Linear Programming Methods for MDPs

Daniel Russo

April 27, 2020

Columbia University

Table of Contents

Primal and dual LP formulations of MDPs

Quick extensions

The LP Approach to ADP

Deriving the primal LP

Lemma If $J \preceq TJ$ then $J \preceq J^*$.
Proof: $J \preceq TJ \preceq T^2 J \preceq \cdots \preceq T^k J \to J^*$.

J^* is the largest vector satisfying $J \preceq TJ$.
That is, J^* solves the multi-objective problem:

$$\begin{aligned} \text{Maximize} \quad & J(s) \quad \forall s \in \mathcal{S} \\ \text{subject to} \quad & J \preceq TJ \end{aligned}$$

Deriving the primal LP (2)

Introduce state-relevance weights $\nu(s) > 0$ with $\sum_{s \in \mathcal{S}} \nu(s) = 1$

- The specific choice of ν does not affect the solution, but it gives an interpretation of the dual.
- Similarly, that $\sum_{s \in \mathcal{S}} \nu(s) = 1$ helps only in interpreting the dual.

J^* is the unique solution to:

$$\begin{aligned} \text{Maximize} \quad & \nu^\top J \\ \text{subject to} \quad & J \preceq TJ \end{aligned}$$

Deriving the primal LP (3)

The $|\mathcal{S}|$ nonlinear inequalities

$$J(s) \leq TJ(s) = \min_{u \in U(s)} g(s,u) + \alpha \sum_{s' \in \mathcal{S}} p_{ss'}(u) J(s') \qquad \forall s \in \mathcal{S}$$

are equivalent to the $\sum_s |U(s)|$ linear inequalities

$$J(s) \leq g(s,u) + \alpha \sum_{s' \in \mathcal{S}} p_{ss'}(u) J(s') \qquad \forall s \in \mathcal{S},\ u \in U(s).$$

This leads to the final LP formulation of computing J^*:

(Primal) Linear program for computing J^*.

$$\begin{aligned} \text{Maximize} \quad & \sum_{s \in \mathcal{S}} \nu(s) J(s) \\ \text{subject to} \quad & J(s) \leq g(s,u) + \alpha \sum_{s' \in \mathcal{S}} p_{ss'}(u) J(s') \quad \forall s \in \mathcal{S},\ u \in U(s) \end{aligned}$$

Dual LP

(Dual) Linear program

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{1-\alpha} \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s,u) g(s,u) \\ \text{subject to} \quad & \sum_{u \in U(s')} x(s',u) = (1-\alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s,u) P_{ss'}(u) \\ & x(s,u) \geq 0 \end{aligned}$$

You will usually see the from below instead, but I find the term above, where we tranform $x \to (1-\alpha)x$, to be more interpretable.

$$\begin{aligned} \text{Minimize} \quad & \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s,u) g(s,u) \\ \text{subject to} \quad & x(s',u) = \nu(s') + \alpha \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s,u) P_{ss'}(u) \\ & x(s,u) \geq 0 \end{aligned}$$

Interpretation of the dual LP

- Rather than discounting, equivalently think of this as a problem with average cost objective, where in every period there is a fixed probability $1 - \alpha$ that the system "restarts" in a state drawn from ν.
- The variable $x(s, u)$ represents the steady state fraction of the spend playing u in state s.
- The constraint

$$\sum_{u \in U(s')} x(s', u) = (1 - \alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s, u) P_{ss'}(u)$$

  is a balance equation required of a stationary distribution:
  - The LHS is the probability of being in state s'.
  - The RHS is the steady-state "flow" into state s'.

Interpretation of the dual LP (2)

(Dual) Linear program

$$\text{Minimize} \quad \frac{1}{1-\alpha}\sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x(s,u)g(s,u)$$

$$\text{subject to} \quad \sum_{u\in U(s')} x(s',u) = (1-\alpha)\nu(s') + \alpha\sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x(s,u)P_{ss'}(u)$$

$$x(s,u) \geq 0$$

The set of feasible solutions to this LP is precisely the set of all discounted state-occupancy measures

$$x_\mu(s,u) = (1-\alpha)\sum_{t=0}^{\infty}\alpha^t\mathbb{P}_\mu(S_t = s, u_t = u | s_0 \sim \nu)$$

induced by stationary randomized policies μ. The dual LP has the intuitive form of minimizing the geometrically averaged cost incurred among feasible state-action frequencies.

