4. Dynamic optimization for non-negative rewards

(also called positive programming)

We are given

$$P : S \times A \longrightarrow \mathrm{Prob}(S)$$

and consider the class of controls $u : S^* \to A$ from §2.
There is also given a reward function

$$r : S \times A \longrightarrow \mathbb{R}^+$$

Define

$$V^u(x) = \mathbb{E}_x^u \sum_{n=0}^{\infty} r(X_n, U_n), \qquad V(x) = \sup_u V^u(x).$$

<u>Notes</u> Taking $c = -r$, this V^u and V are minus the corresponding objects in §2. Since P and r are time-homogeneous, we consider only starting at time $k = 0$.

Value iteration

Consider

$$V_n^u(x) = \mathbb{E}_x^u \sum_{k=0}^{n-1} r(X_k, U_k), \qquad V_n(x) = \sup_u V_n^u(x).$$

Since $r \geq 0$, by monotone convergence, $V_n^u(x) \uparrow V^u(x)$ as $n \to \infty$ for all x and u. So

$$V(x) = \sup_u \sup_n V_n^u(x) = \sup_n \sup_u V_n^u(x) = \sup_n V_n(x).$$

In §3 we showed that

$$V_0(x) \equiv 0, \quad V_{n+1}(x) = \sup_a (r + PV_n)(x,a), \quad x \in S.$$

So this <u>value iteration scheme</u> provides a computation of V.

Proposition 4.1

The optimal reward function V is the minimal non-negative solution of the dynamic optimality equation

$$V(x) = \sup_a (r + PV)(x,a), \quad x \in S.$$

Hence, any control u, for which V^u also satisfies this equation, is optimal, for all starting states x.

Proof By Proposition 2.1, V is a solution. Suppose $F \geq 0$ is another. Note that $F \geq V_0 = 0$ and suppose inductively that $F \geq V_n$. Then

$$F(x) = \sup_a (r + PF)(x,a) \geq \sup_a (r + PV_n)(x,a) = V_{n+1}(x),$$

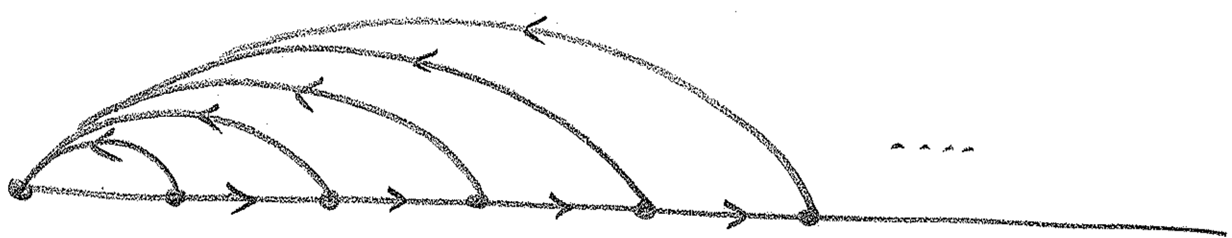
so the induction proceeds. Hence $F \geq \sup_n V_n = V$. $\square$

Example - showing that an optimal control may not exist

Take $S = \mathbb{Z}^+$, $A = \{0, 1\}$

$$f(x,a) = \begin{cases} ax & \text{if } x = 0, \\ a(x+1) & \text{if } x \geq 1, \end{cases} \qquad r(x,a) = (1-a)\left(1 - \frac{1}{x}\right).$$

Thus, in any state $x \geq 1$, we can choose to jump to $x+1$, or to jump to 0 gaining reward $1 - 1/x$. Once at 0, no further reward is gained.



There is no optimal control — why?

This throws some light on the theory

– what it does and does not say.

The optimality equations are

$$V(0) = 0, \qquad V(x) = \max\{1 - \tfrac{1}{x},\ V(x+1)\}, \qquad x \geq 1.$$

For any $\lambda \in [1, \infty)$, $V_\lambda(x) = \lambda 1_{\{x \geq 1\}}$ is a solution,

and these are all the solutions.

By Proposition 4.1, the optimal reward function $V = V_1$,

the minimal (non-negative) solution.

However, the choice of maximizing action in each state

gives $u \equiv 0$ for which $V^u \equiv 0$: it is always

better to wait, but if you wait forever....

Example – prudent gambling

You have £1 and wish to increase this to £N. You can place bets on a sequence of favorable games, each, independently, having a probability $p > \frac{1}{2}$ of success. Your stake must be a whole number of pounds, no greater than your current wealth.

How do you maximize your probability of reaching £N?