Approximate Linear Programming for Average Cost MDPs

Michael H. Veatch

Department of Mathematics

Gordon College

mike.veatch@gordon.edu

Initial version: January 13, 2011

Current revision: June 8, 2011

Abstract

We consider the linear programming approach to approximate dynamic programming with an average cost objective and a finite state space. Using a Lagrangian form of the LP, the average cost error is shown to be a multiple of the best fit differential cost error. This result is analogous to previous error bounds for a discounted cost objective. Second, bounds are derived for average cost error and performance of the policy generated from the LP that involve the mixing time of the MDP under this policy or the optimal policy. These mixing times are infinite in some cases, but shed light on when the methods can be expected to produce a good policy. These results improve on a previous performance bound involving mixing times.

1 Introduction

There has been significant interest recently in linear programming (LP) approaches to approximate dynamic programming (ADP). This approach, called approximate linear programming (ALP), uses the LP form of Bellman's equation and approximates the dynamic programming value function by a linear combination of preselected basis functions. The LP is used to compute the coefficients in the linear combination. A real-time control policy is then obtained from the value function approximation. Empirically, ALP has been used to generate effective control policies for a number of high-dimensional dynamic programs. Compared to other ADP approaches, it also has the simplicity of using only LP, where efficient solvers are available.

Some attractive theoretical guarantees have also been established for finite state, discounted Markov decision problems (MDPs). In particular, de Farias and Van Roy [3] and Desai et al. [7] show that the quality of the approximation to the cost-to-go function obtained by the ALP is proportional, in certain sense, to the quality of the best possible approximation using these basis functions. Most important, the approximation guarantee does not degrade with problem size. Theoretical results for other ADP methods, including approximate value iteration, approximate policy iteration, and temporal-difference methods, are not as strong; see Bertsekas [1, Chapter 6] for an overview.

This paper considers an average cost objective. Average cost dynamic programs are appealing, as a long time horizon is often more realistic and one avoids having to choose a discount rate. However, their theoretical analysis is significantly more involved. Average cost ALP algorithms and their analysis also face two fundamental difficulties. First, the parameters used in the discounted ALP to specify how different states are emphasized (called state-relevance weights) do not appear in the natural average cost formulation. As a result, it is not obvious that the value function approximation from the ALP is useful in the average cost setting. Second, the use of Lyapunov functions in the approximation error bound does not extend to average cost. In the discounted cost setting, the discount factor introduces some slack in the Lyapunov condition, allowing a Lyapunov function to be constructed.

The main contribution of this paper is an average cost error bound. It is the first that relates average cost for an ALP to the quality of the approximation architecture. It avoids the second difficulty by applying a vanishing discount method to the discounted error bound in Desai et al. [7], which does not have a Lyapunov condition. A crucial step in taking the limit is establishing a certain continuity between a (properly scaled) discounted ALP and the average cost ALP.

We also obtain an improved bound on the average cost performance of the ALP policy. This bound uses the policy from a discounted ALP, and involves a mixing time of the MDP under this policy and the optimal policy. These mixing times may be infinite, so the bound is vacuous for some examples. Our bound is stronger than that of de Farias and Van Roy [5]. An average cost error bound using the same ideas is also presented.

The ALP approach was originally proposed by Schweitzer and Seidmann [11]. For discounted MDPs on a finite state space, de Farias and Van Roy [3] provides the error bound described above and a similar bound on performance of the policy implied by the ALP. They use constraint sampling, which is shown to be probabilistically accurate in de Farias and Van Roy [4]. These results are extended in Desai et al. [7] to a Lagrangian form of the ALP that is potentially more accurate. The Lagrangian form is also used in Petrik and Zilberstein [9]. Another discounted performance bound and ALP are given in de Farias and Weber [6]. For average cost problems, two modifications of the ALP approach are proposed in de Farias and Van Roy [2] and [5]. ALP is applied to average cost queueing network control in Veatch [13].

The rest of this paper is organized as follows. The MDP is formulated in Section 2 and the ALPs introduced in Section 3. Section 4 presents bounds on the average cost error using a vanishing discount approach. Section 5 presents bounds on performance of the ALP policy and the average cost error in terms of mixing times. Section 6 discusses the relationship between these and other bound. Some proofs are deferred to the Appendix.

2 Average cost MDPs

Consider a discrete time MDP x_t with a finite state space $\mathcal{S}$ and a finite set of actions $\mathcal{A}(x)$ available in state x. Under a stationary Markov policy $u(x)$, the state process is a Markov chain with transition probability matrix $P_u = \left[p_{u(x)}(x,y)\right]$. A nonnegative cost $g(x,a)$ is incurred when action a is taken in state x. For shorthand, use the notation $g_u(x) = g(x,u(x))$ and

$$(P_u h)(x) = \sum_{y \in \mathcal{S}} p_{u(x)}(x,y)h(y).$$

The average cost under a stationary policy u is

$$\lambda_u = \limsup_{T \to \infty} \frac{1}{T} E_{x,u} \sum_{t=0}^{T-1} g_u(x_t).$$

Here $E_{x,u}$ denotes expectation given the initial state $x_0 = x$ and policy u. Assume that for each u this value is independent of x. This assumption essentially imposes a unichain structure on the resulting Markov chains; see Bertsekas [1, Section 4.2]. The dynamic programming operators are

$$T_u h = g_u + P_u h \text{ and } Th = \min_u T_u h.$$

Bellman's equation for this problem can be written as $Th - h = \lambda \mathbf{1}$. Note that if (λ, h) solves Bellman's equation, then so does $(\lambda, h + k\mathbf{1})$ for any additive constant k. One choice of this constant is made by the bias function, defined for policy u as

$$h_u(x) = \liminf_{T\to\infty} E_{x,u} \sum_{t=0}^{T} [g(x_t) - \lambda_u].$$

Another choice is the differential cost $h_u(x) - h_u(0)$, which is zero in some reference state $x = 0$. There exists an optimal stationary policy with average cost λ^* and bias h^* that satisfy Bellman's equation; see, e.g., Puterman [10, Theorem 8.4.3].

