CE 273
Markov Decision Processes

Lecture 10

Linear Programming Methods

Previously on Markov Decision Processes

Let $\{X_n, n \geq 0\}$ be a DTMC on $S = \mathbb{Z}^+$ with transition matrix P and initial distribution a. For a given n, the marginal distribution of X_n is

$$\begin{aligned} a_j^{(n)} &= \mathbb{P}\big[X_n = j\big] \, \forall \, j \in S \\ &= \sum_{i \in S} \mathbb{P}\big[X_n = j | X_0 = i\big] \mathbb{P}\big[X_0 = i\big] \text{ (Law of Total Probability)} \\ &= \sum_{i \in S} a_i p_{ij}^{(n)} \end{aligned}$$

where $p_{ij}^{(n)}$ is the probability of going from i to j in exactly n steps. Define the n-step transition matrix $P^{(n)}$ as

$$P^{(n)} = \big[p_{ij}^{(n)}\big]_{|S| \times |S|}$$

Hence, to compute the marginal distributions, we need to compute the n-step transition matrices.

Previously on Markov Decision Processes

$$(TJ)(i) = \min_{u \in U(i)} \left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right\} \forall\, i \in X$$

$$(T_\mu J)(i) = \left\{ g(i,\mu(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j) \right\} \forall\, i \in X$$

Lemma (Monotonicity Lemma)

For any $J : X \to \mathbb{R}$ and $J' : X \to \mathbb{R}$ such that $J \leq J'$ and a stationary policy μ,

1. $T^k J \leq T^k J'$
2. $T_\mu^k J \leq T_\mu^k J'$

Lemma (Constant Shift Lemma)

For every k, and $J : X \to \mathbb{R}$ and stationary policy μ

1. $\big(T^k(J + re)\big)(i) = \big(T^k J\big)(i) + \alpha^k r$
2. $\big(T_\mu^k(J + re)\big)(i) = \big(T_\mu^k J\big)(i) + \alpha^k r$

Previously on Markov Decision Processes

Value Iteration

```
Fix a tolerance level $\epsilon > 0$
Select $J_0 \in B(X)$ and $k \leftarrow 0$
$J_1 \leftarrow TJ_0$
while $\|J_{k+1} - J_k\| > \frac{\epsilon(1-\alpha)}{2\alpha}$ do
    $k \leftarrow k + 1$
    $J_{k+1} \leftarrow TJ_k$
end while
Select $\mu_\epsilon$ that satisfies $T_{\mu_\epsilon} J_{k+1} = TJ_{k+1}$
```

In other words, the policy constructed at termination can be written as

$$\mu_\epsilon(i) \in \arg\min_{u \in U(i)} \mathbb{E}\left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J_{k+1}(j) \right\}$$

