Solving Factored MDPs with Exponential-Family Transition Models

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 linear combination of basis functions and optimize it by linear programming. In this paper, we extend the existing HALP paradigm beyond the mixture of beta transition model. As a result, we permit modeling of other transition functions, such as normal and gamma densities, without approximating them. To allow for efficient solutions to the expectation terms in HALP, we identify a rich class of conjugate basis functions. Finally, we demonstrate the generalized HALP framework on a rover planning problem, which exhibits continuous time and resource uncertainty.

Introduction

Space exploration and problems arising in this domain have been a very important source of applied AI research in recent years. The design of a planning module for an autonomous Mars rover is one of the challenging problems. Along these lines, Bresina *et al.* (2002) outlined requirements for such a planning system. These include the ability to plan in continuous time, with concurrent actions, using limited resources, and all these in the presence of uncertainty. In the same paper, Bresina *et al.* (2002) described a simplified rover planning problem, which exhibits some of these characteristics. In this work, we show how to adapt approximate linear programming (ALP) (Schweitzer & Seidmann 1985) to address these types of problems.

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 existing HALP framework (Guestrin, Hauskrecht, & Kveton 2004; Hauskrecht & Kveton 2004) imposes a restriction on solved problems. Every continuous variable must be bounded on the $[0, 1]$ interval and all transition functions are given by a mixture of beta distributions. Different transition models, such as normal distributions, cannot be used directly and have to be approximated. In this work, we alleviate this assumption and allow exponential-family transition models.

The paper is structured as follows. First, we introduce hybrid factored MDPs (Guestrin, Hauskrecht, & Kveton 2004) and extend them by exponential-family transition functions. Second, we generalize HALP to solve the new class of problems efficiently. Third, we propose a rich class of conjugate basis functions that lead to closed-form solutions to the expectation terms in HALP. Finally, we demonstrate the HALP framework on an autonomous rover planning problem.

Generalized 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 section, we introduce a new formalism for representing hybrid factored MDPs with an exponential-family transition model. This formalism is based on the HMDP framework (Guestrin, Hauskrecht, & Kveton 2004) and generalizes its mixture of beta transition model for continuous variables.

A *hybrid factored MDP with an exponential-family transition model (HMDP)* is 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 state 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]

State variables: State variables are either discrete or continuous. The state of the system is observed and described by a vector of value assignments $\mathbf{x} = (\mathbf{x}_D, \mathbf{x}_C)$ which partitions along its discrete and continuous components $\mathbf{x}_D$ and $\mathbf{x}_C$.

Action variables: The action space is distributed and represented by action variables $\mathbf{A}$. The composite action is given by a vector of individual action choices $\mathbf{a} = (\mathbf{a}_D, \mathbf{a}_C)$ which partitions along its discrete and continuous components $\mathbf{a}_D$

[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.

$\mathrm{Dom}(X_i')$	Transition function
$\{0,\ldots,k\}$	**Multinomial distribution** $P(X_i'=j)=\theta_j$ where $\sum_j \theta_j = 1$ and $\theta_j = \phi_{ij}^{\theta}(\mathsf{Par}(X_i'))$
$[0,1]$	**Beta density** $P_{\mathrm{beta}}(X_i'=x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ where $\alpha = \phi_i^{\alpha}(\mathsf{Par}(X_i'))$ and $\beta = \phi_i^{\beta}(\mathsf{Par}(X_i'))$
$(-\infty,\infty)$	**Normal density** $\mathcal{N}(X_i'=x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$ where $\mu = \phi_i^{\mu}(\mathsf{Par}(X_i'))$ and $\sigma = \phi_i^{\sigma}(\mathsf{Par}(X_i'))$
$[0,\infty)$	**Gamma density** $P_{\mathrm{gamma}}(X_i'=x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\exp[-\frac{1}{\beta}x]$ where $\alpha = \phi_i^{\alpha}(\mathsf{Par}(X_i'))$ and $\beta = \phi_i^{\beta}(\mathsf{Par}(X_i'))$

Figure 1: Selecting transition functions based on $\mathrm{Dom}(X_i')$. The functions are parameterized by arbitrary functions $\phi_i(\cdot)$ of their parent sets $\mathsf{Par}(X_i')$.

and $\mathbf{a}_C$. It is natural to assume that each state variable A_i is either of finite cardinality or bounded.