The discounted cost-to-go from initial state x for policy u is

$$J_{\alpha,u}(x) = E_{x,u} \sum_{t=0}^{\infty} \alpha^t g_u(x_t) = \sum_{t=0}^{\infty} \alpha^t (P_u^t g_u)(x)$$

where α is a discount factor in $[0, 1)$. An optimal stationary policy exists for this objective, with cost-to-go function $J_\alpha^*(x) = \min_u J_{\alpha,u}(x)$. Further, J_α^* solves Bellman's equation $J = T_\alpha J$, where the dynamic programming operators are

$$T_{\alpha,u} J = g_u + \alpha P_u J \;\text{ and }\; T_\alpha J = \min_u T_{\alpha,u} J.$$

See Bertsekas [1, Propositions 1.1 and 1.3].

The discounted and average cost problems are related under a vanishing discount rate:

$$\begin{aligned} \lim_{\alpha\uparrow 1} (1-\alpha) J_{\alpha,u}(x) &= \lambda_u & (1)\\ \lim_{\alpha\uparrow 1} (J_{\alpha,u}(x) - J_{\alpha,u}(0)) &= h_u(x) - h_u(0) & (2)\\ \lim_{\alpha\uparrow 1} J_{\alpha,u}(x) - \frac{\lambda_u}{1-\alpha} &= h_u(x) & (3) \end{aligned}$$

for all policies u and states x. See Puterman [10, Corollary 8.2.4].

3 Average cost approximate linear programs

This section introduces the standard ALPs for the average cost and discounted problem. The discounted ALP is reformulated slightly to make a connection with the average cost ALP. In particular, the discounted ALP contains a rescaled variable λ which approximates average cost in the limit. To summarize the notation to follow, λ is a free variable, λ_u, λ^*, $\lambda_{\alpha,u}$, and λ_α^* are average costs under various assumptions, and λ_A^*,

$\lambda^*_{A(\alpha)}$, λ^*_L, $\lambda^*_{L(\alpha)}$ are optimal values of the variable λ in the LP denoted by the subscript.

For discounted problems, Bellman's equation is equivalent to the following linear program for any state relevance weights $c > 0$ (see Bertsekas [1, Section 4.3.3]):

$$\begin{aligned} &\max_J \; c^T J \\ &\text{s.t.} \;\; T_\alpha J \geq J. \end{aligned} \tag{4}$$

Assume, without loss of generality, that c is a distribution ($|c| = 1$). Although the constraint is not linear, it can be written as a linear constraint for each state-action pair, namely,

$$g(x,a) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y) J(y) \geq J(x) \;\text{ for all } a \in \mathcal{A}(x) \text{ for all } x.$$

To create a more tractable LP, the cost-to-go is approximated by

$$(\widetilde{\Phi}\widetilde{r})(x) = \sum_{k=0}^{K} r_k \phi_k(x) \tag{5}$$

using some small set of basis functions ϕ_k and variables r_k. We will make use of the fact that $\mathbf{1}$ is in the span of $\widetilde{\Phi}$, which is implied by the assumptions $\phi_0 = \mathbf{1}$ and $\phi_k(0) = 0$, $k = 1, \ldots, K$. Write $\widetilde{\Phi} = [\phi_0 | \Phi]$ and $\widetilde{r}^T = (r_0, r^T)$. To accommodate a vanishing discount limit, let $\lambda = (1-\alpha) r_0$ so that

$$\widetilde{\Phi}\widetilde{r} = \frac{\lambda}{(1-\alpha)} \mathbf{1} + \Phi r. \tag{6}$$

Multiplying the objective in (4) by $(1-\alpha)$ and using the approximation (5) for J, the approximate LP is

$$\begin{aligned} \text{ALP}(\alpha) \qquad &\max_{\lambda,\, r} \; \lambda + (1-\alpha) c^T \Phi r \\ &\text{s.t.} \;\; T_\alpha \Phi r - \Phi r \geq \lambda \mathbf{1}. \end{aligned}$$

ALP(α) is feasible and, since it is equivalent to (4) with the constraint $J = \widetilde{\Phi}\widetilde{r}$ added, its objective function is bounded above. Let $(\lambda^*_{A(\alpha)}, r^*_{A(\alpha)})$ be an optimal solution (the notation $A(\alpha)$ is short for ALP(α)). Any J satisfying $T_\alpha J \geq J$ is a lower bound, $J \leq J^*_\alpha$; see Puterman [10, Theorem 6.2.2]. Hence, the cost-to-go approximation from ALP(α) is also a lower bound,

$$\widetilde{\Phi}\widetilde{r}^*_{A(\alpha)} \leq J^*_\alpha. \tag{7}$$

An average cost ALP is obtained by setting $\alpha = 1$:

$$\begin{aligned} \text{(ALP)} \quad & \max_{\lambda,\, r} \ \lambda \\ & \text{s.t.} \ \ T\Phi r - \Phi r \geq \lambda \mathbf{1}. \end{aligned}$$

Let (λ^*_A, r^*_A) be an optimal solution. Again, for any (λ, h) satisfying $Th - h \geq \lambda\mathbf{1}$, λ is a lower bound. In particular, $\lambda^*_A \leq \lambda^*$; see Puterman [10, Theorem 8.4.1].

