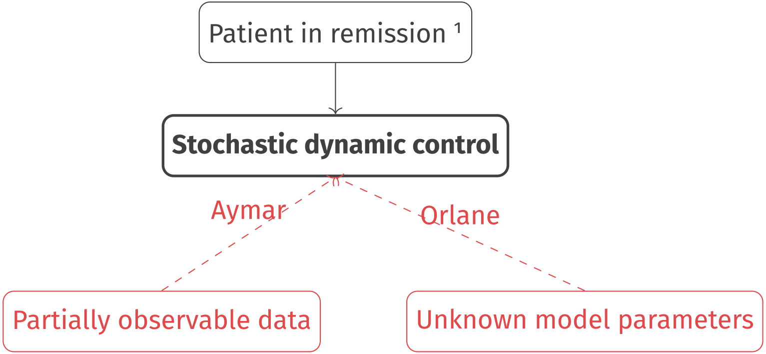


Figure: Patient Data[a]

- Patients who have had cancer are regularly monitored;
- Clonal immunoglobulin concentration is monitored over time;
- The doctor has to make new decisions at each visit.

[a] Data from IUC Oncopole, Toulouse, and CRCT, Toulouse, France

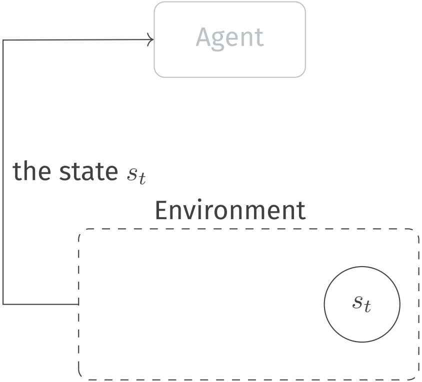
A medical context



[1] Data from IUC Oncopole, Toulouse, and CRCT, Toulouse, France

Contents

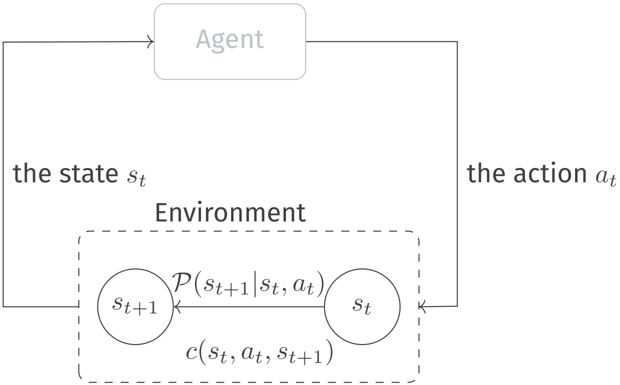
Markov Decision Process (MDP[2])



- $s \in \mathcal{S}$ the state space
- $a \in \mathcal{A}$ the action space
- $\mathcal{P}$ the transition matrix
- $c(s_t, a_t)$ the cost function

[2] ML Puterman (1994). "Finite-horizon Markov decision processes". In: Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley-Interscience, pp. 78–9.

Markov Decision Process (MDP[2])



- $s \in \mathcal{S}$ the state space
- $a \in \mathcal{A}$ the action space
- $\mathcal{P}$ the transition matrix
- $c(s_t, a_t)$ the cost function

[2] ML Puterman (1994). “Finite-horizon Markov decision processes”. In: Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley-Interscience, pp. 78–9.

Solving a MDP

Minimizing a cost

Policy π

Let $f : \mathcal{S} \to \mathcal{A}$ for all $s \in \mathcal{S}$ is a decision rule.
A sequence of decision rules $\pi = (f_0, f_1, \dots, f_{H-1})$ is a policy.
Let Π be the set of all eligible policies.

