Towards Exploiting Duality in Approximate Linear Programming for MDPs

Dmitri Dolgov and Edmund Durfee
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109
(734)763-6648
{ddolgov, durfee}@umich.edu

A weakness of classical methods for solving Markov decision processes is that they scale very poorly because of the flat state space, which subjects them to the *curse of dimensionality*. Fortunately, many MDPs are well-structured, which makes it possible to avoid enumerating the state space. To this end, *factored* MDP representations have been proposed (Boutilier, Dearden, & Goldszmidt 1995; Koller & Parr 1999) that model the state space as a cross product of state features, represent the transition function as a Bayesian network, and assume the rewards can be expressed as sums of compact functions of the state features.

A challenge in creating algorithms for the factored representations is that well-structured problems do not always lead to compact and well-structured solutions (Koller & Parr 1999); that is, an optimal policy does not, in general, retain the structure of the problem. Because of this, it becomes necessary to resort to approximation techniques. Approximate linear programming (ALP) has recently emerged as a very promising MDP-approximation technique (Schweitzer & Seidmann 1985; de Farias & Roy 2003). As such, ALP has received a significant amount of attention, which has led to a theoretical foundation (de Farias & Roy 2003) and efficient solution techniques (e.g., (de Farias & Roy 2004; Guestrin *et al.* 2003; Patrascu *et al.* 2002)). However, this work has focused only on approximating the *primal* LP, and no effort has been invested in approximating the *dual* LP, which is the basis for solving a wide range of constrained MDPs (e.g., (Altman 1999; Dolgov & Durfee 2004)).

Unfortunately, as we demonstrate, linear approximations do not interact with the dual LP as well as they do with the primal LP, because the constraint coefficients cannot be computed efficiently (the operation does not maintain the compactness of the representation). To address this, we propose an LP formulation, which we call a *composite ALP*, that approximates both the primal and the dual optimization coordinates (the value function and the occupation measure), which is equivalent to approximating both the objective functions and the feasible regions of the LPs. This method provides a basis for efficient approximations of constrained MDPs and also serves as a new approach to a widely-discussed problem of dealing with exponentially many constraints in ALPs, which plagues both the primal and the dual ALP formulations alone. As viewed from the latter point of view, the benefit of our composite-ALP approach, which symmetrically approximates the primal and dual coordinates, is that in some domains it might be easier to choose good basis functions for the approximation than it is to find good values for the parameters required by other approaches (e.g., a sampling distribution over the constraint set as used in (de Farias & Roy 2004)).

MDPs and Primal ALP

A standard MDP can be described as $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$, where: $\mathcal{S} = \{i\}$ is a finite set of states, $\mathcal{A} = \{a\}$ is a finite set of actions, $p = [p_{iaj}] : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ defines the transition function, and $r = [r_{ia}] : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ defines the rewards. A solution to such an MDP is a stationary, history-independent, deterministic policy. One way of solving a discounted MDP is to formulate it as the following minimization LP in the value function coordinates v_i:

$$\min \sum_i \alpha_i v_i \;\Big|\; v_i \geq r_{ia} + \gamma \sum_j p_{iaj} v_j, \qquad (1)$$

or, as an equivalent dual:

$$\max \sum_{i,a} r_{ia} x_{ia} \Big| \sum_a x_{ja} - \gamma \sum_{i,a} x_{ia} p_{iaj} = \alpha_j, \;\; x_{ia} \geq 0,$$

where x is called the *occupation measure* (x_{ia} is the expected discounted number of times a is executed in i).

Approximate linear programming (Schweitzer & Seidmann 1985; de Farias & Roy 2003) aims to approximate the primal LP (1) by lowering the dimensionality of the problem by restricting the space of value functions to a linear combination of predefined basis functions h:

$$v(\vec{z}) = \sum_k h_k(\vec{z}_k) w_k \quad \Longleftrightarrow \quad v = Hw, \qquad (2)$$

where $h_k(\vec{z}_k)$ is the k^{th} basis function defined on a subset of the state features $\vec{z}_k \subseteq \vec{z}$, and w are the new optimization variables. Thus, LP (1) can be approximated as:

$$\min \alpha^T H w \;\big|\; AHw \geq r, \qquad (3)$$

where we introduce a constraint matrix $A_{ia,j} = \delta_{ij} - \gamma p_{iaj}$ (where $\delta_{ij} = 1 \Leftrightarrow i = j$). The key property of this approx-

This material is based upon work supported by DARPA/IPTO and the Air Force Research Laboratory under Contract No. FA8750–05–C–0030. The views and conclusions contained in this document are those of the authors, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

imation is that the objective function coefficients $\alpha^T H$ and the constraint coefficients for each state-action pair $(AH)_{ia}$ can be computed efficiently (Guestrin *et al.* 2003). The primal ALP (3) reduces the number of optimization variables from $|\mathcal{S}|$ to $|w|$, but the number of constraints remains exponential at $|\mathcal{S}||\mathcal{A}|$. To mitigate this, several techniques have been proposed, such as sampling (de Farias & Roy 2004) and exploiting problem structure (Guestrin *et al.* 2003).

Dual ALP

As mentioned above, we would like to extend this approach to the dual LP, which is better suited for constrained MDPs. By straightforwardly applying the techniques used in the primal-ALP, we could restrict the optimization to a subset of the occupation measures that are spawned by a certain basis $x = Qy$, yielding the following approximation:

$$\max r^T Qy \Big| A^T Qy = \alpha, \quad Qy \geq 0. \tag{4}$$

At a superficial level, this ALP looks very similar to ALP (3). However, their properties differ significantly.