For any approximation Φr, a greedy discounted policy is

$$u_{\alpha,r}(x) \in \arg\min_{a \in \mathcal{A}(x)} \left\{ g(x,a) + \alpha \sum_y p_a(x,y)(\Phi r)(y) \right\}. \tag{8}$$

The constant function ϕ_0 does not affect the greedy policy and is omitted. Set $\alpha = 1$ for a greedy average cost policy, $u_r = u_{1,r}$.

A disadvantage of (ALP) is that the objective function does not involve Φr or the state relevance weights c. Even if it provides a good average cost bound, the approximation Φr^* might be poor and the policy $u_{\Phi r^*}$ might have poor performance or not even be stabilizing. Modifications have been proposed to improve performance. In de Farias and Van Roy [2], the phase two LP

$$\begin{aligned} \text{(ALP2)} \qquad & \max_r \ c^T\Phi r \\ & \text{s.t.} \ \ (T\Phi r)(x) - (\Phi r)(x) \geq \lambda, \ \ x \neq 0 \end{aligned}$$

is solved with λ fixed. Another possibility is to use the combined objectives, as in ALP(α), and $\max_{\lambda,\, r}$ $\lambda + \eta c^T \Phi r$. One could search over the scalar η to optimize performance. In de Farias and Van Roy [5], a Lagrangian version of (ALP) is considered, with a single additional variable that allows constraints to be violated in amounts specified by a given weighting function.

4 A vanishing discount error bound

The relationship (1) can be used to convert certain $O(\frac{1}{1-\alpha})$ discounted error bounds to an average cost error bound. In particular, we consider error bounds that consist of the "best fit error" within the approximation architecture for some norm, multiplied by a factor that is $O(\frac{1}{1-\alpha})$. Thus, we must show that the best fit error is $O(1)$ as $\alpha \uparrow 1$. We use the following notation for weighted norms. For $\nu > 0$, $\|h\|_{1,\nu} = \sum_x \nu(x)h(x)$ and $\|h\|_{\infty,\nu} = \sup_x \nu(x)h(x)$.

We will need the following lemma regarding convergence of ALP(α) to (ALP) in the objective function. The proof is given in the Appendix.

Lemma 1 *Let $f(\alpha)$ be the optimal objective function value of ALP(α). Then* $\lim_{\alpha\uparrow 1} f(\alpha) = \lambda_A^*$.

Use (6) to write $f(\alpha) = (1-\alpha)c^T\widetilde{\Phi}\widetilde{r}^*_{A(\alpha)}$. Combining Lemma 1 with (1),

$$\begin{aligned}\lambda^* - \lambda_A^* &= \lim_{\alpha\uparrow 1}(1-\alpha)[J_\alpha^*(x) - c^T\widetilde{\Phi}\widetilde{r}^*_{A(\alpha)}] \text{ for all } x \\ &= \lim_{\alpha\uparrow 1}(1-\alpha)c^T(J_\alpha^* - \widetilde{\Phi}\widetilde{r}^*_{A(\alpha)}) \\ &= \lim_{\alpha\uparrow 1}(1-\alpha)\left\|J_\alpha^* - \widetilde{\Phi}\widetilde{r}^*_{A(\alpha)}\right\|_{1,c}. \qquad (9)\end{aligned}$$

The second equality holds because c is a distribution and the third because of (7).

Now consider the best fit cost-to-go error, $\min_{\widetilde{r}}\left\|J_\alpha^* - \widetilde{\Phi}\widetilde{r}\right\|$ for some norm. Let $o(\cdot)$ denote a vector-valued function such that $\lim_{\varepsilon\to 0} o(\varepsilon,x)/\varepsilon = 0$ for all $x \in \mathcal{S}$. Using (3), for any norm,

$$\begin{aligned}\min_{\widetilde{r}}\left\|J_\alpha^* - \widetilde{\Phi}\widetilde{r}\right\| &= \min_{\widetilde{r}}\left\|h^* + \frac{\lambda^*}{1-\alpha} - \widetilde{\Phi}\widetilde{r} + o(1-\alpha)\right\| \\ &= \min_{\widetilde{r}}\left\|h^* - \widetilde{\Phi}\widetilde{r} + o(1-\alpha)\right\| \\ &\le \min_r \|h^* - \Phi r + o(1-\alpha)\|\end{aligned}$$

and

$$\lim_{\alpha\uparrow 1}\min_{\widetilde{r}}\left\|J_\alpha^* - \widetilde{\Phi}\widetilde{r}\right\| \le \min_r \|h^* - \Phi r\|. \qquad (10)$$

The second equality holds because the minimization over r_0 makes any term that is constant over x irrelevant. The inequality is only due to the assumption that $(\Phi r)(0) = 0$. One could include a constant function in Φ and have equality.

Note that the limit in (10) is only finite because the best fit $\widetilde{r}$ is being used. In light of (6) and (3), the cost-to-go error can be written as

$$\begin{aligned} J^*_\alpha - \widetilde{\Phi}\widetilde{r}^*_{A(\alpha)} &= \frac{\lambda^*}{1-\alpha}\mathbf{1} + h^* + o(1-\alpha) - \left(\frac{\lambda^*_{A(\alpha)}}{1-\alpha}\mathbf{1} + \Phi r^*_{A(\alpha)}\right) \\ &= \frac{\lambda^* - \lambda^*_{A(\alpha)}}{1-\alpha}\mathbf{1} + h^* - \Phi r^*_{A(\alpha)} + o(1-\alpha) \end{aligned}$$

and may be unbounded as $\alpha \uparrow 1$.