Transition model: The transition model is given by the conditional probability distribution $P(\mathbf{X}' \mid \mathbf{X}, \mathbf{A})$, where $\mathbf{X}$ and $\mathbf{X}'$ denote the state variables at two successive time steps t and $t+1$. We assume that the transition model can be factored along $\mathbf{X}'$ as $P(\mathbf{X}' \mid \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{n} P(X_i' \mid \mathsf{Par}(X_i'))$ and compactly represented by a *dynamic Bayesian network (DBN)* (Dean & Kanazawa 1989). Typically, the parent set $\mathsf{Par}(X_i') \subseteq \mathbf{X} \cup \mathbf{A}$ is a small subset of state and action variables which allows for a local parameterization of the model.

Parameterization of transition model: One-step dynamics of every state variable is described by its conditional probability distribution $P(X_i' \mid \mathsf{Par}(X_i'))$. These conditionals are chosen from the exponential-family of distributions:

$$P(X_i' = x \mid \mathsf{Par}(X_i')) \;=\; h(x)\exp[\eta^{\intercal} t(x)]/Z(\eta) \qquad (1)$$

based on $\mathrm{Dom}(X_i')$, where η denotes the natural parameters of the distribution, $t(x)$ is a vector of its sufficient statistics, and $Z(\eta)$ is a normalizing function independent of X_i'. The choices that lead to closed-form[2] solutions to the expectation terms in HALP are shown Figure 1. Our work directly generalizes to the mixtures of these transition functions, which provide a very rich class of transition models.

Reward model: The reward function is an additive function $R(\mathbf{x},\mathbf{a}) = \sum_j R_j(\mathbf{x}_j, \mathbf{a}_j)$ of local reward functions defined on the subsets of state and action variables $\mathbf{X}_j$ and $\mathbf{A}_j$.

Optimal value function and policy: The quality of a 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. This optimality criterion guarantees that there always exists an *optimal policy* π^* which is stationary and deterministic (Puterman 1994). The policy is greedy with respect to the *optimal value function* V^*, which is a fixed point of the Bellman equation (Bellman 1957; Bertsekas & Tsitsiklis 1996):

$$V^*(\mathbf{x}) = \sup_{\mathbf{a}} \left[R(\mathbf{x},\mathbf{a}) + \gamma \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[V^*(\mathbf{x}')] \right]. \qquad (2)$$

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

$$\mathcal{T}^* V(\mathbf{x}) = \sup_{\mathbf{a}} \left[R(\mathbf{x},\mathbf{a}) + \gamma \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[V(\mathbf{x}')] \right]. \qquad (3)$$

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

$$\mathrm{E}_{P(\mathbf{x})}[f(\mathbf{x})] = \sum_{\mathbf{x}_D} \int_{\mathbf{x}_C} P(\mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x}_C. \qquad (4)$$

Generalized hybrid ALP

Value iteration, policy iteration, and linear programming are the most fundamental dynamic programming (DP) methods for solving MDPs (Puterman 1994; Bertsekas & Tsitsiklis 1996). Unfortunately, none of these methods is suitable for solving hybrid factored MDPs. First, their complexity grows exponentially in the number of state variables if the variables are discrete. Second, these methods assume a finite support for the optimal value function or policy, which may not exist if continuous variables are present. As a result, any feasible approach to solving arbitrary HMDPs is likely to be approximate. To compute these approximate solutions, Munos and Moore (2002) proposed an adaptive non-uniform discretization of continuous-state spaces and Feng *et al.* (2004) used DP backups of piecewise constant and piecewise linear value functions.

Linear value function model: 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):

$$V^{\mathbf{w}}(\mathbf{x}) = \sum_i w_i f_i(\mathbf{x}). \qquad (5)$$

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).

Generalized 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 5). Therefore, it transforms an initially intractable problem of estimating V^* in the hybrid state space $\mathbf{X}$ into a lower dimensional space $\mathbf{w}$. As shown in the rest of the section, this approach generalizes to HMDPs with exponential-family transition models.

