Controlled Markov chains with observation costs

Jonathan Tam
Supervisor: Prof. Christoph Reisinger

Mathematical Institute
University of Oxford

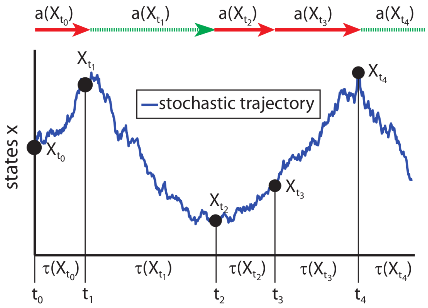
Graduate School: Mathematics of Random Systems
September 2021

Oxford Mathematics

Motivation

- Stochastic control often assumes that a continuous flow of data is available.
- But sometimes it is expensive to obtain data (e.g. medical records of patients).
- Need to optimise timing of measurements, and make decisions based on limited data.
- We would like a version of stochastic control where the state can only be known if a cost is paid.

Motivation



Source: Winkelmann et al. [2]

Figure 1: Schematic realisation of the control framework.

Our information flow should only consist of our past observations.

Setup

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
- Let $X = (X_n)_{n \in \mathbb{N}}$ be a Markov chain with state space $\mathcal{S}$.
- Consider a finite control set $\mathcal{I} := \{1, \ldots, d\}$.
- The transition matrix depends on the control: $P = P_i$.
- Assume the distribution of X is known *a priori*, but the realisation of each X_n comes with an upfront cost c_{obs}.

Setup

Let $\tau = \{\tau_k\}_{k=0}^{\infty}$ be a sequence of strictly increasing random times. Define for all $t \geq 0$,

$$\mathcal{F}_n^{(X,\tau)} := \sigma\Big\{(\tau_0, X_{\tau_0}), (\tau_1, X_{\tau_1}), \ldots, (\tau_k, X_{\tau_k}) : k = \sup\{j : \tau_j < n\}\Big\}.$$

Now we require that τ_k are predictable stopping times (with respect to $\mathcal{F}^{(X,\tau)}$).

Reason? We want τ_k to be $\mathcal{F}_{\tau_{k-1}}^{(X,\tau)}$-measurable.

i.e. at τ_{k-1}, we have our 'decision rule' for the next observation time τ_k.

$\mathcal{F}^{(X,\tau)}$ is our **observation filtration**, and τ is our **observation sequence**.

For each $\tau_k \in \tau$ we associate a random variable ι_k, which is $\mathcal{I}$-valued and $\mathcal{F}^{(X,\tau)}_{\tau_k}$-measurable. This represents the switching locations.

The double sequences $\alpha = (\tau_k, \iota_k)_{k\geq 1}$ form the set of admissible controls (denoted by $\mathcal{A}$) in our observation control problem.

For any time n, define

$$\tilde{\tau}_n = \max\{\tau_k \in \tau : \tau_k < n\}, \;\; \tilde{\iota}_n = \iota_{\tilde{\tau}_n} \tag{1}$$

so the pair $(\tilde{\tau}_n, \tilde{\iota}_n)$ is the most recent observation and switching location.

Define the observations process $\tilde{X}$ by

$$\tilde{X}_n = X_{\tilde{\tau}_n}. \tag{2}$$

Note that $\mathcal{F}^{\tilde{X}} = \mathcal{F}^{(X,\tau)}$. The Markov property of X gives the relation

$$\mathbb{E}[f(X_n)|\mathcal{F}_n^{\tilde{X}}] = \mathbb{E}[f(X_n)|\tilde{X}_n] \tag{3}$$

Observations Process

Consider the measure-valued process μ defined by

$$\begin{aligned}\mu_n(dx) &= \mathbb{P}(X_n \in dx | \mathcal{F}_n^{\tilde{X}}) \\ &\big(= \mathbb{P}(X_n \in dx | \tilde{X}_n) \big).\end{aligned} \tag{4}$$

which is the conditional distribution of X_n given its past observation history.

In fact each realisation of μ_n can be characterised by the values of $(\tilde{\tau}_n, \tilde{X}_n, \tilde{\iota}_n) = (k, x, i)$. We will use the notation $\mu_n^{k,x,i}$ to denote such a realisation.

Reward Functional and Value Function

We want to maximise a reward functional of the form

$$\mathbb{E}\left[\sum_{n=0}^{N} f(n, X_n, \tilde{\iota}_n) - \sum_{\tau_n} c_{\text{obs}}\right] \tag{5}$$

which is equivalent to

$$\mathbb{E}\left[\sum_{n=0}^{N} \mu_n(f(n, \cdot, \tilde{\iota}_n)) - \sum_{\tau_n} c_{\text{obs}}\right] \tag{6}$$

By treating μ as the new state process, we have a fully observable control problem.

