Markov Decision Processes with Applications
Day 1

Nicole Bäuerle

Accra, February 2020
Overview

- Motivation
- Formal Definition of MDP
- Assumptions
- Solution
- Examples
Motivation
Markov Decision Processes (MDPs): Motivation

Let \((X_n)\) be a Markov process (in discrete time) with
- state space \(E\),
- transition probabilities \(Q_n(\cdot|x)\).

Let \((X_n)\) be a controlled Markov process with
- state space \(E\), action space \(A\),
- admissible state-action pairs \(D_n \subset E \times A\),
- transition probabilities \(Q_n(\cdot|x, a)\).

A decision \(A_n\) at time \(n\) is in general \(\sigma(X_1, \ldots, X_n)\)-measurable. However, Markovian structure implies \(A_n = f_n(X_n)\) is sufficient.
General evolution of a Markov Decision Process

- State at stage $n$: $x_n$
- Controller
- Reward at stage $n$: $r_n(x_n, a_n)$
- Random transition with distribution $Q_n(\cdot | x_n, a_n)$
- State at stage $n+1$: $x_{n+1}$
Applications

- Queueing theory (Data transmission, production planning, health care,...)
- Finance (portfolio problems, dividend problems,...)
- Computer science (robotics, shortest path, speech recognition,...)
- Energy (energy mix, real options (gas storage), ...)
- Biology (epidemic processes, cancer treatment,...)
Formal Definition of MDPs
Definition

A Markov Decision Model with planning horizon $N \in \mathbb{N}$ consists of a set of data $(E, A, D_n, Q_n, r_n, g_N)$ with the following meaning for $n = 0, 1, \ldots, N - 1$:

- $E$ is the state space,
- $A$ is the action space,
- $D_n \subset E \times A$ admissible state-action combinations at time $n$,
- $Q_n(\cdot | x, a)$ stochastic transition kernel at time $n$,
- $r_n : D_n \rightarrow \mathbb{R}$ one-stage reward at time $n$,
- $g_N : E \rightarrow \mathbb{R}$ terminal reward at time $N$. 

MDPs: Formal Definition
Remarks

a) If the one-stage reward function \( r'_n \) also depends on the next state, i.e. \( r'_n = r'_n(x, a, x') \), then define

\[
r_n(x, a) := \int r'_n(x, a, x') Q_n(dx'|x, a).
\]

b) In applications often \( D_n \) and \( Q_n \) do not dependent on \( n \) and \( r_n(x, a) := \beta^n r(x, a) \) and \( g_N(x) := \beta^N g(x) \) for a (discount) factor \( \beta \in (0, 1] \). In this case the Markov Decision Model is called \textit{stationary}. 
Policies

- A decision rule at time $n$ is a measurable mapping $f_n : E \to A$ such that $f_n(x) \in D_n(x)$ for all $x \in E$.
- A policy is given by $\pi = (f_0, f_1, \ldots, f_{N-1})$ a sequence of decision rules.
Optimization Problem

For \( n = 0, 1, \ldots, N \), \( \pi = (f_0, \ldots, f_{N-1}) \) define the value functions

\[
V_{n\pi}(x) := \mathbb{E}_{nx}^{\pi} \left[ \sum_{k=n}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N) \right],
\]

\[
V_n(x) := \sup_{\pi} V_{n\pi}(x), \quad x \in E.
\]

A policy \( \pi \) is called *optimal* if \( V_{0\pi}(x) = V_0(x) \) for all \( x \in E \).

Integrability Assumption (\( A_N \)): For \( n = 0, 1, \ldots, N \)

\[
\delta_n^N(x) := \sup_{\pi} \mathbb{E}_{nx}^{\pi} \left[ \sum_{k=n}^{N-1} r_k^+(X_k, f_k(X_k)) + g_N^+(X_N) \right] < \infty, \quad x \in E.
\]
VIPs of MDPs

Figure: Lloyd Shapley (1923 - 2016)
VIPS of MDPs

Figure: Richard Bellman (1920 - 1984)
VIPs of MDPs

Figure: David Blackwell (1912 - 2010)
Literature - Textbooks on MDPs

▶ Shapley (1953)
▶ Bellman (1957, Reprint 2003)
▶ Howard (1960)
▶ Bertsekas and Shreve (1978)
▶ Puterman (1994)
▶ Feinberg and Shwartz (2002)
▶ Powell (2007)
▶ B and Rieder (2011)
Alternative Formulation

The stochastic transition law of a Markov Decision Model is often given by a *transition* or *system function*.

