

4 Dynamic optimization for non-negative rewards

We show how to optimize a time-homogeneous stochastic controllable dynamical system with non-negative rewards over an infinite time-horizon¹⁵.

Let P be a time-homogeneous stochastic controllable dynamical system with state-space S and action-space A . Suppose given a *reward function*

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

Given a control u , define the *expected total reward function*

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

where, as usual, the notation signifies that $(X_n)_{n \geq 0}$ is the controlled process of u , starting from x , and where $U_n = u_n(X_0, \dots, X_n)$. Define also the *optimal reward* or *value function*

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

We are using notation inconsistent with Section 2 because we have defined V as the negative of the corresponding object in Section 2. The optimality equation transforms straightforwardly under this change of notation – one just replaces the infimum by a supremum.

Define for $n \geq 0$

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

By monotone convergence¹⁶, since $r \geq 0$, $V_n^u(x) \uparrow V^u(x)$ as $n \rightarrow \infty$, for all x and u . So

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

The functions V_n are finite-horizon optimal reward functions, which, taking advantage of time-homogeneity, can be computed iteratively using the optimality equation

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

so we have a way to compute V . This procedure is called *value iteration*.

Proposition 4.1. *The optimal reward function 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 .

¹⁵This is also called positive programming.

¹⁶This fundamental result of measure theory states that, for any sequence of measurable functions $0 \leq f_n \uparrow f$, and any measure μ , we have convergence of the corresponding integrals $\mu(f_n) \uparrow \mu(f)$. Here, it justifies the interchange of summation and expectation, for the expectation is a form of integral, and we just take f_n as the partial sum $\sum_{k=0}^{n-1} r(X_k, U_k)$.

Proof. We know that V satisfies the optimality equation by Proposition 2.1. Suppose F is another non-negative solution. Then $F \geq 0 = V_0$. Suppose inductively for $n \geq 0$ 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 $F \geq \sup_n V_n = V$. □

Example (Possible lack of an optimal policy). Consider the controllable dynamical system $f(x, a) = a(x + 1_{\{x \geq 1\}})$, with state-space \mathbb{Z}^+ and action-space $\{0, 1\}$. Take as reward function $r(x, a) = (1 - a)(1 - 1/x)$. Thus, in state $x \geq 1$, we can choose to jump up by 1, or to jump to 0, gaining a reward of $1 - 1/x$. Once at 0, no further reward is gained.

The optimality equations are given by $V(0) = 0$ and

$$V(x) = \max\{1 - 1/x, V(x + 1)\}, \quad x \geq 1.$$

It is straightforward to show that, for any $\lambda \in [1, \infty)$, the function V_λ , given by

$$V_\lambda(x) = \lambda 1_{\{x \geq 1\}},$$

is a solution of the optimality equations, and indeed that there are no other solutions. Then, by the proposition, we can identify the optimal reward function as the smallest of these functions, namely V_1 . *However, there is no optimal control.* If we wait until we get to n , then we gain a reward $1 - 1/n$. But if we wait for ever, we get nothing. Note that waiting forever corresponds to the control $u(x) = 0$ for all x , which has the property that

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

So we see, contrary to the finite-horizon case, that this is not enough to guarantee optimality. We do have for this control that

$$V^u(x) = (r + PV^u)(x, u(x)), \quad x \in S.$$

However, V^u , which is the minimal non-negative solution of *this* equation, is identically zero.

Example (Optimal gambling). A gambler has one pound and wishes to increase it to N pounds. She can place bets on a sequence of favorable games, each independently having probability $p > 1/2$ of success, but her stake must be a whole number of pounds and may not exceed her current fortune. What strategy maximizes her chances of reaching her goal?

We take as state-space $S = \mathbb{Z}^+$. It is natural here to allow a state-dependent action-space¹⁷ $A_x = \{0, 1, \dots, x\}$. The optimality equations are given by

$$V(x) = \max_{a \in A_x} \{pV(x + a) + (1 - p)V(x - a)\}, \quad 1 \leq x \leq N - 1,$$

with $V(0) = 0$ and $V(x) = 1$ for all $x \geq N$. There is no systematic approach to solving these equations, so we guess that the timid strategy of betting one pound each time will

¹⁷See footnote 1.

be optimal. As motivation, we might compare the outcomes, firstly of betting two pounds once, and secondly, of successively betting one pound until we are either two up or two down. So, take $u(x) = 1$ for all x . Then, by a standard Markov chain argument,

$$V^u(x) = pV^u(x+1) + (1-p)V^u(x-1), \quad 1 \leq x \leq N-1,$$

with $V^u(0) = 1$ and $V^u(N) = 1$. These equations have *unique* solution

$$V^u(x) = (1 - \lambda^x)/(1 - \lambda^N),$$

where $\lambda = (1-p)/p \in (0, 1)$. It now follows from the fact that V^u is concave that it satisfies the optimality equations too, so u is optimal.