Reward Functional (finite horizon):

$$J\big(m;\mu_m^{k,x,i};\alpha\big) := \mathbb{E}\left[\sum_{n=m}^{N}\mu_n(f(n,\cdot,\tilde{\iota}_n)) - \sum_{\tau_n\geq m} c_{\mathrm{obs}}\right], \tag{7}$$

$$v\big(m;\mu_m^{k,x,i}\big) := \sup_{\alpha\in\mathcal{A}} J\big(m;\mu_m^{k,x,i};\alpha\big). \tag{8}$$

Dynamic Programming:

$$v(m,\mu_m^{k,x,i}) = \sup_{a\in\mathcal{A}}\Big\{\mathbb{E}\big[f(m,X_m,\tilde{\iota}_m) - \mathbb{1}_{\{\tilde{\tau}_{m+1}=m\}}c_{\mathrm{obs}} + v(m+1,\mu_{m+1})\big|\ \mu_m=\mu_m^{k,x,i}\big]\Big\}. \tag{9}$$

Expanding upon (9), and writing (k, x, i) in place of $\mu_m^{k,x,i}$:

$$v\big(m; (k, x, i)\big) = \sup_{\alpha \in \mathcal{A}} \left\{ \sum_{y \in \mathcal{S}} p_{xy}^{(m-k)}(i) \left[f(m, y, \tilde{\iota}_m) - \mathbb{1}_{\{\tilde{\tau}_{m+1} = m\}} c_{\text{obs}} \right.\right.$$
$$\left.\left. + v\big(m+1; (\tilde{\tau}_{m+1}, \tilde{X}_{m+1}, \tilde{\iota}_m)\big) \right] \right\}. \quad (10)$$

Using finite difference notation, we can write more compactly:

$$\min \left\{ v_{i,x}^{m,k} - v_{i,x}^{m+1,k} - \left(P_i^{(m-k)} f_i\right)_x, \right.$$
$$\left. v_{i,x}^{m,k} - \sup_{j \in \mathcal{I}} \left[\left(P_i^{(m-k)} \big(v_j^{m+1,m} + f_j^m\big)\right)_x - c_{\text{obs}} \right] \right\} = 0. \quad (11)$$

Similarly, for the infinite horizon problem, we have

$$J\left(m;\mu_m^{x,i};\alpha\right) := \mathbb{E}\left[\sum_{n=m}^{\infty}\gamma^{n-m}\mu_n(f(\cdot,\tilde{\imath}_n)) - \sum_{\tau_n\geq m}\gamma^{\tau_n-m}c_{\text{obs}}\right], \tag{12}$$

$$v\left(m;\mu_m^{x,i}\right) := \sup_{\alpha\in\mathcal{A}} J\left(m;\mu_m^{x,i};\alpha\right), \tag{13}$$

which satisfies the quasi-variational inequality (QVI):

$$\min\Bigg\{v_{i,x}^m - \gamma v_{i,x}^{m+1} - \left(P_i^m f_i\right)_x, \\ v_{i,x}^m - \sup_{j\in\mathcal{I}}\left[\left(P_i^m\left(\gamma v_j^1 + f_j\right)\right)_x - c_{\text{obs}}\right]\Bigg\} = 0. \tag{14}$$

In practice, when solving for Equation (14), we have to truncate the domain and impose boundary conditions.

After truncation, can be written more generally as

$$\min\left\{F_i(u_i), u_i - \mathcal{M}u\right\} = 0, \quad u = (u_1, \ldots, u_d) \in \mathbb{R}^{d \times N \times L}, \tag{15}$$

where $\mathcal{M} : \mathbb{R}^{d \times N \times L} \to \mathbb{R}^{N \times L}$ is defined by

$$(\mathcal{M}u)_l^n = \max_{1 \leq j \leq d}\left(\left(A^n u_j^1\right)_l - c_{ij}\right), \tag{16}$$

where A is a non-negative matrix with row sums at most 1, and $F_i : \mathbb{R}^{N \times L} \to \mathbb{R}^{N \times L}$ satisfying the following property: for any $u, v \in \mathbb{R}^{d \times L \times N}$ with $u_{i,l}^n - v_{i,l}^n = \max(u_{j,k}^m - v_{j,k}^m) \geq 0$, we have

$$F_i(u_i)_l^n - F_i(v_i)_l^n \geq \gamma(u_{i,l}^n - v_{i,l}^n). \tag{17}$$

A comparison principle provides uniqueness to the solution.