Now we are ready to state average cost bounds. First, consider the simple error bound of de Farias and Van Roy [3, Theorem 4.1]:

$$\left\| J^*_\alpha - \widetilde{\Phi}\widetilde{r}^*_{A(\alpha)} \right\|_{1,c} \leq \frac{2}{1-\alpha} \min_{\widetilde{r}} \left\| J^*_\alpha - \widetilde{\Phi}\widetilde{r} \right\|_\infty . \tag{11}$$

Theorem 1 *If λ^*_A is an optimal solution to (ALP), then*

$$\lambda^* - \lambda^*_A \leq 2 \min_r \| h^* - \Phi r \|_\infty .$$

Proof. Combining (9), (11), and (10) gives the result. ■

The derivation of Theorem 1 shows that (11) is an $O(\frac{1}{1-\alpha})$ error bound. However, the usefulness of (11) and Theorem 1 is limited because of the L_∞ norm. A second bound in the same paper uses a weighted L_1 norm; however, it relies on a Lyapunov condition that cannot be satisfied as $\alpha \uparrow 1$. We use the improved bound in Desai et al. [7, Theorem 2] to avoid this problem. Introduce a constraint violation budget θ and distribution $\pi > 0$. The *smoothed* ALP is

$$\begin{aligned} \text{SALP}(\alpha) \qquad \max_{\lambda,r,s}\ & \lambda + (1-\alpha)c^T\Phi r \\ \text{s.t.}\ & T_\alpha \Phi r - \Phi r + s \geq \lambda \mathbf{1} \\ & \pi s \leq \theta \\ & s \geq 0. \end{aligned}$$

Error bounds are only available in terms of an idealized distribution. Let $\pi_{\alpha,u}(x) = (1-\alpha)\sum_{t=0}^\infty \alpha^t (c^T P_u^t)(x)$, which can be interpreted as the discounted expected number of visits to state x from the initial distribution c under policy u. The error bound is for the following LP, which has a Lagrangian term replacing the constraint violation budget:

$$\begin{aligned} \text{LSALP}(\alpha) \qquad & \max_{\lambda,r,s} \ \lambda + (1-\alpha)c^T\Phi r - 2\pi_{\alpha,u^*_\alpha} s \\ & \text{s.t.} \ \ T_\alpha \Phi r - \Phi r + s \geq \lambda \mathbf{1} \\ & \qquad s \geq 0. \end{aligned}$$

Note that u^*_α is unknown. Let $(\lambda^*_{L(\alpha)}, r^*_{L(\alpha)}, s^*_{L(\alpha)})$ be an optimal solution LSALP(α) and $\psi \geq 1$ be a function on the state space. The error bound is

$$\left\| J^*_\alpha - \widetilde{\Phi}\widetilde{r}_{L(\alpha)} \right\|_{1,c} \leq \min_{\widetilde{r}} \left\| J^*_\alpha - \widetilde{\Phi}\widetilde{r} \right\|_{\infty,1/\psi} \left[c^T\psi + \frac{2(\pi_{\alpha,u^*_\alpha}\psi)(\alpha\beta(\psi)+1)}{1-\alpha} \right] \tag{12}$$

where

$$\beta(\psi) = \max_{x,a} \frac{\sum_y p_a(x,y)\psi(y)}{\psi(x)}.$$

A good choice of ψ will balance the two factors in (12). On the one hand, ψ should be larger in states where the approximation is less accurate so that the weights $1/\psi$ are small in these states. On the other hand, ψ should not grow quickly or the maximum expected growth rate of ψ in one transition, $\beta(\psi)$, will be large.

The average cost LP corresponding to LSALP(α) is

$$\begin{aligned} \text{(LSALP)} \qquad & \max_{\lambda,r,s} \ \lambda - 2\pi^* s \\ & \text{s.t.} \ \ T\Phi r - \Phi r + s \geq \lambda \mathbf{1} \\ & \qquad s \geq 0 \end{aligned}$$

where $\pi^* = \lim_{\alpha\uparrow 1} \pi_{\alpha,u^*_\alpha}$ is the stationary distribution under the average cost optimal policy. Again, convergence of optimal objective function values is needed. The following lemma is proven in the Appendix.

Lemma 2 *Let $(\lambda^*_{L(\alpha)}, r^*_{L(\alpha)})$ and λ^*_L be optimal solutions to LSALP(α) and (LSALP). Then $\lim_{\alpha\uparrow 1} \lambda^*_{L(\alpha)} + (1-\alpha)c^T\Phi r^*_{L(\alpha)} = \lambda^*_L$.*

Theorem 2 *Let $\psi \geq 1$ be a function on the state space. If λ^*_L is an optimal solution to (LSALP), then*

$$|\lambda^* - \lambda^*_L| \leq \min_r \|h^* - \Phi r\|_{\infty,1/\psi} 2(\pi^*\psi)(\beta(\psi)+1).$$

Proof. Using Lemma 2 in place of Lemma 1, the first two equations in (9) hold with λ_A^* replaced by λ_L^* and $\widetilde{r}^*_{A(\alpha)}$ replaced by $\widetilde{r}^*_{L(\alpha)}$. However, because λ_L^* is not feasible for (ALP), it is not a lower bound and the last equation in (9) is replaced by

$$|\lambda^* - \lambda_L^*| \leq \lim_{\alpha \uparrow 1} (1-\alpha) \left\| J_\alpha^* - \widetilde{\Phi}\widetilde{r}^*_{L(\alpha)} \right\|_{1,c}.$$

Applying (12) and (10),

$$\begin{aligned} |\lambda^* - \lambda_L^*| &\leq \lim_{\alpha \uparrow 1} \min_{\widetilde{r}} \left\| J_\alpha^* - \widetilde{\Phi}\widetilde{r} \right\|_{\infty,1/\psi} \left[(1-\alpha)c^T\psi + 2(\pi_{\alpha,u^*}\psi)(\alpha\beta(\psi)+1)\right] \\ &\leq \min_r \|h^* - \Phi r\|_{\infty,1/\psi} 2(\pi^*\psi)(\beta(\psi)+1) \end{aligned}$$

which completes the proof. ■

Theorem 2 shows that (12) is also an $O(\frac{1}{1-\alpha})$ error bound. The weighted norm makes it potentially much tighter than Theorem 1. Theorem 2 holds for any $\psi \geq 1$. We expect that ψ must be tailored to particular problems to obtain meaningful bounds.

