

## Chapter 9

# Optimization and Stochastic Control for Markov Chains

In many engineering and economic settings, one wishes to optimize the design of the system under consideration. We will consider two different variants of this problem

1. We wish to optimize the system performance over a finite number of decision variables (e.g. we wish to optimize the behavior of a system in which incoming customers get routed to one of two servers; the decision variables here would be the fraction  $p$  of customers routed to server one's queue).
2. We wish to optimize system performance over all possible controls / policies, leading to an "infinite dimensional" optimization problem (e.g. we wish to optimize the behavior of a system fed by  $n$  different classes of customers; at each slot of time, the server needs to select which customer class to serve next. In an infinite time horizon formation, there is an infinite sequence of decisions to be made; each decision can, in principle, depend on the entire observed history of the system up to that time).

### 9.1 Finite-Dimensional Parameter Optimization

Let  $\theta = (\theta_1, \theta_2, \dots, \theta_d)$  be the vector of decision variables over which system performance must be optimized. Suppose that for each  $\theta$ ,  $P(\theta) = (P(\theta, x, y) : x, y \in S)$  is the corresponding transition matrix. We assume that  $|S| < \infty$  and that  $P(\theta)$  is irreducible for each  $\theta$ . We further assume that the performance criterion to be optimized takes the form

$$\alpha(\theta) = \sum_x \pi(\theta, x) r(\theta, x),$$

where  $\pi(\theta) = (\pi(\theta, x) : x \in S)$  is the stationary distribution of  $X = (X_n : n \geq 0)$  under  $P(\theta)$ , and  $r(\theta) = (r(\theta, x) : x \in S)$  is the reward function (i.e.  $r(\theta, x)$  is the reward obtained by spending one unit of time in  $x$  under parameter  $\theta$ ). Typically,  $\alpha(\theta)$  must be optimized numerically. Such numerical algorithms typically require the ability to compute  $\nabla \alpha(\theta)$  efficiently. In this setting

$$\frac{\partial \alpha(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} (\pi(\theta) r(\theta)) = \frac{\partial r(\theta)}{\partial \theta_i} r(\theta) + \pi(\theta) \frac{\partial r(\theta)}{\partial \theta_i}$$

the only tricky term to compute here is the that involving  $\partial_{\theta_i} \pi(\theta)$ . But  $\pi(\theta)$  satisfies  $\pi(\theta) = \pi(\theta) P(\theta)$ . So,

$$\frac{\partial \pi(\theta)}{\partial \theta_i} = \frac{\partial \pi(\theta)}{\partial \theta_i} P(\theta) + \pi(\theta) \frac{\partial P(\theta)}{\partial \theta_i},$$

and hence

$$\frac{\partial \pi(\theta)}{\partial \theta_i} (I - P(\theta)) = \pi(\theta) \frac{\partial P(\theta)}{\partial \theta_i}.$$

The mathematical difficulty that arises here is that  $I - P(\theta)$  is a singular matrix (because 1 is always an eigenvalue of any stochastic matrix).

Because  $\sum_x \pi(\theta, x) = 1$ , for each  $\theta$ , it follows that

$$\sum_x \frac{\partial}{\partial \theta_i} \pi(\theta, x) = 0.$$

It is easily verified that

$$\frac{\partial}{\partial \theta_i} (\pi(\theta) \Pi(\theta)) = 0,$$

where  $\Pi(\theta)$  is the rank one matrix in which all the rows are equal to  $\pi(\theta)$ . Hence

$$\frac{\partial}{\partial \theta_i} \pi(\theta) (I - P(\theta) + \Pi(\theta)) = \pi(\theta) \frac{\partial}{\partial \theta_i} P(\theta). \quad (9.1)$$

If  $X$  is aperiodic under  $P(\theta)$ ,  $P^n(\theta) \rightarrow \Pi(\theta)$  as  $n \rightarrow \infty$ , so that  $(P(\theta) - \Pi(\theta))^n \rightarrow 0$ . Consequently,

$$\sum_{n=0}^{\infty} (P(\theta) - \Pi(\theta))^n$$

converges absolutely and equals  $(I - P(\theta) + \Pi(\theta))^{-1}$ . Hence,  $I - P(\theta)$  is non-singular, and  $\partial_{\theta_i} \pi(\theta)$  is characterized as the unique solution to the linear system 9.1.

**Exercise 9.1:** Prove that if  $P(\theta)$  is a finite irreducible transition matrix, then  $(I - P(\theta) + \Pi(\theta))^{-1}$  exists (i.e. aperiodicity is unnecessary).