The *generalized hybrid ALP (HALP)* formulation is given by a linear program:

$$\text{minimize}_{\mathbf{w}} \; \sum_i w_i \alpha_i \qquad (6)$$

$$\text{subject to:} \; \sum_i w_i F_i(\mathbf{x},\mathbf{a}) - R(\mathbf{x},\mathbf{a}) \geq 0 \quad \forall\, \mathbf{x},\mathbf{a};$$

[2] The term *closed-form* refers to an accepted set of closed-form operations and functions extended by the gamma function.

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

$$\alpha_i = \mathrm{E}_{\psi(\mathbf{x})}[f_i(\mathbf{x})] = \sum_{\mathbf{x}_D} \int_{\mathbf{x}_C} \psi(\mathbf{x}) f_i(\mathbf{x}) \, \mathrm{d}\mathbf{x}_C, \quad (7)$$

$\psi(\mathbf{x})$ 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*:

$$g_i(\mathbf{x}, \mathbf{a}) = \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[f_i(\mathbf{x}')] = \sum_{\mathbf{x}'_D} \int_{\mathbf{x}'_C} P(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}) f_i(\mathbf{x}') \, \mathrm{d}\mathbf{x}'_C. \quad (8)$$

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.

In the remainder of the paper, we address several concerns related to the HALP formulation. First, we analyze the quality of this approximation and relate it to the max-norm error $\|V^* - V^{\mathbf{w}}\|_\infty$, which is a commonly-used metric. Second, we present rich classes of basis functions that lead to closed-form solutions to the expectation terms in the objective function and constraints (Equations 7 and 8). These expectation terms involve sums and integrals over the complete space $\mathbf{X}$, and hence are hard to evaluate. Finally, we discuss approximations to the constraint space in HALP. Note that complete satisfaction of this constraint space may not even be possible since each state-action pair $(\mathbf{x}, \mathbf{a})$ induces a constraint.

Error bounds

For the generalized HALP formulation (6) to be of practical interest, the optimal value function V^* has to lie close to the span of basis functions $f_i(\mathbf{x})$. The following theorem states this intuitive claim formally. In particular, if the reweighted max-norm error $\|V^* - V^{\mathbf{w}}\|_{\infty,1/L}$ can be minimized while the growth rate of $\mathrm{E}_\psi[L]$ is controlled, the optimal solution $\widetilde{\mathbf{w}}$ to the HALP formulation yields a close approximation to the optimal value function V^*. In general, the theorem does not hold in the opposite direction. A low $\mathcal{L}_1$-norm error may not guarantee a low max-norm error.

Theorem 1 *Let $\widetilde{\mathbf{w}}$ be an optimal solution to the generalized HALP formulation (6). Then the quality of the value function $V^{\widetilde{\mathbf{w}}}$ can be bounded as:*

$$\left\| V^* - V^{\widetilde{\mathbf{w}}} \right\|_{1,\psi} \leq \frac{2\mathrm{E}_\psi[L]}{1-\kappa} \min_{\mathbf{w}} \|V^* - V^{\mathbf{w}}\|_{\infty,1/L},$$

where $\|\cdot\|_{1,\psi}$ is an $\mathcal{L}_1$-norm weighted by the state relevance density ψ, $L(\mathbf{x}) = \sum_i w_i^L f_i(\mathbf{x})$ is a Lyapunov function *such that $\kappa L(\mathbf{x}) \geq \gamma \sup_{\mathbf{a}} \mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[L(\mathbf{x}')]$, $\kappa \in [0, 1)$ denotes its* contraction factor, *and $\|\cdot\|_{\infty,1/L}$ is a max-norm weighted by the reciprocal $1/L$.*

Proof: Similarly to Theorem 3 (de Farias & Van Roy 2003), this claim is proved in three steps: finding a point $\overline{\mathbf{w}}$ in the feasible region of the LP, bounding the error of $V^{\overline{\mathbf{w}}}$, which in turn yields a bound on the error of $V^{\widetilde{\mathbf{w}}}$. A comprehensive proof for the discrete-state case was done by de Farias and Van Roy (2003). The proof can be generalized to structured state and action spaces with continuous state variables. ■