Proposition

Suppose $c_{ij} > 0$, *and* $u = (u_i)_{i\in\mathcal{I}}$ *(resp.* $v = (v_i)_{i\in\mathcal{I}}$*) satisfies*

$$\min\{F_i(u_i),\ u_i - \mathcal{M}u\} \le 0 \quad (\textit{resp.} \ge 0), \quad i \in \mathcal{I}; \tag{18}$$

then $u \le v$.

$$\text{QVI:} \quad \min\left\{F_i(u_i), u_i - \mathcal{M}u\right\} = 0. \tag{19}$$

A standard approach to solve is policy iteration. However, no guarantees that resulting matrices are invertible.

We follow the approach taken by Reisinger and Zhang [1].

Consider the penalised problem: find $u^\rho = (u_i^\rho)_{i\in\mathcal{I}} \in \mathbb{R}^{d\times N\times L}$ such that

$$F_i(u_i^\rho)_l^n - \rho \sum_{j\in\mathcal{I}} \pi\left(\left(A^n u_j^{\rho,0}\right)_l - c_{ij} - u_{i,l}^{\rho,n}\right) = 0 \tag{20}$$

where $\rho > 0$ is the penalty parameter and $\pi : \mathbb{R} \to \mathbb{R}$ is continuous, non-decreasing with $\pi|_{(-\infty,0]} = 0$ and $\pi|_{(0,\infty)} > 0$.

Theorem

For any fixed $c_{ij} \geq 0$, the solution to the penalised problem u^ρ converges monotonically from below to a function $u \in \mathbb{R}^{d \times N \times L}$ as $\rho \to \infty$. Moreover u solves the discrete QVI if $c_{ij} > 0$ for all $i, j \in \mathcal{I}$.

To solve for the penalised equation, we can use semismooth Newton methods. Let

$$G^\rho[u] := F_i(u_i)_l^n - \rho \sum_{j \in \mathcal{I}} \pi \left(\left(A^n u_j^0\right)_l - c_{ij} - u_{i,l}^n \right), \tag{21}$$

then given $u^{(k)}$, we obtain the next iterate by solving

$$G^\rho[u^{(k)}] + \mathcal{L}^{(k+1)}[u^{(k)}](u^{(k+1)} - u^{(k)}) = 0, \tag{22}$$

where $\mathcal{L}$ is a generalised derivative of G^ρ.

Numerical Experiment

We apply our framework by extending a HIV-treatment model that appeared in Winklemann et al. [2].

Our model includes scenarios with large sub-optimal observation gaps.

3 regimes: Treatment 1, Treatment 2, no treatment

4 virus types: WT (Wild-type), R1, R2, HR (highly resistant)

State space: $\{0, l, m, h\}^4 \cup *$

$*$ represents eventual death (absorbing state in chain).

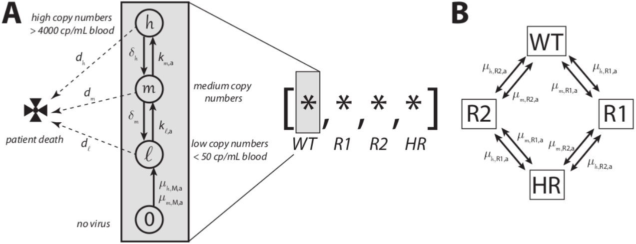


FIG. 3.1. **Simplified HIV Model** *A: Transitions between copy number states n_C. B: Transitions in between viral strains M.*

Source: Winkelmann et al. [2]

Cost function $\tilde{f}(x,i) = c_1(x) + c_2(i)$.

c_1 captures productivity loss from illness, c_2 represents cost of treatment.

Maximisation problem: take $f = -\tilde{f}$.

Reward functional

$$J\big(m;(x,i);\alpha\big) := \mathbb{E}\left[\sum_{n=m}^{\infty} \gamma^{n-m}\, f(\tilde{X}_n, \tilde{\imath}_n) - \sum_{\tau_n \geq m} \gamma^{n-\tau_n} c_{\text{obs}}\right]. \tag{23}$$

