The Linear Programming Approach to Approximate Dynamic Programming. [2, 3]

Presented by: Pablo Samuel Castro and Erick Delage
McGill University

D.P. de Farias
Department of Mechanical Eng., MIT
B. van Roy
Department of Management Science and Eng., Stanford University

March 28, 2006

Outline

Motivation

- Curse of dimensionality makes exact solution of large MDPs intractable
- Interest in approximate DP has grown lately due to some success stories...
- but with significant trial and error and poor generalization
- An LP formulation may hopefully yield theoretical results

Problem formulation

- Finite state space: $\mathcal{S}$, $|\mathcal{S}| = N$
- $\forall x \in \mathcal{S}$ there exists a finite set of actions $\mathcal{A}_x$
- Taking action $a \in \mathcal{A}_x$ yields cost $g_a(x)$
- State transition probabilities: $p_a(x, y).\forall x \in \mathcal{S}.y \in \mathcal{S}$
- With policy u we have: $p_{u(x)}(x, y)$. Consider transition matrix P_u whose (x, y)th entry is $p_{u(x)}(x, y)$

Optimality criterion

- Optimize infinite-horizon discounted cost:

$$J_u(x) = E\left[\sum_{t=0}^{\infty} \alpha^t g_u(x_t) | x_0 = x\right]$$

- Well known there exists a single policy u that minimizes $J_u(x)$ simultaneously for all x
- The goal is to find that single policy

DP Operator T

- Define the DP operators T_u and T as:

$$\begin{aligned} T_u J &= g_u + \alpha P_u J \\ TJ &= \min_u (g_u + \alpha P_u J) \end{aligned} \qquad (1)$$

- The solution of Bellman's equation is $J = TJ$
- The unique solution J^* of (1) is the optimal cost-to-go function: $J^* = \min_u J_u$
- Optimal actions generated by:

$$u(x) = \underset{a \in \mathcal{A}_x}{\text{argmin}} \left(g_a(x) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y) J^*(y) \right)$$

Linear Programming Approach (1/2)

- One approach to solve Bellman's equation:

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

  c is a vector with positive *state-relevance weights*
- Can be shown that any feasible J satisfies $J \leq J^*$
- It follows that for any c, J^* is the unique solution to the above equation

Linear Programming Approach (2/2)

- T is a nonlinear operator
- We can rewrite problem as:

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

- This problem will be referred to as the *exact LP*
- Any realistic problem will have a large number of variables and constraints!

LP Approach to approximate DP (1/2)

- Given pre-selected basis functions $\phi_1, \cdots, \phi_K$, define Φ as:

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \vdots & \phi_K \\ | & & | \end{bmatrix}$$

- Want to compute a weight vector $\tilde{r} \in \mathcal{R}^K$ s.t. $\Phi\tilde{r} \approx J^*$
- Policy defined according to

$$u(x) = \underset{a \in \mathcal{A}_x}{\text{argmin}} \left(g_a(x) + \alpha \sum_{y \in \mathcal{S}} p_a(x, y)(\Phi\tilde{r})(y) \right)$$

would hopefully be near-optimal

LP Approach to approximate DP (2/2)

- As before, can reformulate LP as:

$$\begin{aligned} \max \quad & c^T \Phi r \\ \text{s.t.} \quad & g_a(x) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y)(\Phi r)(y) \geq (\Phi r)(x) \\ & \forall x \in \mathcal{S}. \forall a \in \mathcal{A}_x \end{aligned} \tag{2}$$

- This problem will be referred to as the *approximate LP*
- Number of variables reduced to K, but number of constraints remains as large as before

Importance of state-relevance weights

- In the exact LP, maximizing $c^T J$ yields J^* for any choice of c
- The same is not true for the approximate LP

Lemma

A vector $\tilde{r}$ solves

$$\begin{array}{rc} \max & c^T \Phi r \\ \textit{s.t.} & T\Phi r \geq \Phi r \end{array}$$

if and only if it solves

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

- The algorithm can be lead to generate better approximations in a certain region of the state space by assigning a larger weight to that region!

