Stochastic control for medical treatment optimization

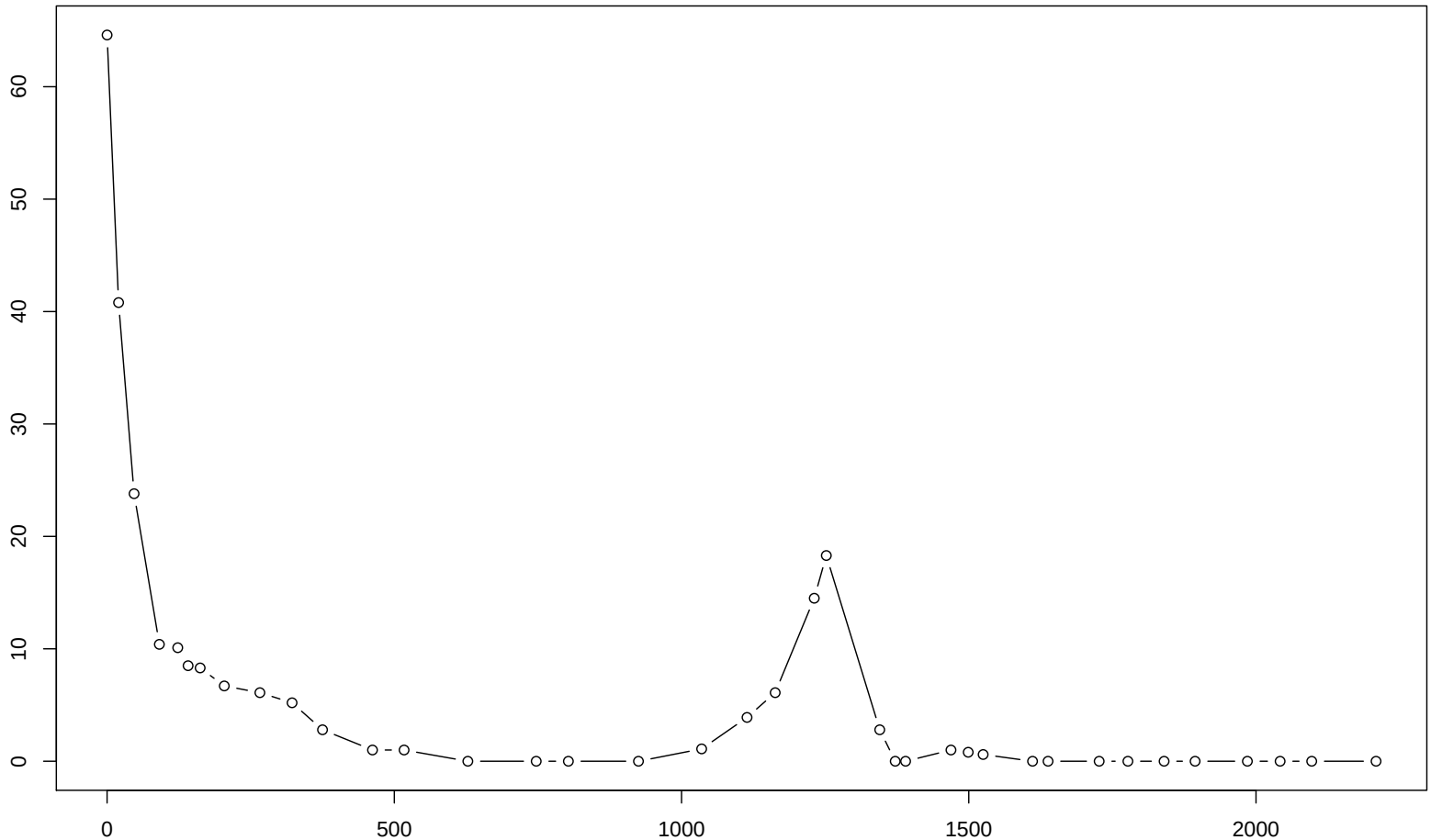
Alice Cleynen, Benoîte de Saporta

Institut Montpelliérain Alexander Grothendieck

CNRS, Univ. Montpellier, France

Motivation : Multiple myeloma

Clinical trial : Intergroupe Francophone du Myélome 2009



Measurement of monoclonal immunoglobulin as a function of time for one patient

Source : Centre de Recherche en Cancérologie de Toulouse

Medical treatment optimization

Use past and present monoclonal immunoglobulin measurements to sequentially choose

- the appropriate treatment to be applied until the next visit to the medical center
- the date of the next visit to the medical center

to improve the patient health status: maintain low levels of monoclonal immunoglobulin while minimizing undesirable side effects and treatment constraints/costs

Difficulties

- relapse date detection: the overall health status is random, not directly observable: use monoclonal immunoglobulin as a marker of the disease, continuous evolution of the marker, but discrete observations dates with low frequency and observations possibly corrupted by noise

Difficulties

- relapse date detection: the overall health status is random, not directly observable: use monoclonal immunoglobulin as a marker of the disease, continuous evolution of the marker, but discrete observations dates with low frequency and observations possibly corrupted by noise
- relapse type detection / treatment choice: several relapse types and treatment types

Difficulties

- relapse date detection: the overall health status is random, not directly observable: use monoclonal immunoglobulin as a marker of the disease, continuous evolution of the marker, but discrete observations dates with low frequency and observations possibly corrupted by noise
- relapse type detection / treatment choice: several relapse types and treatment types
- choice of the next visit date / treatment duration: non trivial compromise between
  - too early heavy treatments with severe side effects
  - too late increased risk of death if the disease is not treated in time

Our approach

- propose a model for the joint evolution of the health status and the marker
- formulate the treatment optimization problem as a stochastic control problem
- propose a numerical method to construct an explicit a policy close to optimality.
- study the performance of this policy on simulated patients with paraemeters calibrated from the clinical trial data

Outline

Variables of interest

- mode m: overall health status
  - $m = 0$: healthy / remission,
  - $m = 1$: disease / type 1 relapse
  - $m = 2$: disease / type 2 relapse
  - $m = 3$: death
- marker ζ: monoclonal immunoglobulin $\zeta \in [\zeta_0, D]$
  - $\zeta = \zeta_0$: nominal value in healthy mode
  - $\zeta = D$: death threshold

