Learning Basis Functions in Hybrid Domains

Branislav Kveton
Intelligent Systems Program
University of Pittsburgh
bkveton@cs.pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
milos@cs.pitt.edu

Abstract

Markov decision processes (MDPs) with discrete and continuous state and action components can be solved efficiently by hybrid approximate linear programming (HALP). The main idea of the approach is to approximate the optimal value function by a set of basis functions and optimize their weights by linear programming. The quality of this approximation naturally depends on its basis functions. However, basis functions leading to good approximations are rarely known in advance. In this paper, we propose a new approach that discovers these functions automatically. The method relies on a class of parametric basis function models, which are optimized using the dual formulation of a relaxed HALP. We demonstrate the performance of our method on two hybrid optimization problems and compare it to manually selected basis functions.

Introduction

Markov decision processes (MDPs) (Bellman 1957; Puterman 1994) provide an elegant mathematical framework for solving sequential decision problems in the presence of uncertainty. However, traditional techniques for solving MDPs are computationally infeasible in real-world domains, which are factored and represented by both discrete and continuous state and action variables. Approximate linear programming (ALP) (Schweitzer & Seidmann 1985) has recently emerged as a promising approach to address these challenges (Kveton & Hauskrecht 2006).

Our paper centers around hybrid ALP (HALP) (Guestrin, Hauskrecht, & Kveton 2004), which is an established framework for solving large factored MDPs with discrete and continuous state and action variables. The main idea of the approach is to approximate the optimal value function by a linear combination of basis functions and optimize it by linear programming (LP). The combination of factored reward and transition models with the linear value function approximation permits the scalability of the approach.

The quality of HALP solutions inherently depends on the choice of basis functions. Therefore, it is often assumed that these are provided as a part of the problem definition, which is unrealistic. The main goal of this paper is to alleviate this assumption and learn basis functions automatically.

In the context of discrete-state ALP, Patrascu *et al.* (2002) proposed a greedy approach to learning basis functions. This method is based on the dual ALP formulation and its scores. Although our approach is similar to Patrascu *et al.* (2002), it is also different in two important ways. First, it is computationally infeasible to build the complete HALP formulation in hybrid domains. Therefore, we rely on its relaxed formulations, which may lead to overfitting of learned approximations on active constraints. To solve this problem, we restrict our search to basis functions with a better state-space coverage. Second, instead of choosing from a finite number of basis function candidates (Patrascu *et al.* 2002), we optimize a class of parametric basis function models. These extensions are nontrivial and pose a number of challenges.

The paper is structured as follows. First, we review hybrid factored MDPs and HALP (Guestrin, Hauskrecht, & Kveton 2004), which are our frameworks for modeling and solving large-scale stochastic decision problems. Second, we show how to improve the quality of relaxed approximations based on their dual formulations. Finally, we demonstrate learning of basis functions on two hybrid MDP problems.

Hybrid factored MDPs

Discrete-state factored MDPs (Boutilier, Dearden, & Goldszmidt 1995) permit a compact representation of stochastic decision problems by exploiting their structure. In this work, we consider hybrid factored MDPs with exponential-family transition models (Kveton & Hauskrecht 2006). This model extends discrete-state factored MDPs to the domains of discrete and continuous state and action variables.

A *hybrid factored MDP with an exponential-family transition model (HMDP)* (Kveton & Hauskrecht 2006) is given by a 4-tuple $\mathcal{M} = (\mathbf{X}, \mathbf{A}, P, R)$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a state space characterized by a set of discrete and continuous variables, $\mathbf{A} = \{A_1, \ldots, A_m\}$ is an action space represented by action variables, $P(\mathbf{X}' \mid \mathbf{X}, \mathbf{A})$ is an exponential-family transition model of state dynamics conditioned on the preceding state and action choice, and R is a reward model assigning immediate payoffs to state-action configurations.[1] In the remainder of the paper, we assume that the quality of a

[1] *General state and action space MDP* is an alternative name for a hybrid MDP. The term *hybrid* does not refer to the dynamics of the model, which is discrete-time.

policy is measured by the *infinite horizon discounted reward* $\mathrm{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0,1)$ is a *discount factor* and r_t is the reward obtained at the time step t. The *optimal policy* π^* can be defined greedily with respect to the *optimal value function* V^*, which is a fixed point of the Bellman equation (Bellman 1957):

$$V^*(\mathbf{x}) = \sup_{\mathbf{a}} \left[R(\mathbf{x}, \mathbf{a}) + \gamma \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[V^*(\mathbf{x}')] \right]. \quad (1)$$

