4 Positive Programming

We address the special theory of maximizing positive rewards, (noting that there may be no optimal policy but that if a policy has a value function that satisfies the optimality equation then it is optimal), and the method of value iteration.

4.1 Example: possible lack of an optimal policy.

Positive programming concerns minimizing non-positive costs, $c(x,u) \leq 0$. The name originates from the equivalent problem of maximizing non-negative rewards, $r(x,u) \geq 0$, and for this section we present results in that setting. The following example shows that there may be no optimal policy.

Suppose the possible states are the non-negative integers and in state x we have a choice of either moving to state $x+1$ and receiving no reward, or moving to state 0, obtaining reward $1-1/i$, and then remaining in state 0 thereafter and obtaining no further reward. The optimality equations is

$$F(x) = \max\{1-1/x, F(x+1)\} \quad x>0\,.$$

Clearly $F(x)=1$, $x>0$, but the policy that chooses the maximizing action in the optimality equation always moves on to state $x+1$ and hence has zero reward. Clearly, there is no policy that actually achieves a reward of 1.

4.2 Characterization of the optimal policy

The following theorem provides a necessary and sufficient condition for a policy to be optimal: namely, its value function must satisfy the optimality equation. This theorem also holds for the case of strict discounting and bounded costs.

Theorem 4.1 *Suppose D or P holds and π is a policy whose value function $F(\pi,x)$ satisfies the optimality equation*

$$F(\pi,x) = \sup_u\{r(x,u) + \beta E[F(\pi,x_1) \mid x_0=x, u_0=u]\}\,.$$

Then π is optimal.

Proof. Let π' be any policy and suppose it takes $u_t(x) = f_t(x)$. Since $F(\pi,x)$ satisfies the optimality equation,

$$F(\pi,x) \geq r(x,f_0(x)) + \beta E_{\pi'}[F(\pi,x_1) \mid x_0=x, u_0=f_0(x)]\,.$$

By repeated substitution of this into itself, we find

$$F(\pi,x) \geq E_{\pi'}\left[\sum_{t=0}^{s-1}\beta^t r(x_t,u_t)\,\middle|\, x_0=x\right] + \beta^s E_{\pi'}[F(\pi,x_s) \mid x_0=x]\,. \qquad (4.1)$$

In case P we can drop the final term on the right hand side of (4.1) (because it is non-negative) and then let $s\to\infty$; in case D we can let $s\to\infty$ directly, observing that this term tends to zero. Either way, we have $F(\pi,x) \geq F(\pi',x)$. ■

4.3 Example: optimal gambling

A gambler has i pounds and wants to increase this to N. At each stage she can bet any fraction of her capital, say $j \leq i$. Either she wins, with probability p, and now has $i+j$ pounds, or she loses, with probability $q=1-p$, and has $i-j$ pounds. Let the state space be $\{0,1,\ldots,N\}$. The game stops upon reaching state 0 or N. The only non-zero reward is 1, upon reaching state N. Suppose $p \geq 1/2$. Prove that the timid strategy, of always betting only 1 pound, maximizes the probability of the gambler attaining N pounds.

Solution. The optimality equation is

$$F(i) = \max_{j,j\leq i}\{pF(i+j) + qF(i-j)\}\,.$$

To show that the timid strategy is optimal we need to find its value function, say $G(i)$, and show that it is a solution to the optimality equation. We have $G(i) = pG(i+1) + qG(i-1)$, with $G(0)=0$, $G(N)=1$. This recurrence gives

$$G(i) = \begin{cases} \dfrac{1-(q/p)^i}{1-(q/p)^N} & p>1/2, \\ \dfrac{i}{N} & p=1/2. \end{cases}$$

If $p=1/2$, then $G(i)=i/N$ clearly satisfies the optimality equation. If $p>1/2$ we simply have to verify that

$$G(i) = \frac{1-(q/p)^i}{1-(q/p)^N} = \max_{j:j\leq i}\left\{\; p\left[\frac{1-(q/p)^{i+j}}{1-(q/p)^N}\right] + q\left[\frac{1-(q/p)^{i-j}}{1-(q/p)^N}\right]\;\right\}.$$

It is a simple exercise to show that $j=1$ maximizes the right hand side. ■

4.4 Value iteration

The infimal cost function F can be approximated by **successive approximation** or **value iteration**. This is important and practical method of computing F. Let us define

$$F_\infty(x) = \lim_{s\to\infty} F_s(x) = \lim_{s\to\infty}\inf_\pi F_s(\pi,x)\,. \qquad (4.2)$$

This exists (by monotone convergence under N or P, or by the fact that under D the cost incurred after time s is vanishingly small.)

Notice that (4.2) reverses the order of $\lim_{s\to\infty}$ and $\inf_\pi$ in (3.6). The following theorem states that we can interchange the order of these operations and that therefore

$F_s(x) \to F(x)$. However, in case N we need an additional assumption:

F (**finite actions**): There are only finitely many possible values of u in each state.

Theorem 4.2 *Suppose that D or P holds, or N and F hold. Then* $F_\infty(x) = F(x)$.