Policy cost and value function

$$J_\pi(s_0) = \mathbb{E}[\sum_{t=0}^{H-1} c(S_t, A_t)|\pi(S_t), S_0 = s_0]$$

Let $\pi^\star$ the optimal policy such that:

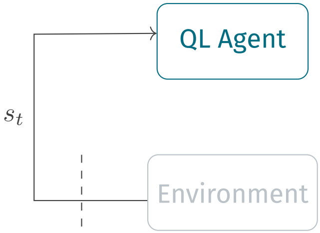
$$V(s_0) = J_{\pi^\star}(s_0) = \min_{\pi \in \Pi} J_\pi(s_0)$$

Optimization criterion

$$V^\star(s_t) = \min_{a_t \in \mathcal{A}} [c(s_t, a_t) + \sum_{s_{t+1} \in \mathcal{S}} \mathcal{P}(s_{t+1}|s_t, a_t) V^\star(s_{t+1})]$$

A model-free method

Q-learning[3,4] algorithm

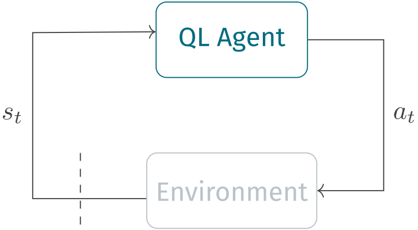


[3] Christopher J. C. H. Watkins and Peter Dayan (May 1992). "Q-learning". In: Mach. Learn. 8.3, pp. 279–292. ISSN: 1573-0565. DOI: 10.1007/BF00992698.

[4] VP Vivek and Dr. Shalabh Bhatnagar (Aug. 2022). "Finite Horizon Q-learning: Stability, Convergence, Simulations and an application on Smart Grids". In: arXiv:2110.15093v3. DOI: 10.48550/arXiv.2110.15093. eprint: 2110.15093v3.

A model-free method

Q-learning[3,4] algorithm



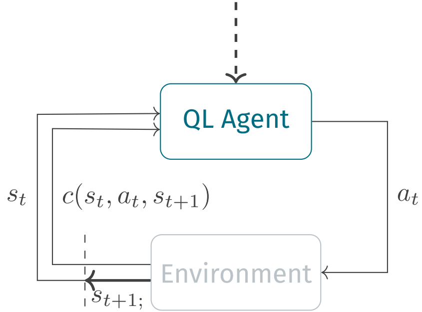
[3] Christopher J. C. H. Watkins and Peter Dayan (May 1992). "Q-learning". In: Mach. Learn. 8.3, pp. 279–292. ISSN: 1573-0565. DOI: 10.1007/BF00992698.

[4] VP Vivek and Dr. Shalabh Bhatnagar (Aug. 2022). "Finite Horizon Q-learning: Stability, Convergence, Simulations and an application on Smart Grids". In: arXiv:2110.15093v3. DOI: 10.48550/arXiv.2110.15093. eprint: 2110.15093v3.

A model-free method

Q-learning[3,4] algorithm

$$Q_n(s_t, a_t) = (1-\alpha)Q_{n-1}(s_t, a_t) + \alpha[c(s_t, a_t) + \min_{a_{t+1}\in\mathcal{A}} Q_{n-1}(s_{t+1}, a_{t+1})]$$

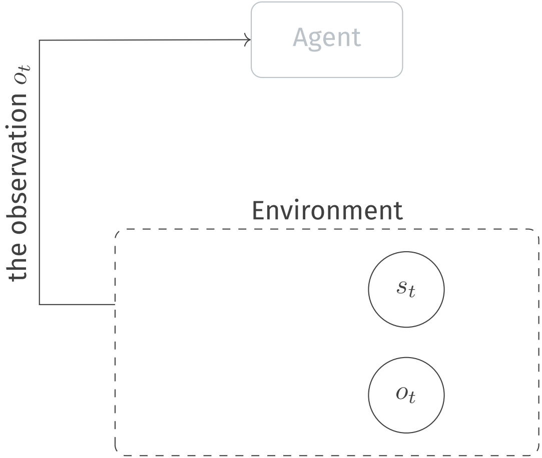


[3] Christopher J. C. H. Watkins and Peter Dayan (May 1992). "Q-learning". In: Mach. Learn. 8.3, pp. 279–292. ISSN: 1573-0565. DOI: 10.1007/BF00992698.

