Markov decision processes

Bo Friis Nielsen[1]

[1]DTU Informatics

02407 Stochastic Processes 9, November 1 2022

Markov decision processes

Today:

- Markov decision processes

Next week

- Brownian motion

Renewal reward processes

Claims $Y_i, i \in \mathbb{N}$ are generated according to a renewal process $\{N(t); t \geq 0\}$. The accumulated $Z(t)$ claim up to time t is

$$Z(t) = \sum_{i=1}^{N(t)} Y_i$$

$N(t)$ Number of claims up to time t

Y_i Size of claim i

$Z(t)$ Accumulated claim up to time t

In general hard to analyse, but

$$\begin{aligned}\mathbb{E}\left(e^{-\theta Z(t)}\right) &= \mathbb{E}\left(\mathbb{E}\left[e^{-\theta Z(t)}\middle| N(t)\right]\right) = \mathbb{E}\left(\mathbb{E}\left[e^{-\theta \sum_{i=1}^{N(t)} Y_i}\middle| N(t)\right]\right)\\ &= \mathbb{E}\left(\mathbb{E}\left[\prod_{i=1}^{N(t)} e^{-\theta Y_i}\middle| N(t)\right]\right) = \mathbb{E}\left(\prod_{i=1}^{N(t)} \mathbb{E}\left[e^{-\theta Y_i}\middle| N(t)\right]\right)\end{aligned}$$

Renewal reward processes

$$\begin{aligned}\mathbb{E}\left(e^{-\theta Z(t)}\right) &= \mathbb{E}\left(\mathbb{E}\left[e^{-\theta Z(t)}\middle| N(t)\right]\right) = \mathbb{E}\left(\mathbb{E}\left[e^{-\theta \sum_{i=1}^{N(t)} Y_i}\middle| N(t)\right]\right)\\ &= \mathbb{E}\left(\mathbb{E}\left[\prod_{i=1}^{N(t)} e^{-\theta Y_i}\middle| N(t)\right]\right) = \mathbb{E}\left(\prod_{i=1}^{N(t)} \mathbb{E}\left[e^{-\theta Y_i}\middle| N(t)\right]\right)\\ &= \mathbb{E}\left(\prod_{i=1}^{N(t)} \mathbb{E}\left[e^{-\theta Y_i}\right]\right) = \mathbb{E}\left(\mathbb{E}\left[e^{-\theta Y_i}\right]^{N(t)}\right) = \phi_{N(t)}\left(\phi_{Y_i}(\theta)\right)\end{aligned}$$

Explicit solution if $N(t)$ is a phase type renewal process and Y_i are phase type distributed.
A renewal reward model of pairs (X_i, Y_i), X_i and Y_i need not be independent.

Markov chain

If we think of a phase type renewal process in terms of the underlying Markov jump process, then rewards are associated with transitions in a Markov process.
If we further assume that the states have some physical meaning, then we could have rewards associated with other transitions (claims) or sojourns (premiums)

Markov chain

A Markov chain $\{X_t; t \in \mathbb{N} \cup \{0\}\}$ in discrete time is characterised by its transition probability matrix/matrices $\boldsymbol{P}$, $\boldsymbol{P}_t$ and the initial probability vector $\boldsymbol{p}_0$.
We could have rewards associated with transitions in the Markov chain.
Suppose we had acces to/control over the transitions mechanism such that we had $P_t(a)$ where the argument a is some action we can take.
See Chapter one of Puterman for examples.
Tamping as an example.

Markov Decision Process set up

- Decision epochs
  $T = \{1, 2, \ldots, N\}$, $T = \{1, 2, \ldots, \infty\}$, $T = [0; \infty[$
- State space S, $S = \cup_t S_t$
- Action sets $A_{s,t}$, $A = \cup_{s \in S} A_{s,t}$
- Rewards $\boldsymbol{R}_{t,A_{s,t}}$
  - Typically expected rewards $r_t(s,a) = \sum_{j \in S_t} r_t(s,a,j) p_t(j|s,a)$
  - For finite horizon $r_N(s)$ - scrap value
- Transition probabilities of $\boldsymbol{P}_{t,A_{s,t}}$
- $\boldsymbol{P}_{A_s}$ could be a Markov transition kernel on a general space
- Even if time is discrete it could be randomised (exponential)

$\{T, S, A_s, p_t(j|s,a), r_t(s,a)\}$ Markov decision process

Process

X_t	State occupied time t
Y_t	Action taken at time t
$r_t(s,a)$	Reward received at time t visiting state s taking action a
Z_t	History process at time t

