6 Average-cost Programming

We address the infinite-horizon average-cost case, the optimality equation for this case and the policy improvement algorithm.

6.1 Average-cost optimization

It can happen that the undiscounted expected total cost is infinite, but the accumulation of cost per unit time is finite. Suppose that for a stationary Markov policy π, the following limit exists:

$$\lambda(\pi, x) = \lim_{t\to\infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \middle| x_0 = x \right] .$$

It is reasonable to expect that there is a well-defined notion of an optimal **average-cost** function, $\lambda(x) = \inf_\pi \lambda(\pi, x)$, and that under appropriate assumptions, $\lambda(x) = \lambda$ should not depend on x. Moreover, one would expect

$$F_s(x) = s\lambda + \phi(x) + \epsilon(s, x)\,,$$

where $\epsilon(s, x) \to 0$ as $s \to \infty$. Here $\phi(x) + \epsilon(s, x)$ reflects a transient due to the initial state. Suppose that the state space and action space are finite. From the optimality equation for the finite horizon problem we have

$$F_s(x) = \min_u \{c(x, u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\}\,. \tag{6.1}$$

So by substituting $F_s(x) \sim s\lambda + \phi(x)$ into (6.1), we obtain

$$s\lambda + \phi(x) \sim \min_u \{c(x, u) + E[(s-1)\lambda + \phi(x_1) \mid x_0 = x, u_0 = u]\}$$

which suggests, what it is in fact, the average-cost optimality equation:

$$\lambda + \phi(x) = \min_u \{c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]\}\,. \tag{6.2}$$

Theorem 6.1 *Let λ denote the minimal average-cost. Suppose there exists a constant λ' and bounded function ϕ such that for all x and u,*

$$\lambda' + \phi(x) \le c(x, u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]\,. \tag{6.3}$$

Then $\lambda' \le \lambda$. This also holds when $\le$ is replaced by $\ge$ and the hypothesis is weakened to: for each x there exists a u such that (6.3) *holds when $\le$ is replaced by $\ge$.*

Proof. Suppose u is chosen by some policy π. By repeated substitution of (6.3) into itself we have

$$\phi(x) \le -t\lambda' + E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \middle| x_0 = x \right] + E_\pi[\phi(x_t) \mid x_0 = x]$$

Divide this by t and let $t \to \infty$ to obtain

$$0 \le -\lambda' + \lim_{t\to\infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \middle| x_0 = x \right] ,$$

where the final term on the right hand side is simply the average-cost under policy π. Minimizing the right hand side over π gives the result. The claim for $\le$ replaced by $\ge$ is proved similarly. ■

Theorem 6.2 *Suppose there exists a constant λ and bounded function ϕ satisfying* (6.2)*. Then λ is the minimal average-cost and the optimal stationary policy is the one that chooses the optimizing u on the right hand side of* (6.2)*.*

Proof. Equation (6.2) implies that (6.3) holds with equality when one takes π to be the stationary policy that chooses the optimizing u on the right hand side of (6.2). Thus π is optimal and λ is the minimal average-cost. ■

The average-cost optimal policy is found simply by looking for a bounded solution to (6.2). Notice that if ϕ is a solution of (6.2) then so is ϕ+(a constant), because the (a constant) will cancel from both sides of (6.2). Thus ϕ is undetermined up to an additive constant. In searching for a solution to (6.2) we can therefore pick any state, say $\bar{x}$, and arbitrarily take $\phi(\bar{x}) = 0$.

6.2 Example: admission control at a queue

Each day a consultant is presented with the opportunity to take on a new job. The jobs are independently distributed over n possible types and on a given day the offered type is i with probability a_i, $i = 1, \ldots, n$. Jobs of type i pay R_i upon completion. Once he has accepted a job he may accept no other job until that job is complete. The probability that a job of type i takes k days is $(1 - p_i)^{k-1} p_i$, $k = 1, 2, \ldots$. Which jobs should the consultant accept?

Solution. Let 0 and i denote the states in which he is free to accept a job, and in which he is engaged upon a job of type i, respectively. Then (6.2) is

$$\begin{aligned} \lambda + \phi(0) &= \sum_{i=1}^{n} a_i \max[\phi(0), \phi(i)], \\ \lambda + \phi(i) &= (1 - p_i)\phi(i) + p_i[R_i + \phi(0)], \quad i = 1, \ldots, n\,. \end{aligned}$$

Taking $\phi(0) = 0$, these have solution $\phi(i) = R_i - \lambda/p_i$, and hence

$$\lambda = \sum_{i=1}^{n} a_i \max[0, R_i - \lambda/p_i]\,.$$

The left hand side is increasing in λ and the right hand side is decreasing λ. Hence there is a root, say λ^*, and this is the maximal average-reward. The optimal policy takes the form: *accept only jobs for which $p_i R_i \ge \lambda^*$.* ■

6.3 Value iteration bounds

Value iteration in the average-cost case is based upon the idea that $F_s(x) - F_{s-1}(x)$ approximates the minimal average-cost for large s.