Treatments

The dynamics of the marker depend on the overall health status m and on the treatment chosen

Possible treatments ℓ

- $\ell = \emptyset$: no treatment
- $\ell = a$: treatment a
  - efficient on type 1 disease
  - slows the evolution of type 2 disease
- $\ell = b$: treatment b
  - efficient on type 2 disease
  - slows the evolution of type 1 disease

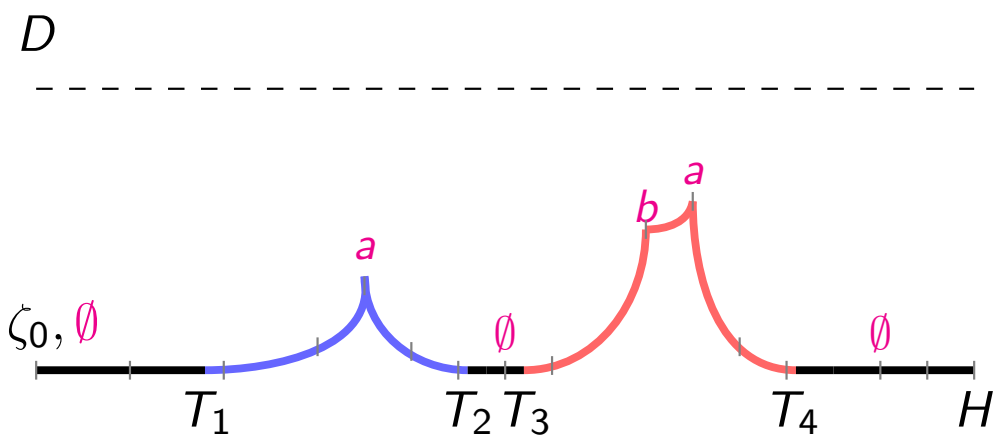
Piecewise deterministic Markov process model

Flow

Conditionally to the current mode m and treatment ℓ values, the dynamics of the marker is deterministic

$$\zeta(t+s) = \zeta(t)\exp(v_m^\ell s)$$

	$\ell = \emptyset$	$\ell = a$	$\ell = b$
$m = 0$	$v_0^\emptyset = 0$	$v_0^a = 0$	$v_0^b = 0$
$m = 1$	$v_1^\emptyset > 0$	$v_1^a < 0$	$0 < v_1^b < v_1^\emptyset$
$m = 2$	$v_2^\emptyset > 0$	$0 < v_2^a < v_2^\emptyset$	$v_2^b < 0$
$m = 3$		$v_3 = 0$	

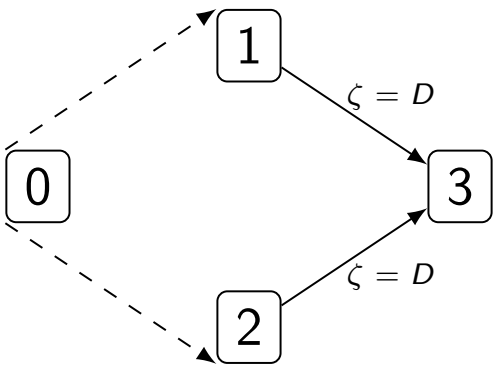


Piecewise deterministic Markov process model

Intensity and jump kernel

The health status is piecewise constant, it changes at deterministic (solid lines) or random (dashes) dates with an intensity depending on

- the marker value ζ and / or the time spent in the current mode m (additional variable u required to keep a Markov process)
- and the treatment applied ℓ



Possibles transitions

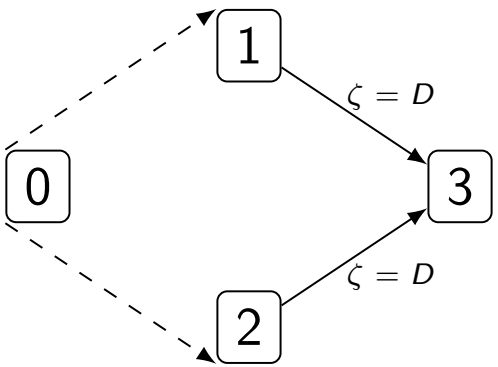
without treatment

Piecewise deterministic Markov process model

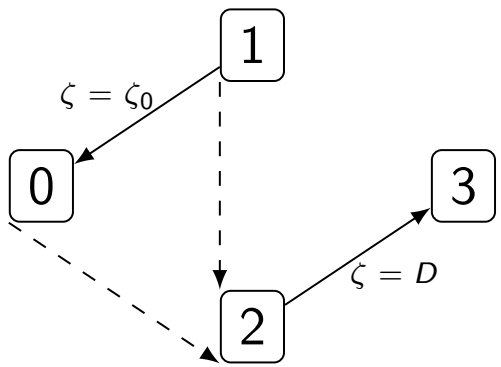
Intensity and jump kernel

The health status is piecewise constant, it changes at deterministic (solid lines) or random (dashes) dates with an intensity depending on

- the marker value ζ and / or the time spent in the current mode m (additional variable u required to keep a Markov process)
- and the treatment applied ℓ



Possibles transitions without treatment



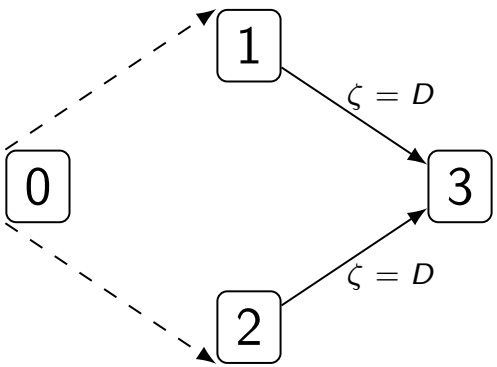
Possibles transitions with treatment a

Piecewise deterministic Markov process model

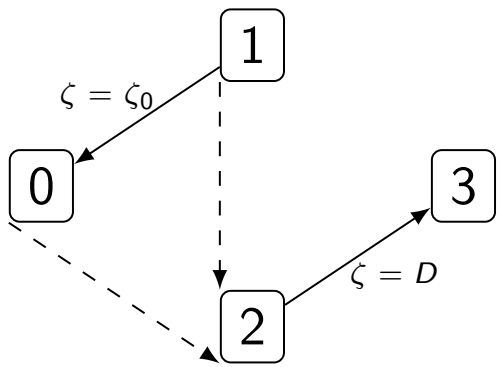
Intensity and jump kernel

The health status is piecewise constant, it changes at deterministic (solid lines) or random (dashes) dates with an intensity depending on

