Lecture Note 16

1 Linear Programming Approach to DP

In previous lectures on dynamic programming, we have studied the value and policy iteration algorithms for solving Bellman's equation. We now introduce a different algorithm, which is based on formulating the dynamic programming problem as a linear program.

Consider the following optimization problem:

$$\begin{aligned} \max_J \quad & c^T J \\ \text{subject to} \quad & TJ \geq J, \end{aligned} \tag{1}$$

and suppose that vector c is strictly positive: $c > 0$. Recall the following lemma from Lecture 3:

Lemma 1 *For any J such that $TJ \geq J$, we have $J \leq J^*$.*

It follows from the previous lemma that, whenever $c > 0$, J^* is the unique solution to problem (1). We refer to this problem as the *exact LP*. Note however that, strictly speaking, this problem is *not* a linear program; in particular, the constraints

$$\begin{aligned} & (TJ)(x) \geq J(x) \\ & \min_a \left\{ g_a(x) + \alpha \sum_y P_a(x,y) J(y) \right\} \geq J(x) \end{aligned} \tag{2}$$

are not linear in the variables J of the LP. However, (1) can easily be converted into an LP by noting that each constraint (2) is equivalent to

$$g_a(x) + \alpha \sum_y P_a(x,y) J(y) \geq J(x), \quad \forall a \in \mathcal{A}_x.$$

Note that the exact LP contains as many variables as the number of states in the system, and as many constraints as the number of state-action pairs.

1.1 Dual Linear Programming

We can also find an optimal policy by solving the dual of the exact LP. For simplicity, we will consider average-cost problems in this section, but the analysis and underlying ideas easily extend to the discounted-cost case. The dual LP has an interesting interpretation, and it can be shown that solving it iteratively via simplex or interior-point methods is equivalent to performing specific forms of policy iteration.

In the average-cost case, it can be shown that the dual LP is given as follows:

$$\begin{aligned} \min_{\mu} \quad & \sum_{x,a} \mu(x,a) g_a(x) && (3) \\ \text{subject to} \quad & \sum_{y} \sum_{a} \mu(y,a) P_a(y,x) = \sum_{a} \mu(x,a), \forall\ x && (4) \\ & \sum_{x,a} \mu(x,a) = 1 \\ & \mu(x,a) \geq 0,\ \forall\ x,a \end{aligned}$$

For simplicity, let us assume that the system is irreducible under every policy.

In order to analyze the dual LP, we consider the notion of *randomized policies*. So far, we have defined a policy to be a mapping from states to actions; in other words, for every state x a policy u prescribes a (deterministic) action $u(x) \in \mathcal{A}_{\S}$. Alternatively, we can consider an extended definition where, for any state action, a policy u prescribes a *probability* $u(x,a)$ for taking each action $a \in \mathcal{A}_x$. Each policy u is now associated with a transition matrix P_u such that

$$P_u(x,y) = \sum_{a} u(x,a) P_a(x,y),$$

and yields a stationary state distribution π_u, which is the unique solution to

$$\begin{aligned} & \pi_u^T P_u = \pi_u^T \\ & \sum_{x} \pi_u(x) = 1 \\ & \pi_u(x) \geq 0. \end{aligned}$$

We can also verify that the state costs associated with policy u are given by

$$g_u(x) = \sum_{a} u(x,a) g_a(x).$$

With these notions in mind, it can be shown that the variables $\mu(x,a)$ for any feasible solution to the dual LP can be interpreted as state-action frequencies for a randomized policy. Indeed, let

$$\pi(x) = \sum_{a} \mu(x,a), \tag{5}$$

and

$$u(x,a) = \frac{\mu(x,a)}{\pi(x)}, \tag{6}$$

if $\pi(x) > 0$, and $u(x,\cdot)$ be an arbitrary distribution over $\mathcal{A}_x$, otherwise. Note that in either case we have

$$\mu(x,a) = \pi(x) u(x,a),$$

and $u(x,a)$ is a randomized policy. Then we can show that π corresponds to the stationary state distribution π_u, and $\sum_{x,a} \mu(x,a) g_a(x,a)$ corresponds to the average cost λ_u of policy u.