## 9.2 Stochastic Control

### 9.2.1 Infinite Horizon Stochastic Control

The formulation here is that, for each  $x \in S$ , there exists a set  $\mathcal{A}(x)$  of actions available to the decision-maker / controller in  $x$ . If action  $a \in \mathcal{A}(x)$  is taken in state  $x$ , this produces an immediate reward of  $r(x, a)$ . In addition, the choice of  $a$  affects the dynamics of the system, so that there exists a transition probability  $P_a(x, y)$  (depending on  $a$ ) describing the probability that the next state visited is  $y$ . We start with the problem of finding an optimal control  $A^* = (A_n^* : n \geq 0)$  (i.e. an optimal sequence of actions  $(A_n^* : n \geq 0)$  for which  $A_n^* \in \mathcal{A}(X_n)$ ) that maximizes the infinite horizon discounted reward:

$$\max_{A_n : n \geq 0} \mathbb{E}_x \left[ \sum_{n=0}^{\infty} e^{-\alpha n} r(X_n, A_n) \right]$$

for  $\alpha > 0$ . The key idea here is to use an approach similar to that used earlier in the course in the setting of “first transition analysis”. Put

$$v(x) = \max_{A_n : n \geq 0} \mathbb{E}_x \left[ \sum_{n=0}^{\infty} e^{-\alpha n} r(X_n, A_n) \right].$$

The function  $v = (v(x) : x \in S)$  is called the value function of the control problem. The function  $v$  satisfies the Hamilton-Jacobi-Bellman (HJB) equation (also called the optimality equation)

$$v(x) = \max_{a \in \mathcal{A}(x)} \left[ r(x, a) + e^{-\alpha} \sum_y P_a(x, y) v(y) \right] = \max_{a \in \mathcal{A}(x)} [r(x, a) + e^{-\alpha} \mathbb{E}[v(X_1) | X_0 = x, A_0 = a]] \quad (9.2)$$

Given a solution  $v$  to the non-linear equation 9.2, the optimal control ( $A_n^* : n \geq 0$ ) is given by  $A_n^* = a^*(X_n)$ , where  $a^*(x)$  is any maximizing action of the right-hand side of equation 9.2. So the key is to compute the value function  $v$ .

One approach is to use “value iteration”. Note that equation 9.2 establishes that  $v$  is a “fixed point” of the operator  $\mathcal{R}$ , where

$$(\mathcal{R}w)(x) = \max_{a \in \mathcal{A}(x)} [r(x, a) + e^{-\alpha} \mathbb{E} [w(X_1) | X_0 = x, A_0 = a]]$$

i.e.  $v$  satisfies  $v = \mathcal{R}v$ . Value iteration is just the method of successive approximation applied to this fixed point problem. Let  $v_0$  be an initial guess of  $v$ , and recursively, define ( $v_n : n \geq 0$ ) via  $v_n = \mathcal{R}v_{n-1}$  for  $n \geq 1$ . With the presence of the discounting factor  $e^{-\alpha}$ , it can be shown that  $\mathcal{R}$  is a contraction operator (i.e.  $\|\mathcal{R}w_1 - \mathcal{R}w_2\| \leq \beta \|w_1 - w_2\|$  for  $\beta \in (0, 1)$ ), which establishes that  $v_n \rightarrow v$  as  $n \rightarrow \infty$  (and exponentially fast in  $n$ ).

The value function  $v$  can also be computed as the solution to the following linear program (LP):

$$\begin{aligned} \min_v \quad & \sum_x v(x) \\ \text{s.t.} \quad & v(x) \geq r(x, a) + e^{-\alpha} \\ & \sum_y P_a(x, y)v(y), \quad x \in S, \quad a \in \mathcal{A}(x) \end{aligned}$$

### 9.2.2 Finite Horizon Stochastic Control

The set is similar to the above section, however, we no longer have a discounting factor, and we are only interested in the reward over a given (finite) time. Our objective is

$$\max_{A_j: 0 \leq j \leq n} \mathbb{E}_x \left[ \sum_{j=0}^n r(X_j, A_j) \right].$$

Here we let

$$v(i, x) = \max_{A_j: i \leq j \leq n} \mathbb{E}_x \left[ \sum_{j=i}^n r(X_j, A_j) \right].$$

Note that

$$v(n, x) = \max_{a \in \mathcal{A}(x)} r(x, a),$$

and

$$v(i, x) = \max_{a \in \mathcal{A}(x)} \left[ r(x, a) + \sum_y P_a(x, y)v(i+1, y) \right] \quad (9.3)$$

In this setting, we compute the value function by first computing  $v_n$  and then computing  $v_{n-1}, v_{n-2}, \dots, v_0$  through backwards recursion of (9.3). The optimal action  $A_i^* = a^*(i, X_i)$ , where  $a^*(i, x)$  is the action maximizing the right-hand side of equation 9.3.