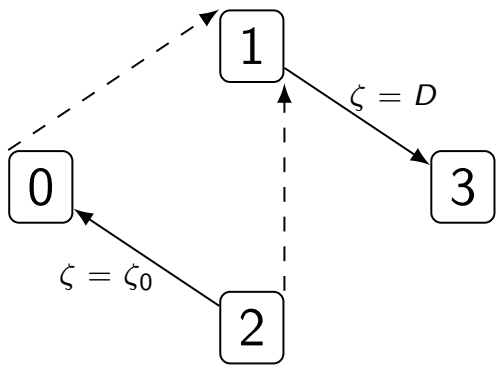
- the marker value ζ and / or the time spent in the current mode m (additional variable u required to keep a Markov process)
- and the treatment applied ℓ



Possibles transitions without treatment



Possibles transitions with treatment a



Possibles transitions with treatment b

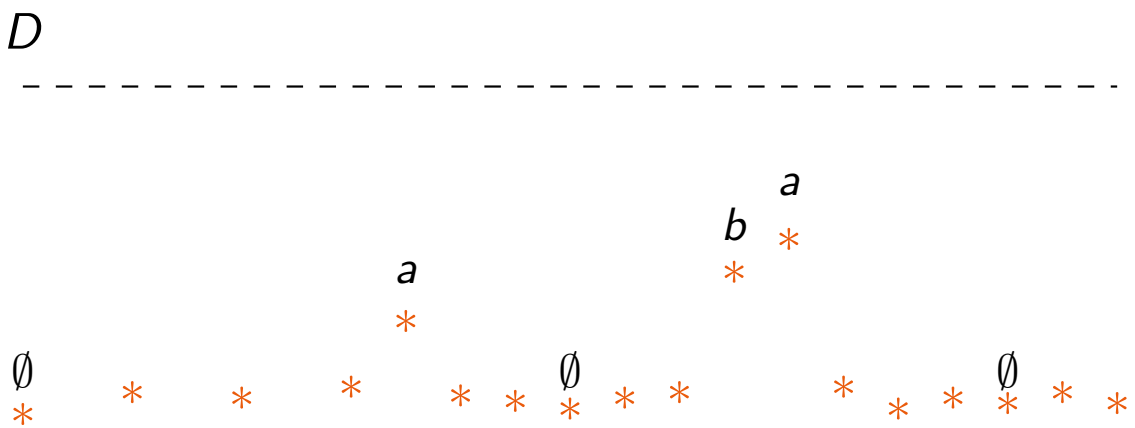
Stochastic impulse control problem

Sequential decision-making

- which treatment?
- for how long?

to keep the marker as close as possible to the nominal value ζ_0, while

- the mode is hidden
- the jump dates are hidden
- the marker is observed through noise and with low frequency



State of the art of impulse control for PDMPs

PDMP continuous time modelling

- close to biological reality
- allows the problem to be modeled with a small number of parameters, all of which are interpretable
- theoretical and numerical framework of impulse control for PDMPs well-defined under perfect observation at all times [Davis 93],[dS, Dufour, Zhang 14] or under partial observation when the jump dates are observed [dSDZ 14],[Bäuerle, Lange 17]
  - choose impulse dates
  - choose new process location after the impulse
- explicit theoretical and numerical construction of ϵ-optimal policies very difficult if the jump dates are not observed

State of the art of impulse control for PDMPs

PDMP continuous time modelling

- close to biological reality
- allows the problem to be modeled with a small number of parameters, all of which are interpretable
- theoretical and numerical framework of impulse control for PDMPs well-defined under perfect observation at all times [Davis 93],[dS, Dufour, Zhang 14] or under partial observation when the jump dates are observed [dSDZ 14],[Bäuerle, Lange 17]
  - choose impulse dates
  - choose new process location after the impulse
- explicit theoretical and numerical construction of ϵ-optimal policies very difficult if the jump dates are not observed

Simplify the problem: drastically limit the number of available options for treatment durations

Outline

Time between visits

δ minimum time between two observations

- δ is not small: typically 15 days for multiple myeloma (control horizon $H = 2400$ days)
- choice of the time r until the next visit restricted to some multiples of δ : $r \in \{15, 30, 60\}$ days

Time between visits

δ minimum time between two observations

- δ is not small: typically 15 days for multiple myeloma (control horizon $H = 2400$ days)
- choice of the time r until the next visit restricted to some multiples of δ : $r \in \{15, 30, 60\}$ days

Only transitions from t to $t + 15$, $t + 30$ and $t + 60$ of the continuous process are required: we only need to know the state of the process at discrete dates all multiples of δ.

