RL for MDP
Chap. 1 : Optimal control of MDP

Idris Kharroubi

LPSM
Sorbonne Université

DUFE

Outline

Deterministic case

Consider a "system" evolving through discrete time.
Time represented by the variable $n = 0, 1, \ldots$.
Sate of the system at time n described by x_n in E.
This system is said to be dynamic if there exists fonctions $f_n : E \to E$ such that

$$x_{n+1} = f_n(x_n).$$

Examples

- mecanic
- populations
- assets

Controlled case

Suppose that we can act on the dynamics of this system.
Then

$$x_{n+1} = f_n(x_n, u_n)$$

where $u_n \in C$ is the control chosen at time n.

Extension to the random case ?

Markov chains

Fix a measurable space $(E, \mathcal{E})$.

Definition

A probability transition on E is a family $(P(x, A))_{(x,A)\in E\times\mathcal{E}}$ such that

- $A \mapsto P(x, A)$ is a probability measure on $\mathcal{E}$ for all $x \in E$.
- $x \mapsto P(x, A)$ is a measurable for all $A \in \mathcal{E}$

We denote by $P(x, dy)$ the measure $A \mapsto P(x, A)$.

Definition

A process $(X_n)_{n\geqslant 0}$ is a Markov chain with transitions $(P_n)_{n\geqslant 0}$ if

$$\mathbb{P}(X_{n+1} \in A | X_0, \ldots, X_n) \quad = \quad P_n(X_n, A)$$

for any $A \in \mathcal{E}$.

A Markov chain $(X_n)_n$ is said homogenous if its transition P does not depend on n.

Example of a MC

Fix $(\varepsilon_n)_n$ r.v. valued in some set W.

$(X_n)_n$, process valued in E defined by X_0 and the following dynamics

$$X_{n+1} \quad = \quad \varphi_n(X_n, \varepsilon_n), \ \forall n \in \mathbb{N} \tag{1}$$

where $\varphi_n : E \times W \to E$.

Suppose that the r.v. $(\varepsilon_n)_n$ are IID with law μ and indep of X_0

Then X_n is a Markov chain with transition

$$P_n(x, A) = \mathbb{P}(\varphi_n(x, \varepsilon_n) \in A) = \mu\{w \in W; \varphi_n(x, w) \in A\}. \tag{2}$$

for all $(x, A) \in E \times \mathcal{E}$.

Functional description

Introduce the following notation: if P is a probability transition on E and $f : E \to \mathbb{R}^+$ measurable, we set

$$Pf(x) \quad = \quad \int_E f(y)\, P(x, dy),$$

for all $x \in E$.

Proposition

$(X_n)_n$ is a MC with transition $(P_n)_n$ iff

$$\mathbb{E}(f(X_{n+1})|X_0, \ldots X_n) = P_n f(X_n),$$

for any measurable $f : E \mapsto \mathbb{R}^+$ and $n \in \mathbb{N}$.

Proof. Holds for $f = \mathbb{1}_A$ with $A \subset E$ by definition. By linerarity,holds for $f = \sum_{k=1}^n a_k \mathbb{1}_{A_k}$. Then pass to the limit. $\square$

Outline

Control and Controlled transition

Fix a space of controls C and a σ-algebra $\mathcal{C}$ on C.

Definition

A controlled probability transition on E is a family $(P^u(x,A))_{(x,A,u)\in E\times\mathcal{E}\times C)}$ such that

- $(P^u(x,A))_{(x,A)\in E\times\mathcal{E}}$ is a probability transition on E for all $u\in U$,
- $(x,u)\mapsto P^u(x,A)$ is a measurable for all $A\in\mathcal{E}$.

We now define the notion of feedback (randomized) control.

Definition

- A control is a sequence $\pi=(a_n)_n$ of measurable functions from E to C.
- A randomized control is a sequence $\pi=(a_n)_n$ of measurable functions from E to $\mathcal{P}(C)$ the set of probability measures on C.

Constraints on the control

We may assume that there are some constraints on the choice of the control.

In this case, we consider that for a position x at time n the constraint is represented by a subset $D_n(x)$ of C.

We then denote by F_n the set of measurable functions $f_n : \ E \to C$ such that $f_n(x) \in D_n(x)$ for all $x \in E$.