### 9.2.3 Long-Run Average Reward

Suppose we wish to maximize

$$\max_{A_j: j \geq 0} \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_x \left[ \sum_{j=0}^{n-1} r(X_j, A_j) \right].$$

We solve this by taking a limit (as  $n \rightarrow \infty$ ) in the finite horizon control problem discussed earlier. It seems plausible that  $v(i, x)$  takes the approximate form

$$v(i, x) \approx (n - i)c + v(x),$$

where  $c$  is the maximizing long-run average reward per unit time. This suggests that

$$(n - i + 1)c + v(x) \approx \max_{a \in \mathcal{A}(x)} \left[ r(x, a) + \sum_y P_a(x, y) ((n - i)c + v(y)) \right]$$

sending  $n \rightarrow \infty$ , we find that  $c$  and  $v = (v(x) : x \geq 0)$  should satisfy the HJB equation

$$c + v(x) + \max_{a \in \mathcal{A}(x)} \left[ r(x, a) + \sum_y P_a(x, y)v(y) \right] \quad (9.4)$$

The optimal control  $(A_n^* : n \geq 0)$  then takes the form  $A_n^* = a^*(X_n)$ , where  $a^*(x)$  is the maximizing action on the right-hand side of (9.4). To compute the solution  $v$  of (9.4), one approach is via linear programming:

$$\begin{aligned} \min_{c, v} \quad & c + \sum_x v(x) \\ \text{s.t.} \quad & c + v(x) \geq r(x, a) + \sum_y P_a(x, y)v(y), \quad x \in S, \quad a \in \mathcal{A}(x) \end{aligned}$$

It is interesting to study the dual linear program here:

$$\begin{aligned} \max_{\pi(x, a)} \quad & \sum_{x, a} \pi(x, a)r(x, a) \\ \text{s.t.} \quad & \sum_{a \in \mathcal{A}(y)} \pi(y, a) = \sum_{x, a \in \mathcal{A}(x)} P_a(x, y), \quad \forall y \in S \\ & \sum_{a \in \mathcal{A}(y), y \in S} \pi(y, a) = 1 \\ & \pi(y, a) \geq 0, \quad y \in S, \quad a \in \mathcal{A}(y) \end{aligned}$$

Given the solution to the dual linear program, we can read off the optimal control: when in state  $x$ , choose action  $a \in \mathcal{A}(x)$  with probability

$$\pi(a|x) = \frac{\pi(x, a)}{\sum_{a' \in \mathcal{A}(x)} \pi(x, a')}.$$

**Remark 9.1:** Because linear programs tend to seek optimal solutions at extreme points of the feasible region,  $\pi(a|x)$  is typically positive (and hence equal to one) for only one action  $a \in \mathcal{A}(x)$ .

**Remark 9.2:** The dual linear programming formulation is convenient for problem settings in which one wishes to add an expectation constraint on the control problem (e.g. find the optimal control subject to  $\sum_{x, a} \pi(x, a)g(x, a) \geq b$ ).

### 9.3 Optimal Stopping

The use of value function based methods can also be used to solve so called “optional stopping” problems. One example of such an optimal stopping problem arises in the context of American Options. Such an option has an expiration time, call it  $n$ . (The expiration time  $n$  can be equal to infinity, in which case it is a “perpetual option”.) An American option can be exercised at any time  $i \in \{0, 1, \dots, n\}$ . The option holder has the option, at the exercise time, to buy a share of the underlying security at price  $c$ . If we model the price of the security over time as a Markov chain  $X = (X_j : j \geq 0)$ , with transition matrix  $P$ , the goal is to find the exercise time  $T$  that maximizes

$$\mathbb{E}_x \left[ [X_T - c]^+ \right].$$

More generally, the goal is to find a time  $T$  maximizing

$$\mathbb{E}_x [r(X_T)]$$

for some reward function  $r = (r(x) : x \in S)$ . If the time  $T$  is constrained to a finite horizon  $\{0, 1, \dots, n\}$ , put

$$v(i, x) = \max_{i \leq T \leq n} \mathbb{E}_x [r(X_T)].$$

Here

$$v(i, x) = \max \left\{ r(x), \sum_y P(x, y) v(i+1, y) \right\}$$

subject to  $v(n, x) = r(x)$ . Again, the value function can be computed through a backwards recursion. When  $v(i, x) = r(x)$ , the optimal policy is to stop when in state  $x$  at time  $i$ ; otherwise, the optimal policy is to continue.

If  $n = \infty$ , the HJB equation is

$$v(x) = \max \left\{ r(x), \sum_y P(x, y) v(y) \right\}.$$

One way to solve for  $v$  is to compute it as the solution to a linear program.