5 Error and performance bounds using mixing times

Even discounted bounds that are not $O(\frac{1}{1-\alpha})$ can be converted to average cost bounds using the method of de Farias and Van Roy [5]. The idea is to perturb the average cost MDP in a way that makes it equivalent to a discounted MDP. The average cost bound is then the sum of a discounted bound, scaled by $1-\alpha$, and a bound on the impact of the perturbation on average cost. The perturbation term contains a "mixing time" of the system under the optimal or, in the case of performance bounds, the greedy policy. For some MDPs the mixing time is infinite and no bound is obtained.

The perturbed MDP with parameter α has transition probability matrix $P_{\alpha,u} = \alpha P_u + (1-\alpha)\mathbf{1}c^T$ and stationary distribution $\pi_{\alpha,u}$. At each transition there is a restart with probability $1-\alpha$ to the restart distribution c. The expected cost until a reset from state x is the same as the expected discounted cost in the original system, $J_{\alpha,u}$. Let $\lambda_{\alpha,u}$ and λ_α^* be the average cost in the perturbed system under policy u and the optimal policy, respectively. Then

$$\lambda_{\alpha,u} = (1-\alpha)c^T J_{\alpha,u}. \tag{13}$$

If $c > 0$, the cost relationship (13) implies that the same policies are optimal in the discounted and the perturbed MDPs. Define the mixing time of policy u as

$$z_u = \inf\left\{ z : \left| \frac{1}{t}\sum_{\tau=0}^{t-1} c^T P_u^\tau g_u - \lambda_u \right| < \frac{z}{t} \text{ for all } t \right\}.$$

Also let $z^* = z_{u^*}$. Note that z_u depends on c and measures how quickly expected average cost over a finite horizon, starting from the restart distribution, approaches average cost. Their Theorem 4.1 states that

$$|\lambda_{\alpha,u} - \lambda_u| \leq z_u(1-\alpha) \tag{14}$$

bounding the impact of the perturbation on average cost.

For a performance bound, write

$$\begin{aligned} \lambda_u - \lambda^* &= \lambda_{\alpha,u} - \lambda^*_\alpha + (\lambda_u - \lambda_{\alpha,u}) + (\lambda^*_\alpha - \lambda^*) \\ &\leq \lambda_{\alpha,u} - \lambda^*_\alpha + z_u(1-\alpha) + z^*(1-\alpha) \\ &= (1-\alpha)\left\|J_{\alpha,u} - J^*_\alpha\right\|_{1,c} + (z_u + z^*)(1-\alpha). \end{aligned} \tag{15}$$

Any discounted performance bound that uses a weighted L_1 norm can be combined with (15) to obtain an average cost performance bound; the restart distribution c must be used as the weights in the norm. We will use the discounted performance bound in Desai et al. [7].

Recall that $r^*_{L(\alpha)}$ is an optimal solution to LSALP(α) and $u_{\alpha,r}$ is the greedy policy with respect to Φr in the discounted problem. Define $\nu(\eta, \Phi r)$ to be the discounted expected frequency of visits to each state under this policy from an initial distribution η, i.e.,

$$\nu(\eta, \Phi r) = (1-\alpha)\eta^T \sum_{t=0}^{\infty} (\alpha P_{u_{\alpha,r}})^t = (1-\alpha)\eta^T (I - \alpha P_{u_{\alpha,r}})^{-1}.$$

Assume that the state relevance weights c in LSALP(α) satisfy

$$c = \nu(\eta, \Phi r^*_{L(\alpha)}) \tag{16}$$

and let $u_{L(\alpha)} = u_{\alpha, r^*_{L(\alpha)}}$. The performance bound analogous to (12) is

$$\left\|J_{\alpha,u_{L(\alpha)}} - J^*_\alpha\right\|_{1,c} \leq \frac{1}{1-\alpha} \min_{\widetilde{r}} \left\| J^*_\alpha - \widetilde{\Phi}\widetilde{r} \right\|_{\infty,1/\psi} \left[c^T\psi + \frac{2(\pi_{\alpha,u^*_\alpha}\psi))(\alpha\beta(\psi)+1)}{1-\alpha} \right]. \tag{17}$$

Combining (15) and (17) gives the following theorem.

Theorem 3 *Let $\psi \geq 1$ be a function on the state space and c satisfy (16). If $u_{L(\alpha)}$ is a greedy policy associated with LSALP(α) for some $\alpha < 1$, then*

$$\lambda_{u_{L(\alpha)}} - \lambda^* \leq \min_{\widetilde{r}} \left\| J_\alpha^* - \widetilde{\Phi}\widetilde{r} \right\|_{\infty,1/\psi} \left[c^T\psi + \frac{2(\pi_{\alpha,u_\alpha^*}\psi))(\alpha\beta(\psi)+1)}{1-\alpha} \right] + (z_{u_{L(\alpha)}} + z^*)(1-\alpha). \tag{18}$$

It is not clear how to choose state relevance weights c to satisfy (16) because $r_{L(\alpha)}^*$ depends on c. However, even without this idealized choice of c, (18) can be viewed as an approximate bound. Note that the bound holds for any α, not just $\alpha \uparrow 1$. One must choose α and a function ψ to obtain a specific bound. While the first term is directly related to the strength of the approximation architecture Φ, the last term depends on the mixing times under the greedy and optimal policies.