Lemma 2 *For every feasible solution $\mu(x,a)$ of the dual LP, let $\pi(x)$ and $u(x,a)$ be given by (??) and (6). Then $\pi = \pi_u$, and $\sum_{x,a} \mu(x,a) g_a(x)$ corresponds to the average cost λ_u of policy u.*

Proof: Consider constraints (4). Then we have

$$\sum_y \sum_a \mu(y,a) P_a(y,x) = \sum_a \mu(x,a)$$
$$\sum_y \sum_a \pi(y) u(y,a) P_a(y,x) = \pi(x)$$
$$\sum_y \pi(y) P_u(y,x) = \pi(x).$$

We also have $\pi(x) \geq 0$ for all x, and

$$\begin{aligned} \sum_x \pi(x) &= \sum_x \sum_a \mu(x,a) \\ &= 1. \end{aligned}$$

We conclude that π is a stationary distribution associated with policy u. Since by assumption the system is irreducible under every policy, each policy has a unique stationary distribution, and we have $\pi = \pi_u$. We now have

$$\begin{aligned} \sum_{x,a} \mu(x,a) g_a(x) &= \sum_x \sum_a \pi_u(x) u(x,a) g_a(x) \\ &= \sum_x \pi_u(x) g_u(x) \\ &= \lambda_u. \end{aligned}$$

□

From the previous lemma, we conclude that each feasible solution of the dual LP is identified with a policy u, and the variables $\mu(x,a)$ correspond to the probability of observing state x and action a, in steady state. Consider using simplex or an interior-point method for solving the dual LP. Either method will generate a sequence of feasible solutions $\mu_0, \mu_1, \ldots$, with decreasing value of the objective function. Interpreting this sequence with Lemma 2, we see that this is equivalent to generating a sequence of policies $u_0, u_1, \ldots$, with decreasing average cost, and solving the LP corresponds to performing policy iteration.

2 Approximate Linear Programming

With an aim of computing a weight vector $\tilde{r} \in \Re^K$ such that $\Phi\tilde{r}$ is a close approximation to J^*, in view of the exact LP one might pose the following optimization problem:

$$\begin{aligned} \max_r \quad & c^T \Phi r \\ \text{subject to} \quad & T\Phi r \geq \Phi r \end{aligned} \tag{7}$$

As with the case of exact dynamic programming, the optimization problem (**??**) can be recast as a linear program

$$\begin{aligned} \max \quad & c^T \Phi r \\ \text{s.t.} \quad & g_a(x) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y)(\Phi r)(y) \geq (\Phi r)(x), \quad \forall x \in S, a \in \mathcal{A}_x. \end{aligned}$$

We will refer to this problem as the *approximate LP*. Note that the approximate LP involves a potentially much smaller number of variables — it has one variable for each basis function. However, the number of constraints remains as large as in the exact LP. Fortunately, most of the constraints become inactive, and solutions to the linear program can be approximated efficiently, as we will show in future lectures.

2.1 State-Relevance Weights

In the exact LP, for any vector c with positive components, maximizing $c^T J$ yields J^*. In other words, the choice of state-relevance weights does not influence the solution. The same statement does not hold for the approximate LP. In fact, the choice of state-relevance weights may bear a significant impact on the quality of the resulting approximation.

To motivate the role of state-relevance weights, let us start with a lemma that offers an interpretation of their function in the approximate LP.

Lemma 3 *A vector $\tilde{r}$ solves*

$$\begin{aligned} \max \quad & c^T \Phi r \\ \text{s.t.} \quad & T\Phi r \geq \Phi r, \end{aligned}$$

if and only if it solves

$$\begin{aligned} \min \quad & \|J^* - \Phi r\|_{1,c} \\ \text{s.t.} \quad & T\Phi r \geq \Phi r. \end{aligned}$$

Proof: It is clear that the approximate LP is equivalent to minimizing $c^T(J^* - \Phi r)$ over all feasible r. For all Φr such that $T\Phi r \geq \Phi r$, we have $\Phi r \leq J^*$, and $c^T(J^* - \Phi r) = \|J^* - \Phi r\|_{1,c}$. □

Lemma 3 suggests that the state-relevance weights may be used to control the quality of the approximation to the cost-to-go function over different portions of the state space. Recall that ultimately we are interested in generating good *policies*, rather than good approximations to the cost-to-go function, and ideally we would like to choose c to reflect that objective. The following theorem, from Lecture 12, suggests certain choices for state-relevance weights. Recall that

$$\mu_{\nu,J}^T = c^T (I - \alpha P_{u_J})^{-1}.$$

Theorem 1 *Let $J : \mathcal{S} \mapsto \Re$ be such that $TJ \geq J$. Then*

$$\|J_{u_J} - J^*\|_{1,\nu} \leq \frac{1}{1-\alpha} \|J - J^*\|_{1,\mu_{\nu,J}}. \tag{8}$$