Theorem 6.3 *Define*

$$m_s = \min_x \{F_s(x) - F_{s-1}(x)\}\,, \qquad M_s = \max_x \{F_s(x) - F_{s-1}(x)\}\,. \tag{6.4}$$

Then $m_s \le \lambda \le M_s$, where λ is the minimal average-cost.

Proof. (*starred*) Suppose that the first step of a s-horizon optimal policy follows Markov plan f. Then

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

Hence

$$F_{s-1}(x) + m_s \le c(x,u) + E[F_{s-1}(x_1) \mid x_0 = x, u_0 = u]\,,$$

for all x, u. Applying Theorem 6.1 with $\phi = F_{s-1}$ and $\lambda' = m_s$, implies $m_s \le \lambda$. The bound $\lambda \le M_s$ is established in a similar way. ■

This justifies the following **value iteration algorithm**. At termination the algorithm provides a stationary policy that is within $\epsilon \times 100\%$ of optimal.

(0) Set $F_0(x) = 0$, $s = 1$.
(1) Compute F_s from

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

(2) Compute m_s and M_s from (6.4). Stop if $M_s - m_s \le \epsilon m_s$. Otherwise set $s := s+1$ and goto step (1).

6.4 Policy improvement

Policy improvement is an effective method of improving stationary policies.

Policy improvement in the average-cost case.

In the average-cost case a policy improvement algorithm can be based on the following observations. Suppose that for a policy $\pi = f^\infty$, we have that λ, ϕ is a solution to

$$\lambda + \phi(x) = c(x, f(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f(x_0)]\,,$$

and suppose for some policy $\pi_1 = f_1^\infty$,

$$\lambda + \phi(x) \ge c(x, f_1(x_0)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_1(x_0)]\,, \tag{6.5}$$

with strict inequality for some x. Then following the lines of proof in Theorem 6.1

$$\lim_{t\to\infty} \frac{1}{t} E_\pi \left[\sum_{s=0}^{t-1} c(x_s, u_s) \middle| x_0 = x \right] = \lambda \ge \lim_{t\to\infty} \frac{1}{t} E_{\pi_1} \left[\sum_{s=0}^{t-1} c(x_s, u_s) \middle| x_0 = x \right]\,.$$

If there is no π_1 for which (6.5) holds then π satisfies (6.2) and is optimal. This justifies the following **policy improvement algorithm**

(0) Choose an arbitrary stationary policy π_0. Set $s = 1$.
(1) For a given stationary policy $\pi_{s-1} = f_{s-1}^\infty$ determine ϕ, λ to solve

$$\lambda + \phi(x) = c(x, f_{s-1}(x)) + E[\phi(x_1) \mid x_0 = x, u_0 = f_{s-1}(x)]\,.$$

This gives a set of linear equations, and so is intrinsically easier to solve than (6.2).
(2) Now determine the policy $\pi_s = f_s^\infty$ from

$$\begin{aligned} c(x, f_s(x)) &+ E[\phi(x_1) \mid x_0 = x, u_0 = f_s(x)] \\ &= \min_u \{c(x,u) + E[\phi(x_1) \mid x_0 = x, u_0 = u]\}\,, \end{aligned}$$

taking $f_s(x) = f_{s-1}(x)$ whenever this is possible. By applications of Theorem 6.1, this yields a strict improvement whenever possible. If $\pi_s = \pi_{s-1}$ then the algorithm terminates and π_{s-1} is optimal. Otherwise, return to step (1) with $s := s+1$.

If both the action and state spaces are finite then there are only a finite number of possible stationary policies and so the policy improvement algorithm will find an optimal stationary policy in finitely many iterations. By contrast, the value iteration algorithm can only obtain more and more accurate approximations of λ^*.

Policy improvement in the discounted-cost case.

In the case of strict discounting, the following theorem plays the role of Theorem 6.1. The proof is similar, by repeated substitution of (6.6) into itself.

Theorem 6.4 *Suppose there exists a bounded function G such that for all x and u,*

$$G(x) \le c(x,u) + \beta E[G(x_1) \mid x_0 = x, u_0 = u]. \tag{6.6}$$

Then $G \le F$, where F is the minimal discounted-cost function. This also holds when $\le$ is replaced by $\ge$ and the hypothesis is weakened to: for each x there exists a u such that (6.6) *holds when $\le$ is replaced by $\ge$.*

The policy improvement algorithm is similar. E.g., step (1) becomes

(1) For a given stationary policy $\pi_{s-1} = f_{s-1}^\infty$ determine G to solve

$$G(x) = c(x, f_{s-1}(x)) + \beta E[G(x_1) \mid x_0 = x, u_0 = f_{s-1}(x)]\,.$$