Controlled process

Given a control $(a_n)_n$ and a controlled probability transition P, we define the controlled process as a process $(X_n)_n$ satisfying

$$\mathbb{E}(f(X_{n+1})|X_0, \ldots, X_n) \quad = \quad P_n^{a_n(X_n)} f(X_n)$$

for any $n \geqslant 0$ and any bounded (or nonnegative) measurable function f.

We notice that

- the law of such a process is unique,
- the controlled process is a Markov chain.

In the sequel, we shall denote $\alpha_n = a_n(X_n)$

Optimal control problem

We fix

- a terminal time N,
- N intermediary reward functions $r_0, \ldots, r_{N-1} : \ E \times C \to \mathbb{R}$,
- a terminal reward function $g : \ E \to \mathbb{R}$.

Our goal is to compute the function V defined by

$$V(x) \quad = \quad \sup_{\pi} V_{\pi}(x)$$

where

$$V_{\pi}(x) \quad = \quad \mathbb{E}_{x,\pi}\Big[\sum_{n=0}^{N-1} r_n(X_n, \alpha_n) + g(X_N)\Big]$$

$\mathbb{E}_{x,\pi}$ stands for the expectation operator under the condition $X_0 = x$ and the strategy π.

Outline

We introduce the functions $V_{0,\pi}, \dots V_{N,\pi}$ defined by

$$V_{n,\pi}(x) \quad = \quad \mathbb{E}^{\pi}_{n,x}\Big[\sum_{k=n}^{N-1} r_k(X_k, \alpha_k) + g(X_N)\Big]$$

and

$$V_n(x) \quad = \quad \sup_{\pi} V_{n,\pi}(x)$$

for $n = 0, \dots, N$ and $x \in E$.

Observe that $V_N = V_{N,\pi} = g$.

Reward operators

We first introduce

$$\mathbb{M}(E) \quad := \quad \{v : \; E \to [-\infty, \infty) \mid v \text{ is measurable}\}.$$

We have $V_{n,\pi} \in \mathbb{M}(E)$ for all π and n. We do not know about V_n.

We next define the following operators for $n = 0, 1, \ldots, N-1$.

- For $v \in \mathbb{M}(E)$ define

$$(L_n v)(x,a) \quad := \quad r_n(x,a) + \int v(y) P_n(dy|x,a), \quad (x,a) \in E \times A$$

  whenever the integral exists.

- For $v \in \mathbb{M}(E)$ and $f \in F_n$ define

$$(T_{n,f} v)(x) \quad := \quad (L_n v)(x, f(x)), \quad x \in E.$$

- For $v \in \mathbb{M}(E)$ define

$$(T_n v)(x) \quad := \quad \sup_{a \in A} (L_n v)(x,a), \quad x \in E.$$

  T_n is called the maximal reward operator at time n.

Monotony of the reward operator

Lemma

The operators L_n, $T_{n,f}$ and T_n are monotone, that is for $v, w \in \mathbb{M}(E)$ such that $v(x) \leqslant w(x)$ for all $x \in E$ we have

(i) $L_n v(x,a) \leqslant L_n w(x,a)$ *for all* x, a,

(ii) $T_{n,f} v(x) \leqslant T_{n,f} w(x)$ *for all* x, f,

(iii) $T_n v(x) \leqslant T_n w(x)$ *for all* x.

Proof. Suppose that $v(x) \leqslant w(x)$ for all $x \in E$. Then

$$\int v(y) P_n(dy|x,a) \quad \leqslant \quad \int w(y) P_n(dy|x,a).$$

Thus, $L_n v(x,a) \leqslant L_n w(x,a)$ which implies the first and second statement. Taking the supremum over all $a \in C$ implies the third statement.

Theorem (Reward Iteration)

Let $\pi = (f_0, \dots, f_{N-1})$ *be an N-stage policy. We have*

$$V_{N,\pi} \quad = \quad g_N$$

and

$$V_{n,\pi} \quad = \quad T_{n,f_n} V_{n+1,\pi}$$

for $n = 0, \dots, N-1$*. In particular, we have*

$$V_{n,\pi} \quad = \quad T_{n,f_n} \cdots T_{N-1,f_{N-1}} g_N.$$

for $n = 0, \dots, N-1$*.*

Proof of the Reward Iteration

For $x \in E$ we have

$$
\begin{aligned}
V_{n,\pi}(x) &= \mathbb{E}^{\pi}_{n,x}\Big[\sum_{k=n}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N)\Big] \\
&= r_n(x, f_n(x)) + \mathbb{E}^{\pi}_{n,x}\Big[\mathbb{E}^{\pi}_{n,x}\Big[\sum_{k=n+1}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N) | X_{n+1}\Big]\Big] \\
&= r_n(x, f_n(x)) \\
&\quad + \int \mathbb{E}^{\pi}_{n+1,y}\Big[\sum_{k=n+1}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N)\Big] P_n(dy|x, f_n(x)) \\
&= r_n(x, f_n(x)) + \int V_{n+1,\pi}(y) P_n(dy|x, f_n(x)) \quad = \quad T_{n,f_n} V_{n+1,\pi}(x)\ .
\end{aligned}
$$