Accordingly, the *hybrid Bellman operator* $\mathcal{T}^*$ is given by:

$$\mathcal{T}^* V(\mathbf{x}) = \sup_{\mathbf{a}} \left[R(\mathbf{x}, \mathbf{a}) + \gamma \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[V(\mathbf{x}')] \right]. \quad (2)$$

In the remainder of the paper, we denote expectation terms over discrete and continuous variables in a unified form:

$$\mathrm{E}_{P(\mathbf{x})}[f(\mathbf{x})] = \sum_{\mathbf{x}_D} \int_{\mathbf{x}_C} P(\mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x}_C. \quad (3)$$

Hybrid ALP

Since a factored representation of an MDP may not guarantee a structure in the optimal value function or policy (Koller & Parr 1999), we resort to *linear value function approximation* (Bellman, Kalaba, & Kotkin 1963; Van Roy 1998):

$$V^{\mathbf{w}}(\mathbf{x}) = \sum_i w_i f_i(\mathbf{x}). \quad (4)$$

This approximation restricts the form of the value function $V^{\mathbf{w}}$ to the linear combination of $|\mathbf{w}|$ basis functions $f_i(\mathbf{x})$, where $\mathbf{w}$ is a vector of tunable weights. Every basis function can be defined over the complete state space $\mathbf{X}$, but often is restricted to a subset of state variables $\mathbf{X}_i$ (Bellman, Kalaba, & Kotkin 1963; Koller & Parr 1999). Refer to Hauskrecht and Kveton (2004) for an overview of alternative methods to solving hybrid factored MDPs.

HALP formulation

Similarly to the discrete-state ALP (Schweitzer & Seidmann 1985), *hybrid ALP (HALP)* (Guestrin, Hauskrecht, & Kveton 2004) optimizes the linear value function approximation (Equation 4). Therefore, it transforms an initially intractable problem of estimating V^* in the hybrid state space $\mathbf{X}$ into a lower dimensional space $\mathbf{w}$. The HALP formulation is given by a linear program[2]:

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \ \sum_i w_i \alpha_i \quad (5) \\ &\text{subject to:} \ \sum_i w_i F_i(\mathbf{x}, \mathbf{a}) - R(\mathbf{x}, \mathbf{a}) \geq 0 \quad \forall \, \mathbf{x}, \mathbf{a}; \end{aligned}$$

where $\mathbf{w}$ represents the variables in the LP, α_i denotes *basis function relevance weight*:

$$\begin{aligned} \alpha_i &= \mathrm{E}_{\psi(\mathbf{x})}[f_i(\mathbf{x})] \quad (6) \\ &= \sum_{\mathbf{x}_D} \int_{\mathbf{x}_C} \psi(\mathbf{x}) f_i(\mathbf{x}) \, \mathrm{d}\mathbf{x}_C, \end{aligned}$$

[2] In particular, the HALP formulation (5) is a *linear semi-infinite optimization* problem with infinitely many constraints. The number of basis functions is finite. For brevity, we refer to this optimization problem as linear programming.

$\psi(\mathbf{x})$ is a *state relevance density function* weighting the approximation, and $F_i(\mathbf{x}, \mathbf{a}) = f_i(\mathbf{x}) - \gamma g_i(\mathbf{x}, \mathbf{a})$ is the difference between the basis function $f_i(\mathbf{x})$ and its discounted *backprojection*:

$$\begin{aligned} g_i(\mathbf{x}, \mathbf{a}) &= \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[f_i(\mathbf{x}')] \quad (7) \\ &= \sum_{\mathbf{x}'_D} \int_{\mathbf{x}'_C} P(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}) f_i(\mathbf{x}') \, \mathrm{d}\mathbf{x}'_C. \end{aligned}$$

Vectors $\mathbf{x}_D$ ($\mathbf{x}'_D$) and $\mathbf{x}_C$ ($\mathbf{x}'_C$) are the discrete and continuous components of value assignments $\mathbf{x}$ ($\mathbf{x}'$) to all state variables $\mathbf{X}$ ($\mathbf{X}'$). The HALP formulation is feasible if the set of basis functions contains a constant function $f_0(\mathbf{x}) \equiv 1$. We assume that such a basis function is always present.

The quality of the approximation was studied by Guestrin *et al.* (2004) and Kveton and Hauskrecht (2006). These results give a justification for minimizing our objective function $\mathrm{E}_{\psi}[V^{\mathbf{w}}]$ instead of the max-norm error $\|V^* - V^{\mathbf{w}}\|_{\infty}$. Expectation terms in the objective function (Equation 6) and constraints (Equation 7) are efficiently computable if the basis functions are *conjugate* to the transition model and state relevance density functions (Guestrin, Hauskrecht, & Kveton 2004; Kveton & Hauskrecht 2006). For instance, normal and beta transition models are complemented by normal and beta basis functions. To permit the conjugate choices when the transition models are mixed, we assume that every basis function $f_i(\mathbf{x})$ decouples as a product:

$$f_i(\mathbf{x}_i) = \prod_{X_j \in \mathbf{X}_i} f_{ij}(x_j) \quad (8)$$

of univariate basis function factors $f_{ij}(x_j)$. This seemingly strong assumption can be partially relaxed by considering a linear combination of basis functions.