- Suppose $Z_1, Z_2, \ldots, Z_N$ are random variables with values in a measurable space $\mathcal{Z}$.
- These random variables are called *disturbances*.
- The distribution $Q^Z_n$ of $Z_{n+1}$ may depend on the current state and action at time $n$ such that $Q^Z_n(\cdot|x, a)$ is a stochastic kernel for $(x, a) \in D_n$.
- The new state at time $n + 1$ can be described by a *transition function* $T_n : D_n \times \mathcal{Z} \rightarrow E$ such that

$$x_{n+1} = T_n(x_n, a_n, z_{n+1}).$$
Example: Consumption Problem

- Suppose there is an investor with given initial capital.
- At the beginning of each of \( N \) periods she can decide how much of the capital she consumes and how much she invests into a risky asset.
- The amount she consumes is evaluated by a utility function \( U \) (strictly increasing, concave) as well as the terminal wealth.
- The remaining capital is invested into a risky asset.

How should she consume/invest in order to maximize the sum of her expected utility?
Example: Consumption Problem

The reward is given by $U(a)$ and the terminal reward by $U(x)$. Hence the aim is to maximize

$$
\mathbb{E}^\pi_x \left[ \sum_{k=0}^{N-1} U(f_k(X_k)) + U(X_N) \right]
$$

where the maximization is over all policies $\pi = (f_0, \ldots, f_{N-1})$.

- $Z_{n+1}$ is the random return of the risky asset in $[n, n + 1)$.
- $Z_1, \ldots, Z_N$ are non-negative, independent r.v.
- $X_{n+1} = (X_n - f_n(X_n))Z_{n+1}$. 
Example: Consumption Problem

Data is given by

- $E := \mathbb{R}_+$ where $x_n \in E$ is the wealth of the investor at $n$,
- $A := \mathbb{R}_+$ where $a_n \in A$ is the wealth consumed at time $n$,
- $D_n(x) := [0, x]$ for all $x \in E$.
- $Z := \mathbb{R}_+$ where $z_n \in Z$ is the return of the asset in $[n, n+1)$,
- $T_n(x_n, a_n, z_{n+1}) := (x_n - a_n)z_{n+1}$ is the transition function,
- $Q_n^Z(\cdot | x, a) :=$ distribution of $Z_{n+1}$ (independent of $(x, a)$),
- $r_n(x, a) := U(a)$ is the one-stage reward,
- $g_N(x) := U(x)$. 

Notation

Let $\mathbb{M}(E) := \{v : E \rightarrow [-\infty, \infty) | v \text{ is measurable}\}$ and define the following operators for $v \in \mathbb{M}(E)$:

**Definition**

a) $(L_n v)(x, a) := r_n(x, a) + \int v(x')Q_n(dx'|x, a), \ (x, a) \in D_n,$

b) $(T_{nf} v)(x) := (L_n v)(x, f(x)), \ x \in E,$

c) $(T_n v)(x) := \sup_{a \in D_n(x)}(L_n v)(x, a).

Note $T_n v \notin \mathbb{M}(E)$ in general.

A decision rule $f_n$ is called *maximizer* of $v$ at time $n$ if $T_{nf_n} v = T_n v.$
Properties of Operators

Lemma

All three operators are monotone, i.e. for \( v, w \in \mathbb{M}(E) \) with \( v(x) \leq w(x) \) for all \( x \in E \) it holds:

a) \( L_n v(x, a) \leq L_n w(x, a) \) for all \( (x, a) \in D_n \),

b) \( T_{nf} v(x) \leq T_{nf} w(x) \) for all \( x \in E, f \in F_n \),

c) \( T_n v(x) \leq T_n w(x) \) for all \( x \in E \).
Theorem (Reward Iteration)

For a policy $\pi = (f_0, \ldots, f_{N-1})$ and $n = 0, 1, \ldots, N - 1$:

a) $V_{N\pi} = g_N$ and $V_{n\pi} = T_{nf_n} V_{n+1,\pi}$,

b) $V_{n\pi} = T_{nf_n} \cdots T_{N-1f_{N-1}} g_N$. 
Proof

Part a): For \( x \in E \) we have

\[
V_{n\pi}(x) = \mathbb{E}_{n\pi}^\pi \left[ \sum_{k=n}^{N-1} r_k(x_k, f_k(x_k)) + g_N(x_N) \right]
\]

\[
= \mathbb{E}_{n\pi}^\pi [r_n(x, f_n(x))] + \mathbb{E}_{n\pi}^\pi \left[ \sum_{k=n+1}^{N-1} r_k(x_k, f_k(x_k)) + g_N(x_N) \right]
\]