Decision rules and policies

- d_t Decision rule. Action taken at time t. $d_t : S_t \to A_{s,t}$
- Z_t History, $Z_t = (s_1, a_1, s_2, a_2, \dots, s_N)$
- Π Policy, complete collection of decision rules $\Pi = (d_1, d_2, \dots)$

Policy types/decision rules 2.1.4

$K \in \{\mathrm{SD}, \mathrm{SR}, \mathrm{MR}, \mathrm{MD}, \mathrm{HD}, \mathrm{HR}, \mathrm{MR}, \mathrm{MD}\}$

- **SD** Stationary deterministic policy, (sometimes *pure* policy)
- **SR** Stationary random policy
- **MD** Markovian deterministic
- **MR** Markovian random
- **HD** History dependent deterministic
- **HR** Hisory random

Relations

$$\Pi^{\mathrm{SD}} \subset \Pi^{\mathrm{SR}} \subset \Pi^{\mathrm{MR}} \subset \Pi^{\mathrm{HR}}$$
$$\Pi^{\mathrm{SD}} \subset \Pi^{\mathrm{MD}} \subset \Pi^{\mathrm{MR}} \subset \Pi^{\mathrm{HR}}$$
$$\Pi^{\mathrm{SD}} \subset \Pi^{\mathrm{MD}} \subset \Pi^{\mathrm{HD}} \subset \Pi^{\mathrm{HR}}$$

Given a policy we have a stochastic process - typically a Markov reward process

One step MDP

$$r_1(s,a') + \mathbb{E}_s^{\Pi}(v(X_2)) = r_1(s,a') + \sum_{j\in S} p_1(j|s,a')v(j)$$

$$\max_{a'\in A_s}\left\{r_1(s,a') + \sum_{j\in S} p_1(j|s,a')v(j)\right\} =$$
$$r_1(s,a^*) + \sum_{j\in S} p_1(j|s,a^*)v(j)$$

$$\operatorname*{argmax}_{x\in X} = \{x' \in X | \forall x \in X : g(x') \geq g(x)\}$$

$$a_s^* = \operatorname*{argmax}_{a'\in A_{s,t}}\left\{r_1(s,a') + \sum_{j\in S} p_1(j|s,a')v(j)\right\}$$

Expected total reward criterion 4.1.2

Definition

$$v_N^{\pi}(s) = \mathbb{E}_s^{\pi}\left\{\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_n)\right\} \quad (1)$$

$$v_{N,\lambda}^{\pi}(s) = \mathbb{E}_s^{\pi}\left\{\sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, Y_t) + \lambda^{N-1} r_N(X_n)\right\} \quad (2)$$

Optimal policies

$$
\begin{aligned}
v_N^{\pi^*}(s) &\geq v_N^{\pi}(s), && s \in S \\
v_N^{\pi_\epsilon^*}(s) + \epsilon &\geq v_N^{\pi}(s), && s \in S \\
v_N^{\star}(s) &= \sup_{\pi \in \Pi^{HR}} v_N^{\pi}(s) && \\
\left(v_N^{\star}(s)\right. &= \left.\max_{\pi \in \Pi^{HR}} v_N^{\pi}(s)\right) && \\
v_N^{\pi^*}(s) &= v_N^{\star}(s), && s \in S \\
v_N^{\pi_\epsilon^*}(s) + \epsilon &> v_N^{\star}(s), && s \in S
\end{aligned}
$$

Optimal stopping

Decision epochs: $T = \{1, 2, \ldots, N\}, \quad N \leq \infty$

States: $S = S' \cup \{\Delta\}$

Actions: $A_s = \begin{cases} \{C, Q\} & s \in S' \\ \{C\} & s = \Delta \end{cases}$

Rewards: $r_t(s, a) = \begin{cases} -f_t(s) & s \in S' \quad a = C \\ g_t(s) & s \in S' \quad a = Q \\ 0 & s = \Delta \end{cases}$

$r_N(s) = h(s)$

Transition probabilities

$$
p_t(j|s, a) = \begin{cases} p_t(j|s) & s \in S' \quad j \in S' \quad a = C \\ 1 & s \in S' \quad j = \Delta \quad a = Q \\ 1 & s = j = \Delta \quad a = C \\ 0 & \text{otherwise} \end{cases}
$$

Secretary problem setup

- N candidates apply for a position
- Objective is to hire the best person
- A decision needs to be taken immediately after the interview

$$
s = \begin{cases} 0 & \text{Current candidate not best so far} \\ 1 & \text{Current candidate best so far} \\ \Delta & \text{Interview process stopped} \end{cases}
$$

