3 Dynamic Programming over the Infinite Horizon

We define the cases of discounted, negative and positive dynamic programming and establish the validity of the optimality equation for an infinite horizon problem.

3.1 Discounted costs

For a discount factor, $\beta \in (0, 1]$, the **discounted-cost criterion** is defined as

$$\mathbf{C} = \sum_{t=0}^{h-1} \beta^t c(x_t, u_t, t) + \beta^h \mathbf{C}_h(x_h)\,. \tag{3.1}$$

This simplifies things mathematically, particularly when we want to consider an infinite horizon. If costs are uniformly bounded, say $|c(x, u)| < B$, and discounting is strict ($\beta < 1$) then the infinite horizon cost is bounded by $B/(1-\beta)$. In economic language, if there is an interest rate of $r\%$ per unit time, then a unit amount of money at time t is worth $\rho = 1 + r/100$ at time $t+1$. Equivalently, a unit amount at time $t+1$ has present value $\beta = 1/\rho$. The function, $F(x, t)$, which expresses the minimal present value at time t of expected-cost from time t up to h is

$$F(x,t) = \inf_{u_t,\ldots,u_{h-1}} E\left[\sum_{\tau=t}^{h-1} \beta^{\tau-t} c(x_\tau, u_\tau, \tau) + \beta^{h-t}\mathbf{C}_h(x_h) \,\middle|\, x_t = x\right]. \tag{3.2}$$

The DP equation is now

$$F(x,t) = \inf_u \left[c(x,u,t) + \beta E F(a(x,u,t), t+1)\right], \quad t < h\,, \tag{3.3}$$

where $F(x, h) = \mathbf{C}_h(x)$.

3.2 Example: job scheduling

A collection of n jobs is to be processed in arbitrary order by a single machine. Job i has processing time p_i and when it completes a reward r_i is obtained. Find the order of processing that maximizes the sum of the discounted rewards.

Solution. Here we take 'time k' as the point at which the $n-k$ th job has just been completed and the state at time k as the collection of uncompleted jobs, say S_k. The dynamic programming equation is

$$F_k(S_k) = \max_{i \in S_k} [r_i \beta^{p_i} + \beta^{p_i} F_{k-1}(S_k - \{i\})]\,.$$

Obviously $F_0(\emptyset) = 0$. Applying the method of dynamic programming we first find $F_1(\{i\}) = r_i\beta^{p_i}$. Then, working backwards, we find

$$F_2(\{i,j\}) \;=\; \max[r_i\beta^{p_i} + \beta^{p_i+p_j} r_j\,,\; r_j\beta^{p_j} + \beta^{p_j+p_i} r_i]\,.$$

There will be 2^n equations to evaluate, but with perseverance we can determine $F_n(\{1, 2, \ldots, n\})$. However, there is a simpler way.

An interchange argument. Suppose that jobs are scheduled in the order $i_1, \ldots, i_k, i, j, i_{k+3}, \ldots, i_n$. Compare the reward of this schedule to one in which the order of jobs i and j are reversed: $i_1, \ldots, i_k, j, i, i_{k+3}, \ldots, i_n$. The rewards under the two schedules are respectively

$$R_1 + \beta^{T+p_i} r_i + \beta^{T+p_i+p_j} r_j + R_2 \quad \text{and} \quad R_1 + \beta^{T+p_j} r_j + \beta^{T+p_j+p_i} r_i + R_2\,,$$

where $T = p_{i_1} + \cdots + p_{i_k}$, and R_1 and R_2 are respectively the sum of the rewards due to the jobs coming before and after jobs i, j; these are the same under both schedules. The reward of the first schedule is greater if $r_i\beta^{p_i}/(1-\beta^{p_i}) > r_j\beta^{p_j}/(1-\beta^{p_j})$. Hence a schedule can be optimal only if the jobs are taken in decreasing order of the indices $r_i\beta^{p_i}/(1-\beta^{p_i})$. This type of reasoning is known as an **interchange argument**.

There are a couple points to note. (i) An interchange argument can be useful for solving a decision problem about a system that evolves in stages. Although such problems can be solved by dynamic programming, an interchange argument – when it works – is usually easier. (ii) The decision points need not be equally spaced in time. Here they are the points at which a number of jobs have been completed.

3.3 The infinite-horizon case

In the finite-horizon case the cost function is obtained simply from (3.3) by the backward recursion from the terminal point. However, when the horizon is infinite there is no terminal point and so the validity of the optimality equation is no longer obvious.

Let us consider the time-homogeneous Markov case, in which costs and dynamics do not depend on t, i.e., $c(x, u, t) = c(x, u)$. Suppose also that there is no terminal cost, i.e., $\mathbf{C}_h(x) = 0$. Define the *s-horizon cost under policy* π as

$$F_s(\pi, x) = E_\pi\left[\sum_{t=0}^{s-1} \beta^t c(x_t, u_t) \,\middle|\, x_0 = x\right],$$

where E_π denotes expectation over the path of the process under policy π. If we take the infimum with respect to π we have the *infimal s-horizon cost*

$$F_s(x) = \inf_\pi F_s(\pi, x).$$

Clearly, this always exists and satisfies the optimality equation

$$F_s(x) = \inf_u \left\{c(x,u) + \beta E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\right\}, \tag{3.4}$$

with terminal condition $F_0(x) = 0$.

The *infinite-horizon cost under policy* π is also quite naturally defined as

$$F(\pi, x) = \lim_{s\to\infty} F_s(\pi, x). \tag{3.5}$$

This limit need not exist, but it will do so under any of the following scenarios.