[4] VP Vivek and Dr. Shalabh Bhatnagar (Aug. 2022). "Finite Horizon Q-learning: Stability, Convergence, Simulations and an application on Smart Grids". In: arXiv:2110.15093v3. DOI: 10.48550/arXiv.2110.15093. eprint: 2110.15093v3.

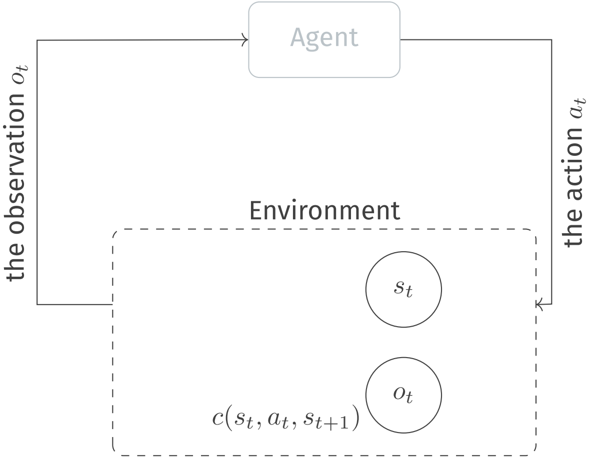
Contents

Partially observable Markov Decision Process (POMDP)



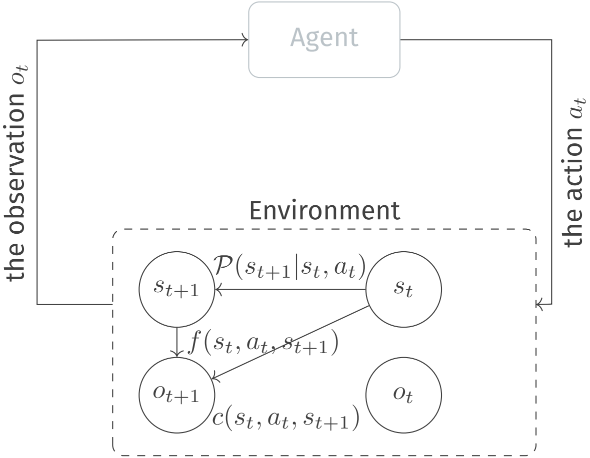
- $s \in \mathcal{S}$ the state space
- $o \in \mathcal{O}$ the observation space
- $a \in \mathcal{A}$ the action space
- $\mathcal{P}$ the transition matrix
- f a mesurable function
- $c(s_t, a_t, s_{t+1})$ the cost function

Partially observable Markov Decision Process (POMDP)



- $s \in \mathcal{S}$ the state space
- $o \in \mathcal{O}$ the observation space
- $a \in \mathcal{A}$ the action space
- $\mathcal{P}$ the transition matrix
- f a mesurable function
- $c(s_t, a_t, s_{t+1})$ the cost function

Partially observable Markov Decision Process (POMDP)



- $s \in \mathcal{S}$ the state space
- $o \in \mathcal{O}$ the observation space
- $a \in \mathcal{A}$ the action space
- $\mathcal{P}$ the transition matrix
- f a mesurable function
- $c(s_t, a_t, s_{t+1})$ the cost function

The *history* is defined as a sequence of actions and observations.

A history

$$h_t = \{o_0, a_0, o_1, a_1, \cdots, o_{t-1}, a_{t-1}, o_t\}$$

The *history* is defined as a sequence of actions and observations.