Expectation terms

Since our basis functions are often restricted to small subsets of state variables, expectation terms (Equations 7 and 8) in the generalized HALP formulation (6) should be efficiently computable. Before we justify this claim, let us assume the state relevance density function $\psi(\mathbf{x})$ factors along $\mathbf{X}$ as:

$$\psi(\mathbf{x}) = \prod_{i=1}^{n} \psi_i(x_i), \quad (9)$$

where $\psi_i(x_i)$ is an exponential-family distribution over the state variable X_i. As a consequence, we can view both types of expectation terms, $\mathrm{E}_{\psi(\mathbf{x})}[f_i(\mathbf{x})]$ and $\mathrm{E}_{P(\mathbf{x}'|\mathbf{x},\mathbf{a})}[f_i(\mathbf{x}')]$, as being instances of $\mathrm{E}_{P(\mathbf{x})}[f_i(\mathbf{x})]$, where $P(\mathbf{x}) = \prod_j P(x_j)$ is a factored probability distribution. Therefore, to compute the expectation terms in HALP, it suffices to address a more general problem of estimating $\mathrm{E}_{P(\mathbf{x})}[f_i(\mathbf{x})]$. Since our work extends to the mixtures of state relevance densities, the independence assumption in Equation 9 can be partially relaxed.

Before computing the expectation term $\mathrm{E}_{P(\mathbf{x})}[f_i(\mathbf{x})]$ over the complete state space $\mathbf{X}$, we realize that the basis function $f_i(\mathbf{x})$ is defined on a subset of state variables $\mathbf{X}_i$. Therefore, we know that $\mathrm{E}_{P(\mathbf{x})}[f_i(\mathbf{x})] = \mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)]$, where $P(\mathbf{x}_i)$ denotes a factored distribution on a lower dimensional space $\mathbf{X}_i$. If no further assumptions are made, the expectation term $\mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)]$ may be still hard to compute. Although it can be estimated by a variety of numerical methods, for instance Monte Carlo (Andrieu *et al.* 2003), these techniques are imprecise if the sample size is small, and quite computationally expensive if a high precision is needed. Therefore, we try to avoid such an approximation step. Instead, we introduce an appropriate form of basis functions that leads to closed-form solutions to $\mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)]$.

First, let us assume that every basis function $f_i(\mathbf{x}_i)$ factors along the state variables $\mathbf{X}$ as:

$$f_i(\mathbf{x}_i) = \prod_{X_j \in \mathbf{X}_i} f_{ij}(x_j). \quad (10)$$

Then the expectation term $\mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)]$ decomposes as a product:

$$\mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)] = \prod_{X_j \in \mathbf{X}_i} \mathrm{E}_{P(x_j)}[f_{ij}(x_j)] \quad (11)$$

of expectations over individual variables X_j. As a result, an efficient solution to $\mathrm{E}_{P(\mathbf{x}_i)}[f_i(\mathbf{x}_i)]$ is guaranteed by efficient solutions to its univariate components $\mathrm{E}_{P(x_j)}[f_{ij}(x_j)]$.

Exponential-family distributions

To obtain closed-form solutions to $\mathrm{E}_{P(x_j)}[f_{ij}(x_j)]$, we consider univariate basis function factors:

$$f(x) = h(x) \exp[\eta^\intercal t(x)] / Z(\eta), \quad (12)$$

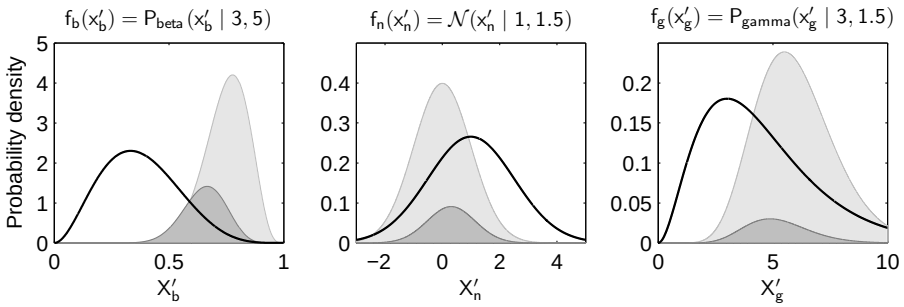

Figure 2: Expectation of three basis functions $f(x')$, denoted by thick black lines, with respect to three transition densities from Example 1, shown in a light gray color. Darker gray lines are given by the product $P(x')f(x')$. The area below corresponds to the expectation terms $\mathrm{E}_{P(x')}[f(x')]$.