Contrasting Lemma 3 with the bound on the increase in costs (8) given by Theorem 1, we may want to choose state-relevance weights c that capture the (discounted) frequency with which different states are expected to be visited. Note that the frequency with which different states are visited in general depends on the policy being used. One possibility is to have an iterative scheme, where the approximate LP is solved multiple times with state-relevance weights adjusted according to the intermediate policies being generated. Alternatively, a plausible conjecture is that some problems will exhibit structure making it relatively easy

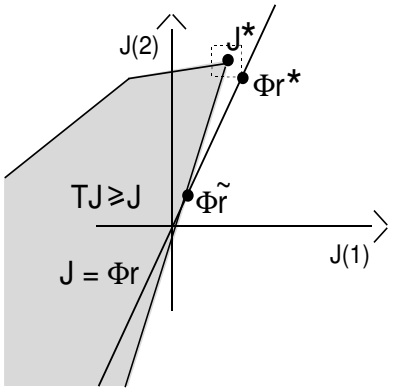

Figure 1: Graphical interpretation of approximate linear programming

to make guesses about which states are desirable and therefore more likely to be visited often by reasonable policies, and which ones are typically avoided and rarely visited. We expect structures enabling this kind of procedure to be reasonably common in large-scale problems, in which desirable policies often exhibit some form of "stability," guiding the system to a limited region of the state space and allowing only infrequent excursions from this region.

3 Approximation Error Analysis

For the approximate LP to be useful, it should deliver good approximations when the cost-to-go function is near the span of selected basis functions. Figure 1 illustrates the issue. Consider an MDP with states 1 and 2. The plane represented in the figure corresponds to the space of all functions over the state space. The shaded area is the feasible region of the exact LP, and J^* is the pointwise maximum over that region. In the approximate LP, we restrict attention to the subspace $J = \Phi r$.

In Figure 1, the span of the basis functions comes relatively close to the optimal cost-to-go function J^*; if we were able to perform, for instance, a maximum-norm projection of J^* onto the subspace $J = \Phi r$, we would obtain the reasonably good approximate cost-to-go function Φr^*. At the same time, the approximate LP yields the approximate cost-to-go function $\Phi\tilde{r}$. In this section, we develop bounds guaranteeing that $\Phi\tilde{r}$ is not too much farther from J^* than Φr^* is.

Note that, in general, we cannot guarantee that $\Phi\tilde{r}$ will be close to J^*, regardless of how close Φr^* is; for instance, Figure 2 illustrates a worst-case scenario, where even though Φr^* is close to J^*, the approximate LP does not even have a feasible solution. However, the following theorem shows that, with some mild conditions on the basis functions, a bound relating the distance between J^* and Φr^* to the distance between J^* and $\Phi\tilde{r}$ can be developed.

Theorem 2 *If $\Phi v = e$[1] for some v, then we have*

$$\|J^* - \Phi\tilde{r}\|_{1,c} \leq \frac{2}{1-\alpha} \min_r \|J^* - \Phi r\|_\infty.$$