$f_t(s) = 0, \quad g_t(0) = 0, \quad g_t(1) = \frac{t}{N}$

Backward induction algorithm

1. Set $t = N$ and $u_N^*(s_N) = r_N(s_N), \quad \forall s_N \in S$
2. t^{--} For each $s_t \in S$

$$
\begin{aligned}
u_t^*(s_t) &= \max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right\} \\
A_{s,t}^* &= \operatorname*{argmax}_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right\}
\end{aligned}
$$

3. If $t = 1$ stop, otherwise return to step 2

Secretary problem backward induction

$$u_N^*(s) = \begin{cases} 0 & s=0\\ 1 & s=1 \\ 0 & s=\Delta\end{cases}$$

$$u_t^*(s) =$$

$$\begin{cases} \max\left(0, \frac{1}{1+t}u_{t+1}^*(1)+\frac{t}{t+1}u_{t+1}^*(0)\right) & s=0\\ \max\left(g(t)+u_{t+1}^*(\Delta), \frac{1}{1+t}u^*_{t+1}(1)+\frac{t}{t+1}u^*_{t+1}(0)\right) & s=1\\ u^*_{t+1}(\Delta) & s=\Delta\end{cases}$$

Secretary problem backward induction - cont

$$u_t^*(s) = \begin{cases} \frac{1}{1+t}u_{t+1}^*(1)+\frac{t}{t+1}u_{t+1}^*(0) & s=0\\ \max\left(\frac{t}{N}, \frac{1}{1+t}u^*_{t+1}(1)+\frac{t}{t+1}u^*_{t+1}(0)\right) & s=1\\ 0 & s=\Delta\end{cases}$$

$$u_t^*(s) = \begin{cases} \frac{1}{1+t}u_{t+1}^*(1)+\frac{t}{t+1}u_{t+1}^*(0) & s=0\\ \max\left(\frac{t}{N}, u^*_t(0)\right) & s=1\\ 0 & s=\Delta\end{cases}$$

$$A^*_{s,t}(0) = \begin{cases} C & s=0\\ Q : \frac{t}{N} > u_t^*(0) & s=1\\ C & s=\Delta\end{cases}$$

Bo Friis Nielsen Markov decision processes

Assume $u_t^*(1) \geq \frac{\tau}{N}$ then $u_t^*(1) = u_t^*(0) \geq \frac{\tau}{N}$

So: $u_{t-1}^*(1) = \max\left(\frac{\tau-1}{N}, u^*_{t-1}(0)\right)$

with $u^*_{t-1}(0) = \frac{1}{\tau-1+1}u_t^*(1) + \frac{\tau-1}{\tau-1+1}u_t^*(0) = u_t^*(0) \geq \frac{\tau}{N} > \frac{\tau-1}{N}$

Bo Friis Nielsen Markov decision processes

Calculation of $u_t^*(0)$

$u_t^*(0) = \frac{1}{1+t}u^*_{t+1}(1) + \frac{t}{t+1}u^*_{t+1}(0) = \frac{1}{1+t}\frac{1+t}{N} + \frac{t}{t+1}u^*_{t+1}(0) =$
$\frac{1}{N} + \frac{t}{t+1}u^*_{t+1}(0)$
$u_N^*(0) = 0$
$u^*_{N-1}(0) = \frac{1}{N} + \frac{N-1}{N}\cdot 0 = \frac{1}{N}$
$u^*_{N-2}(0) = \frac{1}{N} + \frac{N-2}{N-1}\cdot\frac{1}{N} = \frac{N-2}{N}\left(\frac{1}{N-2}+\frac{1}{N-1}\right)$
$u^*_{N-3}(0) = \frac{1}{N} + \frac{N-3}{N-2}\cdot\frac{N-2}{N}\left(\frac{1}{N-2}+\frac{1}{N-1}\right) =$
$\frac{N-3}{N}\left(\frac{1}{N-3}+\frac{1}{N-2}+\frac{1}{N-1}\right)$
So $u_t^*(0) = \frac{t}{N}\sum_{k=t}^{N-1}\frac{1}{k}$ and
$\tau = \max_t\left(\sum_{k=t}^{N-1}\frac{1}{k} > 1\right)$
$\int_1^N \frac{1}{x} = \log(N) : \sum_{k=t}^{N-1}\frac{1}{k} \simeq \log\left(\frac{N-1}{t}\right)$,
$\log\left(\frac{N}{\tau}\right) \simeq= 1 \Rightarrow \tau = Ne^{-1}$

Bo Friis Nielsen Markov decision processes