Measuring quality of policies (1/2)

- If ν is the initial state distribution, a measure of the quality of policy u is: $E_{X\sim\nu}\left[J_u(X) - J^*(X)\right] = \|J_u - J^*\|_{1,\nu}$
- Define a measure $\mu_{u,\nu}$ over state space associated with policy u and distribution ν given by

$$\begin{aligned} \mu_{u,\nu}^T &= (1-\alpha)\nu^T \sum_{t=0}^{\infty} \alpha^t P_u^t \\ &= (1-\alpha)\nu^T (I - \alpha P_u)^{-1} \end{aligned}$$

- $\mu_{u,\nu}$ captures expected frequency of visits to each state when system runs under policy u, conditioned on initial state distributed according to ν
- It turns out that $\mu_{u,\nu}$ is a probability distribution

Measuring quality of policies (2/2)

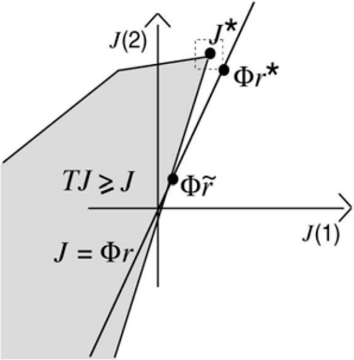
Theorem

Let $J : \mathcal{S} \mapsto \mathcal{R}$ be such that $TJ \geq J$. Then

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

- The above theorem says that if the approximate cost-to-go function J is close to J^*, the performance of the policy generated by J should also be close to the performance of the optimal policy
- We may want to choose c so that it captures frequency with which different states are visited (which in general depends on policy being used)

Error bounds for the approximate LP



- Would like to guarantee that $\Phi\tilde{r}$ is not too much farther from J^* than Φr^* is

A Simple Bound

Theorem

Let e be in the span of the columns of Φ and c be a probability distribution. Then, if $\tilde{r}$ is an optimal solution to the approximate LP,

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

- 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.
- However, $\min_r \|J^* - \Phi r\|_\infty$ is typically huge in practice
- Also, the above bound doesn't take the choice of c into account

Lyapunov Functions

- Introduce operator H for all $V : \mathcal{S} \mapsto \mathcal{R}$ as:

$$(HV)(x) = \max_{a \in \mathcal{A}_x} \sum_y P_a(x,y)V(y)$$

- For each $V : \mathcal{S} \mapsto \mathcal{R}$, define a scalar β_V by

$$\beta_V = \max_x \frac{\alpha(HV)(x)}{V(x)}$$

- Denote $V : \mathcal{S} \mapsto \mathcal{R}^+$ a **Lyapunov function** if $\beta_V < 1$
- Equivalent to condition that there exist $V > 0$ and $\beta < 1$ s.t. $\alpha(HV)(x) \leq \beta V(x), \quad \forall x \in \mathcal{S}$
- β_V conveys a degree of ”stability”, with stronger values representing stronger stability

An Improved Bound

Theorem

Let $\tilde{r}$ be a solution of the approximate LP. Then, for any $v \in \mathcal{R}^K$ such that $(\Phi v)(x) > 0$ for all $x \in \mathcal{S}$ and $\alpha H \Phi v < \Phi v$,

$$\|J^* - \Phi \tilde{r}\|_{1,c} \leq \frac{2c^T \Phi v}{1 - \beta_{\Phi v}} \min_r \|J^* - \Phi r\|_{\infty, 1/\Phi v}$$

- With introduction of $\|\cdot\|_{\infty,1/\Phi v}$, the error at each state is now weighted by the reciprocal of the Lyapunov function value.
- The Lyapunov function should take on large values in undesirable regions of state space (where J^* is large
- State relevance weights are now factored into new bound

The Constraint Sampling Strategy - I

Consider the approximate LP:

$$\begin{aligned} \text{maximize} \quad & c^T \Phi r, \\ \text{subject to} \quad & T\Phi r \geq \Phi r. \end{aligned} \tag{3}$$

Problems remaining:

- Objective $c^T \Phi r$ is hard to evaluate.
- Number of constraints is large.