Previously on Markov Decision Processes

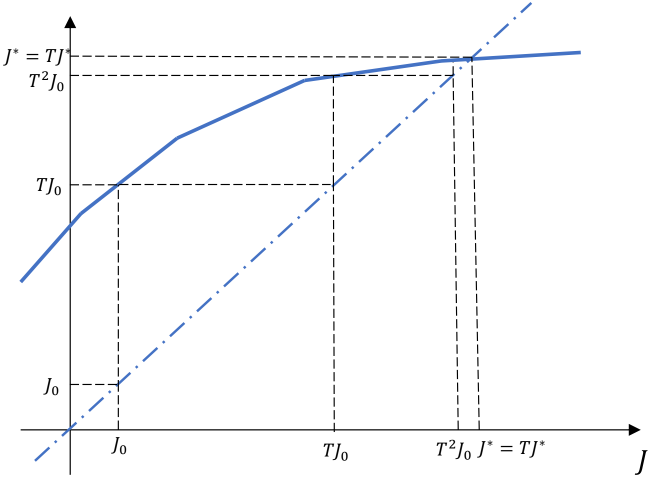


Figure: Value Iteration

Previously on Markov Decision Processes

```
POLICY ITERATION
  Pick an initial policy $\mu_0$ (say a Greedy policy)
  Set $\mu_1$ such that $T_{\mu_1} J_{\mu_0} = T J_{\mu_0}$ and $k \leftarrow 0$
  while $\mu_{k+1} \neq \mu_k$ do
      $k \leftarrow k + 1$
      Compute $J_{\mu_k}$ by solving $J_{\mu_k} = T_{\mu_k} J_{\mu_k}$, i.e.,          ▷ Policy Evaluation
          $J_{\mu_k} = (I - \alpha P_{\mu_k})^{-1} g_{\mu_k}$
      Compute a new policy $\mu_{k+1}$ that satisfies          ▷ Policy Improvement
          $T_{\mu_{k+1}} J_{\mu_k} = T J_{\mu_k}$
  end while
  $\mu^* \leftarrow \mu_k$ and $J^* \leftarrow J_{\mu_k}$
```

Since the termination criteria in the above algorithm compares policies between consecutive iterations, breaking ties arbitrarily can slow convergence.

Hence, we set $\mu_{k+1}(i) = \mu_k(i)$ whenever possible or stop when $J_{\mu_k} = TJ_{\mu_k}$

Previously on Markov Decision Processes

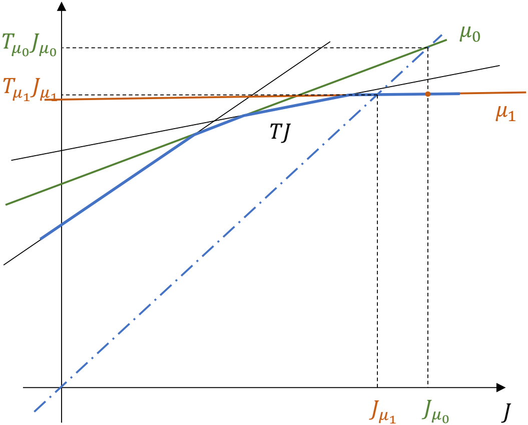


Figure: Policy Iteration

Lecture Outline

1. Linear Programming Review
2. LP Methods
3. Advantages and Disadvantages

Lecture Outline

Linear Programming Review

Linear Programming Review

Introduction

Linear programs are a special class of optimization problems in which the objective and the constraints are linear.

They can be written in the following canonical form

$$\begin{aligned} \min\ & c^T x \\ \text{s.t.}\ & Ax \leq b \\ & x \geq 0 \end{aligned}$$

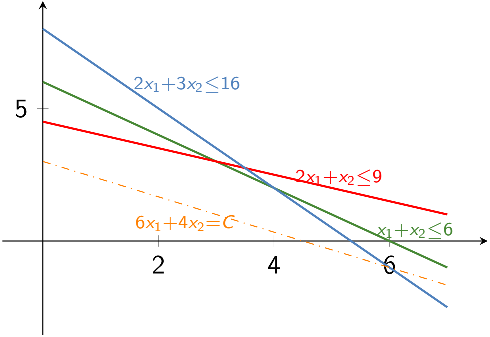
where c, x, and b are vectors of dimensions $n \times 1$, $n \times 1$, and $m \times 1$ respectively. A is a $m \times n$ matrix.

Linear Programming Review

Graphical Solutions

Consider the following LP which involves two decision variables:

$$
\begin{aligned}
&\max\ 6x_1 + 4x_2\\
\text{s.t. }\ & x_1 + x_2 \le 6\\
& 2x_1 + x_2 \le 9\\
& 2x_1 + 3x_2 \le 16\\
& x_1 \ge 0\\
& x_2 \ge 0
\end{aligned}
$$



Lines of the form $6x_1 + 4x_2 = C$ are also called isoprofit or level curves.

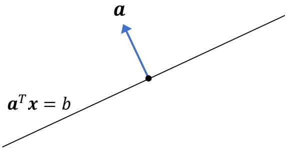
The feasible region of LPs is a polyhedron and the optimum occurs at an extreme/corner point. How many corner points are possible? In general, we would have at most $\binom{m+n}{n}$ corner solutions.