D (**discounted programming**): $0 < \beta < 1$, and $|c(x,u)| < B$ for all x, u.

N (**negative programming**): $0 < \beta \leq 1$ and $c(x,u) \geq 0$ for all x, u.

P (**positive programming**): $0 < \beta \leq 1$ and $c(x,u) \leq 0$ for all x, u.

Notice that the names 'negative' and 'positive' appear to be the wrong way around with respect to the sign of $c(x,u)$. However, the names make sense if we think of equivalent problems of maximizing rewards. Maximizing positive rewards (P) is the same thing as minimizing negative costs. Maximizing negative rewards (N) is the same thing as minimizing positive costs. In cases N and P we usually take $\beta = 1$.

The existence of the limit (possibly infinite) in (3.5) is assured in cases N and P by monotone convergence, and in case D because the total cost occurring after the sth step is bounded by $\beta^s B/(1-\beta)$.

3.4 The optimality equation in the infinite-horizon case

The *infimal infinite-horizon cost* is defined as

$$F(x) = \inf_{\pi} F(\pi, x) = \inf_{\pi} \lim_{s\to\infty} F_s(\pi, x)\,. \tag{3.6}$$

The following theorem justifies our writing an optimality equation.

Theorem 3.1 *Suppose D, N, or P holds. Then $F(x)$ satisfies the optimality equation*

$$F(x) = \inf_u \{c(x,u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u)]\}\,. \tag{3.7}$$

Proof. We first prove that '$\geq$' holds in (3.7). Suppose π is a policy, which chooses $u_0 = u$ when $x_0 = x$. Then

$$F_s(\pi, x) = c(x,u) + \beta E[F_{s-1}(\pi, x_1) \mid x_0 = x, u_0 = u]\,. \tag{3.8}$$

Either D, N or P is sufficient to allow us to takes limits on both sides of (3.8) and interchange the order of limit and expectation. In cases N and P this is because of monotone convergence. Infinity is allowed as a possible limiting value. We obtain

$$\begin{aligned} F(\pi, x) &= c(x,u) + \beta E[F(\pi, x_1) \mid x_0 = x, u_0 = u] \\ &\geq c(x,u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u] \\ &\geq \inf_u \{c(x,u) + \beta E[F(x_1) \mid x_0 = x, u_0 = u]\}\,. \end{aligned}$$

Minimizing the left hand side over π gives '$\geq$'.

To prove '$\leq$', fix x and consider a policy π that having chosen u_0 and reached state x_1 then follows a policy π^1 which is suboptimal by less than ϵ from that point, i.e., $F(\pi^1, x_1) \leq F(x_1) + \epsilon$. Note that such a policy must exist, by definition of F, although π^1 will depend on x_1. We have

$$\begin{aligned} F(x) &\leq F(\pi, x) \\ &= c(x,u_0) + \beta E[F(\pi^1, x_1) \mid x_0 = x, u_0] \\ &\leq c(x,u_0) + \beta E[F(x_1) + \epsilon \mid x_0 = x, u_0] \\ &\leq c(x,u_0) + \beta E[F(x_1) \mid x_0 = x, u_0] + \beta\epsilon\,. \end{aligned}$$

Minimizing the right hand side over u_0 and recalling ϵ is arbitrary gives '$\leq$'. ■

3.5 Example: selling an asset

A spectulator owns a rare collection of tulip bulbs and each day has one opportunity to sell it, which he may either accept or reject. The potential sale prices are independently and identically distributed with probability density function $g(x)$, $x \geq 0$. Each day there is a probability $1-\beta$ that the market for tulip bulbs will collapse, making his bulb collection completely worthless. Find the policy that maximizes his expected return and express it as the unique root of an equation. Show that if $\beta > 1/2$, $g(x) = 2/x^3$, $x \geq 1$, then he should sell the first time the sale price is at least $\sqrt{\beta/(1-\beta)}$.

Solution. There are only two states, depending on whether he has sold the collection or not. Let these be 0 and 1 respectively. The optimality equation is

$$\begin{aligned} F(1) &= \int_{y=0}^{\infty} \max[y, \beta F(1)]\, g(y)\, dy \\ &= \beta F(1) + \int_{y=0}^{\infty} \max[y - \beta F(1), 0]\, g(y)\, dy \\ &= \beta F(1) + \int_{y=\beta F(1)}^{\infty} [y - \beta F(1)]\, g(y)\, dy \end{aligned}$$

Hence

$$(1-\beta)F(1) = \int_{y=\beta F(1)}^{\infty} [y - \beta F(1)]\, g(y)\, dy\,. \tag{3.9}$$

That this equation has a unique root, $F(1) = F^*$, follows from the fact that left and right hand sides are increasing and decreasing in $F(1)$ respectively. Thus he should sell when he can get at least βF^*. His maximal reward is F^*.

Consider the case $g(y) = 2/y^3$, $y \geq 1$. The left hand side of (3.9) is less that the right hand side at $F(1) = 1$ provided $\beta > 1/2$. In this case the root is greater than 1 and we compute it as

$$(1-\beta)F(1) = 2/\beta F(1) - \beta F(1)/[\beta F(1)]^2\,,$$

and thus $F^* = 1/\sqrt{\beta(1-\beta)}$ and $\beta F^* = \sqrt{\beta/(1-\beta)}$.

If $\beta \leq 1/2$ he should sell at any price.

Notice that discounting arises in this problem because at each stage there is a probability $1-\beta$ that a 'catastrophe' will occur that brings things to a sudden end. This characterization of a manner in which discounting can arise is often quite useful.