Time between visits

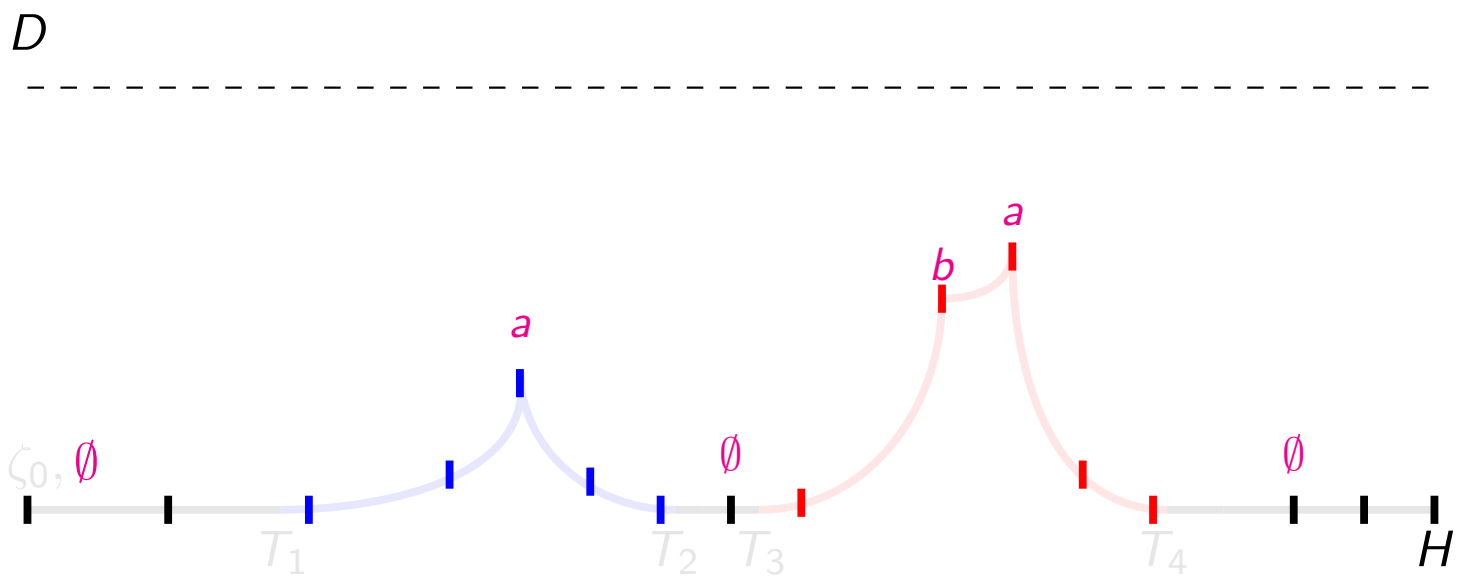
δ minimum time between two observations

- δ is not small: typically 15 days for multiple myeloma (control horizon $H = 2400$ days)
- choice of the time r until the next visit restricted to some multiples of δ : $r \in \{15, 30, 60\}$ days

Only transitions from t to $t + 15$, $t + 30$ and $t + 60$ of the continuous process are required: we only need to know the state of the process at discrete dates all multiples of δ.

It is not a discrete approximation of the continuous time process, we look at the real continuous time process process at discrete dates. In particular, the process can change mode between two observation dates.

Discrete time process



Markov decision process

Initialization

- $X_0 = (m(0), \zeta(0)) = (0, \zeta_0)$ health status, marker + additional variables
- $Y_0 = \zeta(0) + \epsilon_0$ initial observation, with ϵ_0 random noise + additional variables
- $n \leftarrow 0$ current step, $t \leftarrow 0$ current date

Markov decision process

Initialization

- $X_0 = (m(0), \zeta(0)) = (0, \zeta_0)$ health status, marker + additional variables
- $Y_0 = \zeta(0) + \epsilon_0$ initial observation, with ϵ_0 random noise + additional variables
- $n \leftarrow 0$ current step, $t \leftarrow 0$ current date

Iterations Given the current value of n, t, X_n and Y_n

- using only the available observations $Y_0, \ldots, Y_n$, choose the next decision $A_n = (\ell, r)$ ℓ = treatment, r = time until the next visit
- generate the next (hidden) marker value: $X_{n+1} = (m(t + r), \zeta(t + r))$ with the continuous time dynamics, conditionally to $X_n = (m(t), \zeta(t))$ and ℓ
- generate the next observation $Y_{n+1} = \zeta(t + r) + \epsilon_{n+1}$
- $n \leftarrow n + 1$, $t \leftarrow t + r$

Value function

- an admissible policy π is a sequence of decision rules based only on the observations available at each time point. Let Π be the set of admissible policies
- the optimization horizon is $H = N\delta$
- let c be the running cost function, and C the terminal cost function

$$V(x,y) = \inf_{\pi\in\Pi} \mathbb{E}^{\pi}_{(x,y)}\Big[\sum_{n=0}^{N-1} c(X_n, Y_n, A_n) + C(X_N, Y_N)\Big]$$

We are searching for a numerically tractable approximation of the value function V and an explicit policy π^* close to the optimum

Outline

Problem 1: partial observations

The process has hidden components X_n and observed ones Y_n: POMPD (partially observed MDP)

Problem 1: partial observations

The process has hidden components X_n and observed ones Y_n: POMPD (partially observed MDP)

Classical solution: convert the POMDP into a fully observed MDP on a larger space containing the belief space, using the belief process or filter [Bäuerle, Rieder 11], [Cleynen, dS 18]

$$\Theta_n(\cdot) = \mathbb{P}(X_n \in \cdot \mid Y_0, Y_1, \ldots, Y_n)$$

Explicit recursive formula (numerically intractable) linking Θ_{n+1} to Θ_n and Y_{n+1}

Equivalence between the POMDP and the belief MDP

$$V(x,y) = V'(\delta_x, y) = \inf_{\pi \in \Pi} \mathbb{E}^{\pi}_{(\delta_x, y)} \Big[\sum_{n=0}^{N-1} c'(\Theta_n, Y_n, A_n) + C'(\Theta_N, Y_N) \Big]$$

with

- $c'(\theta, y, , a) = \int c(x, y, a)\theta(dx)$
- $C'(\theta, y) = \int C(x, y)\theta(dx)$

Dynamic programming

V' can be computed by backward iterations using dynamic programming: set

$$V'_N(\theta, y) = C'(\theta, y)$$
$$V'_n(\theta, y) = \inf_a \{c'(\theta, y, a) + R' V'_{n+1}(\theta, y, a)\}$$

then $V'_0 = V'$ and an optimal policy is obtained by taking the arginf at each step

R' is the controlled transition kernel of the chain (Θ_n, Y_n):

$$R' f(\theta, y', a) = \mathbb{E}[f(\Theta_{n+1}, Y_{n+1}) | \ (\Theta_n, Y_n) = (\theta, y), A_n = a]$$

Outline

Problem 2 : continuous state space / infinite dimension

Θ_n lives in an infinite dimensional space, one cannot integrate analytically nor numerically against the kernel R'