\[
= r_n(x, f_n(x)) + \mathbb{E}_{n\pi}^\pi \left[ \mathbb{E}_{n\pi}^\pi \left[ \sum_{k=n+1}^{N-1} r_k(x_k, f_k(x_k)) + g_N(x_N) \middle| x_{n+1} \right] \right]
\]
Proof

\[ V_{n\pi}(x) = \mathbb{E}_n^{\pi} \left[ \sum_{k=n}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N) \right] \]

\[ = r_n(x, f_n(x)) \]

\[ + \int \mathbb{E}_{n+1, x'}^{\pi} \left[ \sum_{k=n+1}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N) \right] Q_n(dx'|x, f_n(x)) \]

\[ = r_n(x, f_n(x)) + \int V_{n+1, \pi}(x') Q_n(dx'|x, f_n(x)). \]

Part b) follows from a) by induction.
Example: Consumption Problem

- For $f_n \in F_n$ the $T_{nf_n}$ operator in this example reads

$$T_{nf_n}v(x) = U_n(f_n(x)) + \mathbb{E} v((x - f_n(x))Z_{n+1}).$$

- Assume that $U_n(x) := \log x$ for all $n$ and $g_N(x) := \log x$.
- The return distribution is independent of $n$ and has finite expectation $\mathbb{E} Z$. Then $(A_N)$ is satisfied

- If we choose the $N$-stage policy $\pi = (f_0, \ldots, f_{N-1})$ with $f_n(x) = cx$ and $c \in [0, 1]$ then the Reward Iteration yields

$$V_{0\pi}(x) = (N+1) \log x + N \log c + \frac{(N+1)N}{2} \left( \log(1-c) + \mathbb{E} \log Z \right).$$

Hence $\pi^* = (f_0^*, \ldots, f_{N-1}^*)$ with $f_n^*(x) = c^* x$ and $c^* = \frac{2}{N+3}$ maximizes the expected log-utility (among all linear consumption policies).
Assumptions and Solution
Structure Assumption (SA$_N$)

There exist sets $\mathbb{M}_n \subset \mathbb{M}(E)$ and $\Delta_n \subset F_n$ such that for all $n = 0, 1, \ldots, N - 1$:

(i) $g_N \in \mathbb{M}_N$.

(ii) If $v \in \mathbb{M}_{n+1}$ then $T_n v$ is well-defined and $T_n v \in \mathbb{M}_n$.

(iii) For all $v \in \mathbb{M}_{n+1}$ there exists a maximizer $f_n$ of $v$ with $f_n \in \Delta_n$.

If $E$ and $A$ are finite, then (SA$_N$) is satisfied with $\mathbb{M}_n = \mathbb{M}(E)$ and $\Delta_n = F_n$. 
Main Theorem

Theorem (Structure Theorem)

Let \((SA_N)\) be satisfied. Then it holds:

a) \(V_n \in \mathbb{M}_n\) and the sequence \((V_n)\) satisfies the Bellman equation, \(i.e.\) for \(n = 0, 1, \ldots, N - 1\)

\[
V_N(x) = g_N(x),
\]

\[
V_n(x) = \sup_{a \in D_n(x)} \left\{ r_n(x, a) + \int V_{n+1}(x')Q_n(dx'|x, a) \right\}.
\]

b) \(V_n = T_n T_{n+1} \ldots T_{N-1} g_N\).

c) For all \(n\) there exist maximizers \(f_n\) of \(V_{n+1}\) with \(f_n \in \Delta_n\), and every sequence of maximizers \(f_n^*\) of \(V_{n+1}\) defines an optimal policy \((f_0^*, f_1^*, \ldots, f_{N-1}^*)\) for the \(N\)-stage MDP.
Proof

Since b) follows from a) it suffices to prove a) and c).

We show by induction on $n$ (backwards) that $V_n \in \mathbb{M}_n$ and

$$V_{n\pi^*} = T_n V_{n+1} = V_n$$

where $\pi^* = (f_0^*, \ldots, f_{N-1})$ is the policy generated by the maximizers of $V_1, \ldots, V_N$ and $f_n^* \in \Delta_n$.

We know $V_N = g_N \in \mathbb{M}_N$ by (SA$_N$) (i).

Suppose the statement is true for $N-1, \ldots, n+1$. Since $V_k \in \mathbb{M}_k$ for $k = N, \ldots, n+1$, the maximizers $f_n^*, \ldots, f_{N-1}^*$ exist and we obtain

$$V_{n\pi^*} = T_{nf_n^*} V_{n+1,\pi^*} = T_{nf_n^*} V_{n+1} = T_n V_{n+1}.$$

Hence $V_n \geq T_n V_{n+1}$. 
Proof

- On the other hand for an arbitrary policy \( \pi \)

\[
V_{n\pi} = T_{nf_n} V_{n+1, \pi} \leq T_{nf_n} V_{n+1} \leq T_n V_{n+1}.
\]

- Taking the supremum over all policies yields \( V_n \leq T_n V_{n+1} \).