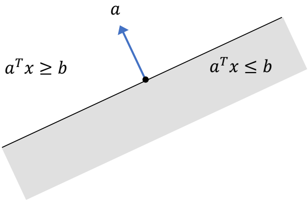
Linear Programming Review

Feasible Region

Definition (Hyperplane)

Sets of the form $\{x \in \mathbb{R}^n \,|\, a^T x = b\}$, where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$ are called hyperplanes.





Definition (Halfspace)

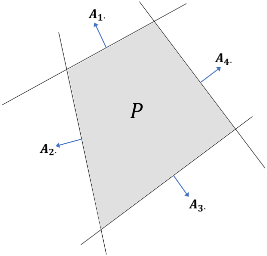
Sets of the form $\{x \in \mathbb{R}^n \,|\, a^T x \leq b\}$, where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$ are called halfspaces.

Linear Programming Review

Feasible Region

Definition (Polyhedron)

A polyhedron is a set of the form $P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$.



The feasible region is a convex set and a point $x \in X$ is said to be an extreme point if it cannot be expressed as a strict convex combination of two distinct points in X.

Mathematically, x is an extreme point if $\nexists\ x_1, x_2 \in X$ with $x_1 \neq x_2$ and $\lambda \in (0,1)$ such that $x = \lambda x_1 + (1-\lambda)x_2$.

Linear Programming Review

Augmented LP

Definition

A collection of vectors $a_1, a_2, \ldots, a_n$ is linearly independent if $\sum_{j=1}^{n} \lambda_j a_j = 0 \Rightarrow \lambda_j = 0 \, \forall \, j$

Definition

A basic solution of $Ax = b$ is a solution that only uses linearly independent columns of A.

In other words, the x values corresponding to all other columns are 0s.

Proposition

Let $x' \in X = \{x : Ax = b, x \geq 0\}$. Then x' is an extreme point of $X \Leftrightarrow x'$ is a non-negative basic solution of $Ax = b$.

Linear Programming Review

Augmented LP

One can write the $Ax \leq b$ constraints in equality form by adding slack variables as follows:

$$\begin{bmatrix} A \mid I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = b$$

Thus, the constraints of the LP can be reformulated as $A'x' = b$, $x' \geq 0$. A' has m rows and $m+n$ columns. To get a basic solution, we need to select m columns, which can be done in $\binom{m+n}{m}$ ways.

For each such choice of m columns, we get a solution x' and we check if it is ≥ 0. Mathematically,

$$A' = \begin{bmatrix} A'_B \mid A'_N \end{bmatrix} \begin{bmatrix} x'_B \\ x'_N \end{bmatrix} = b$$

and $A'_B x'_B = b$ and $x'_N = 0$. The simplex method used to solve LPs enumerates corner points by moving from one basic solution to another.

Linear Programming Review

Duality

Given a LP, a closely related formulation called the dual LP can be written as follows:

Primal LP

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

Dual LP

$$\begin{aligned} \max \quad & b^T y \\ \text{s.t.} \quad & A^T y \leq c \\ & y \leq 0 \end{aligned}$$

The dual has the following features:

- If the primal is a minimization problem, the dual has a maximization objective
- The dual has as many variables as the constraints of the primal
- The dual has as many constraints as the number of variables in the primal
- The dual of the dual is same as the primal LP

Linear Programming Review

Duality

- The dual variables have neat economic interpretation as shadow prices and reflect the sensitivity of the objective to the RHS of the constraints.
- For $\geq$ constraints, the RHS can be viewed as requirements and hence increasing them will worsen the objective. On the other hand, $\leq$ constraints can be interpreted as resource constraints. Increasing resources will improve the objective. This logic can be used to determine the sign of the dual variables.
- When there are equality constraints (can be written as ≥ 0 and ≤ 0 constraints), the associated variables in the dual is unconstrained and vice versa.

Note that in the above discussion, when counting constraints, we consider only the structural ones and not non-negativity constraints.