A history

$$h_t = \{o_0, a_0, o_1, a_1, \cdots, o_{t-1}, a_{t-1}, o_t\}$$

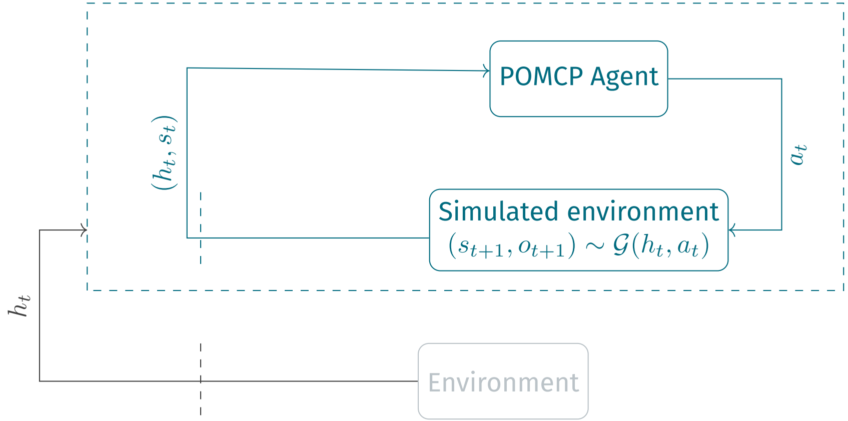
Optimization criterion

$$V^\star(h) = \min_{a_t \in \mathcal{A}} [c(s_t, a_t) + \sum_{o_{t+1} \in \mathcal{O}} \mathcal{P}(o_{t+1} | h_{t+1}, a_t) V^\star(h_{t+1})]$$

A model-based method

Partially Observable Monte-Carlo Planning (POMCP[5])

with $h_t = (o_0, a_0, o_1, \dots, o_{t-1}, a_{t-1}, o_t)$,

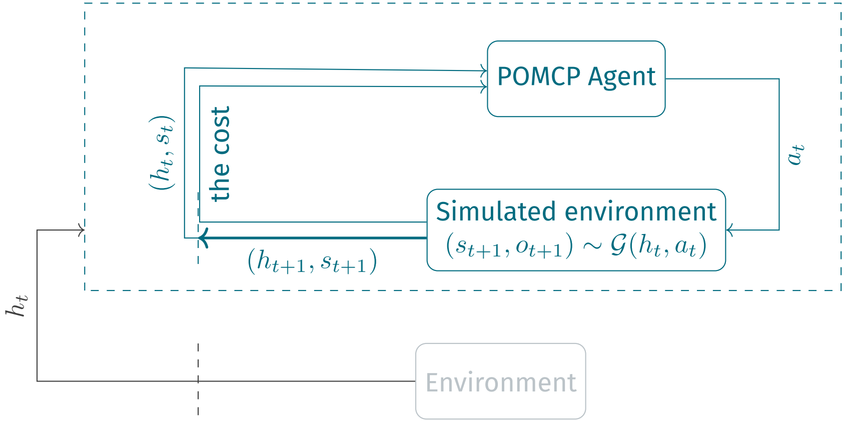


[5] David Silver and Joel Veness (2010). "Monte-Carlo Planning in Large POMDPs". In: Advances in Neural Information Processing Systems 23. URL: https://papers.nips.cc/paper_files/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html

A model-based method

Partially Observable Monte-Carlo Planning (POMCP[5])

with $h_t = (o_0, a_0, o_1, \ldots, o_{t-1}, a_{t-1}, o_t)$,

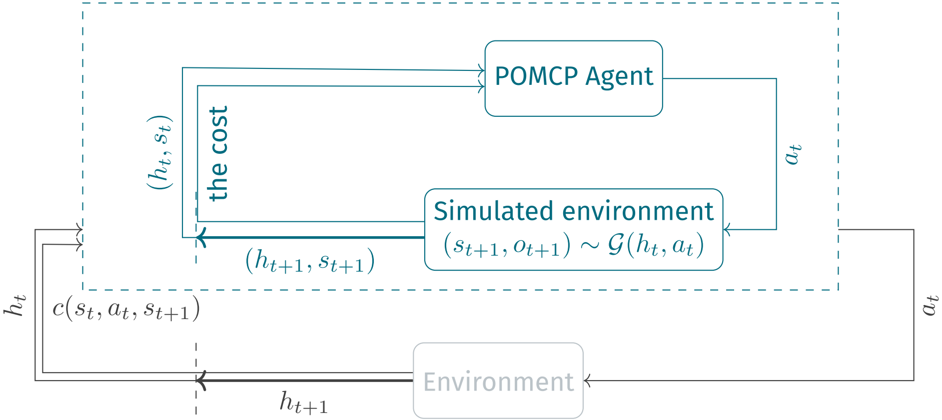


[5] David Silver and Joel Veness (2010). "Monte-Carlo Planning in Large POMDPs". In: Advances in Neural Information Processing Systems 23. URL: https://papers.nips.cc/paper_files/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html