- one cannot solve the dynamic programming equations
- the filter cannot be simulated

Problem 2 : continuous state space / infinite dimension

Θ_n lives in an infinite dimensional space, one cannot integrate analytically nor numerically against the kernel R'

- one cannot solve the dynamic programming equations
- the filter cannot be simulated

Solution: two-step discretization

- discretize the marker state space to obtain anapproximate filter $\bar{\Theta}$
  - $\bar{\Theta}_n$ has finite support
  - the recurrence relation between $\bar{\Theta}_n$ is $\bar{\Theta}_{n+1}$ is computable
  - $\bar{\Theta}_n$ can be simulated
- discretize the probabilities at each point in the support of the filter

The integral against R' reduces to a calculable weighted sum and if R' is sufficiently regular, we obtain a good approximation of V

Problem 3 : kernel regularity

Because of the boundary jumps when the marker reaches ζ_0 or D, the kernel P of X_n is not regular over the entire space E, but only locally Lipschitz on a specific partition of E

Problem 3 : kernel regularity

Because of the boundary jumps when the marker reaches ζ_0 or D, the kernel P of X_n is not regular over the entire space E, but only locally Lipschitz on a specific partition of E

Solution: be extra careful when creating the first discretization grid Ω. A point and its projection must always belong to the same subspace in the partition: place points symmetrically with respect to certain threshold values.

Approximation of the value function

Theorem

Under regularity assumptions for the parameters and compatibility assumptions between the grids and boundaries, one has

first discretization error

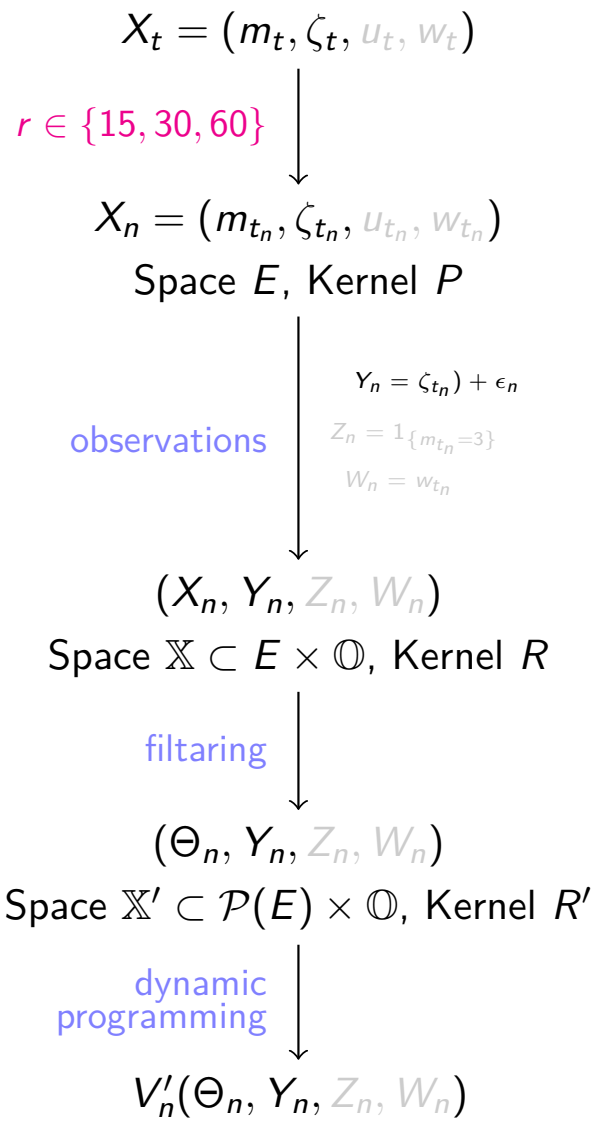
$$|V'_N(\bar{\theta}, \bar{y}) - \bar{V}'_N(\bar{\theta}, \bar{y})| = 0$$
$$|V'_n(\bar{\theta}, \bar{y}) - \bar{V}'_n(\bar{\theta}, \bar{y})| \leq C_{v'_n} \max \mathcal{D}_j, \quad 0 \leq n < N,$$