From stationary policies to feasible solutions

Lemma For a stationary (possibly randomized) policy μ, let

$$x_\mu(s,u) = (1-\alpha)\sum_{k=0}^{\infty}\alpha^k \mathbb{P}(s_k = s, u_k = u | s_0 \sim \nu).$$

Then x_μ is a feasible solution.
proof: Define $x_\mu(s) = \sum_{u\in U(s)} x_\mu(s,u)$. Our goal is to show feasibility:

$$\begin{aligned}
x_\mu(s') &= (1-\alpha)\nu(s') + \alpha\sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x_\mu(s,u)P_{ss'}(u) \\
&= (1-\alpha)\nu(s') + \alpha\sum_{s\in\mathcal{S}} x_\mu(s)P_{ss'}(\mu)
\end{aligned}$$

To show this, we will show,

$$\alpha\sum_{s\in\mathcal{S}} x_\mu(s)P_{ss'}(\mu) = x_\mu(s') - (1-\alpha)\nu(s')$$

From stationary policies to feasible solutions (2)

$$\begin{aligned}
\alpha \sum_{s \in \mathcal{S}} x_\mu(s) P_{ss'}(\mu) &= \alpha(1-\alpha) \sum_{s \in \mathcal{S}} \left(\sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu(s_k = s | s_0 \sim \nu) \right) P_{ss'}(\mu) \\
&= \alpha(1-\alpha) \sum_{k=0}^{\infty} \alpha^k \left(\sum_{s \in \mathcal{S}} \mathbb{P}_\mu(s_k = s | s_0 \sim \nu) P_{ss'}(\mu) \right) \\
&= \alpha(1-\alpha) \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu(s_{k+1} = s' | s_0 \sim \nu) \\
&= (1-\alpha) \sum_{k=0}^{\infty} \alpha^{k+1} \mathbb{P}_\mu(s_{k+1} = s' | s_0 \sim \nu) \\
&= (1-\alpha) \sum_{k=1}^{\infty} \alpha^k \mathbb{P}_\mu(s_k = s' | s_0 \sim \nu) \\
&= \underbrace{(1-\alpha) \sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu(s_k = s' | s_0 \sim \nu)}_{x_\mu(s')} - (1-\alpha)\nu(s').
\end{aligned}$$

From feasible solutions to stationary policies

Lemma: Let $x(s,u)$ be a feasible solution to the dual LP. Let $\mu(s,u) = \frac{x(s,u)}{\sum_{u' \in U(s)} x(s,u')}$. Then,

$$x(s,u) = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k \mathbb{P}_\mu(s_k = s, u_k = u | s_0 \sim \nu) := x_\mu(s,u)$$

Proof: Define $x(s) = \sum_{u \in U(s)} x(s,u)$ and $x_\mu(s) = \sum_{u \in U(s)} x_\mu(s,u)$.

$$\begin{aligned}
x(s') &= (1-\alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} \sum_{u \in U(s)} x(s,u) P_{ss'}(u) \\
&= (1-\alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} x(s) \sum_{u \in U(s)} \left(\frac{x(s,u)}{\sum_{u' \in U(s)} x(s,u')} \right) P_{ss'}(u) \\
&= (1-\alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} x(s) \sum_{u \in U(s)} \mu(s,u) P_{ss'}(u) \\
&= (1-\alpha)\nu(s') + \alpha \sum_{s \in \mathcal{S}} x(s) P_{ss'}(\mu)
\end{aligned}$$

In vector notation, $x^\top = (1-\alpha)\nu^\top + \alpha x^\top P_\mu$, or $x = (1-\alpha)\nu^\top (I - \alpha P_\mu)^{-1} = x_\mu \in \mathbb{R}^{|\mathcal{S}|}$.

Table of Contents

Primal and dual LP formulations of MDPs

Quick extensions

The LP Approach to ADP

Extension: Dual LP for average cost objectives

In a finite state, average cost problem, the goal is to minimize

$$\limsup_{K\to\infty} \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K} g(s_k, u_k)\right]$$

For unichain MDPs, where each state is reachable from each other state, the optimal objective value is $\lambda^* = \lim_{\alpha\to 1}(1-\alpha)J^*(s)$.