Numerical Experiment

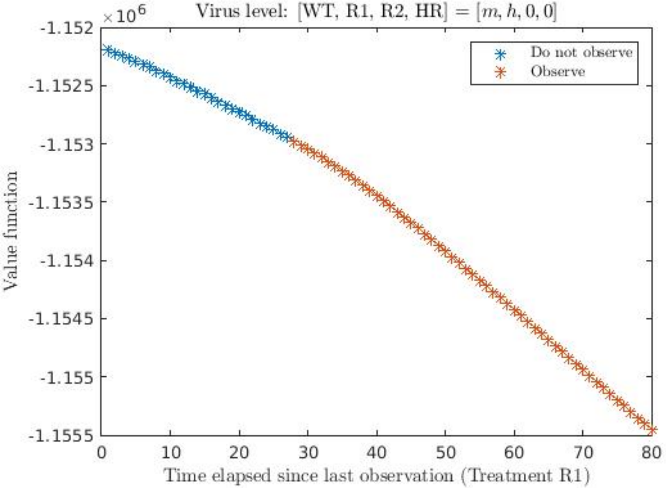
	ρ	10^3	2×10^3	4×10^3	8×10^3	16×10^3	32×10^3
$N = 150,\, c = 200$	(a)	10	10	10	10	10	10
	(b)	47.09	23.56	11.79	5.89	2.95	1.47
$N = 150,\, c = 400$	(a)	13	13	13	13	13	13
	(b)	47.17	23.60	11.81	5.90	2.95	1.48
$N = 150,\, c = 800$	(a)	17	17	17	17	17	17
	(b)	47.34	23.69	11.85	5.93	2.96	1.48
$N = 300,\, c = 400$	(a)	13	13	13	13	13	13
	(b)	81.29	40.68	20.35	10.17	5.09	2.54
$N = 600,\, c = 400$	(a)	13	13	13	13	13	13
	(b)	136.00	68.05	34.04	17.02	8.51	4.26

Figure 2: (a) Number of Newton iterations; (b) $\|v^\rho - v^{2\rho}\|$.

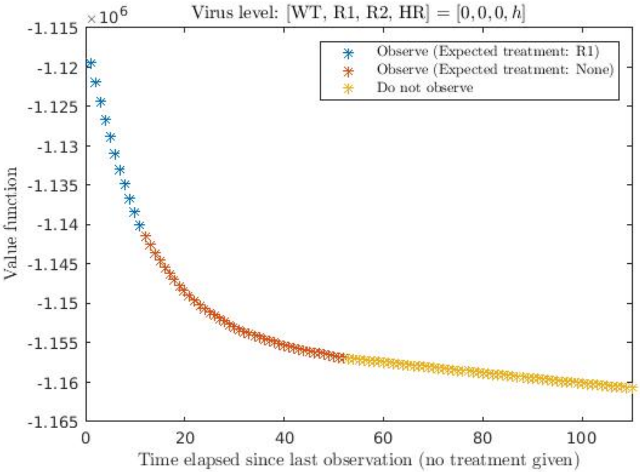
First-order convergence with respect to penalty parameter.

Number of Newton iterations remain constant across the size of ρ.

Numerical Experiment



Numerical Experiment



Further work: Bayesian parameter estimation

If transition matrix P depends on unknown parameters θ, can establish DPP involving the 'prior' and 'posterior' distributions.

$$v\big(m;(k,x,i);\rho\big) = \sup_{\alpha\in\mathcal{A}}\Bigg\{\sum_{y\in\mathcal{S}} p_{xy}^{\rho,(m-k)}(i)\Bigg[f(m,y,\tilde{\iota}_m) - \mathbb{1}_{\{\tilde{\tau}_{m+1}=m\}}c_{\text{obs}} + v\big(m+1;(\tilde{\tau}_{m+1},\tilde{X}_{m+1},\tilde{\iota}_m);\rho'\big)\Bigg]\Bigg\}, \tag{24}$$

$$p_{xy}^{\rho,(m-k)}(i) = \int_{\Theta} p_{xy|\theta}^{(m-k)}(i)\ \rho(d\theta), \tag{25}$$

$$\rho'(d\theta) = \begin{cases}\rho(d\theta), & \tilde{\tau}_{m+1}=k;\\ \rho(d\theta)\cdot p_{xy|\theta}^{(m-k)}(i)/p_{xy}^{\rho,(m-k)}(i), & \tilde{\tau}_{m+1}=m.\end{cases} \tag{26}$$

If X is a diffusion (e.g. solution to an SDE), we expect the corresponding value function to be a solution of

$$\min\Big\{ -\partial_s v_i(s,x) + \gamma v_i(s,x) - \mathbb{E}_i\left[f_i(X_s^x)\right], \quad v_i(s,x) - \max_{j\in\mathcal{I}}\Big(\mathbb{E}_i[v_j(0,X_s^x)] - c_{\mathrm{obs}}\Big)\Big\} = 0. \tag{27}$$

The general framework is similar, but technicalities with viscosity solutions have to be dealt with.

C. Reisinger and Y. Zhang.
A penalty scheme for monotone systems with interconnected obstacles: Convergence and error estimates.
SIAM J. Numer. Anal., 57(4):1625–1648, 2019.

S. Winkelmann, C. Schütte, and M. v. Kleist.
Markov control processes with rare state observation: Theory and application to treatment scheduling in HIV–1.
Commun. Math. Sci., 12(5):859–877, 2014.