The derivation of an error bound is similar. Given the cost-to-go estimate from LSALP(α), estimate average cost using (13) as

$$\begin{aligned} \lambda_{L(\alpha)} &= (1-\alpha)c^T\widetilde{\Phi}\widetilde{r}_{L(\alpha)}^* \\ &= \lambda_{L(\alpha)}^* + (1-\alpha)c^T\Phi r_{L(\alpha)}^*. \end{aligned} \tag{19}$$

Then

$$\begin{aligned} \lambda^* - \lambda_{L(\alpha)} &= \lambda_\alpha^* - \lambda_{L(\alpha)} + (\lambda^* - \lambda_\alpha^*); \\ |\lambda^* - \lambda_{L(\alpha)}| &\leq (1-\alpha)\left\| J_\alpha^* - \widetilde{\Phi}\widetilde{r}_{L(\alpha)}^* \right\|_{1,c} + z^*(1-\alpha). \end{aligned} \tag{20}$$

Any discounted error bound that uses a weighted L_1 norm can be combined with (20) to obtain an average cost error bound; the restart distribution c must be used as the weights in the norm. Using (12) yields the following theorem.

Theorem 4 *Let $\psi \geq 1$ be a function on the state space, c satisfy (16), and $\lambda_{L(\alpha)}$ satisfy (19). Then*

$$|\lambda^* - \lambda_{L(\alpha)}| \leq \min_{\widetilde{r}} \left\| J_\alpha^* - \widetilde{\Phi}\widetilde{r} \right\|_{\infty,1/\psi} \left[(1-\alpha)c^T\psi + 2(\pi_{\alpha,u_\alpha^*}\psi)(\alpha\beta(\psi)+1) \right] + z^*(1-\alpha). \tag{21}$$

Again, one must choose α and a function ψ to obtain a specific bound and the first term is directly related to the strength of the approximation architecture Φ. The additional term only depends on the mixing time under the optimal policy.

6 Relation to other bounds

Theorem 2 is the first result that relates average cost error for an ALP directly to the quality of the approximation architecture. More specifically, the average cost error bound is equal to a constant times the best fit differential cost error, in the sense of a weighted max-norm. The constant depends on the weights, which can be selected for individual problems to obtain good bounds. Although the constant depends on the unknown optimal stationary distribution, it does not depend on the approximation architecture and is expected to scale well to large state spaces. It shares these properties with the discounted cost-to-go error bound in Desai et al. [7] from which it is derived. The error bound in Theorem 4 contains an additional term that depends on a mixing time under the optimal policy. The average cost results in de Farias and Van Roy [2] are quite different and do not constitute an error bound.

Theorem 3 and de Farias and Van Roy [5, Corollary 4.1] bound average cost performance of the ALP policy. Both bounds contain the same mixing time term and require one to choose the discount factor and state-relevance weights, since a discounted ALP is used, and the weights in the norm. However, the discounted performance bound used to derive their average cost bound has several disadvantages. The bound for $\lambda_{\alpha.u}$ has a factor of $1/(1-\alpha)$, so that the bound for $J_{\alpha.u}$ has a factor of $1/(1-\alpha)^2$. It also contains two other quantities (their β, θ) which cannot be computed and depend on algorithm parameters. The bound is for a Lagrangian version of ALP(α) and requires choosing the Lagrange multiplier (η) and weights ($\psi(x)$). Although they also present a performance bound for a queueing example, it is not based on Corollary 4.1.

It would be of interest to obtain an average cost performance bound using the vanishing discount approach of Section 4. The discounted performance bounds cited above are all $O(1/(1-\alpha)^2)$ and cannot be used. However, it might be possible to modify one of them to obtain the desired $O(\frac{1}{1-\alpha})$ dependence on α.

7 Countable State Spaces

It would be desirable to extend the results above to countable state spaces. Because Theorem 1 uses a max-norm, it is generally not applicable. The bounds Theorems 2 - 4 should apply with some additional conditions, if ψ is chosen so that the weighted best fit errors $\|h^* - \Phi r\|_{\infty,1/\psi}$ and $\left\|J^*_\alpha - \widetilde{\Phi}\widetilde{r}\right\|_{\infty,1/\psi}$ appearing there are finite. Of the results used to obtain them, only the bound involving mixing times (15) was derived for countable state spaces. Several technical issues would need to be addressed.

Average cost problems on countable state spaces may not have an optimal stationary policy and (λ^*, h^*) that satisfy Bellman's equation. One set of conditions to guarantee existence of an optimal policy is (SEN) in Sennott [12, Theorem 7.2.3]: For $\alpha \in (0,1)$ and $u = u_\alpha^*$, the left side of (1) is bounded by functions of x and the left side of (2) is bounded above by a function of x and below by a constant. A mild additional condition (see Theorem 7.5.6) guarantees solutions to Bellman's equation. This theory establishes (1) - (3) for $u = u_\alpha^*$ (and $u = u^*$ for the average cost quantities), which is all that is needed in (9) and (10). Note that the average cost λ_u is defined as a $\limsup$ and that the limit might not exist for infinite state spaces without stronger conditions; see Puterman [10, Corollary 8.10.8].