where η denotes their natural parameters, $t(x)$ is a vector of their sufficient statistics, and $Z(\eta)$ is a normalizing function independent of x. The following proposition offers a recipe for choosing univariate *conjugate* factors $f_{ij}(x_j)$ to complement transition functions $P(x_j)$.

Proposition 1 *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 $h(x) \equiv 1$, $\mathrm{E}_{P(x)}[f(x)]$ has a closed-form solution:

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

Proof: The proposition is proved in a sequence of steps:

$$\begin{aligned} \mathrm{E}_{P(x)}[f(x)] &= \int_x P(x)f(x)\,\mathrm{d}x \\ &= \int_x \frac{\exp[\eta_P^\intercal t(x)]}{Z(\eta_P)}\frac{\exp[\eta_f^\intercal t(x)]}{Z(\eta_f)}\,\mathrm{d}x \\ &= \frac{Z(\eta_P + \eta_f)}{Z(\eta_P)Z(\eta_f)}\underbrace{\int_x \frac{\exp[(\eta_P + \eta_f)^\intercal t(x)]}{Z(\eta_P + \eta_f)}\,\mathrm{d}x}_{1}. \end{aligned}$$

The final step is a consequence of integrating a distribution. Since integration is a distributive operation, the proof generalizes to the mixtures of $P(x)$ and $f(x)$. ■

Corollary 1 *Let $P(x)$ and $f(x)$ be mixtures of exponential-family densities over X in the same canonical form. Assuming that $h(x) \equiv 1$, $\mathrm{E}_{P(x)}[f(x)]$ has a closed-form solution.*

Proposition 1 and Corollary 1 have an important implication for selecting an appropriate class of basis functions. For instance, normal and beta transition models are complemented by normal and beta basis functions. These in turn guarantee closed-form solutions to Equation 8.

Example 1 *To illustrate closed-form solutions to the expectation terms in HALP, we consider a transition model:*

$$\begin{aligned} P(\mathbf{x}') &= P_{\mathrm{beta}}(x'_b \mid 15, 5)\mathcal{N}(x'_n \mid 0, 1) \\ &\quad P_{\mathrm{gamma}}(x'_g \mid 12, 0.5) \end{aligned}$$

on the state space with three variables $\mathbf{X} = \{X_b, X_n, X_g\}$ such that $\mathrm{Dom}(X_b) = [0, 1]$, $\mathrm{Dom}(X_n) = (-\infty, \infty)$, and $\mathrm{Dom}(X_g) = [0, \infty)$. Following Proposition 1, we may conclude that basis functions of the form:

$$\begin{aligned} f(\mathbf{x}') &= f_b(x'_b)f_n(x'_n)f_g(x'_g) \\ &= P_{\mathrm{beta}}(x'_b \mid \alpha_b, \beta_b)\mathcal{N}(x'_n \mid \mu_n, \sigma_n) \\ &\quad P_{\mathrm{gamma}}(x'_g \mid \alpha_g, \beta_g) \end{aligned}$$

permit closed-form solutions to $\mathrm{E}_{P(\mathbf{x}')}[f(\mathbf{x}')]$. A graphical interpretation of this computation is given in Figure 2. Brief inspection verifies that univariate products $P(x')f(x')$ have the same canonical form as $P(x')$ and $f(x')$.

Hauskrecht and Kveton (2004) have recently identified polynomial basis functions as a conjugate choice for the mixture of beta transition model. Since any polynomial can be written as a linear combination of the products of beta densities, this result follows from Corollary 1. Similarly to our conjugate choices, piecewise constant functions establish another well-behaving category of basis functions. The expectations of these functions are computed as a weighted sum of cumulative distribution functions corresponding to the transitions.