(Dual) Linear program for average cost, unichain, MDPs

$$\begin{aligned}
\text{Minimize} \quad & \sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x(s,u)g(s,u) \\
\text{subject to} \quad & \sum_{u\in U(s')} x(s',u) = \sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x(s,u)P_{ss'}(u) \\
& \sum_{s\in\mathcal{S}}\sum_{u\in U(s)} x(s,u) = 1 \\
& x(s,u) \geq 0
\end{aligned}$$

Extension: Adding constraints

A central advantage of the LP approach is that it is easy to add constraints on steady-state behavior, as in the following LP. Here c represents an alternative cost measure whose average must lie below a threshold $\bar{c}$.

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{s\in\mathcal{S}} \sum_{u\in U(s)} x(s,u) g(s,u) \\
\text{subject to} \quad & \sum_{u\in U(s')} x(s',u) = \sum_{s\in\mathcal{S}} \sum_{u\in U(s)} x(s,u) P_{ss'}(u) \\
& \sum_{s\in\mathcal{S}} \sum_{u\in U(s)} x(s,u) c(s,u) \le \bar{c} \\
& \sum_{s\in\mathcal{S}} \sum_{u\in U(s)} x(s,u) = 1 \\
& x(s,u) \ge 0
\end{aligned}
$$

Table of Contents

Primal and dual LP formulations of MDPs

Quick extensions

The LP Approach to ADP

The LP Approach to Approximate DP

We can approximate MDP solutions by approximating:

1. The primal LP.
   - Since the decision variable is the cost-to-go function, this is a kind of cost-to-go approximation.
2. The dual LP
   - This is more akin to policy approximation. But to my knowledge, current approaches require approximating the set of feasible occupancy-measures.

We'll look at approach 1 as introduced by Schweitzer and Seidmann [1985] and follow the understanding developed in De Farias and Van Roy [2003, 2004].

Cost-to-go approximation in the primal LP

We face the problem

$$\max\{\nu^\top J : TJ \succeq J\}$$

which can be formulated as an LP. Believing $J^* \approx \Phi\theta$, we introduce the auxiliary problem

$$\begin{aligned} \text{Maximize}_\theta \quad & \nu^\top \Phi\theta \\ \text{subject to} \quad & T\Phi\theta \succeq \Phi\theta \end{aligned}$$

We can write this a an LP with decision variable $\theta \in \mathbb{R}^d$:

$$\begin{aligned} \text{Maximize} \quad & (\nu^\top \Phi)\theta \\ \text{Subject to} \quad & g(s,u) + \alpha \sum_{s' \in \mathcal{S}} p_{ss'}(u) \Phi\theta(s') \geq (\Phi\theta)(s) \quad \forall s \in \mathcal{S},\, u \in U(s). \end{aligned}$$

Constraint sampling

- A first issue is that the LP has $|\mathcal{S}| \times |U|$ constraints, and we have in mind problems where $|\mathcal{S}|$ is enormous.
- De Farias and Van Roy [2004] provide theoretical analysis of a constraint sampling approach, where we impose only a randomly sampled subset of the constraints.
- The intuition is that most constraints are irrelevant, and under some conditions, many provide nearly redundant information. Recall that basic feasible solution is described by $d << |\mathcal{S}|$ active linearly independent constraints.

Approximation guarantee and feature selection

Assume we can exactly solve the problem:

$$\begin{aligned} \mathrm{Maximize}_{\theta} \quad & \nu^{\top}\Phi\theta \\ \text{subject to} \quad & T\Phi\theta \succeq \Phi\theta \end{aligned}$$

In what sense are we approximating J^*?

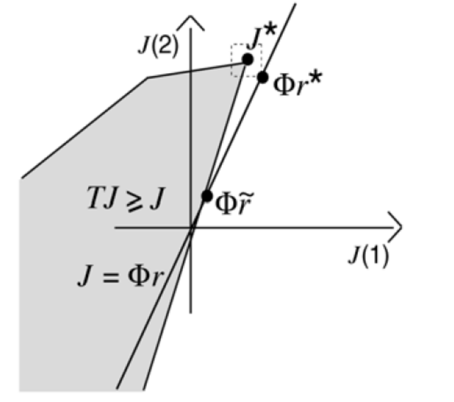
First Lemma: A vector $\tilde{\theta}$ solves the optimization problem above if and only if it solves

$$\begin{aligned} \mathrm{Minimize}_{\theta} \quad & \|J^* - \Phi\theta\|_{1,\nu} \\ \text{subject to} \quad & T\Phi\theta \succeq \Phi\theta \end{aligned}$$

Why? Subtract $\nu^{\top} J^*$ from the objective. The constraint implies $\Phi\theta \preceq J^*$, so $\nu^{\top}(\Phi\theta - J^*) = -\|J^* - \Phi\theta\|_{1,\nu}$.