Linear Programming Review

Duality

Table: Quick Reference for Writing the Dual

min problem		max problem
ith constraint $\geq$	$\leftrightarrow$	ith variable ≥ 0
ith constraint $\leq$	$\leftrightarrow$	ith variable ≤ 0
ith constraint $=$	$\leftrightarrow$	ith variable is unrestricted
jth variable ≥ 0	$\leftrightarrow$	jth constraint $\leq$
jth variable ≤ 0	$\leftrightarrow$	jth constraint $\geq$
jth variable is unrestricted	$\leftrightarrow$	jth constraint $=$

Linear Programming Review

Duality

Theorem (Weak Duality Theorem)

If x is feasible to the primal and y is feasible to the dual, then $c^T x \geq b^T y$

Theorem (Strong Duality Theorem)

If the primal and the dual are feasible, then there exists x^* and y^* such that $c^T x^* = b^T y^*$

Thus, one can solve the primal or dual and get the same optimum objective!

Another useful result is the complementary slackness condition according to which the optimal primal-dual pair satisfies

$$y_i^*(A_{i.}x^* - b_i) = 0 \, \forall, i = 1, \ldots, m$$

Lecture Outline

LP Methods

LP Methods

Introduction

In addition to VI and PI, one can use linear programming to solve an MDP.

From the monotonicity lemma,

$$J \le TJ \Rightarrow J \le J^*$$

Also, from the Bellman equations $J^* = TJ^*$. Thus, among all functions J that satisfy $J \le TJ$, J^* is the "largest".

$$J \le TJ$$

$$\Rightarrow J(i) \le g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \, \forall \, i = 1, \ldots, n, u \in U(i)$$

Thus, if we treat $J(i)$s as the decision variables, these form a linear set of constraints. What should the objective be?

LP Methods

Alternate LP

Since J^* must be "largest" component wise, we can think of J^* as the solution to the linear program,

$$\max_i \sum_{i=1}^{n} y(i)$$

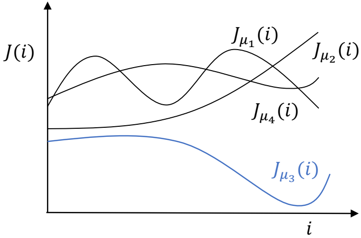
$$\text{s.t. } y(i) \leq g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) y(j) \qquad \forall\, i = 1, \ldots, n, u \in U(i)$$

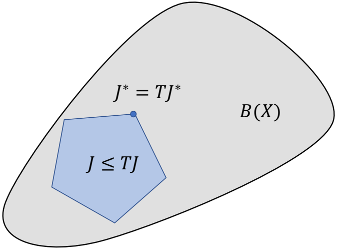
If each state has m actions, the above LP has n variables and mn constraints.

LP Methods

Geometric Interpretation

Note that the new optimization model has a maximization objective but we are still minimizing the total expected discounted cost.

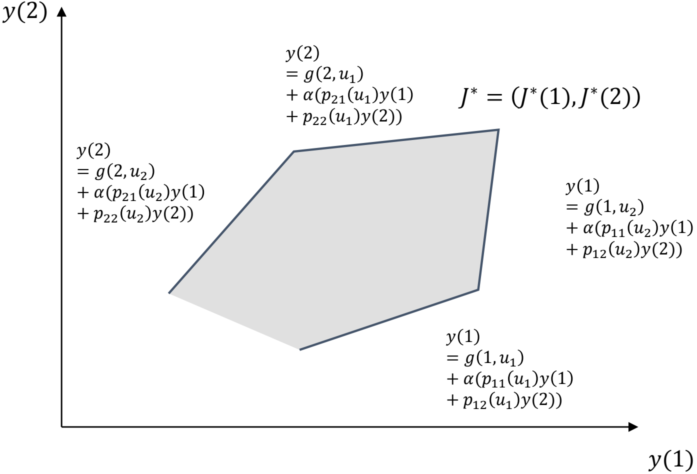




The minimization objective we had earlier was across the set of policies (a finite set when the states and actions are finite). The LP on the other hand operates in the space of value functions.

