Chapter 11

Linear Programming Methods

[1]In this chapter we consider the linear programming approach to dynamic programming. First, Bellman's equation can be reformulated as a linear program whose solution is the optimal value function, offering a computational alternative to the value and policy iteration methods. Next, the dual of this linear program gives a new point of view on the optimal control problem, where one solves a dynamic optimization problem by optimizing over the vectors of "state-action frequencies", which correspond to Markov randomized stationary policies. This formulation is particularly useful in treating multi-objective or constrained dynamic programming problems, as well as Markov games. LP methods are also useful for sensitivity analysis. Finally, they can be combined with approximation methods as discussed in chapter 20.

11.1 Linear Programming Formulations

We consider in this chapter finite Markov decision processes, i.e., problems with a finite number of states and controls. We discuss first consider discounted cost problems as in chapter 6. Denote the state space $X = \{1, \ldots, n\}$. We then know that starting with a cost vector $J \in \mathbb{R}^n$, we have $\lim_{k\to\infty} T^k J = J^*$, where J^* is the optimal value function, see theorem 6.5.1. Suppose $J \leq TJ$. Then by the monotonicity of the dynamic programming operator T, we get

$$J \leq TJ \leq T^2 J \leq \ldots \leq \lim_{k\to\infty} T^k J = J^* = TJ^*.$$

Hence $J \leq TJ \Rightarrow J \leq J^*$, and so J^* is the largest vector that satisfies the constraint $J \leq TJ$. Moreover, the nonlinear constraints

$$J(i) \leq \min_{u\in \mathsf{U}(i)} \left\{ c(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right\}, \; i = 1, \ldots, n,$$

[1]This version: November 16 2009.

can be rewritten as a system of *linear* inequalities on the variables $J(i), i = 1, \ldots, n$

$$J(i) \leq c(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j),\ i = 1, \ldots, n,\ u \in \mathsf{U}(i).$$

Let the $\alpha \in (0,1)$. From the preceding discussion we see immediately that the vector J^* is the solution of any linear program of the form

$$\begin{aligned} &\text{maximize} \quad (1-\alpha) \sum_{i=1}^{n} \nu_i J_i \qquad (11.1) \\ &\text{subject to } J_i \leq c(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j),\ i = 1, \ldots, n,\ u \in \mathsf{U}(i), \end{aligned}$$

where $\nu = [\nu_1, \ldots, \nu_n]^T$ is a set of positive weights $\nu_i > 0, i = 1, \ldots, n$, and the decision variables are $J_i, i = 1, \ldots, n$. The normalization factor $(1-\alpha)$ can be omitted but it is convenient for our purpose. Note that if in addition we choose ν to belong to the interior of the probability simplex

$$\nu \in \Delta_{n-1} = \{x \in \mathbb{R}^n : x_i > 0, \sum_{i=1}^{n} x_i = 1\},$$

then ν_i can be interpreted as a probability of starting in state $i \in \mathsf{X}$, and $\nu^T J^* = \sum_{i=1}^{n} \nu_i J^*(i)$ is the optimal expected cost for the optimal control problem when the initial state is probabilistically distributed with distribution ν. If we wanted to compute $J^*(i)$ for a subset of states $\tilde{\mathsf{X}} \subset \mathsf{X}$, we could take some of the coefficients ν_i equal to 0 for $i \notin \tilde{\mathsf{X}}$, but it is not clear why one would want to do this, since the computational complexity of the problem is not reduced. The linear program (11.1) has n variables and up to $\sum_{i=1}^{n} |\mathsf{U}(i)|$ constraints, which makes the method impractical for large state and control spaces, even when some large scale linear programming techniques are used. In this case, the linear program (11.1) can be combined with approximation architectures, see chapter 20.

Dual Linear Program

The dual program of (11.1) is perhaps more interesting for applications. It can be written

$$\min \sum_{x \in \mathsf{X}} \sum_{u \in \mathsf{U}(x)} c(x,u) \rho(x,u) \qquad (11.2)$$

$$\text{s.t.} \sum_{u \in \mathsf{U}(x)} \rho(x,u) - \sum_{x' \in \mathsf{X}} \sum_{u \in \mathsf{U}(x')} \alpha\, p_{x'x}(u) \rho(x',u) = (1-\alpha)\nu_x, \quad \forall x \in \mathsf{X} \qquad (11.3)$$

$$\rho(x,u) \geq 0, \quad \forall x \in \mathsf{X}, u \in \mathsf{U}(x).$$

It is sometimes stated that this dual program is computationally preferable because it has only $|\mathsf{X}|$ constraints (not counting the nonnegativity constraints) [Put94, p.224], but since now we have more variables than in (11.1), this statement does not seem very clear.