Solving HALP

An optimal solution $\widetilde{\mathbf{w}}$ to the HALP formulation (5) is given by a finite set of *active constraints* at a vertex of the feasible region. However, identification of this active set is a computational problem. In particular, it requires searching through an exponential number of constraints, if the state and action variables are discrete, and infinitely many constraints, if any of the variables are continuous. As a result, it is in general infeasible to find the optimal solution $\widetilde{\mathbf{w}}$. Therefore, we resort to finite approximations to the constraint space in HALP whose optimal solution $\widehat{\mathbf{w}}$ is close to $\widetilde{\mathbf{w}}$. This notion of an approximation is formalized as follows.

Definition 1 *The HALP formulation is* relaxed*:*

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \ \sum_i w_i \alpha_i \quad (9) \\ &\text{subject to:} \ \sum_i w_i F_i(\mathbf{x}, \mathbf{a}) - R(\mathbf{x}, \mathbf{a}) \geq 0 \quad (\mathbf{x}, \mathbf{a}) \in \mathcal{C}; \end{aligned}$$

if only a subset $\mathcal{C}$ of its constraints is satisfied.

The HALP formulation (5) is solved approximately by solving its relaxed formulations (9). Several methods for building and solving these approximate LPs have been proposed

(Hauskrecht & Kveton 2004; Guestrin, Hauskrecht, & Kveton 2004; Kveton & Hauskrecht 2005). The quality of these formulations can be measured by the δ-infeasibility of their solutions $\widehat{\mathbf{w}}$. This metric represents the maximum violation of constraints in the complete HALP.

Definition 2 *Let $\widehat{\mathbf{w}}$ be a solution to a relaxed HALP formulation (9). The vector $\widehat{\mathbf{w}}$ is* δ-infeasible *if $V^{\widehat{\mathbf{w}}} - \mathcal{T}^* V^{\widehat{\mathbf{w}}} \geq -\delta$ for all $\mathbf{x} \in \mathbf{X}$, where $\mathcal{T}^*$ is the hybrid Bellman operator.*

Learning basis functions

The quality of HALP approximations depends on the choice of basis functions. However, basis functions leading to good approximations are rarely known a priori. In this section, we describe a new method that learns these functions automatically. The method starts from an initial set of basis functions and adds new functions *greedily* to improve the current approximation. Our approach is based on a class of parametric basis function models that are optimized on preselected domains of state variables. These domains represent our initial preference between the quality and complexity of solutions. In the rest of this section, we describe in detail how to score and optimize basis functions in hybrid domains.

Optimization of relaxed HALP

De Farias and Van Roy (2003) analyzed the quality of ALP. Based on their work, we may conclude that optimization of the objective function $\mathrm{E}_\psi[V^{\mathbf{w}}]$ in HALP is identical to minimizing the $\mathcal{L}_1$-norm error $\|V^* - V^{\mathbf{w}}\|_{1,\psi}$. This equivalence can be proved from the following proposition.

Proposition 1 *Let $\widetilde{\mathbf{w}}$ be a solution to the HALP formulation (5). Then $V^{\widetilde{\mathbf{w}}} \geq V^*$.*

Proof: The Bellman operator $\mathcal{T}^*$ is a contraction mapping. Based on its monotonicity, $V \geq \mathcal{T}^* V$ implies $V \geq \mathcal{T}^* V \geq \cdots \geq V^*$ for any value function V. Since constraints in the HALP formulation (5) enforce $V^{\widetilde{\mathbf{w}}} \geq \mathcal{T}^* V^{\widetilde{\mathbf{w}}}$, we conclude $V^{\widetilde{\mathbf{w}}} \geq V^*$. ■

Therefore, the objective value $\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right]$ is a natural measure for evaluating the impact of added basis functions. Unfortunately, the computation of $\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right]$ is in general infeasible. To address this issue, we optimize a surrogate metric, which is represented by the relaxed HALP objective $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$. The next proposition relates the objective values of the complete and relaxed HALP formulations.