second discretization error

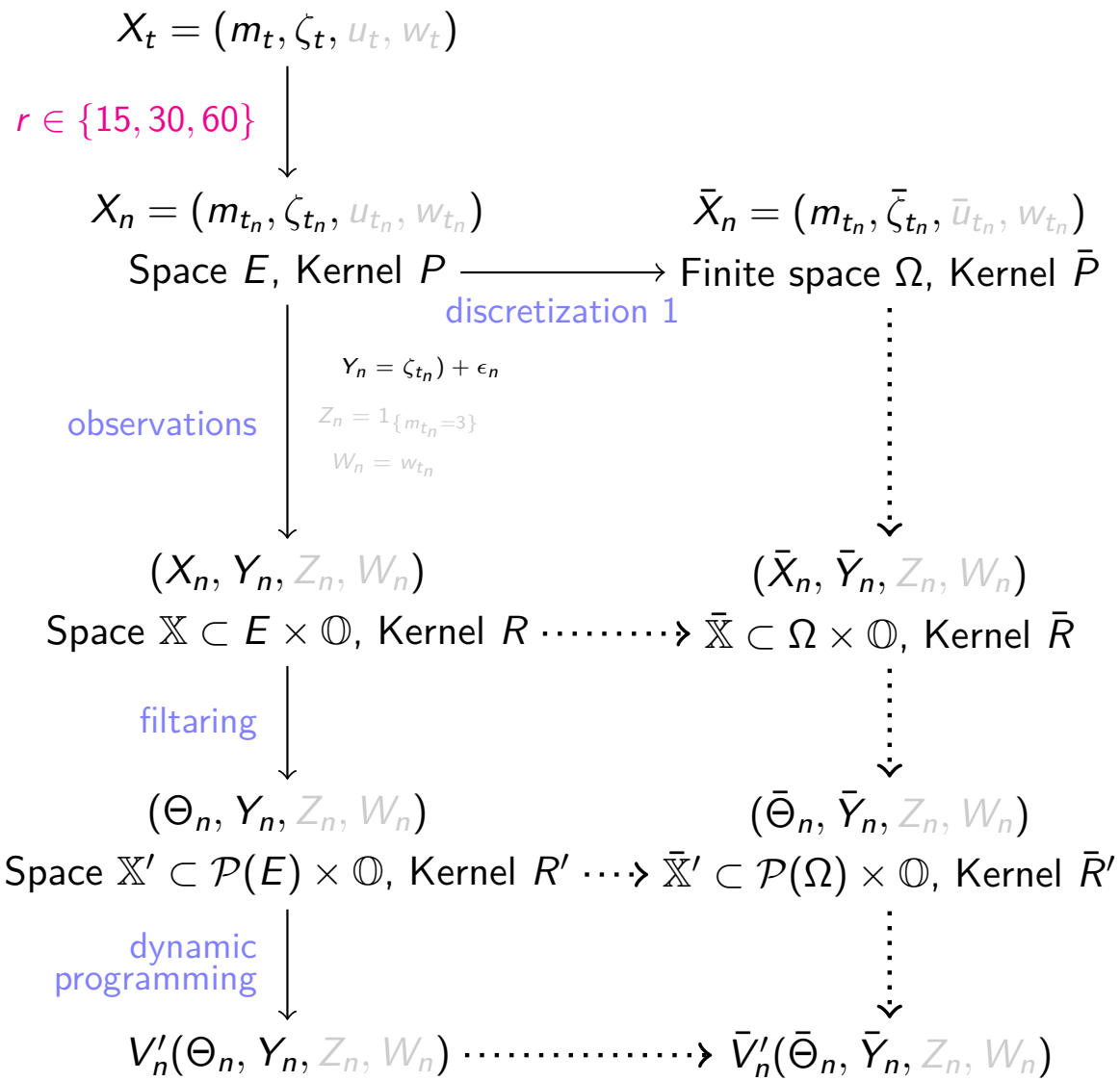
$$|\hat{V}'_N(\hat{\theta}, \bar{y}) - \bar{V}'_N(\hat{\theta}, \bar{y})| = 0,$$
$$|\hat{V}'_n(\hat{\theta}, \bar{y}) - \bar{V}'_n(\hat{\theta}, \bar{y})| \leq C_{\bar{v}'_n} \max \bar{\mathcal{D}}_j, \quad 0 \leq n < N,$$

where $C_{v'_n}$ and $C_{\bar{v}'_n}$ only depend on n, N, δ and the regularity parameters and $\mathcal{D}_j$ is the diameter of the j-th cell of the first grid, $\bar{\mathcal{D}}_j$ that of the j-th cell of the second grid

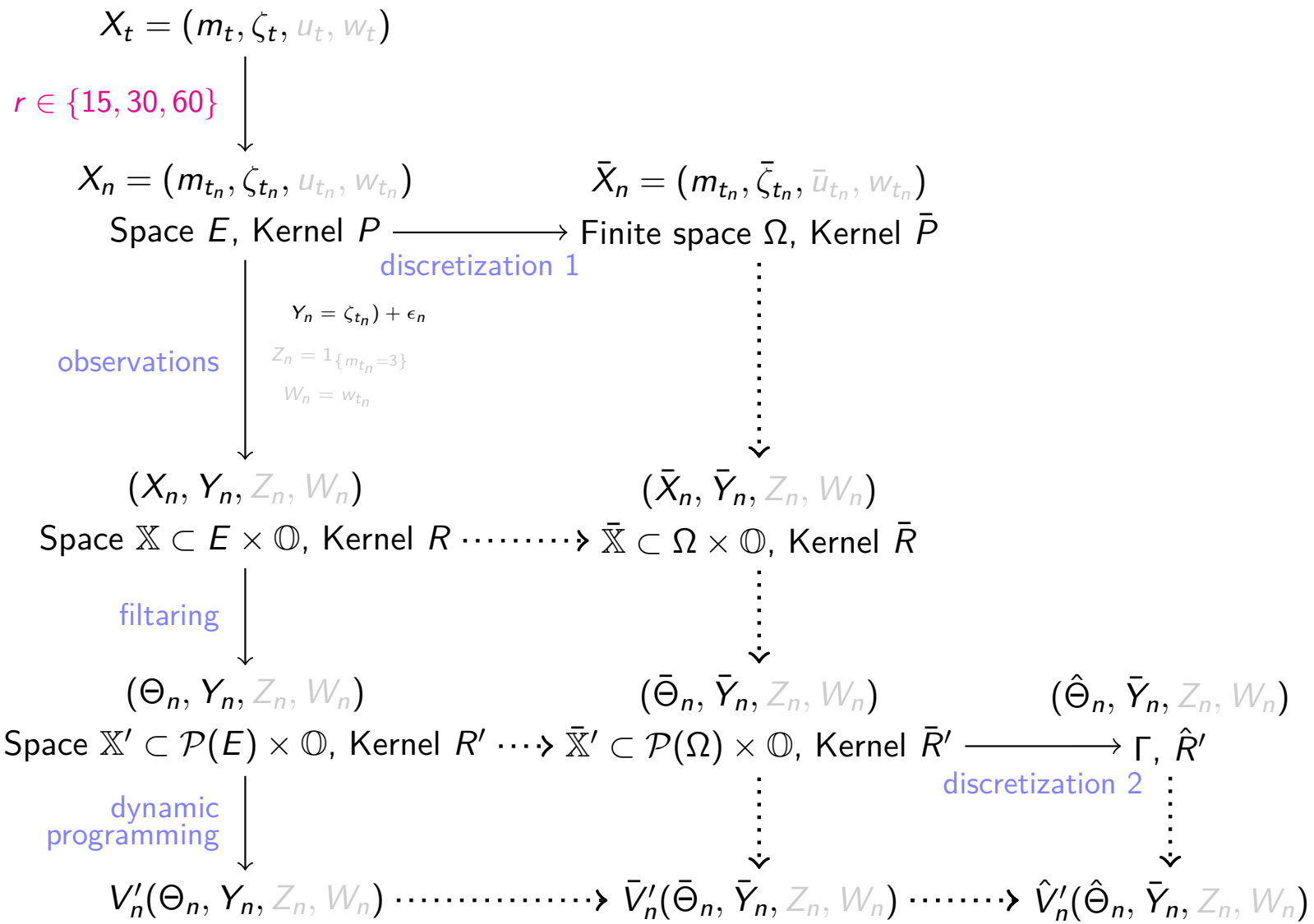
Graphical sketch of the proof



Graphical sketch of the proof



Graphical sketch of the proof



Candidate strategy

Initialization

- Filter initialized at $\bar{\theta} \leftarrow \delta_{(0,\zeta_0)}$
- initial observation y
- current time $t \leftarrow 0$, current step $n \leftarrow 0$