[1] $\Phi v = e$ means

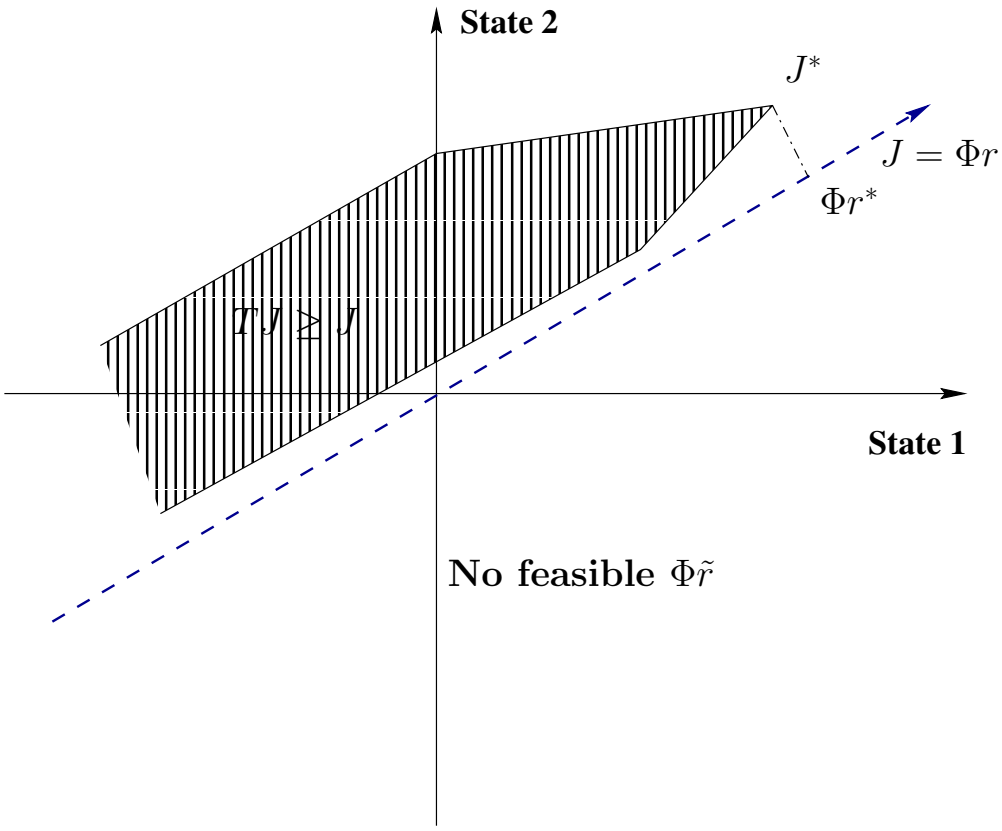

Figure 2: A Special Case of ALP

Before proving Theorem 2, we state and prove the following auxiliary lemma:

Lemma 4 *For all J, let*

$$\tilde{J} = J - \frac{1+\alpha}{1-\alpha}\|J - J^*\|_\infty e$$

Then, we have

$$T\tilde{J} \geq \tilde{J}$$

Proof: Let $\epsilon = \|J - J^*\|_\infty$. Thus we have

$$T\tilde{J} = T(J - \frac{1+\alpha}{1-\alpha}\epsilon e)$$
$$\|TJ - TJ^*\|_\infty \leq \alpha\|J - J^*\|_\infty = \alpha\epsilon$$

Then

$$\begin{aligned} T\tilde{J} &\geq J^* - \alpha\epsilon - \frac{\alpha(1+\alpha)}{1-\alpha}\epsilon e \\ &\geq J - (1+\alpha)\epsilon e - \frac{\alpha(1+\alpha)}{1-\alpha}\epsilon e \\ &= \tilde{J} + \frac{1+\alpha}{1-\alpha}\epsilon e - (1+\alpha)\epsilon e - \frac{\alpha(1+\alpha)}{1-\alpha}\epsilon e \\ &= \tilde{J}. \end{aligned}$$

□

We are now ready to finish the proof of Theorem 2. Let $r^* = \arg\min_r \|J^* - \Phi r\|_\infty$. Let $\epsilon = \|J^* - \Phi r^*\|_\infty$. Then by Lemma 4, we have

$$\Phi\bar{r} = \Phi r^* - \frac{1+\alpha}{1-\alpha}\epsilon e$$

is a feasible solution for the ALP. From Lemma 3, we have

$$\begin{aligned} \|J^* - \Phi\tilde{r}\|_{1,c} &\leq \|J^* - \Phi\bar{r}\|_{1,c} \\ &= \|J^* - \Phi r^* - \frac{1+\alpha}{1-\alpha}\epsilon e\|_{1,c} \\ &\leq \|J^* - \Phi r^*\|_{1,c} + \frac{1+\alpha}{1-\alpha}\epsilon \\ &\leq \|J^* - \Phi r^*\|_\infty + \frac{1+\alpha}{1-\alpha}\epsilon \\ &\leq \epsilon + \frac{1+\alpha}{1-\alpha}\epsilon \\ &= \frac{2}{1-\alpha}\epsilon \end{aligned}$$

□

Theorem 2 establishes that when the optimal cost-to-go function lies close to the span of the basis functions, the approximate LP generates a good approximation. In particular, if the error $\min_r \|J^* - \Phi r\|_\infty$ goes to zero (e.g., as we make use of more and more basis functions) the error resulting from the approximate LP also goes to zero.

Though the above bound offers some support for the linear programming approach, there are some significant weaknesses:

1. The bound calls for an element of the span of the basis functions to exhibit uniformly low error over all states. In practice, however, $\min_r \|J^* - \Phi r\|_\infty$ is typically huge, especially for large-scale problems.

2. The bound does not take into account the choice of state-relevance weights. As demonstrated in the previous section, these weights can significantly impact the approximation error. A sharp bound should take them into account.

In the next lecture, we will show how the previous analysis can be generalized to take into account structure about the underlying MDP and address the aforementioned issues.