The rest of the proof follows by a backward induction.

Verification

Natural approach to find optimality: at each time $n = 0, \dots, N-1$ find a maximizer of V_{n+1}.
This leads to consider the so-called Bellman equations given by

$$\begin{aligned} V_N &= g_N & (3)\\ V_n &= T_n V_{n+1}\ , \quad n = 0, \dots, N-1\ . & (4) \end{aligned}$$

Theorem (Verification Theorem)

Let $(v_n)_{n=0}^N \in \mathbb{M}(E)^{N+1}$ be a solution of Bellman equations (3)-(4).

(i) *We have $v_n \geqslant V_n$ for $n = 0, 1, \dots, N$.*

(ii) *If f_n^* is a maximizer of v_{n+1} for $n = 0, 1, \dots, N-1$, then $v_n = V_n$ and the policy $\pi^* = (f_0^*, f_1^*, \dots, f_{N-1}^*)$ is optimal for the N-stage Markov Decision Problem.*

We prove both assertions by a backward induction on $n = 0, \dots, N$.

(i) For $n = N$ we have $v_N = g_N = V_N$. Suppose $v_{n+1} \geqslant V_{n+1}$, then for all $\pi = (f_0, \dots, f_{N-1})$ we have

$$v_n \quad = \quad T_n v_{n+1} \quad \geqslant \quad T_n V_{n+1} \quad \geqslant \quad T_{nf_n} V_{n+1,\pi} \quad = \quad V_{n,\pi}.$$

(ii) We show recursively that $v_n = V_n = V_{n,\pi^*}$. For $n = N$ this is obvious. Suppose the statement is true for $n+1$, then

$$V_n \quad \leqslant \quad v_n \quad = \quad T_{n,f_n^*} v_{n+1} \quad = \quad T_{n,f_n^*} V_{n+1} \quad = \quad V_{n,\pi^*} \quad \leqslant \quad V_n \; .$$

Outline

Structure Assumption

In general existence of an optimal policy is not guaranteed.
Need to make further assumptions about the structure of the problem

Assumption (Structure Assumption)

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* $f_n \in \Delta_n$.

Rk: if E and C are finite, Structure Assumption is satisfied.

Structure Theorem

Theorem (Structure Theorem)

Suppose that Structure Assumption holds.

(i) $V_n \in \mathbb{M}_n$ *and the sequence* $(V_n)_{n=0}^N$ *satisfies the Bellman equations* (3)-(4) *i.e.*

$$\begin{aligned} V_N &= g_N \\ V_n &= \sup_{a \in A} \left\{ r_n(x,a) + \int V_{n+1}(y) P_n(dy|x,a) \right\}, \quad n = 0, \ldots, N-1 . \end{aligned}$$

(ii) $V_n = T_n T_{n+1} \ldots T_{N-1} g_N$ *for* $n = 0, \ldots, N-1$.

(iii) *For* $n = 0, 1, \ldots, N-1$ *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 Markov Decision Problem.*

Proof Structure Theorem

Since (ii) follows directly from (i) it suffices to prove (i) and (ii).

Proof by backward induction on $n = N-1, \dots, 0$ that $V_n \in \mathbb{M}_n$ and that

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

where $\pi^* = (f_0^*, f_1^*, \dots, f_{N-1}^*)$ is the policy generated by the maximizers of $V_1, \dots, V_N$ and $f_n^* \in \Delta_n$.

Corollary

Let Structure Assumption hold. For $0 \leqslant n \leqslant m \leqslant N$ *we have*

$$V_n(x) \quad = \quad \sup_{\pi} \mathbb{E}^{\pi}_{n,x}\Big[\sum_{k=n}^{m-1} r_k(X_k, f_k(X_k)) + V_m(X_m)\Big]$$

for all $x \in E$.