Candidate strategy

Initialization

- Filter initialized at $\bar{\theta} \leftarrow \delta_{(0,\zeta_0)}$
- initial observation y
- current time $t \leftarrow 0$, current step $n \leftarrow 0$

While the horizon or death is not reached

- project the current filter $\bar{\theta}$ onto the second grid Γ to obtain $\hat{\theta}$
- choose the action $a^* = (\ell, r)$ given by dynamic programming for $(\hat{\theta}, y, n)$
- give treatment ℓ until time $t + r$
- collect the new observation y on date $t + r$
- update the filter with the discretized operator from $\bar{\theta}$ and y
- $t \leftarrow t + r$, $n \leftarrow n + 1$

Outline

Problem 4: Decisions influence the dynamics

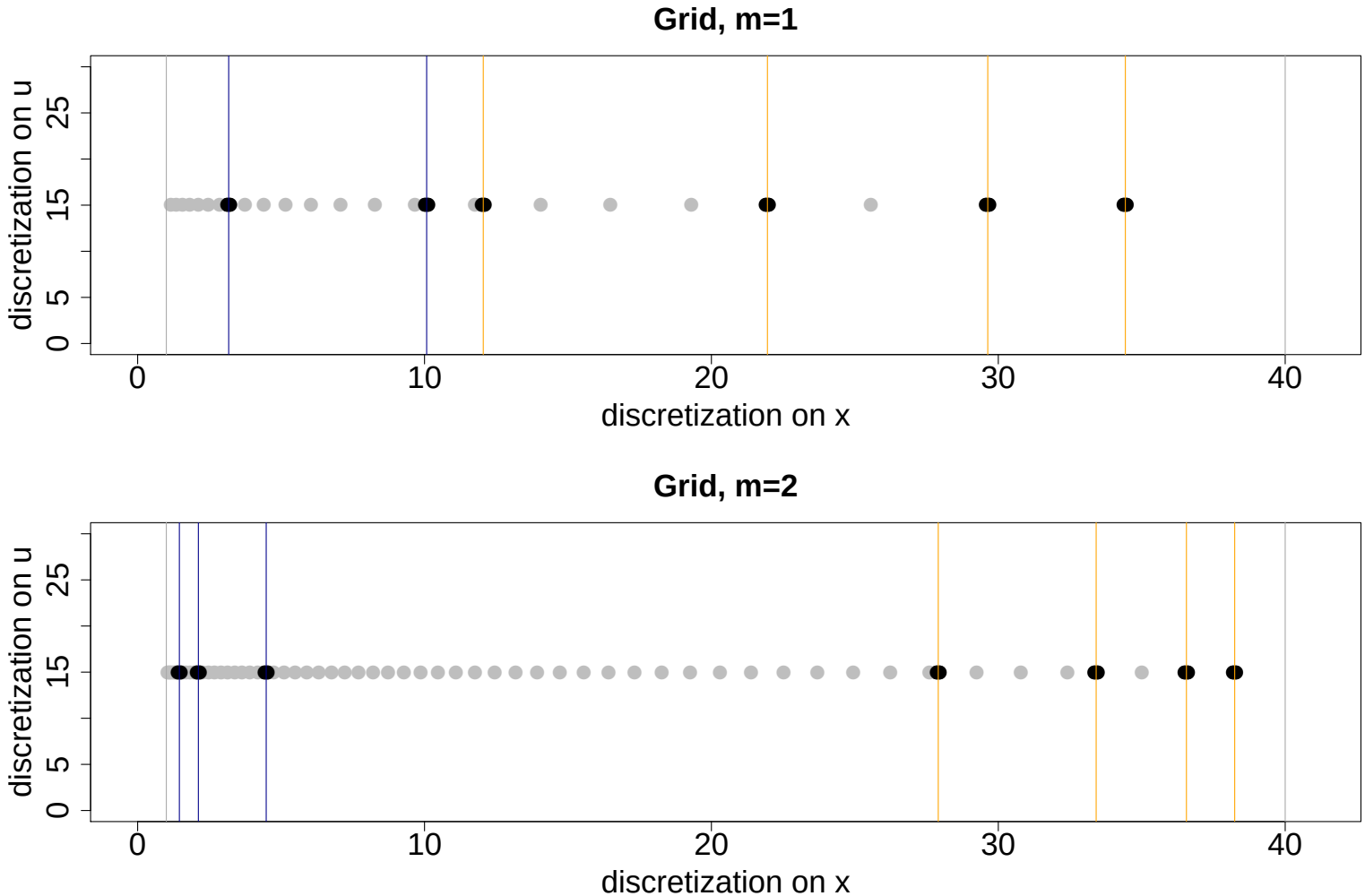
As the decisions significantly influence the dynamics, we cannot explore all the possible trajectories by simulation ($\sim 10^{152}$ policies) and the use of simulations is thus very limited to construct the grids

Problem 4: Decisions influence the dynamics

As the decisions significantly influence the dynamics, we cannot explore all the possible trajectories by simulation ($\sim 10^{152}$ policies) and the use of simulations is thus very limited to construct the grids

Solution: take advantage of the process rigidity
For a PDMP, the only source of randomness comes from jumps: until the first jump time, the process remains constant at ζ_0. The process will not visit all areas of the state space. Use this a priori information as best as possible so select the first grid points.
Enrich the second grid with simulations under the candidate policy.