Proposition 2 *Let $\widetilde{\mathbf{w}}$ be a solution to the HALP formulation (5) and $\widehat{\mathbf{w}}$ be a solution to its relaxed formulation (9) that is δ-infeasible. Then the objective value $\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right]$ is bounded as:*

$$\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right] \leq \mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right] + \frac{\delta}{1-\gamma}.$$

Proof: The claim is proved in two steps. First, we construct a point $\overline{\mathbf{w}}$ in the feasible region of the HALP such that $V^{\overline{\mathbf{w}}}$ is within $O(\delta)$ distance from $V^{\widehat{\mathbf{w}}}$. The point $\overline{\mathbf{w}}$ is given by:

$$\overline{\mathbf{w}} = \widehat{\mathbf{w}} + \frac{\delta}{1-\gamma} e,$$

where $e = (1, 0, \ldots, 0)$ is an indicator of the constant basis function $f_0(\mathbf{x}) \equiv 1$. This point satisfies all requirements and its feasibility can be handily verified by solving:

$$\begin{aligned} V^{\overline{\mathbf{w}}} - \mathcal{T}^* V^{\overline{\mathbf{w}}} &= V^{\widehat{\mathbf{w}}} + \frac{\delta}{1-\gamma} - \left(\mathcal{T}^* V^{\widehat{\mathbf{w}}} + \frac{\gamma\delta}{1-\gamma}\right) \\ &= V^{\widehat{\mathbf{w}}} - \mathcal{T}^* V^{\widehat{\mathbf{w}}} + \delta \\ &\geq 0, \end{aligned}$$

where $V^{\widehat{\mathbf{w}}} - \mathcal{T}^* V^{\widehat{\mathbf{w}}} \geq -\delta$ holds from the δ-infeasibility of $\widehat{\mathbf{w}}$. Since $\overline{\mathbf{w}}$ is feasible in the complete HALP, we conclude $\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right] \leq \mathrm{E}_\psi\left[V^{\overline{\mathbf{w}}}\right]$, which leads to our final result. ■

Proposition 2 has an important implication. Optimization of the objective function $\mathrm{E}_\psi\left[V^{\widetilde{\mathbf{w}}}\right]$ is possible by minimizing a relaxed objective $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$.

Scoring basis functions

To minimize the relaxed objective $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$, we use the dual formulation of a relaxed HALP.

Definition 3 *Let every variable in the relaxed HALP formulation (9) be subject to the constraint $w_i \geq 0$. Then the* dual relaxed HALP *is given by a linear program:*

$$\begin{aligned} \text{maximize}_\omega \quad & \sum_{(\mathbf{x},\mathbf{a}) \in \mathcal{C}} \omega_{\mathbf{x},\mathbf{a}} R(\mathbf{x},\mathbf{a}) \qquad (10) \\ \text{subject to:} \quad & \sum_{(\mathbf{x},\mathbf{a}) \in \mathcal{C}} \omega_{\mathbf{x},\mathbf{a}} F_i(\mathbf{x},\mathbf{a}) - \alpha_i \leq 0 \quad \forall\, i \\ & \omega_{\mathbf{x},\mathbf{a}} \geq 0; \end{aligned}$$

where $\omega_{\mathbf{x},\mathbf{a}}$ represents the variables in the LP, one for each constraint in the primal relaxed HALP, and the scope of the index i are all basis functions $f_i(\mathbf{x})$.

Based on the duality theory, we know that the primal (9) and dual (10) formulations have the same objective values. Thus, minimizing the objective of the dual minimizes the objective of the primal. Since the dual formulation is a maximization problem, its objective value can be lowered by adding a new constraint, which corresponds to a basis function $f(\mathbf{x})$ in the primal. Unfortunately, to evaluate the true impact of adding $f(\mathbf{x})$ on decreasing $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$, we need to resolve the primal with the added basis function. This step is computationally expensive and would significantly slow down any procedure that searches in the space of basis functions.

Similarly to Patrascu *et al.* (2002), we consider a different scoring metric. We define *dual violation magnitude* $\tau^{\widehat{\omega}}(f)$:

$$\tau^{\widehat{\omega}}(f) = \sum_{(\mathbf{x},\mathbf{a}) \in \mathcal{C}} \widehat{\omega}_{\mathbf{x},\mathbf{a}} [f(\mathbf{x}) - \gamma g_f(\mathbf{x},\mathbf{a})] - \alpha_f, \qquad (11)$$

which measures the amount by which the optimal solution $\widehat{\omega}$ to a dual relaxed HALP violates the constraint corresponding to the basis function $f(\mathbf{x})$. This score can be interpreted as evaluating the quality of cutting planes in the dual. Therefore, if $\tau^{\widehat{\omega}}(f)$ is nonnegative, a higher value of $\tau^{\widehat{\omega}}(f)$ is often correlated with a large decrease in the objective $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$ when the basis function $f(\mathbf{x})$ is added to the primal. Based

on the empirical evidence (Patrascu *et al.* 2002), this criterion prefers meaningful basis functions and it is significantly cheaper than resolving the primal.