Independence assumptions

Since the expectation terms in the generalized HALP are of the form $\mathrm{E}_{P(x)}[f(x)]$, we can extend the current set of basis functions by their linear combinations due to the identity:

$$\begin{aligned} &\mathrm{E}_{P(x)}[w_f f(x) + w_g g(x)] = \\ &\quad w_f \mathrm{E}_{P(x)}[f(x)] + w_g \mathrm{E}_{P(x)}[g(x)]\,. \end{aligned} \tag{13}$$

As a result, we can partially correct for the independence assumption in Equation 10. Furthermore, if we assume that the univariate factors $f_{ij}(x_j)$ are polynomials, the linear combination of basis functions $f_i(\mathbf{x}_i)$ is a polynomial. Following the Weierstrass approximation theorem (Jeffreys & Jeffreys 1988), this polynomial is sufficient to approximate any continuous basis function over $\mathbf{X}_i$ with any precision.

Constraint space approximations

An optimal solution $\widetilde{\mathbf{w}}$ to the generalized HALP formulation (6) 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 components 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 $\widetilde{\mathbf{w}}$. Thus, we resort to constraint space approximations 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*:*

$$\text{minimize}_{\mathbf{w}} \sum_i w_i \alpha_i \tag{14}$$

$$\text{subject to: } \sum_i w_i F_i(\mathbf{x}, \mathbf{a}) - R(\mathbf{x}, \mathbf{a}) \geq 0 \quad (\mathbf{x}, \mathbf{a}) \in \mathcal{C};$$

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

The HALP formulation (6) is solved approximately by solving its relaxed formulations (14). Several methods for building and solving these approximate LPs have been proposed: Monte Carlo sampling of constraints (Hauskrecht & Kveton 2004), ε-grid discretization (Guestrin, Hauskrecht, & Kveton 2004), and an adaptive MCMC search for a violated constraint (Kveton & Hauskrecht 2005). Each technique comes with its advantages and limitations.

Monte Carlo methods approximate the constraint space in HALP by its sample. Unfortunately, their efficiency depends on an appropriate choice of sampling distributions. The ones that yield good approximations and polynomial sample size bounds are closely related to the optimal solutions and rarely known in advance (de Farias & Van Roy 2004). On the other hand, constraint sampling is easily applied in continuous domains and its space complexity is proportional to the number of variables. The ε-HALP formulation relaxes the continuous portion of the constraint space to an ε-grid by discretizing continuous variables $\mathbf{X}_C$ and $\mathbf{A}_C$. Since the discretized constraint space preserves its factored structure, we can satisfy it compactly (Guestrin, Koller, & Parr 2001). Although this relaxation guarantees $\widehat{\mathbf{w}} \to \widetilde{\mathbf{w}}$ if $\varepsilon \to 0$, it is impractical for small ε (Kveton & Hauskrecht 2005). In addition, ε-grid discretization cannot be used for unbounded state variables. Finally, construction of relaxed formulations can be viewed as an incremental search for violated constraints. The search can be performed intelligently based on the structure of factored MDPs. An example of such an approach is the MCMC method of Kveton and Hauskrecht (2005).

Experiments

The goal of our experiments is to demonstrate the quality of generalized HALP approximations rather than the scale-up potential of the framework. Therefore, we use a realistic but low-dimensional autonomous rover problem (Bresina *et al.* 2002). For scale-up studies in hybrid spaces, please refer to Guestrin *et al.* (2004) and Kveton and Hauskrecht (2005). These conclusions fully extend to the generalized HALP.

Generalized HALP

Basis configurations	Reward	Time	OV
2×2 (41)	27.2 ± 34.5	5	60.5
3×3 (131)	27.2 ± 34.5	18	56.3
5×5 (381)	27.2 ± 34.5	75	49.9
9×9 (1 191)	27.2 ± 34.5	560	42.8

Grid-based VI

Grid configurations	Reward	Time
5×5 (250)	25.5 ± 33.4	< 1
9×9 (810)	26.2 ± 32.9	2
17×17 (2 890)	27.2 ± 33.8	20
33×33 (10 890)	27.4 ± 34.1	281

Figure 3: Comparison of two approaches to solving the rover problem (Bresina *et al.* 2002). The methods are compared by the objective value of a relaxed HALP (OV), the expected discounted reward of a corresponding policy, and their computation time (in seconds). The expected reward is estimated by the Monte Carlo simulation of 5000 trajectories starting at the initial exploration stage s_1 (Figure 4). The variance of our estimates is due to the natural variance of policies at s_1. The results are reported for different configurations of basis functions (grid points). The value in parentheses denotes the total number of basis functions (grid points).