Quality of the approximation

- In general, we cannot (even nearly) achieve $\min_\theta \|J^* - \Phi\theta\|_{1,\nu}$.
- There may be no *feasible* solution close to the best-approximate cost-to-go function.



A simple error bound and feature selection

Theorem Suppose e is in the span of the columns of Φ. Then, the approximate LP is feasible. If $\tilde{\theta}$ is an optimal solution to the approximate LP,

$$\|J^* - \Phi\tilde{\theta}\|_{1,\nu} \leq \frac{2}{1-\alpha} \min_{\theta} \|J^* - \Phi\theta\|_\infty$$

Proof of feasibility: Let $\Phi\hat{\theta} = e$. For any θ consider a new $\theta_\gamma = \theta - \gamma\hat{\theta}$. We have

$$\Phi\theta_\gamma = \Phi\theta - \gamma e$$

$$T\Phi\theta_\gamma = T(\Phi\theta - \gamma e) = T\Phi\theta - \alpha\gamma e$$

For $\gamma \geq \|\Phi\theta - T\Phi\theta\|_\infty/(1-\alpha)$.

$$T\Phi\theta_\gamma - \Phi\theta_\gamma = (\Phi\theta - T\Phi\theta) - (1-\alpha)\gamma e \succeq 0$$

Proof

Take $\theta^* = \text{argmin}_\theta \|\Phi\theta - J^*\|_\infty$ and $\epsilon = min_\theta \|\Phi\theta - J^*\|_\infty$.

Show

$$\|T\Phi\theta^* - J^*\|_\infty \leq \alpha\|\Phi\theta - J^*\|_\infty = \alpha\epsilon$$

Then,

$$\begin{aligned} T\left(\Phi(\theta^* - \gamma\hat{\theta})\right) &= T\Phi\theta^* - \alpha\gamma e \\ &\geq J^* - \alpha\epsilon e - \alpha\gamma e \\ &\geq \Phi\theta^* - \epsilon e - \alpha\epsilon e - \alpha\gamma e. \quad = \Phi\left(\theta^* - \gamma\hat{\theta}\right) + (1-\alpha)\gamma e \end{aligned}$$

This ensures $\theta^* - \frac{1+\alpha}{1-\alpha}\hat{\theta}$ feasible. The triangle inequality can be used to show

$$\|\Phi\left(\theta^* - \frac{1+\alpha}{1-\alpha}\hat{\theta}\right) - J^*\|_\infty \leq \frac{2}{1-\alpha}\epsilon.$$

Since there exists a feasible solution close to J^* in infinity norm, the optimal solution in $\|\cdot\|_{1,\nu}$ must be at least that close.

Choice of basis and state-relevance weights

What is special about e? It acts a Lyapunov function: $e \succeq 0$ and

$$Te \succeq \kappa e \qquad \text{where } \kappa < 1.$$

The natural choice of state-relevance weights are uniform. Other choices work better in other problems.

Warmup Example: Autonomous Queue

An N state autonomous queue has state dynamics

$$s_{t+1} = \begin{cases} \min\{x_t + 1, N-1\} & \text{with prob. } p \\ \max\{x_t - 1\,, 0\} & \text{otherwise} \end{cases}$$

and no decision-variables. The stationary probabilities are $\pi(x) = \pi(0)\left(\frac{p}{1-p}\right)^x$ for $x = 0, \cdots N-1$. If $V(X) = x^2 + 2/(1-\alpha)$, then

$$TV \preceq \frac{1+\alpha}{2} V$$

De Farias and Van Roy [2003] suggest to choose $\nu = \pi$ and to include the basis vectors $\phi(x) = x^2$ and $\phi(x) = 1$. See the paper for full theory.

Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003.

Daniela Pucci De Farias and Benjamin Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of operations research*, 29 (3):462–478, 2004.

Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of mathematical analysis and applications*, 110(2):568–582, 1985.