First discretization



Second discretization

- Construct an initial grid in dimension $|\Omega|$: for each point ω in Ω take the probability distribution which loads ω with proba 0.95, and the rest of the points according to a Dirichlet distribution, estimate of the kernel $\hat{R}$ by Monte Carlo simulations
- Calculate the candidate policy for this grid
- Iterations
  - simulate *nsim* trajectories controlled with the candidate policy of the current grid
  - for each trajectory and each instant, calculate the distance between the estimated filter and its projection on the current grid and add the estimated filter in the next grid if this distance exceeds a certain threshold
  - Estimate $\hat{R}$ by Monte Carlo on the new grid and restart dynamic programming to update the candidate policy

Outline

Competing policies

- Gold Standard (unreachable) decisions based on perfect observation of the mode at each measurement instant
  - assign the correct treatment
  - visits with fixed step
- Filter (only the first discretization is used) use filter to estimate the mode
  - assign the treatment adapted to the most probable mode
  - visits with fixed step
- Standard (reference hospital protocol) based on thresholds s_{rel} for relapse and s_{rem} for remission.
  - As long as $y \leq s_{rel}$, $\ell = \emptyset$, $r = 60$ days
  - If $y > s_{rel}$, $\ell = b$ (corresponding to the most frequent type of relapse 2) and $r = 15$ days
  - If at the next visit y has decreased, treatment b is maintained with visits every 15 days until s_{rem} is reached
  - Otherwise, $\ell = a$ with visits every 15 days

Cost functions

The running cost $c(x, a)$ has the form

$$c(x, a) = \mathbb{E}[\tilde{c}(X_0, A_0, X_1)|X_0 = x, A_0 = a]$$

with

$$\tilde{c}(x, a, x') = C_V + \kappa|\zeta' - \zeta_0|r + \beta r \mathbb{1}_{\{m=0, \ell \neq \emptyset\}}.$$

if $m \neq 3$, where

- $x = (m, \zeta)$, $a = (\ell, r)$, $x' = (m', \zeta')$
- C_V: fixed cost per visit emotional burden + medical costs
- $\beta > 0$: penalty for applying a treatment without disease side effects
- $\kappa|\zeta' - \zeta_0|r$: approximation of the time spent in the disease and the severity of the disease
- M: death cost (paid only once if $m = 3$)

Performance

	Visit dates	cost (stand. dev.)	filtered cost (std)
candidate policy	optimal choice	136.23 (3.91)	134.74 (0.82)
	15 days	213.92 (1.66)	215.16 (0.75)
	60 days	145.37 (4.94)	140.58 (0.99)
Filter	15 days	209.96 (2.38)	210.2 (0.72)
	60 days	169.39 (6.76)	170.56 (2.15)
Gold Standard	15 days	161.51 (0.04)	
	60 days	52.31 (0.82)	
Standard		438.92 (20.42)	

500 simulated patients, ~ 100 grid points for Ω, ~ 1000 grid points for Γ other parameters calibrated on the clinical trial data

Outline

Summary

- first constructive result of an ϵ-optimal strategy for an impulse control problem for PDMP with hidden jump times

Summary

- first constructive result of an ϵ-optimal strategy for an impulse control problem for PDMP with hidden jump times
- theoretical guarantees on the approximation of the value function

Summary

- first constructive result of an ϵ-optimal strategy for an impulse control problem for PDMP with hidden jump times
- theoretical guarantees on the approximation of the value function
- good numerical performance

Summary

- first constructive result of an ϵ-optimal strategy for an impulse control problem for PDMP with hidden jump times
- theoretical guarantees on the approximation of the value function
- good numerical performance
- numerically intensive, and highly problem-dependent

Summary

- first constructive result of an ϵ-optimal strategy for an impulse control problem for PDMP with hidden jump times
- theoretical guarantees on the approximation of the value function
- good numerical performance
- numerically intensive, and highly problem-dependent
- generalizable to a certain extent:
  - not too many modes / variables, stay in low dimension
  - not too many possible jumps between two observations
  - generic deterministic flow between jumps (with a minimum regularity)
  - generic jump intensity (with a minimum regularity), possible addition of other boundary jumps

Ongoing work : ANR HSMM-INCA

with Alice Cleynen (CNRS) and Régis Sabbadin (Inrae)

Key step: estimate/simulate/discretize the filter

- exploration of other simulation-based methods: Monte Carlo Tree Search, Particle Filter Aymar Thierry d'Argenlieu's internship 2022

Ongoing work : ANR HSMM-INCA

with Alice Cleynen (CNRS) and Régis Sabbadin (Inrae)

Key step: estimate/simulate/discretize the filter

- exploration of other simulation-based methods: Monte Carlo Tree Search, Particle Filter Aymar Thierry d'Argenlieu's internship 2022

Learning parameters while optimizing

- reinforcement learning framework PhD thesis of Orlane Le Quellennec 2022-2025

Ongoing work : ANR HSMM-INCA

with Alice Cleynen (CNRS) and Régis Sabbadin (Inrae)

Key step: estimate/simulate/discretize the filter

- exploration of other simulation-based methods: Monte Carlo Tree Search, Particle Filter Aymar Thierry d'Argenlieu's internship 2022

Learning parameters while optimizing

- reinforcement learning framework PhD thesis of Orlane Le Quellennec 2022-2025

Towards more realistic models

- minimum duration of treatment once a treatment has started
- adapt the parameters to the number of relapses: resistance to treatment
- allow patient-specific parameters
- ...

References

[BL 17] N. Bäuerle, D. Lange *Markov decision processes with applications to finances*

[BR 11] N. Bäuerle, U. Rieder *Optimal control of partially observed PDMPs*

[CdS 18] A. Cleynen, B. de Saporta *Change-point detection for PDMPs*

[CdS 23] A. Cleynen, B. de Saporta *Numerical method to solve impulse control problems for partially observed piecewise deterministic Markov processes*

`https://github.com/acleynen/PDMP-control`

[Davis 93] M. Davis, *Markov models and optimization*

[dSDZ 14] B. de Saporta, F. Dufour, H. Zhang *Numerical methods for simulation and optimization of PDMPs: application to reliability*