Experimental setup

The autonomous rover problem was recently presented as a challenging real-world task, which involves continuous time and resource uncertainty. We represent the problem as a generalized HMDP with three state variables S, T, and E, and one binary action variable A. The state variable S is discrete and denotes 10 stages of rover exploration, the variable T is continuous and represents remaining time to achieve a goal, and the variable E is continuous and stores the energy level of the rover. The transition functions of T and E are given by normal distributions conditioned on the action choice a, exploration stage s, time t, and energy level e (Bresina *et al.* 2002). Three branches of the exploration plan yield rewards of 10, 55, and 100. The optimization problem is to choose one of these branches with respect to the remaining time and energy. A complete description of the example can be found in Bresina *et al.* (2002). The state relevance density function $\psi(\mathbf{x})$ is uniform. The discount factor γ is 0.95.

An approximate solution to this problem is obtained from a relaxed HALP whose constraints are restricted to an ε-grid ($\varepsilon = 1/16$). The optimal value function V^* is approximated by various configurations of bivariate normal basis functions over the state variables T and E. The functions are centered at vertices of uniform $n \times n$ grids over the state variables T and E. The grids are replicated for each exploration stage s. For all univariate basis function factors, the variance parameter is given by $(|\mathrm{Dom}(X_i)|/n)^2$, where $|\mathrm{Dom}(T)| = 4200$ and $|\mathrm{Dom}(E)| = 20$. As a baseline for our approximations, we consider value functions computed by value iteration on uniformly discretized variables T and E (Chow & Tsitsiklis 1991; Rust 1997). The algorithm converges after 5 iterations due to the dynamics of the problem.