Now consider the LPs. The assumption that c is a distribution is now restrictive. de Farias and Van Roy [5, Theorem 2.1] use the following finiteness condition on the variables in the discounted LP to guarantee a unique solution. If u_α^* is an optimal policy and $\pi_{\alpha,u_\alpha^*}^T J_\alpha^* < \infty$, then J_α^* is the unique optimal solution to (4) with the additional constraint

$$\pi_{\alpha,u_\alpha^*}^T |J| < \infty.$$

Verifying that (λ^*, h^*) is the unique solution to the average cost LP (or Bellman's equation) requires much stronger assumptions and is used in the proof of Lemma 4. Lemma 4 also uses Blackwell optimality, which is not guaranteed for infinite state spaces; however, it should be possible to use limit points of the discounted policies as in Sennott [12]. The LP continuity result (Lemma 3) must also be extended to infinite dimensions.

A third issue is extending the discounted error and performance bounds to countable state spaces. We comment on just one step, where an LP lower bound property is needed. Although (7) does not hold for LSALP(α), the similar lower bound

$$\widetilde{\Phi}\widetilde{r}_{A(\alpha)}^* \leq J_\alpha^* + \left(\sum_{t=0}^{\infty} (\alpha P_{u_\alpha^*})^t \right) s_{L(\alpha)}^*$$

is used to derive (12) and (17). Here $s_{L(\alpha)}^*$ is the vector of constraint violations found by LSALP(α). It may be possible to show that this bound still holds using the LP theory for countable state spaces. See Veatch [13, Appendix A] for a lower bound property for queueing networks.

Appendix: Proof of Lemmas 1 and 2

Lemmas 1 and 2 are statements about continuity of the optimal objective function value of an LP. Continuity with respect to changes in the objective function coefficients is straightforward and is used in parametric linear programming. Continuity with respect to constraint coefficients requires some assumptions. We will use the following theorem, a more general form of which is in Martin [8].

Lemma 3 *If the LP*

$$\begin{aligned} \omega(A,b,c) \quad &= \quad \max \quad c^T x \\ & \qquad s.t. \quad Ax \le b \end{aligned}$$

and its dual both have bounded optimal solutions for all (A,b,c) *in a set* S *sufficiently close to* (A^*,b^*,c^*) *then* $\omega(A,b,c)$ *is continuous at* (A^*,b^*,c^*) *relative to* S*.*

We will show that ALP(α) has a bounded optimal solution by relating it to (4). Note that scaling is essential here, as $r_0 = \lambda/(1-\alpha)$ is not bounded.

Lemma 4 *There exist optimal solutions* $(\lambda^*_{A(\alpha)}, r^*_{A(\alpha)})$ *to ALP(α) that are bounded over* $\alpha \in [\alpha_0, 1]$ *for some* $\alpha_0 < 1$*.*

Proof. The LP (4) has a unique optimal solution because it is equivalent to Bellman's equation. Change to the scaled variables λ and h, with $J = \frac{\lambda}{1-\alpha}\mathbf{1} + h$ and $h(0) = 0$, and multiply the objective by $1-\alpha$, giving

$$\begin{aligned} \text{LP}(\alpha) \qquad & \max_{\lambda,\, h} \; \lambda + (1-\alpha)c^T h \\ & \text{s.t.} \;\; T_\alpha h - h \ge \lambda \mathbf{1} \\ & \qquad h(0) = 0. \end{aligned}$$

This transformation is 1-to-1, so LP(α) also has a unique optimal solution, say $(\lambda^*_{LP(\alpha)}, h^*_\alpha)$. Note that LP(1) is the linear program for the average cost MDP and also has a unique optimal solution. Equation (2) implies that

$$\lim_{\alpha \uparrow 1} h^*_\alpha(x) - h^*_\alpha(0) = \lim_{\alpha \uparrow 1} J^*_\alpha(x) - J^*_\alpha(0) = h^*(x) - h^*(0)$$

and (1) implies that

$$\lim_{\alpha \uparrow 1} \lambda^*_{LP(\alpha)} = \lim_{\alpha \uparrow 1} (1-\alpha) J^*_\alpha(x) = \lambda^*$$

for all x. In particular, $\|h^*_\alpha\|_2 \le M$ and $|\lambda^*_{LP(\alpha)}| \le M$ for some M for all $\alpha \in [\alpha_1, 1]$ for some $\alpha_1 < 1$. Here $\|\cdot\|_2$ is the Euclidean norm.

We will consider the rate at which the objective of LP(α) changes as one moves from the optimal solution. Let u be a Blackwell optimal policy, i.e., it is optimal for all $\alpha \in [\alpha_2, 1]$ for some $\alpha_2 \in [\alpha_1, 1)$. If there are multiple optimal actions in some state, eliminate the ones other than $u(x)$ from the LP. Then LP(α) has the same binding constraints for all $\alpha \in [\alpha_2, 1]$, namely, the constraint for the optimal action in each state. Since there are n such constraints, n variables not counting $h(0)$, and a unique optimal solution, these constraints are linearly independent for $\alpha \in [\alpha_2, 1]$.

Let V_α be the set of feasible directions from $(\lambda^*_{LP(\alpha)}, h^*_\alpha)$, $z_\alpha = \lambda + (1-\alpha)c^T h$ the objective function, $D_v z_\alpha$ the directional derivative of z_α in the direction v at the optimal point $(\lambda^*_{LP(\alpha)}, h^*_\alpha)$, and $m_\alpha = \max_{v \in V_\alpha} D_v z_\alpha$. Derivatives of the objective function in feasible directions from a unique optimal solution must be negative, so $m_\alpha < 0$. The maximum is achieved on an extreme direction of V_α, i.e., where $n-1$ of the binding constraints hold with equality. Let $v(\alpha)$ be a unit vector in one of the extreme directions. The coefficients of the constraints are linear and therefore continuous in α. By the continuous mapping theorem, $v(\alpha)$ and $D_{v(\alpha)} z_\alpha$ are also continuous on $\alpha \in [\alpha_2, 1]$. Since m_α is the maximum over a finite set of extreme directions, it is also continuous on $\alpha \in [\alpha_2, 1]$. But at $\alpha = 1$, $m_1 < 0$, so $m_\alpha < m$ for $\alpha \in [\alpha_0, 1]$ for some $m < 0$ and $\alpha_0 \in [\alpha_2, 1)$.