- Altogether it follows that

\[
V_{n\pi^*} = T_n V_{n+1} = V_n
\]

and in view of (SA\(_N\)), \( V_n \in \mathbb{M}_n \).
Backward Induction Algorithm

1. Set $n = N$ and for $x \in E$:

$$V_N(x) := g_N(x).$$

2. Set $n := n - 1$ and compute for all $x \in E$

$$V_n(x) = \sup_{a \in D_n(x)} \left\{ r_n(x, a) + \int V_{n+1}(x') Q_n(dx'|x, a) \right\}.$$

Compute a maximizer $f^*_n$ of $V_{n+1}$.

3. If $n = 0$, then the value function $V_0$ is computed and the optimal policy $\pi^*$ is given by $\pi^* = (f^*_0, \ldots, f^*_N)$. Otherwise, go to step 2.
Examples
Example: Consumption Problem

First suppose that \((SA_N)\) is satisfied and we can apply the algorithm. We obtain \(V_N(x) = \log x\) and

\[
V_{N-1}(x) = T_{N-1} V_N(x) = \sup_{a \in [0,x]} \left\{ \log a + \log(x - a) + \mathbb{E} \log Z \right\}
\]

\[
= 2 \log x + 2 \log 0.5 + \mathbb{E} \log Z
\]

where the maximizer is given by \(f^*_{N-1}(x) = 0.5x\).

By induction:

\[
V_n(x) = (N - n + 1) \log x + d_n, \quad 0 \leq n \leq N
\]

where \(d_N = 0\) and

\[
d_n = d_{n+1} + (N-n) \mathbb{E} \log Z - \log (N - n + 1) + (N-n) \log \left( \frac{N-n}{N-n+1} \right),
\]

and the maximizer is \(f^*_n(x) = \frac{1}{N-n+1} x\).
Example: Consumption Problem

Finally it remains to show that \((SA_N)\) is satisfied. This can be verified by choosing

\[ \mathcal{M}_n := \{ v \in \mathcal{M}(E) \mid v(x) = b \log x + d \text{ for constants } b, d \in \mathbb{R} \} \]

\[ \Delta_n := \{ f \in F_n \mid f(x) = cx \text{ for } c \in \mathbb{R} \}. \]

The necessary calculations are the same as we have performed before.
Stationary MDP

Suppose the data does not depend on $n$ and is given by $(E, A, D, Q, r_n, g_N)$ with $r_n := \beta^n r$, $g_N := \beta^N g$ and $\beta \in (0, 1]$. The reward over $n$ stages under a policy $\pi \in F^n$ is given by

$$J_{n\pi}(x) := \mathbb{E}_x^n \left[ \sum_{k=0}^{n-1} \beta^k r(X_k, f_k(X_k)) + \beta^ng(X_n) \right], \quad x \in E.$$ 

The maximal expected discounted reward over $n$ stages is

$$J_0(x) := g(x)$$

$$J_n(x) := \sup_{\pi \in F^n} J_{n\pi}(x), \quad x \in E, \ 1 \leq n \leq N.$$
There exist sets $\mathcal{M} \subset \mathcal{M}(E)$ and $\Delta \subset \mathcal{F}$ such that:

(i) $g \in \mathcal{M}$.

(ii) If $v \in \mathcal{M}$ then $Tv(x)$ is well-defined and $Tv \in \mathcal{M}$.

(iii) For all $v \in \mathcal{M}$ there exists a maximizer $f \in \Delta$ of $v$, i.e.

$$T_f v(x) = Tv(x), \quad x \in E.$$
**Stationary MDP: Main Theorem**

**Theorem (Structure Theorem)**

Let \((SA_N)\) be satisfied.

a) Then \(J_n \in \mathbb{M}\) and the Bellman equation \(J_n = TJ_{n-1}\) holds:

\[
J_0(x) = g(x) \\
J_n(x) = \sup_{a \in D(x)} \left\{ r(x, a) + \beta \int J_{n-1}(x')Q(dx'|x, a) \right\}.
\]

Moreover, \(J_n = T^n g\).

b) For all \(n\) there exist maximizers \(f_n^*\) of \(J_{n-1}\) with \(f_n^* \in \Delta\), and every sequence of maximizers \(f_n^*\) of \(J_{n-1}\) defines an optimal policy \((f_N^*, \ldots, f_1^*)\) for the stationary \(N\)-stage MDP.