Theorem (Dynamic programming principle)

Suppose that Assumption 9 holds. Then we have

$$V_{n,\pi^*}(x) = V_n(x) \quad \Rightarrow \quad V_{m,\pi^*}(X_m) = V_m(X_m) \qquad \mathbb{P}^{\pi^*}_{n,x} - a.s.,$$

for $x \in E$ and $0 \leqslant n \leqslant m \leqslant N$, i.e. if $(f_n^*, \ldots, f_{N-1}^*)$ is optimal for the time period $[n, N]$ then $(f_m^*, \ldots, f_{N-1}^*)$ is optimal for $[m, N]$.

Proof. From **Reward Iteration** Theorem and the definition of V_m we have

$$\begin{aligned} V_n(x) \quad &= \quad V_{n,\pi^*}(x) \quad = \quad T_{nf_n^*} \ldots T_{m-1,f_{m-1}^*} V_{m,\pi^*}(x) \\ &= \quad \mathbb{E}^{\pi^*}_{n,x}\Big[\sum_{k=n}^{m-1} r_k(X_k, \alpha_k) + V_{m,\pi^*}(X_m)\Big] \\ &\leqslant \quad \mathbb{E}^{\pi^*}_{n,x}\Big[\sum_{k=n}^{m-1} r_k(X_k, \alpha_k) + V_m(X_m)\Big] \quad \leqslant \quad V_n(x) \end{aligned}$$

where we have used Previous Corollary for the last inequality. This implies

$$\mathbb{E}^{\pi^*}_{n,x}\Big[V_{m,\pi^*}(X_m) - V_m(X_m)\Big] \quad = \quad 0$$

which gives the result.

Example: consumption problem

An investor decides at each time $0, \ldots, N-1$, how much of the capital is consumed and how much is invested into a risky asset.

The consumed amount is evaluated by a utility function U as well as the terminal wealth.

The remaining capital is invested into a risky asset

The state x of the system is here the available capital.
The action $a = f(x)$ is the consumed amount of money, constraint $0 \leqslant a \leqslant x$.

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

$$\mathbb{E}_x^\pi\Big[\sum_{k=0}^{N-1} U\big(f_k(X_k)\big) + U(X_N)\Big]$$

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

Denote by Z_{n+1} the random return of our risky asset over period $[n, n+1)$.

We suppose that $Z_1, \ldots, Z_N$ are non-negative, independent L^1 random variables.

The utility function U is increasing and concave. Thus

$$U(x) \quad \leqslant \quad c_1 + c_2 x \,, \quad x \in \mathbb{R} \,.$$

Then

$$\begin{aligned} V_n(x)^+ \quad & \leqslant \quad \sup \mathbb{E}^{\pi}_{nx}\Big[\sum_{k=n}^{N-1} U^+(f_k(X_k)) + U^+(X_N) \Big] \\ & \leqslant \quad (N-n+1)c_1 + xc_2 \sum_{k=n}^{N} \mathbb{E}[Z_n] \ldots \mathbb{E}[Z_k] \quad < \quad +\infty, \end{aligned}$$

Now let us assume that $U(x) := \log(x)$ for all n and $g_N(x) := \log(x)$.

Moreover, we assume that the return distribution does not depend on n and has finite expectation $\mathbb{E}[Z]$.

Suppose that Structure Assumption is satisfied. We then apply Structure Theorem, and we get $V_N(x) = \log(x)$ and

$$\begin{aligned} V_{N-1}(x) &= T_{N-1}V_N(x) \\ &= \sup_{a\in[0,x]}\Big\{\log(a)+\log(x-a)+\mathbb{E}(\log(Z))\Big\} \\ &= 2\log(x) - 2\log(2) + \mathbb{E}(\log(Z)) \end{aligned}$$

where the maximizer is given by $f^*_{N-1}(x) = \frac{1}{2}x$. Now by induction we obtain

$$V_n(x) = (N-n+1)\log(x) + d_n, \quad 0 \leqslant n \leqslant N$$

where $d_N = 0$ and

$$d_n = d_{n+1} + (N-n)\mathbb{E}(\log(Z)) - \log(N-n+1) + (N-n)\log\Big(\frac{N-n}{N-n+1}\Big),$$

and we obtain the maximizer $f^*_n(x) = \frac{1}{N-n+1}x$.

Finally we notice that Structure Assumption is satisfied by taking

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

and

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