Proof. First we prove '$\leq$'. Given any $\bar{\pi}$,

$$F_\infty(x) = \lim_{s\to\infty} F_s(x) = \lim_{s\to\infty} \inf_\pi F_s(\pi, x) \leq \lim_{s\to\infty} F_s(\bar{\pi}, x) = F(\bar{\pi}, x).$$

Taking the infimum over $\bar{\pi}$ gives $F_\infty(x) \leq F(x)$.

Now we prove '$\geq$'. In the positive case, $c(x,u) \leq 0$, so $F_s(x) \geq F(x)$. Now let $s \to \infty$. In the discounted case, with $|c(x,u)| < B$, imagine subtracting $B > 0$ from every cost. This reduces the infinite-horizon cost under any policy by exactly $B/(1-\beta)$ and $F(x)$ and $F_\infty(x)$ also decrease by this amount. All costs are now negative, so the result we have just proved applies. [Alternatively, note that

$$F_s(x) - \beta^s B/(1-\beta) \leq F(x) \leq F_s(x) + \beta^s B/(1-\beta)$$

(can you see why?) and hence $\lim_{s\to\infty} F_s(x) = F(x)$.]

In the negative case,

$$\begin{aligned} F_\infty(x) &= \lim_{s\to\infty} \min_u \{c(x,u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\ &= \min_u \{c(x,u) + \lim_{s\to\infty} E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\} \\ &= \min_u \{c(x,u) + E[F_\infty(x_1) \mid x_0 = x, u_0 = u]\}, \end{aligned} \tag{4.3}$$

where the first equality follows because the minimum is over a finite number of terms and the second equality follows by Lebesgue monotone convergence (since $F_s(x)$ increases in s). Let π be the policy that chooses the minimizing action on the right hand side of (4.3). This implies, by substitution of (4.3) into itself, and using the fact that N implies $F_\infty \geq 0$,

$$\begin{aligned} F_\infty(x) &= E_\pi \left[\sum_{t=0}^{s-1} c(x_t, u_t) + F_\infty(x_s) \,\middle|\, x_0 = x \right] \\ &\geq E_\pi \left[\sum_{t=0}^{s-1} c(x_t, u_t) \,\middle|\, x_0 = x \right]. \end{aligned}$$

Letting $s \to \infty$ gives $F_\infty(x) \geq F(\pi, x) \geq F(x)$. ■

4.5 Example: pharmaceutical trials

A doctor has two drugs available to treat a disease. One is well-established drug and is known to work for a given patient with probability p, independently of its success for other patients. The new drug is untested and has an unknown probability of success θ, which the doctor believes to be uniformly distributed over $[0,1]$. He treats one patient per day and must choose which drug to use. Suppose he has observed s successes and f failures with the new drug. Let $F(s,f)$ be the maximal expected-discounted number of future patients who are successfully treated if he chooses between the drugs optimally from this point onwards. For example, if he uses only the established drug, the expected-discounted number of patients successfully treated is $p + \beta p + \beta^2 p + \cdots = p/(1-\beta)$. The posterior distribution of θ is

$$f(\theta \mid s, f) = \frac{(s+f+1)!}{s!f!}\theta^s(1-\theta)^f, \quad 0 \leq \theta \leq 1,$$

and the posterior mean is $\bar{\theta}(s,f) = (s+1)/(s+f+2)$. The optimality equation is

$$F(s,f) = \max\left[\frac{p}{1-\beta}, \frac{s+1}{s+f+2}\left(1 + \beta F(s+1,f)\right) + \frac{f+1}{s+f+2}\beta F(s,f+1) \right].$$

It is not possible to give a nice expression for F, but we can find an approximate numerical solution. If $s+f$ is very large, say 300, then $\bar{\theta}(s,f) = (s+1)/(s+f+2)$ is a good approximation to θ. Thus we can take $F(s,f) \approx (1-\beta)^{-1}\max[p, \bar{\theta}(s,f)]$, $s+f = 300$ and work backwards. For $\beta = 0.95$, one obtains the following table.

f \ s	0	1	2	3	4	5
0	.7614	.8381	.8736	.8948	.9092	.9197
1	.5601	.6810	.7443	.7845	.8128	.8340
2	.4334	.5621	.6392	.6903	.7281	.7568
3	.3477	.4753	.5556	.6133	.6563	.6899
4	.2877	.4094	.4898	.5493	.5957	.6326

These numbers are the greatest values of p for which it is worth continuing with at least one more trial of the new drug. For example, with $s = 3$, $f = 3$ it is worth continuing with the new drug when $p = 0.6 < 0.6133$. At this point the probability that the new drug will successfully treat the next patient is 0.5 and so the doctor should actually prescribe the drug that is least likely to cure! This example shows the difference between a **myopic policy**, which aims to maximize immediate reward, and an optimal policy, which forgets immediate reward in order to gain information and possibly greater rewards later on. Notice that it is worth using the new drug at least once if $p < 0.7614$, even though at its first use the new drug will only be successful with probability 0.5.