As with any approximate method, an important question is how close it is to the exact solution. It is known (de Farias & Roy 2003) that the primal ALP is equivalent to a program that minimizes an L_1-norm of the difference between the approximation and the optimal value function. We can similarly show that (4) is equivalent to an LP that minimizes $\sum_{i,a} r_{ia} \left[x^*_{ia} - (Qy)_{ia}\right]$, a measure of distance to the optimal occupation measure x^*. However, this is not a norm (just a weighted sum of signed errors) and thus does not provide the same degree of comfort about convergence results (large positive and negative errors might cancel out).

The second important question is whether the objective-function and the constraint coefficients in (4) can be computed efficiently. It turns out that the former can, but the latter cannot, and therein lies the biggest problem of the dual ALP. This is due to the difference between the left-hand-side operator $A(\cdot)$, as used in the primal ALP (AH), and the right-hand-side operator $(\cdot)A$, as used in the dual ALP ($Q^T A$). The former can be computed efficiently, because $\sum_a P(a|b) = 1$ and a product of such terms drops out, while the latter cannot, since a product of terms of the form $\sum_b P(a|b)$ is not as easy to compute efficiently.

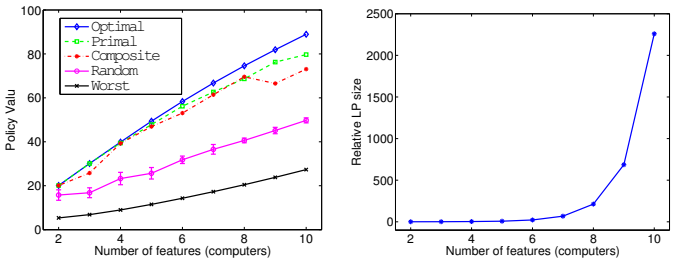
Composite ALP

The primal ALP approximates the primal variables v, which is equivalent to approximating the feasible region of the dual; the dual does the opposite. We can combine the two approximations, by substituting $x = Qy$ into the dual of (3):

$$\max r^T Qy \Big| H^T A^T Qy = H^T \alpha, \quad Qy \geq 0. \tag{5}$$

This ALP still has $|\mathcal{S}||\mathcal{A}|$ constraints (in $Qy \geq 0$), but this can be dealt with in several ways: for example, using the constraint-reformulation method of (Guestrin *et al.* 2003) or by only considering non-negative bases Q, and replacing the constraints with stricter $y \geq 0$ (which introduces another source of approximation error).

The benefit of the composite ALP is that it combines the efficiency gains of the primal and the dual LPs: besides the $Qy \geq 0$ constraints that can be dealt with as above, it has only $|y|$ variables and $|w|$ constraints. Furthermore, the constraint coefficients in $H^T A^T Q$ can be computed efficiently



(a) (b)

Figure 1: Composite ALP: quality (a), efficiency (b).

(despite the problematic dual approximation $A^T Q$) by first applying the primal approximation and then the dual to the result: $(H^T A^T)Q$. The composite ALP can be viewed as an alternative method to constraint sampling (de Farias & Roy 2004): instead of throwing out all but a small subset of the constraints, we are applying a more general transformation of the feasible region into a smaller-dimensional space.

We implemented the composite ALP and evaluated it on several unconstrained MDPs, including the "SysAdmin" problem from (Guestrin *et al.* 2003), the results for which we report here. To ensure feasibility of (5), we introduced free variables $|e| \leq \epsilon$, changed the equality constraints to $H^T A^T Qy + e = H^T \alpha$, and iteratively relaxed ϵ until the ALP became feasible. Figure 1a shows the value of policies obtained by the composite and the primal ALPs, compared to the exact solution (with random and worst policies shown for comparison); the mean value of policies produced by the composite ALP was 88% of the optimal-random gap, and the difference between the primal and the composite ALPs was around 7%. The number of primal and dual basis functions used scaled linearly with the number of world features; Figure 1b shows the relative efficiency gains, expressed as the ratio of the size of the constraint matrix in exact LP (1) to the one in the composite ALP (5). The main benefit of the composite ALP comes from its applicability to constrained MDPs; our preliminary experiments indicate that it might perform competitively on unconstrained MDPs as well.

References

Altman, E. 1999. *Constrained Markov Decision Processes*.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 1995. Exploiting structure in policy construction. In *IJCAI-95*, 1104–1111.

de Farias, D., and Roy, B. V. 2003. The linear programming approach to approximate dynamic programming. *OR* 51(6).

de Farias, D., and Roy, B. V. 2004. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. of OR* 29(3):462–478.

Dolgov, D., and Durfee, E. 2004. Optimal resource allocation and policy formulation in loosely-coupled Markov decision processes. In *ICAPS-04*, 315–324.

Guestrin, C.; Koller, D.; Parr, R.; and Venkataraman, S. 2003. Efficient solution algorithms for factored MDPs. *JAIR* 19.

Koller, D., and Parr, R. 1999. Computing factored value functions for policies in structured MDPs. In *IJCAI-99*, 1332–1339.

Patrascu, R.; Poupart, P.; Schuurmans, D.; Boutilier, C.; and Guestrin, C. 2002. Greedy linear value-approximation for factored markov decision processes. In *AAAI-02*.

Schweitzer, P., and Seidmann, A. 1985. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications* 110:568 582.