Remark. For the initial distribution, in the following we use the notation ν_x and $\nu(x)$ interchangeably. Also, an equivalent way of writing (11.3) is

$$\sum_{x' \in \mathsf{X}} \sum_{u \in \mathsf{U}(x')} (\delta_x(x') - \alpha\, p_{x'x}(u))\rho(x', u) = (1-\alpha)\nu(x).$$

If we sum these constraints over x, we see that

$$\sum_{x' \in \mathsf{X}} \sum_{u \in \mathsf{U}(x')} \rho(x', u) = 1, \tag{11.4}$$

and so the set of feasible solutions is a polytope (i.e., *bounded* polyhedron).

11.2 State-Action Frequencies

The quantities $\{\rho(x,a)\}_{x,a}$, if they satisfy the constraints (11.3), can be interpreted as a *state-action frequencies*, i.e., $\rho(x,a)$ is the proportion of time for which the system is in state x and action u is taken, under a certain stationary policy. In fact, randomized Markov[2] stationary policies are in correspondence with the feasible solutions of (11.2). Note that the quantity

$$\rho(x) := \sum_{u \in \mathsf{U}(x)} \rho(x, u)$$

appearing in the LP (11.2) has then the interpretation of state frequencies. Note that by (11.4) we see that $\{\rho(x)\}_x$ forms a probability distribution over X, i.e., $\sum_{x \in \mathsf{X}} \rho(x) = 1$. Similarly $\{\rho(x,u)\}_{x,u}$ is a probability measure on the space of state-action pairs by (11.4).

Consider a Markov policy π, not necessarily stationary, and possibly randomized. Together with the initial distribution ν, this defines a probability measure $\mathbb{P}^\pi_\nu$ on the sample paths the state-action pairs. We define the state-action frequencies corresponding to ν and π as

$$\begin{aligned} \rho^\pi_\nu(x,u) &:= (1-\alpha) \sum_{x' \in \mathsf{X}} \nu(x') \sum_{k=0}^{\infty} \alpha^k\, \mathbb{P}^\pi(X_k = x, U_k = u | X_0 = x') \qquad (11.5) \\ &= (1-\alpha) \sum_{k=0}^{\infty} \alpha^k\, \mathbb{P}^\pi_\nu(X_k = x, U_k = u). \end{aligned}$$

[2] A Markov policy is just a policy that depends only on the current state. We haven't encountered any other policy in this course, nor the need for them. Here we are adding the possibility of randomizing controls.

Clearly ρ_ν^π forms a probability measure on the space of state-action pairs. It turns out that there is always a *stationary* policy μ which achieves the same vector of state-action frequencies. To see this, consider the stationary randomized policy μ defined by

$$\mu(u|x) = \frac{\rho_\nu^\pi(x,u)}{\rho_\nu^\pi(x)},$$

whenever the denominator is non-zero. When it is zero, chose $\mu(\cdot|x)$ arbitrarily. Here $\mu(u|x)$ represents the probability of choosing control $u \in \mathsf{U}(x)$ when the state is x. We have

$$\begin{aligned}
\rho_\nu^\pi(x) &= (1-\alpha)\nu(x) + (1-\alpha)\alpha \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\nu^\pi(X_{k+1} = x) \\
&= (1-\alpha)\Big\{\nu(x) \\
&+ \alpha \sum_{k=0}^{\infty} \alpha^k \sum_{x' \in \mathsf{X}} \sum_{u \in \mathsf{U}(x')} \mathbb{P}(X_{k+1} = x | X_k = x', U_k = u) \mathbb{P}_\nu^\pi(X_k = x', U_k = u)\Big\} \\
&= (1-\alpha)\nu(x) + \sum_{x' \in \mathsf{X}} \sum_{u \in \mathsf{U}(x')} \alpha p_{x'x}(u) \rho_\nu^\pi(x', u) \qquad (11.6) \\
&= (1-\alpha)\nu(x) + \alpha \sum_{x' \in \mathsf{X}} \rho_\nu^\pi(x') \sum_{u \in \mathsf{U}(x')} p_{x'x}(u) \mu(u|x') \\
&= (1-\alpha)\nu(x) + \alpha \sum_{x' \in \mathsf{X}} \rho_\nu^\pi(x') P_\mu(x', x).
\end{aligned}$$