Now consider ALP(α). It is equivalent to LP(α) with the constraint $h = \Phi r$ added. We will show that adding this constraint only shifts the optimal solution a bounded distance. The optimal value of ALP(α) is nonnegative. By (1), the optimal value of LP(α) is bounded, i.e., $z^* = \sup_{\alpha \in [0,1]} z^*_\alpha < \infty$. Using the lower bound $|m|$ on the rate of change and upper bound z^* on the difference in optimal values, $\left\| \Phi r^*_{A(\alpha)} - h^*_\alpha \right\|_2 \leq z^* / |m|$ for $\alpha \in [\alpha_0, 1]$. Combining this bound with the bound on h^*_α, $\left\| \Phi r^*_{A(\alpha)} \right\|_2 \leq z^*/m + M$ for $\alpha \in [\alpha_0, 1]$. Since $\Phi r^*_{A(\alpha)}$ is bounded, there exists an $r^*_{A(\alpha)}$ that is bounded on the same interval. Similarly, $\left| \lambda^*_{A(\alpha)} - \lambda^*_{LP(\alpha)} \right| < z^*/m$ and $|\lambda^*_{A(\alpha)}| \leq z^*/m + M$ on this interval. ■

Proof of Lemma 1. Since (ALP) is just ALP(α) with $\alpha = 1$, the statement to be proved is continuity of the optimal objective function value at $\alpha = 1$. By Lemma 3, ALP(α) has bounded optimal solutions. The dual of ALP(α) shares with the dual of LP(α) the constraint, associated with the variable λ, that the dual variables are a distribution. Thus, the conditions of Lemma 3 hold. ■

Lemma 5 *There exist optimal solutions $(\lambda^*_{L(\alpha)}, r^*_{L(\alpha)})$ and λ^*_L to LSALP(α) and (LSALP) that are bounded over $\alpha \in [\alpha_0, 1]$ for some $\alpha_0 < 1$.*

Proof. Define

$$\begin{aligned} \text{SLP}(\alpha) \qquad \max_{\lambda,\, h, s} \ & \lambda + (1-\alpha)c^T h \\ \text{s.t.} \ & T_\alpha h - h + s \geq \lambda \mathbf{1} \\ & \pi s \leq \theta \\ & s \geq 0 \\ & h(0) = 0. \end{aligned}$$

For fixed s, SLP(α) can be interpreted as LP(α) for an MDP with cost $g(x, a) + s(x)$ and similarly for SALP(α). By Lemma 4, there are optimal solutions to SALP(α) for fixed s that are bounded over $\alpha \in [\alpha_1, 1]$ for some $\alpha_1 < 1$. Furthermore, they are uniformly bounded over feasible s. Therefore, there are bounded optimal solutions to SALP(α). For some θ dependent on α, any optimal solution of SALP(α) is an optimal solution of LSALP(α). Hence, there are also bounded optimal solutions of LSALP(α). ■

Acknowledgements

I would like to thank Vivek Farias for his helpful suggestions.

References

[1] D.P. Bertsekas, *Dynamic programming and optimal control*, 3rd ed., vol. 2, Athena Scientific, Belmont, MA, 2007.

[2] D.P. de Farias and B. Van Roy, *Approximate linear programming for average-cost dynamic programming*, Advances in Neural Information Processing Systems 15, MIT Press, 2003.

[3] ______, *The linear programming approach to approximate dynamic programming*, Oper. Res. **51** (2003), no. 6, 850–865.

[4] ______, *On constraint sampling for the linear programming approach to approximate dynamic programming*, Math. Oper. Res. **29** (2004), no. 3, 462–478.

[5] ______, *A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees*, Math. Oper. Res. **31** (2006), no. 3, 597–620.

[6] D.P. de Farias and T. Weber, *Choosing the cost vector of the linear programming approach to approximate dynamic programming*, CDC, 2008, http://dx.doi.org/10.1109/CDC.2008.4739452, pp. 67–72.

[7] V.V. Desai, V.F. Farias, and C.C. Moallemi, *Approximate dynamic programming via a smoothed linear program*, Working paper, Graduate School of Business, Columbia University, 2009.

[8] D.H. Martin, *On the continuity of the maximum in parametric linear programming*, J. Optim. Theory Appl. **17** (1975), no. 3, 205–210.

[9] M. Petrik and S. Zilberstein, *Constraint relaxation in approximate linear programs*, Proceedings of the 26th International Conference on Machine Learning (Montreal) (L. Bottou and M. Littman, eds.), Omnipress, 2009, pp. 809–816.

[10] M.L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley and Sons, Inc., New York, 1994.

[11] P. Schweitzer and A. Seidmann, *Generalized polynomial approximations in Markovian decision processes*, J. of Mathematical Analysis and Applications **110** (1985), 568–582.

[12] L.I. Sennott, *Stochastic dynamic programming and the control of queueing systems*, Wiley, New York, 1999.

[13] M.H. Veatch, *Approximate linear programming for networks: Average cost bounds*, Working paper, Gordon College, Dept. of Math. Available at http://faculty.gordon.edu/ns/mc/mike-veatch, 2010.