A model-based method

Partially Observable Monte-Carlo Planning (POMCP[5])

with $h_t = (o_0, a_0, o_1, \ldots, o_{t-1}, a_{t-1}, o_t)$,



[5] David Silver and Joel Veness (2010). "Monte-Carlo Planning in Large POMDPs". In: Advances in Neural Information Processing Systems 23. URL: https://papers.nips.cc/paper_files/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html

Contents

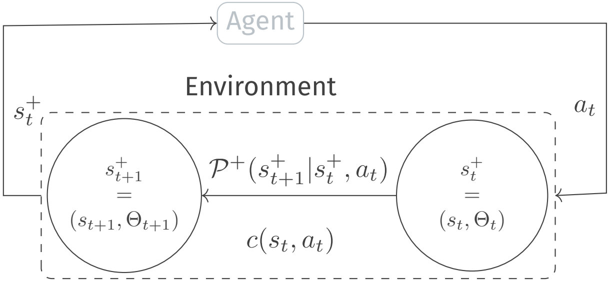
A bayesian approach

$s_t \backslash s_{t+1}$	$(0,0,0)$	$(1,0,1)$	$(1,0,2)$	$(1,1,1)$	$(1,1,2)$	$(1,2,1)$	$(1,2,2)$	$(1,3,1)$	$(1,3,2)$	$(2,4,0)$
$(0,0,0)$	$p^{\emptyset}_{(0,0,0)}$	$p^{\emptyset}_{(1,0,1)}$	$p^{\emptyset}_{(1,0,2)}$	0	0	0	0	0	0	0

Remark:

- $P(.|s=(0,0,0), a=\emptyset) \sim \mathcal{M}(p^{\emptyset}_{(0,0,0)}, p^{\emptyset}_{(1,0,1)}, p^{\emptyset}_{(1,0,2)})$
- Conjugate distribution : $f(p^{\emptyset}|\Theta^{\emptyset}) \sim \mathcal{D}(\theta^{\emptyset}_{(0,0,0)}, \theta^{\emptyset}_{(1,0,1)}, \theta^{\emptyset}_{(1,0,2)})$

Bayes-Adaptive Markov Decision Process (BAMDP[6])

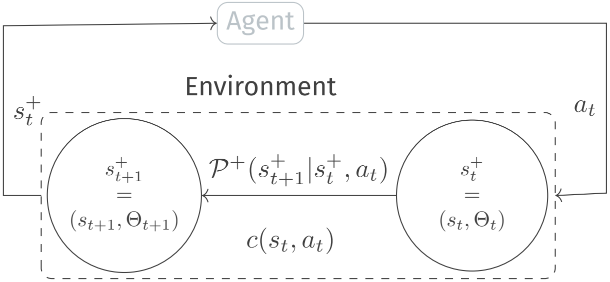


- $s^+ \in \mathcal{S}^+$ the hyper-state space
- $\mathcal{P}^+$ the transition matrix
- $\Theta_{t+1} = \Theta_t + \Delta_{s_{t+1}}^{a_t}$, with

$$\Delta_{s_{t+1}}^{a_t} = \begin{cases} 1 & \text{if } (s = (0, 0, 0), a_t, s_{t+1}), \\ 0 & \text{else.} \end{cases}$$

[6] Michael O'Gordon Duff (2002). "Optimal learning: Computational procedures for Bayes -adaptive Markov decision processes". PhD thesis. University of Massachusetts Amherst.