Dual violation magnitude $\tau^{\widehat{\omega}}(f)$ can be evaluated very efficiently. Scoring of a basis function $f(\mathbf{x})$ requires computation of $g_f(\mathbf{x}, \mathbf{a})$ and $f(\mathbf{x})$ in all state-action pairs $(\mathbf{x}, \mathbf{a}) \in \mathcal{C}$. The number of computed terms can be significantly reduced since the only nonzero scalars $\widehat{\omega}_{\mathbf{x},\mathbf{a}}$ in Equation 11 are those that correspond to active constraints in the primal. Based on the duality theory, we conclude that the dual solution $\widehat{\omega}$ can be expressed in terms of the primal solution $\widehat{\mathbf{w}}$. As a result, we do not even need to formulate the dual to obtain $\widehat{\omega}$.

Optimization of basis functions

Dual violation magnitude $\tau^{\widehat{\omega}}(f)$ scores basis functions $f(\mathbf{x})$ and our goal is to find one with a high score. To allow for a systematic search among all basis functions, we assume that they factor along the state variables $\mathbf{X}$ (Equation 8). Moreover, their univariate factors $f_j(x_j)$ are from the exponential family of distributions:

$$f_j(x_j) = h_j(x_j) \exp[\eta_{f_j}^{\intercal} t_j(x_j)] / Z_j(\eta_{f_j}), \qquad (12)$$

where η_{f_j} denotes their natural parameters, $t_j(x_j)$ is a vector of sufficient statistics, and $Z_j(\eta_{f_j})$ is a normalizing function independent of X_j.

Any optimization method can be used to maximize $\tau^{\widehat{\omega}}(f)$ with respect to the natural parameters of $f(\mathbf{x})$. In this work, we employ the gradient method. Based on the independence assumption in Equation 8, we may conclude that the gradient $\nabla \tau^{\widehat{\omega}}(f)$ can be expressed by the derivatives of the univariate terms $f_j(x_j)$ and $\mathrm{E}_{P(x'_j|\mathbf{x},\mathbf{a})}\big[f_j(x'_j)\big]$. The derivatives have analytical forms for conjugate basis function choices.

Proposition 3 *Let:*

$$f(x) = h(x) \exp[\eta_f^{\intercal} t(x)] / Z(\eta_f)$$

be an exponential-family density over X, where η_f denotes its natural parameters, $t(x)$ is a vector of sufficient statistics, and $Z(\eta_f)$ is a normalizing function independent of X. Then:

$$\frac{\partial f(x)}{\partial \eta_{fk}} = f(x) \left[t_k(x) + \frac{1}{Z(\eta_f)} \frac{\partial Z(\eta_f)}{\partial \eta_{fk}} \right]$$

has a closed-form solution, where η_{fk} and $t_k(x)$ denote the k-th elements of the vectors η_f and $t(x)$.

Proof: Direct application of basic differentiation laws. ■

Proposition 4 *Let:*

$$\begin{aligned} P(x) &= h(x) \exp[\eta_P^{\intercal} t(x)] / Z(\eta_P) \\ f(x) &= h(x) \exp[\eta_f^{\intercal} t(x)] / Z(\eta_f) \end{aligned}$$

be exponential-family densities over X in the same canonical form, where η_P and η_f denote their natural parameters, $t(x)$ is a vector of sufficient statistics, and $Z(\cdot)$ is a normalizing function independent of X. If $h(x) \equiv 1$, then:

$$\frac{\partial \mathrm{E}_{P(x)}[f(x)]}{\partial \eta_{fk}} = \frac{1}{Z(\eta_P) Z(\eta_f)} \frac{\partial Z(\eta_P + \eta_f)}{\partial \eta_{fk}} - \frac{Z(\eta_P + \eta_f)}{Z(\eta_P) Z(\eta_f)^2} \frac{\partial Z(\eta_f)}{\partial \eta_{fk}}$$

has a closed-form solution, where η_{fk} denotes the k-th element of the vector η_f.