LP Methods

Geometric Interpretation



$$y(2) = g(2, u_1) + \alpha(p_{21}(u_1)y(1) + p_{22}(u_1)y(2))$$

$$J^* = (J^*(1), J^*(2))$$

$$y(2) = g(2, u_2) + \alpha(p_{21}(u_2)y(1) + p_{22}(u_2)y(2))$$

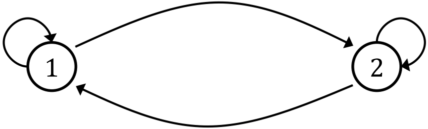
$$y(1) = g(1, u_2) + \alpha(p_{11}(u_2)y(1) + p_{12}(u_2)y(2))$$

$$y(1) = g(1, u_1) + \alpha(p_{11}(u_1)y(1) + p_{12}(u_1)y(2))$$

LP Methods

Example

Formulate the linear program for the following MDP with two states 1 and 2. Let $a_1 = a_2 = 0.5$. Assume that the discount factor is 0.9.



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 0.5$
- $p_{1j}(u_1) = [3/4 \;\; 1/4]$
- $p_{1j}(u_2) = [1/4 \;\; 3/4]$

- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = 1, g(2, u_2) = 3$
- $p_{2j}(u_1) = [3/4 \;\; 1/4]$
- $p_{2j}(u_2) = [1/4 \;\; 3/4]$

LP Methods

Alternate Formulation

The coefficients of the objective can be augmented by positive scalars $a_1, a_2, \ldots, a_n$ (Why?). We will set $\sum_{i=1}^{n} a_i = 1$ so that the a's can be interpreted as the initial distribution over the state space just like in DTMCs.

$$\max_{i} \sum_{i=1}^{n} a_i y(i)$$

$$\text{s.t. } y(i) \leq g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) y(j) \qquad \forall\, i = 1, \ldots, n, u \in U(i)$$

which can be rewritten in the canonical form as

$$\max_{i} \sum_{i=1}^{n} a_i y(i)$$

$$\text{s.t. } y(i) - \alpha \sum_{j=1}^{n} p_{ij}(u) y(j) \leq g(i,u) \qquad \forall\, i = 1, \ldots, n, u \in U(i)$$

Formulate the dual LP of the above primal.

LP Methods

Dual LP

- The number of dual variables equal to the number of constraints in the primal.
- Since the primal constraints are of the $\leq$ form, the dual variables must be ≥ 0.
- Since the primal variables are unconstrained, the dual will have equality constraints

Hence, define dual variables $z(i, u)$ where $i \in X, u \in U(i)$.

$$\min \sum_{i=1}^{n} \sum_{u \in U(i)} g(i, u) z(i, u)$$

$$\text{s.t.} \quad \sum_{u \in U(i)} z(i, u) - \sum_{j=1}^{n} \sum_{u \in U(j)} \alpha p_{ji}(u) z(j, u) = a_i \qquad \forall\, i = 1, \dots, n$$

$$z(i, u) \geq 0 \qquad \forall\, u \in U(i), i = 1, \dots, n$$

LP Methods

Dual LP

Solving the primal directly gives us J^*. How do we get the optimal policy from the primal? Use $T_\mu J^* = TJ^*$. It can be shown that the optimal policy does not depend on the 'a' vector.

How can we get the optimal policy and value functions if we solve the dual? The variables in the dual $z(i,u)$ require some interpretation.

Notice that if $z(i,u)$ is a feasible solution to the dual LP, then for each $i = 1, \ldots, n$, $\sum_{u \in U(i)} z(i,u) > 0$. (Why?)

From the constraint of the dual LP, for a state i,

$$\sum_{u \in U(i)} z(i,u) = a_i + \sum_{j=1}^{n} \sum_{u \in U(j)} \alpha p_{ji}(u) z(j,u)$$

and α and a_i are strictly positive. Thus, $(\sum_{u \in U(i)} z(i,u))^{-1}$ is well defined. We will use this observation to define randomized policies.