Bayes-Adaptive Markov Decision Process (BAMDP[6])



- $s^+ \in \mathcal{S}^+$ the hyper-state space
- $\mathcal{P}^+$ the transition matrix
- $\Theta_{t+1} = \Theta_t + \Delta^{a_t}_{s_{t+1}}$, with

$$\Delta^{a_t}_{s_{t+1}} = \begin{cases} 1 & \text{if } (s = (0, 0, 0), a_t, s_{t+1}), \\ 0 & \text{else.} \end{cases}$$

Optimization criterion

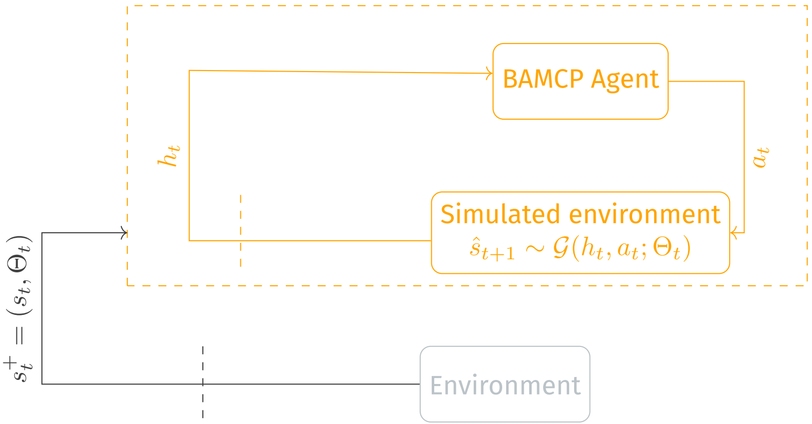
$$V^\star(s_t, \Theta_t) = \min_{a_t \in \mathcal{A}} [c(s_t, a_t) + \sum_{s_{t+1}^+ \in \mathcal{S}^+} \mathcal{P}^+(s_{t+1}^+ | s_t^+, a_t) V^\star(s_{t+1}, \Theta_{t+1})]$$

[6] Michael O'Gordon Duff (2002). "Optimal learning: Computational procedures for Bayes -adaptive Markov decision processes". PhD thesis. University of Massachusetts Amherst.

A model-based method

Bayes-Adaptive Monte-Carlo Planning (BAMCP[7])

with $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$,

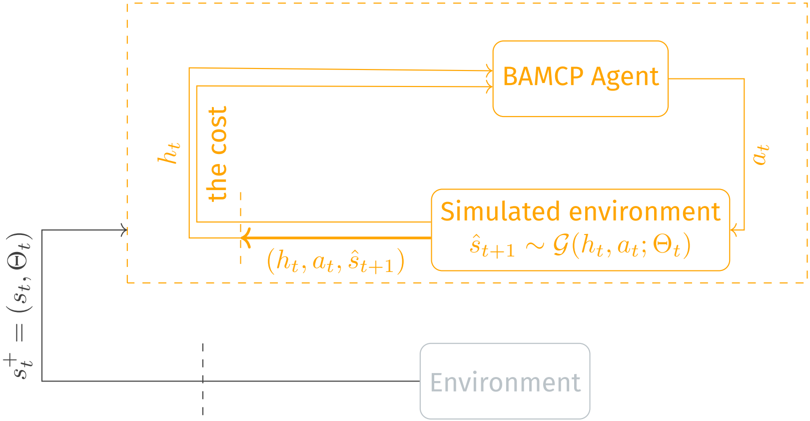


[7] Arthur Guez, David Silver, and Peter Dayan (2012). "Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search". In: Advances in Neural Information Processing Systems. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc.

A model-based method

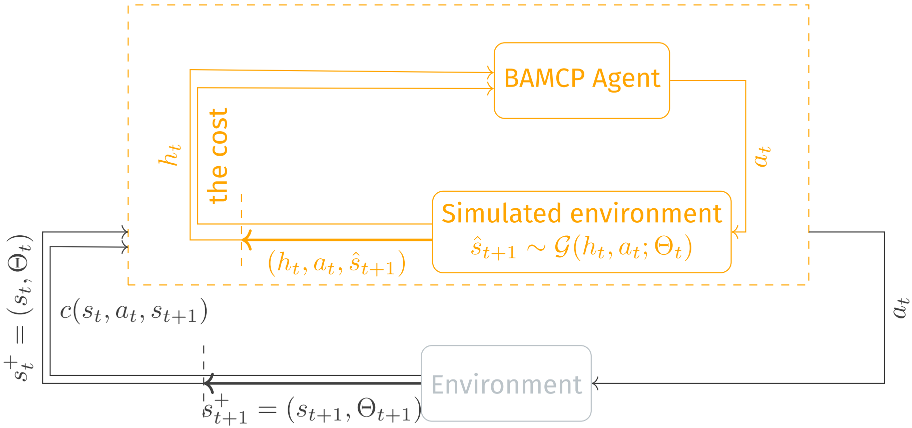
Bayes-Adaptive Monte-Carlo Planning (BAMCP[7])