The Constraint Sampling Strategy - II

Approximation:

$$\begin{array}{lll} \text{maximize} & \tilde{c}^T \Phi \hat{r} & \\ \text{subject to} & (T\Phi r)(x) \geq (\Phi r)(x) & \text{for all } x \in \{x_1, \ldots, x_N\} \\ & r \in \mathcal{N} & \end{array} \qquad (4)$$

- $\tilde{c}^T \Phi r$ can be obtained by sampling according to the distribution c (c is positive by definition and can be made to sum to 1 without changing the problem).
- **If we sample some reasonable number of constraints, then "almost all" others will be satisfied.**
- **The constraints that are not satisfied don't distort the solution too much.**

Main Theorem - I

Given:

$$\begin{array}{ll} \text{maximize} & c^T x, \\ \text{subject to} & Ax \leq b, \end{array} \tag{5}$$

and a probability distribution μ over the rows of A.

Define $\hat{x}_N$ as the optimal solution of the following LP:

$$\begin{array}{ll} \text{maximize} & c^T x, \\ \text{subject to} & A_{i_j} x \leq b_{i_j}, \text{ for } j = 1, 2, \ldots, N, \end{array} \tag{6}$$

where A_{i_j} is the i_jth row of the matrix A, i_j are sampled IID according to a distribution μ.

Main Theorem - II

Theorem

For arbitrary $\epsilon, \delta > 0$, if $N \geq n/(\epsilon\delta) - 1$, then

$$\mathbb{P}\left\{\mu(\{i | A_i \hat{x}_N > b_i\}) \leq \epsilon\right\} \geq 1 - \delta, \quad (7)$$

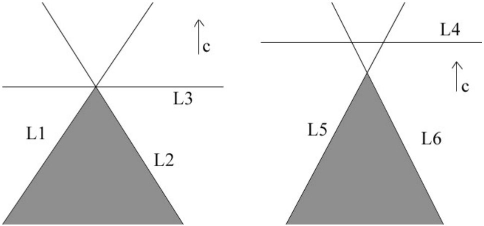
where the probability is taken over the random sampling of constraints.

- ϵ represents a tolerance or control on how many constraints are allowed to be violated.
- $1 - \delta$ represents a confidence level.
- The theorem states that given an ϵ and δ, the number of constraints we need for (7) to hold is linear in n, and does not depend on m.

Proof - I

Definition

Given an LP, a constraint is called a support constraint if the optimal objective value is changed if the constraint is relaxed.



Theorem

If there are n variables in an LP, which is bounded and feasible, then there are at most n support constraints.

Proof - II

Theorem

If $\hat{x}_N$ is the solution to the sampled LP (6)*, then*

$$\mathrm{E}\left[\mu\left(\{i : A_i \hat{x}_N > b_i\}\right)\right] \leq \frac{n}{N+1},$$

where the expectation above is taken over the random sampling of constraints.

Proof.

Considering solving problem 6 with $N+1$ constraints.

$$\mathbb{P}\left\{A_{i_{N+1}} \hat{x}_N > b_{i_{N+1}}\right\} \leq \frac{n}{N+1}.$$

It is easy to show that:

$$\mathbb{P}\left\{A_{i_{N+1}} \hat{x}_N > b_{i_{N+1}}\right\} = \mathrm{E}\left[\mu\left(\{i : A_i \hat{x}_N > b_i\}\right)\right].$$

Proof - III

From Markov inequality:

$$\mathbb{P}\left\{\mu(\{i|A_i\hat{x}_N > b_i\}) > \epsilon\right\} \leq \frac{1}{\epsilon}\mathrm{E}\left[\mu(\{i|A_i\hat{x}_N > b_i\})\right] \leq \frac{n}{\epsilon(N+1)} \leq \delta.$$

- Proof is true for any convex constraints [1]
- Proof can also be done using PAC-learning bounds of the linear classifier $x^T\tilde{a} \leq 0$ for samples $\tilde{a}$ drawn according to a fix distribution.(Vapnik-Chervonenkis [4])

How close is the solution to the relaxed problem?

Instead of finding $\tilde{r}$ that optimizes:

$$\begin{array}{lll} \text{maximize} & c^T \Phi r & \\ \text{subject to} & (T\Phi r)(x) \geq (\Phi r)(x) & \text{for all } x \in S \end{array} \tag{8}$$

We want to use $\hat{r}$ that optimizes:

$$\begin{array}{lll} \text{maximize} & c^T \Phi r & \\ \text{subject to} & (T\Phi r)(x) \geq (\Phi r)(x) & \text{for all } x \in \{x_1, \ldots, x_N\} \\ & r \in \mathcal{N} & \end{array} \tag{9}$$

where $\mathcal{N}$ is a bounded convex set which will prevent the optimization from taking too much advantage of excluded constraints.