In matrix notation, this can be written

$$(\rho_\nu^\pi)^T = (1-\alpha)\nu^T (I - \alpha P_\mu)^{-1} = (1-\alpha)\nu^T \left(\sum_{k=0}^{\infty} \alpha^k P_\mu^k \right).$$

In other words,

$$\rho_\nu^\pi(x) = (1-\alpha) \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\nu^\mu(X_k = x) = \rho_\nu^\mu(x). \qquad (11.7)$$

Moreover, by definition of $\mu(u|x)$, we have

$$\rho_\nu^\mu(x,u) = (1-\alpha) \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\nu^\mu(X_k = x) \mu(u|x).$$

Using the relation (11.7), we get

$$\rho_\nu^\mu(x,u) = (1-\alpha) \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\nu^\mu(X_k = x) \frac{\rho_\nu^\pi(x,u)}{\rho_\nu^\mu(x)} = \rho_\nu^\pi(x,u).$$

Note that the constraints (11.6) satisfied by the state-action frequencies (of any Markov policy) are exactly the constraints (11.3) of the linear program. We have the following theorem.

Theorem 11.2.1. *1. For each Markov stationary randomized policy μ and positive initial distribution ν, $\{\rho_\nu^\mu(x,u)\}_{x,u}$ is a feasible solution to the linear program (11.2).*

2. Suppose $\{\rho(x,u)\}_{x,u}$ is a feasible solution to the problem (11.2). Then, for each $x \in \mathsf{X}$, we have $\sum_{u\in\mathsf{U}(x)} \rho(x,u) > 0$. Define the Markov randomized stationary policy μ by

$$\mu(u|x) = \frac{\rho(x,u)}{\sum_{u'\in\mathsf{U}(x)} \rho(x,u')} = \frac{\rho(x,u)}{\rho(x)}. \tag{11.8}$$

Then for this policy μ, we can define the state-action frequencies ρ_ν^μ by (11.5), and we have $\rho_\nu^\mu(x,u) = \rho(x,u)$ for all $x \in \mathsf{X}$ and $u \in \mathsf{U}(x)$.

Proof. The first statement was proved above, see (11.6). Consider now the statement 2. The positivity of $\rho(x)$ follows from that of $\nu(x)$ and the non-negativity of $\rho(x,u)$ in (11.3). Then note that $\rho(x,a)$ satisfies the constraint (11.6) and so we have immediately $\rho(x,u) = \rho_\nu^\mu(x,u)$ for all x,u by following the steps above. □

Recall now that our objective is to optimize a function of the form

$$J_\pi(\nu) := (1-\alpha)\sum_{x\in\mathsf{X}} \nu(x) J_\pi(x) = \mathbb{E}_\nu^\pi\Big[\sum_{k=0}^{\infty} \alpha^k (1-\alpha) c(X_k, U_k)\Big], \tag{11.9}$$

where $\mathbb{E}_\nu^\pi$ is the expectation operator over paths of the Markov chain obtained once the policy π and the initial distribution ν are fixed. Here we generalize our earlier notation, with $J_\pi(\nu)$ being the cost of the policy π when the initial state is distributed according to ν. We can rewrite this objective as

$$\begin{aligned} J_\pi(\nu) &= \mathbb{E}_\nu^\pi\Big[\sum_{k=0}^{\infty}\sum_{x\in\mathsf{X}}\sum_{u\in\mathsf{U}(x)} \alpha^k(1-\alpha)c(x,u)1\{X_k = x, U_k = u\}\Big], \\ &= \sum_{x\in\mathsf{X}}\sum_{u\in\mathsf{U}(x)} c(x,u)\rho_\nu^\pi(x,u), \end{aligned}$$

where the interchange of summation and expectation is allowed by the dominated convergence theorem. This last expression is exactly the objective of the linear program (11.2). By theorem 11.2.1, assuming ν is positive for simplicity, there is a bijection between Markov stationary randomized policies and feasible solutions of the LP (11.2) (which maps a stationary policy to its state-action frequencies, and in the reverse direction, defines a stationary policy by (11.8))[3]. Recall that a basic feasible solution of an LP is a solution that cannot be expressed as a nontrivial convex combination of any other feasible solutions of the

[3] $\nu(x) > 0$ guarantees $\rho(x) > 0$. If $\rho(x) = 0$, we have a choice in the definition of the stationary policy for the states that are almost surely never visited, see the beginning of this section.