Proof: Based on Kveton and Hauskrecht (2006), we know:

$$\mathrm{E}_{P(x)}[f(x)] = \frac{Z(\eta_P + \eta_f)}{Z(\eta_P) Z(\eta_f)}.$$

The rest follows from basic differentiation laws. ■

Overfitting on active constraints

Optimization of the violation magnitude $\tau^{\widehat{\omega}}(f)$ easily *overfits* on active constraints in the primal relaxed HALP. To describe the phenomenon, let us assume that the basis function $f(\mathbf{x})$ is of unit magnitude, unimodal, and centered at an active constraint $(\mathbf{x}', \mathbf{a}')$. If $f(\mathbf{x}'') = 0$ at the remaining active constraints $(\mathbf{x}'', \mathbf{a}'')$, $\widehat{\omega}_{\mathbf{x}',\mathbf{a}'} f(\mathbf{x}')$ is the only positive term in Equation 11. Therefore, maximization of $\tau^{\widehat{\omega}}(f)$ can be performed by keeping $f(\mathbf{x}')$ fixed and minimizing the negative terms $g_f(\mathbf{x}, \mathbf{a})$ and α_f. Since the terms can be bounded from above as:

$$g_f(\mathbf{x}, \mathbf{a}) \;\leq\; \mathrm{E}[f(\mathbf{x})] \max_{\mathbf{x}'} P(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}) \qquad (13)$$

$$\alpha_f(\mathbf{x}, \mathbf{a}) \;\leq\; \mathrm{E}[f(\mathbf{x})] \max_{\mathbf{x}} \psi(\mathbf{x}), \qquad (14)$$

Equation 11 can be locally maximized by lowering the mass $\mathrm{E}[f(\mathbf{x})]$ corresponding to the function $f(\mathbf{x})$.

Although a peaked basis function may lower the relaxed objective $\mathrm{E}_{\psi}\big[V^{\widehat{\mathbf{w}}}\big]$, it is unlikely that it lowers the objective $\mathrm{E}_{\psi}\big[V^{\tilde{\mathbf{w}}}\big]$ in HALP. This observation can be understood from Proposition 2. Peaked basis functions have a high Lipschitz constant, which translates into a high δ-infeasibility of their relaxed solutions $\widehat{\mathbf{w}}$. If δ is high, the bound in Proposition 2 becomes loose, and the minimization of $\mathrm{E}_{\psi}\big[V^{\widehat{\mathbf{w}}}\big]$ no longer guarantees a low objective value $\mathrm{E}_{\psi}\big[V^{\tilde{\mathbf{w}}}\big]$.

To compensate for this behavior, we propose a modification to the gradient method. Instead of returning an arbitrary basis function that maximizes $\tau^{\widehat{\omega}}(f)$, we *restrict* our attention to those that adhere to a certain Lipschitz factor K. This parameter regulates the smoothness of our approximations.

Experiments

Experimental setup

Our approach to learning basis functions is demonstrated on two hybrid optimization problems: 6-ring irrigation network (Guestrin, Hauskrecht, & Kveton 2004) and a rover planning problem (Bresina *et al.* 2002). The irrigation network problems are challenging for state-of-the-art MDP solvers due to the factored state and action spaces. The goal of an irrigation network operator is to select discrete water-routing actions $\mathbf{A}_D$ to optimize continuous water levels $\mathbf{X}_C$ in multiple interconnected irrigation channels. The transition model is parameterized by beta distributions and represents water flows conditioned on the operation modes of regulation devices. The reward function is additive and described by a mixture of two normal distributions for each channel except for the outflow channel. The 6-ring network involves 10 continuous state and 10 discrete action variables. On the other hand, the rover problem is represented by only a single action variable

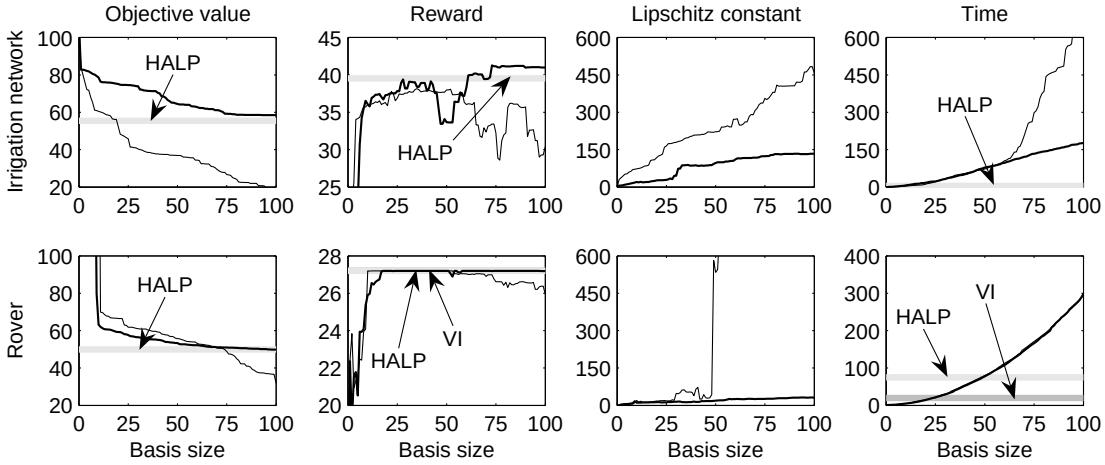

Figure 1: Comparison of the greedy (thin black lines) and restricted greedy (thick black lines) methods on the 6-ring irrigation network and rover problems. The methods are compared by the objective value of a relaxed HALP, the expected reward of a corresponding policy, upper bound on the Lipschitz constant of $V^{\widehat{\mathbf{w}}}$, and computation time (in seconds). This upper bound is computed as $\sum_i \widehat{w}_i \sum_{X_j \in \mathbf{X}_i} K_{ij}$, where K_{ij} represents the Lipschitz constant of the univariate basis function factor $f_{ij}(x_j)$. Thick gray lines denote our baselines.