LP Methods

Dual LP

Proposition

1. *For each randomized policy μ, define $z_\mu(i,u)$ as*

$$z_\mu(i,u) = \sum_{j=1}^{n} a_j \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu\left[x_k = i, u_k = u | x_0 = j\right]$$

Then, $z_\mu(i,u)$ is a feasible solution to the dual LP

2. *Suppose $z(i,u)$ is a dual feasible solution. Then, a randomized stationary policy μ constructed such that*

$$\mathbb{P}\left[\mu(i) = u\right] = \frac{z(i,u)}{\sum_{u' \in U(i)} z(i,u')}$$

Then $z_\mu(i,u)$ defined using part (1) equals $z(i,u)$ for all $i \in X, u \in U(i)$.

Note that the above proposition deals with only dual feasible solutions. We will get to optimality later.

LP Methods

Dual LP

Proof of (1).

We need to show $\sum_{u\in U(i)} z_\mu(i,u) - a_i = \sum_{j=1}^n \sum_{u\in U(j)} \alpha p_{ji}(u) z_\mu(j,u)$

$$\sum_{j=1}^{n} \sum_{u\in U(j)} \alpha p_{ji}(u) z_\mu(j,u)$$

$$= \sum_{j=1}^{n} \sum_{u\in U(j)} \alpha p_{ji}(u) \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu \left[x_k = j, u_k = u | x_0 = l\right]$$

$$= \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^{k+1} \sum_{j=1}^{n} \sum_{u\in U(j)} p_{ji}(u) \mathbb{P}_\mu \left[x_k = j, u_k = u | x_0 = l\right]$$

$$= \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^{k+1} \sum_{j=1}^{n} \sum_{u\in U(j)} \mathbb{P}_\mu \left[x_{k+1} = i | x_k = j, u_k = u\right] \mathbb{P}\left[x_k = j, u_k = u | x_0 = l\right]$$

$$= \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^{k+1} \mathbb{P}_\mu \left[x_{k+1} = i | x_0 = l\right] = \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^{k+1} P_\mu^{(k+1)}(l,i)$$

LP Methods

Dual LP

Proof of (1).

Now consider

$$
\begin{aligned}
&\sum_{u\in U(i)} z_\mu(i,u) - a_i \\
&= \sum_{u\in U(i)} \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu[x_k = i, u_k = u | x_0 = l] - a_i \\
&= \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^k \sum_{u\in U(j)} \mathbb{P}_\mu[x_k = i, u_k = u | x_0 = l] - a_i \\
&= \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu[x_k = i | x_0 = l] - a_i = \sum_{l=1}^{n} a_l \left(\sum_{k=0}^{\infty} \alpha^k P_\mu^k(l,i) - I(l,i) \right) \\
&= \sum_{l=1}^{n} a_l \left(\sum_{k=0}^{\infty} \alpha^k P_\mu^{(k)}(l,i) - \alpha^0 P_\mu^{(0)}(l,i) \right) = \sum_{l=1}^{n} a_l \sum_{k=0}^{\infty} \alpha^{k+1} P_\mu^{(k+1)}(l,i)
\end{aligned}
$$

■

LP Methods

Dual LP

Thus, the dual variable $z(i, u)$ can be interpreted as "total discounted joint probability of the system occupying a state i and choosing action u" assuming that the initial state distribution is a_i.

$$z_\mu(i, u) = \sum_{j=1}^{n} a_j \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu \left[x_k = i, u_k = u | x_0 = j \right]$$

When multiplied by $g(i, u)$ and summed over all state-action pairs, i.e., $\sum_{i=1}^{n} \sum_{u \in U(i)} g(i, u) z(i, u)$, we get the total discounted cost (starting with initial distribution a).

This interpretation applies to any feasible dual solution. In addition, if we have a basic feasible solution (corner point), the z values have an interesting property.

Note that the with the assumption of bounded costs, the dual has a feasible solution.

Proposition

Let z be a basic feasible solution, then $z(i,u) > 0$ for exactly one $U(i)$ for all $i = 1,\ldots,n$

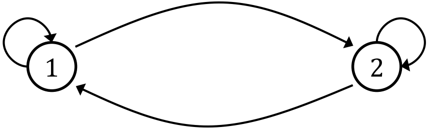
Since the optimal solution is a basic feasible solution, this is true for the optimal solution too. One can use the complementary slackness condition to prove this for the optimal solution.