with $h_t = (s_0, a_0, s_1, \dots, s_{t-1}, a_{t-1}, s_t)$,



[7] Arthur Guez, David Silver, and Peter Dayan (2012). "Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search". In: Advances in Neural Information Processing Systems. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc.

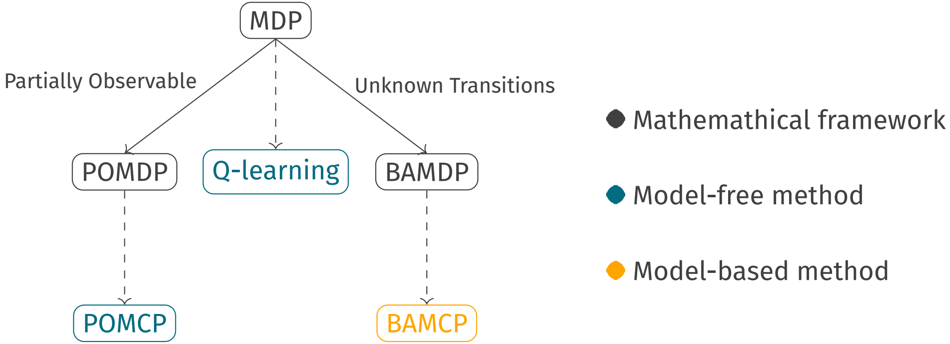
with $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$,



[7] Arthur Guez, David Silver, and Peter Dayan (2012). "Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search". In: Advances in Neural Information Processing Systems. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc.

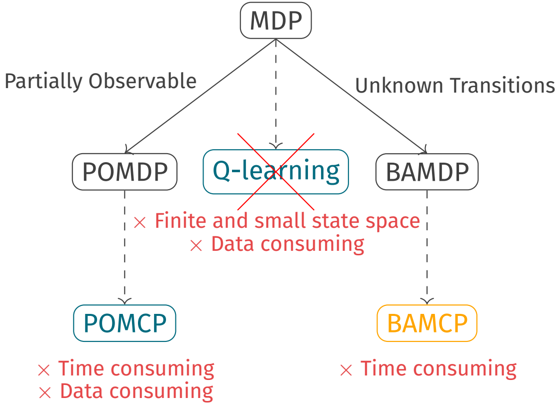
Contents

Conclusion



Unlike model-free methods and deep reinforcement learning, **model-based approaches** do not require as much interaction with the environment.

Conclusion



- Mathematical framework
- Model-free method
- Model-based method

Unlike model-free methods and deep reinforcement learning, **model-based approaches** do not require as much interaction with the environment.

Perspectives

A real-life problem

Modelling

Controled PDMP

× partially observable
× partially unkown model
× semi-Markov

Gymnasium

POMDP

× partially observable

BAMDP

× partially unkown model

BAPOMDP

× partially unkown model
× partially observable

MDP

× large state space
× continuous state space

Resolution

Exact resolution by DP is no longer possible.
Resolution by **simulations** must be applied.

Perspectives

Modelling

Controled PDMP

POMDP | BAMDP | BAPOMDP

MDP
× large state space
× continuous state space

Resolution

Exact resolution by DP is no longer possible.
Resolution by **simulations** must be applied.

POMCP	BAMCP	BAPOMCP	Deel RL
✓ convergence guarantees	✓ convergence guarantees	✓ convergence guarantees	× no convergence guarantees
× applicabilty limited ?	× applicabilty limited ?	× applicabilty limited ?	✓ always applicable
× data consuming	× data consuming	× data consuming	× data consuming
× on-line time consuming	× on-line time consuming	× on-line time consuming	✓ on-line time

Any questions ?