Bound on Approximation - Theorem

Letting the constraints in problem (9) be sampled according to π_α, the "expected distribution of the initial $c^T P^t_{\mu^*}$ weighted by the value of α^t":

$$\pi_\alpha = (1-\alpha)c^T(I-\alpha P_{\mu^*})^{-1} = (1-\alpha)\sum_{t=0}^{\infty}\alpha^t c^T P^t_{\mu^*}$$

We get the following result:

Theorem

If $N \geq \frac{4K}{(1-\alpha)\epsilon\delta}\frac{\sup_{r\in\mathcal{N}}\|J^*-\Phi r\|_\infty}{c^T J^*}$ *then*

$$\|J^*-\Phi\hat{r}\|_{1,c} \leq \|J^*-\Phi\tilde{r}\|_{1,c} + \epsilon\|J^*\|_{1,c} \qquad \textit{with probability } 1-\delta$$

Bound on Approximation - Proof

$$
\begin{aligned}
\|J^* - \Phi\hat{r}\|_{1,c} &= c^T |J^* - \Phi\hat{r}| \\
&\leq c^T (I - \alpha P_{\mu^*})^{-1} |g - (I - \alpha P_{\mu^*})\Phi\hat{r}| \\
&= c^T (I - \alpha P_{\mu^*})^{-1} \big((g - (I - \alpha P_{\mu^*})\Phi\hat{r}) \\
&+ 2(g - (I - \alpha P_{\mu^*})\Phi\hat{r})^- \big) \\
&= c^T (J^* - \Phi\hat{r}) + 2c^T (I - \alpha P_{\mu^*})^{-1} (T_{\mu^*}\Phi\hat{r} - \Phi\hat{r})^- \\
&\leq c^T (J^* - \Phi\tilde{r}) + 2c^T (I - \alpha P_{\mu^*})^{-1} (T_{\mu^*}\Phi\hat{r} - \Phi\hat{r})^- \\
&\leq \|J^* - \Phi\tilde{r}\|_{1,c} + \frac{2}{1-\alpha} \pi (T_{\mu^*}\Phi\hat{r} - \Phi\hat{r})^- \\
&\leq \|J^* - \Phi\tilde{r}\|_{1,c} + \frac{2}{1-\alpha} \mu(\{i | A_i \hat{x}_N > b_i\}) \sup_{r \in \mathcal{N}} \|T\Phi r - \Phi r\|_\infty \\
&\leq \|J^* - \Phi\tilde{r}\|_{1,c} + \epsilon \|J^*\|_{1,c} \qquad \text{with probability } 1 - \delta
\end{aligned}
$$

Overall Bound on Approximation

Corollary

If $N \geq \frac{4K}{(1-\alpha)\epsilon\delta} \frac{\sup_{r\in\mathcal{N}} \|J^* - \Phi r\|_\infty}{c^T J^*}$ *and* $\Phi r = e$ *for some* r*, then:*

$$\|J^* - \Phi \hat{r}\|_{1,c} \leq \frac{2}{1-\alpha} \min_r \|\Phi r - J^*\|_\infty + \epsilon \|J^*\|_{1,c} \qquad \textit{with probability } 1-\delta$$

Remaining issues:

- Does approximating $\tilde{c}^T x$ affect the solution?
- Where to get π_α, the "expected distribution of the initial $c^T P_{\mu^*}^t$ weighted by the value of α^t"?
- How to chose the basis functions?
- $\frac{2}{1-\alpha} \min_r \|\Phi r - J^*\|_\infty$ is quite loose, can we expect better results in practice?

Bibliography

- G. Calafiore and M. C. Campi.
  Uncertain convex programs: randomized solutions and confidence levels.
  Math. Program., 102(1):25–46, 2005.
- D. de Farias and B. V. Roy.
  The linear programming approach to approximate dynamic programming, 2001.
- D. P. de Farias and B. V. Roy.
  On constraint sampling in the linear programming approach to approximate dynamic programming.
  Math. Oper. Res., 29(3):462–478, 2004.
- M. J. Kearns and U. V. Vazirani.
  An introduction to computational learning theory.
  MIT Press, Cambridge, MA, USA, 1994.