Using the above proposition, given an optimal dual solution z^*, we can construct a deterministic optimal policy by looking at the actions for which $z(i,u) > 0$.

LP Methods

Example

Formulate the dual linear program for the following MDP with two states 1 and 2. Assume that the discount factor is 0.9. Let $a_1 = a_2 = 0.5$.



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 0.5$
- $p_{1j}(u_1) = [3/4 \;\; 1/4]$
- $p_{1j}(u_2) = [1/4 \;\; 3/4]$

- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = 1, g(2, u_2) = 3$
- $p_{2j}(u_1) = [3/4 \;\; 1/4]$
- $p_{2j}(u_2) = [1/4 \;\; 3/4]$

Lecture Outline

Advantages and Disadvantages

Advantages and Disadvantages

Disadvantages

There is both good news and bad news with the LP methods. The bad news first:

- Generating the simplex tableau takes time
- Structural results such as convexity and monotonicity cannot be proven

Advantages and Disadvantages

Advantages

The good news with LPs is that we can

- Use existing LP solvers
- Perform sensitivity analysis
- Add side constraints

Advantages and Disadvantages

MDPs with Side Constraints

One could have a secondary objective say the cost of choosing a certain control in a state. For instance:

- You are navigating a drone and your objective is to match a certain trajectory (one-step costs are the deviations in the current and target position), but you may have a constraint on the fuel consumption.
- You might want to reach a particular destination in the shortest possible time but you may also have an upper bound on the operating costs/tolls.

Advantages and Disadvantages

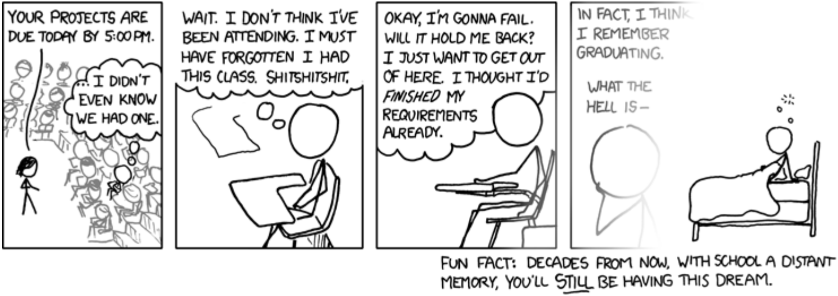
MDPs with Side Constraints

Such models can be formulated as

$$
\begin{aligned}
&\min \sum_{i=1}^{n} \sum_{u \in U(i)} g(i,u) z(i,u) \\
\text{s.t.}\quad &\sum_{u \in U(i)} z(i,u) - \sum_{j=1}^{n} \sum_{u \in U(j)} \alpha p_{ji}(u) z(j,u) = a_i && \forall\, i = 1, \dots, n \\
&\sum_{i=1}^{n} \sum_{u \in U(i)} c(i,u) z(i,u) \le C \\
&z(i,u) \ge 0 && \forall\, u \in U(i), i = 1, \dots, n
\end{aligned}
$$

The optimal policy in these problems is usually randomized!

Your Moment of Zen


YOUR PROJECTS ARE DUE TODAY BY 5:00 PM.
...I DIDN'T EVEN KNOW WE HAD ONE.
WAIT. I DON'T THINK I'VE BEEN ATTENDING. I MUST HAVE FORGOTTEN I HAD THIS CLASS. SHITSHITSHIT.
OKAY, I'M GONNA FAIL. WILL IT HOLD ME BACK? I JUST WANT TO GET OUT OF HERE. I THOUGHT I'D *FINISHED* MY REQUIREMENTS ALREADY.
IN FACT, I THINK I REMEMBER GRADUATING.
WHAT THE HELL IS—
FUN FACT: DECADES FROM NOW, WITH SCHOOL A DISTANT MEMORY, YOU'LL STILL BE HAVING THIS DREAM.