LP. It corresponds to extreme points of the feasible region. A key property of basic feasible solutions is that when an LP with m rows has a bounded optimal solution, then any basic feasible solution has at most m positive components. The next proposition establishes a one-to-one correspondence between stationary deterministic policies and extreme points (basic feasible solutions) of the LP (11.2). Since we know that the optimal solution of (11.2) is attained at one of these extreme points, this gives us a proof that to find optimal policies, it is sufficient to consider deterministic policies. Moreover, the discussion at the beginning of this section tells us that non-stationary policies do not provide better solutions. Finally, the LP geometry shows that in certain cases, it is possible that several deterministic policies are optimal, in which case randomizing between these policies gives a convex set of optimal randomized policies.

Proposition 11.2.2. *Let ρ be a basic feasible solution to the LP (11.2). Then μ defined by (11.8) is deterministic. Conversely, the state-action frequency vector of a Markov deterministic policy is a basic feasible solution to the LP (11.2).*

Proof. We know that for all x, $\sum_{u \in \mathsf{U}(x)} \rho(x,u) > 0$. Moreover, the LP (11.2) has $|\mathsf{X}|$ rows. Hence in a basic feasible solution, we must have $\rho(x,u) > 0$ for exactly one u, and we conclude that μ defined by (11.8) is deterministic.

For the converse, assume ρ_ν^μ is feasible but not basic. Then there are two distinct basic feasible solutions ρ_1 and ρ_2 and $0 \leq \theta \leq 1$ such that $\rho_\nu^\mu = (1-\theta)\rho_1 + \theta\rho_2$. Because ρ_1 and ρ_2 are distinct, there is some $x \in \mathsf{X}$ such that $\rho_1(x,u_1) > 0$ and $\rho_1(x,u_2) > 0$ for $u_1 \neq u_2$. But then $\rho_\nu^\mu(x,u_1) > 0$ and $\rho_\nu^\mu(x,u_2) > 0$, so μ must be randomized. □

We summarize the results in the following theorem.

Theorem 11.2.3.

- *There exists a bounded optimal basic feasible solution ρ^* to the LP (11.2).*
- *If ρ^* is an optimal solution to (11.2), then μ^* defined by (11.8) is an optimal policy.*
- *If ρ^* is an optimal basic solution to (11.2), then μ^* defined by (11.8) is an optimal deterministic policy.*
- *If μ is an optimal policy, then ρ_ν^μ is an optimal solution for (11.2).*
- *If μ is an optimal deterministic policy, then ρ_ν^μ is an optimal basic solution for (11.2).*
- *For any positive vector ν, the LP (11.2) has the same optimal basis (columns of the constraint matrix that determine the basic feasible solution). Hence the optimal policy does not depend on ν.*

Note that if ρ^* and J^* are optimal solutions of the LPs (11.2) and (11.1), we have (strong duality holds)

$$J^*(\nu) = (1-\alpha)\sum_{x\in\mathsf{X}} \nu(x)J^*(x) = \sum_{x\in\mathsf{X}}\sum_{u\in\mathsf{U}(x)} \rho^*(x,u)c(x,u).$$

Remark. Solving the LP (11.2) by the simplex method with block pivoting is equivalent to policy iteration.

11.3 Constrained Dynamic Programming

In view of (11.9), it is easy to add constraints to the model. Let $(x,u) \mapsto d(x,u)$ be another cost function, and suppose we wish to solve the initial optimal control problem for $c(x,u)$ with the additional constraint that the expected total discounted cost for d does not exceed some constant D

$$\mathbb{E}_\nu^\pi\Big[\sum_{k=0}^{\infty} \alpha^k d(X_k,U_k)\Big] \le D.$$

This constraints can be rewritten

$$\sum_{x\in\mathsf{X}}\sum_{u\in\mathsf{U}(x)} \rho_\nu^\mu(x,a)d(x,a) \le D,$$

and can be directly added to the LP (11.2). This cuts the polytope corresponding to the feasible region of the unconstrained DP problem, and in general, optimal policies must then include some randomization, with more randomization necessary as we add more constraints, see the proof of proposition 11.2.2. To deal with such constraints in the more standard dynamic programming formulation, we would have to introduce Lagrange multipliers.

11.4 Markov Games

For a future version.