A and three state variables S (exploration stage), T (remaining time to achieve a goal), and E (energy level) (Kveton & Hauskrecht 2006). Three branches of the rover exploration plan yield rewards 10, 55, and 100. The optimization problem is to choose one of these branches given the remaining time and energy. The state relevance density function $\psi(\mathbf{x})$ is uniform in both optimization problems. The discount factor γ is 0.95.

An optimal solution to both problems is approximated by a relaxed HALP whose constraints are restricted to an ε-grid ($\varepsilon = 1/8$). We compare two methods for learning new basis functions: *greedy*, which optimizes the dual violation magnitude $\tau^{\widehat{\omega}}(f)$, and *restricted greedy*, where the optimization is controlled by the Lipschitz threshold K. Both methods are evaluated for up to 100 added basis functions. The threshold K is regulated by an increasing logarithmic schedule from 2 to 8, which corresponds to the resolution of our ε-grid.

In the 6-ring irrigation network problem, we optimize univariate basis functions of the form:

$$f(x_i) = P_{\text{beta}}(x_i \mid \alpha, \beta). \tag{15}$$

Their parameters i, α, and β are initialized randomly. Our baseline is represented by 40 univariate basis functions suggested by Guestrin *et al.* (2004). In the rover planning problem, we optimize unimodal basis functions:

$$\begin{aligned} f(s,t,e) \;=\; & P(s \mid \theta_1, \ldots, \theta_{10}) \\ & \mathcal{N}(t \mid \mu_t, \sigma_t)\mathcal{N}(e \mid \mu_e, \sigma_e), \end{aligned} \tag{16}$$

where $P(s \mid \theta_1, \ldots, \theta_{10})$ is a multinomial distribution over 10 stages of rover exploration. All parameters are initialized randomly. Our baselines are given by value iteration, where the continuous variables S and T are discretized on the 17×17 grid, and a relaxed ε-HALP formulation ($\varepsilon = 1/16$) with 381 basis functions (Kveton & Hauskrecht 2006).

Experiments are performed on a Dell Precision 380 workstation with 3.2GHz Pentium 4 CPU and 2GB RAM. Linear programs are solved by the dual simplex method in CPLEX. Our experimental results are reported in Figures 1 and 2.

Experimental results

Figure 1 demonstrates the benefits of automatic basis function learning. On the 6-ring irrigation network problem, we learn better policies than the existing baseline in a very short time (150 seconds). On the rover problem, we learn as good policies as our baselines and this in comparable computation time. These results are even more encouraging since we may achieve additional several-fold speedup by caching relaxed HALP formulations.

Figure 1 also confirms our hypothesis that the minimization of the relaxed objective $\mathrm{E}_\psi\left[V^{\widehat{\mathbf{w}}}\right]$ without restricting the search yields suboptimal policies. As the number of learned basis functions grows, we can observe a correlation between dropping objectives and rewards, and growing upper bound on the Lipschitz factor of the approximations.

Finally, Figure 2 illustrates value functions learned on the 6-ring irrigation network problem. We can observe the phenomenon of overfiting (the second row from the top) or the gradual improvement of approximations constructed by the restricted greedy search (the last two rows).

Conclusions

Learning of basis functions in hybrid spaces is an important step towards applying MDPs to real-world problems. In this work, we presented a greedy method that achieves this goal. This method performs very well on two tested hybrid MDP problems and surpasses existing baselines by the quality of policies and computation time. An interesting open research question is the combination of our greedy search with a state space analysis (Mahadevan 2005; Mahadevan & Maggioni 2006).

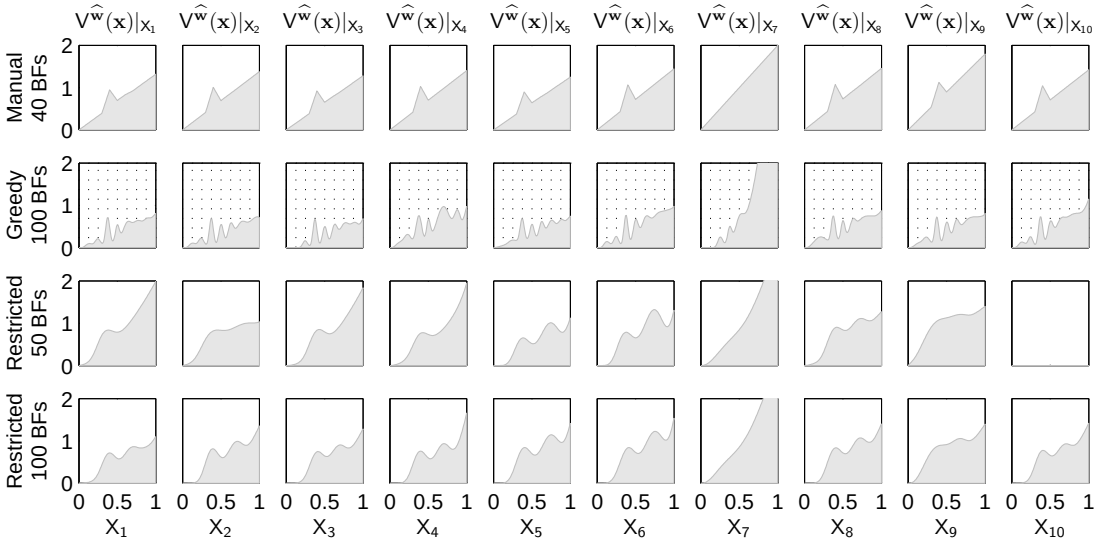

Figure 2: Univariate projections $V^{\hat{\mathbf{w}}}(\mathbf{x})|_{X_j} = \sum_{i:X_j=X_i} \widehat{w}_i f_i(x_i)$ of approximate value functions $V^{\hat{\mathbf{w}}}$ on the 6-ring irrigation network problem. From top down, we show value functions learned from 40 manually selected basis functions (BFs) (Guestrin, Hauskrecht, & Kveton 2004), 100 greedily learned BFs, and 50 and 100 BFs learned by the restricted greedy search. Note that the greedy approximation overfits on the ε-grid ($\varepsilon = 1/8$), which is represented by dotted lines.

Acknowledgment

We thank anonymous reviewers for comments that led to the improvement of the paper. This research was supported by Andrew Mellon Predoctoral Fellowships awarded in the academic years 2004-06 to Branislav Kveton and by the National Science Foundation grant ANI-0325353. The first author recognizes support from Intel Corporation in the summer 2005.

References

Bellman, R.; Kalaba, R.; and Kotkin, B. 1963. Polynomial approximation – a new computational technique in dynamic programming: Allocation processes. *Mathematics of Computation* 17(82):155–161.

Bellman, R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 1995. Exploiting structure in policy construction. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1104–1111.

Bresina, J.; Dearden, R.; Meuleau, N.; Ramakrishnan, S.; Smith, D.; and Washington, R. 2002. Planning under continuous time and resource uncertainty: A challenge for AI. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 77–84.

de Farias, D. P., and Van Roy, B. 2003. The linear programming approach to approximate dynamic programming. *Operations Research* 51(6):850–856.

Guestrin, C.; Hauskrecht, M.; and Kveton, B. 2004. Solving factored MDPs with continuous and discrete variables. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 235–242.

Hauskrecht, M., and Kveton, B. 2004. Linear program approximations for factored continuous-state Markov decision processes. In *Advances in Neural Information Processing Systems 16*, 895–902.

Koller, D., and Parr, R. 1999. Computing factored value functions for policies in structured MDPs. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1332–1339.

Kveton, B., and Hauskrecht, M. 2005. An MCMC approach to solving hybrid factored MDPs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1346–1351.

Kveton, B., and Hauskrecht, M. 2006. Solving factored MDPs with exponential-family transition models. In *Proceedings of the 16th International Conference on Automated Planning and Scheduling*.

Mahadevan, S., and Maggioni, M. 2006. Value function approximation with diffusion wavelets and Laplacian eigenfunctions. In *Advances in Neural Information Processing Systems 18*, 843–850.

Mahadevan, S. 2005. Samuel meets Amarel: Automating value function approximation using global state space analysis. In *Proceedings of the 20th National Conference on Artificial Intelligence*, 1000–1005.

Patrascu, R.; Poupart, P.; Schuurmans, D.; Boutilier, C.; and Guestrin, C. 2002. Greedy linear value-approximation for factored Markov decision processes. In *Proceedings of the 18th National Conference on Artificial Intelligence*, 285–291.

Puterman, M. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons.

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

Van Roy, B. 1998. *Planning Under Uncertainty in Complex Structured Environments*. Ph.D. Dissertation, Massachusetts Institute of Technology.