Experiments are performed on a Dell Precision 380 work-

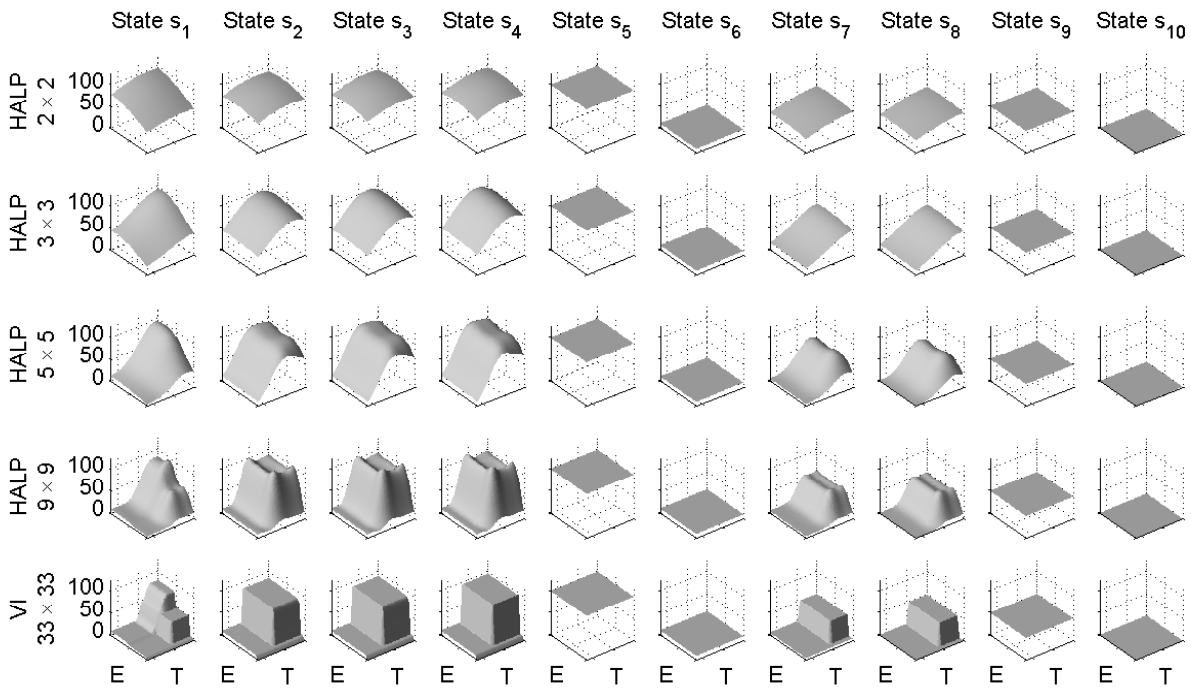

Figure 4: Value function approximations corresponding to the results in Figure 3. The approximations are presented as functions of time (T) and energy (E) for each exploration stage $S = \{s_1, \ldots, s_{10}\}$. Note that the most elaborate HALP approximation (9×9 basis configuration) closely resembles the optimal value function (Bresina *et al.* 2002).

station 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 3 and 4.

Experimental results

Based on our results, we can draw the following conclusion. The generalized HALP is a conceptually valid and practical way of solving hybrid optimization problems. Although our basis functions are placed blindly without any prior knowledge, the simplest HALP approximation (2×2 basis configuration) yields a close-to-optimal policy. The approximation is obtained even faster than a corresponding grid approximation of the same quality. This result is even more encouraging since we may achieve additional several-fold speedup by considering the locality of basis functions in HALP.

Interestingly, even if the quality of HALP approximations improves with a larger number of basis functions, the quality of policies stays the same. Since the optimal value function (Bresina *et al.* 2002) is monotonically increasing in T and E, we believe that capturing this behavior is crucial for obtaining a close-to-optimal policy. The simplest HALP approximation (2×2 basis configuration) exhibits this trend.

Conclusions

Development of efficient methods for solving large factored MDPs is a challenging task. In this work, we demonstrated a non-trivial extension to the HALP framework, which significantly expands the class of solvable problems. Furthermore, we applied the framework to an autonomous rover planning problem and found a close-to-optimal policy with a small set of blindly constructed basis functions. Unfortunately, such a naive approach to choosing basis functions would be infeasible if the number of state variables was larger. The objective of our future research is to learn good sets of basis functions automatically. In the context of discrete-state ALP, Patrascu *et al.* (2002) proposed a greedy approach to learn an appropriate class of linear approximators. Mahadevan (2005) and Mahadevan and Maggioni (2006) used a state space analysis to discover the set of plausible basis functions.

Acknowledgment

During the academic years 2004-06, the first author was supported by two Andrew Mellon Predoctoral Fellowships. The first author recognizes support from Intel Corporation in the summer 2005. This research was also partially supported by two National Science Foundation grants CMS-0416754 and ANI-0325353. We thank anonymous reviewers for providing comments that led to the improvement of the paper.

References

Andrieu, C.; de Freitas, N.; Doucet, A.; and Jordan, M. 2003. An introduction to MCMC for machine learning. *Machine Learning* 50:5–43.

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.

Bertsekas, D., and Tsitsiklis, J. 1996. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

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.

Chow, C.-S., and Tsitsiklis, J. 1991. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control* 36(8):898–914.

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

de Farias, D. P., and Van Roy, B. 2004. On constraint sampling for the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research* 29(3):462–478.

Dean, T., and Kanazawa, K. 1989. A model for reasoning about persistence and causation. *Computational Intelligence* 5:142–150.

Feng, Z.; Dearden, R.; Meuleau, N.; and Washington, R. 2004. Dynamic programming for structured continuous Markov decision problems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 154–161.

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.

Guestrin, C.; Koller, D.; and Parr, R. 2001. Max-norm projections for factored MDPs. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, 673–682.

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.

Jeffreys, H., and Jeffreys, B. 1988. *Methods of Mathematical Physics*. Cambridge, United Kingdom: Cambridge University Press.

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.

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.

Munos, R., and Moore, A. 2002. Variable resolution discretization in optimal control. *Machine Learning* 49:291–323.

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.

Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65(3):487–516.

Schuurmans, D., and Patrascu, R. 2002. Direct value-approximation for factored MDPs. In *Advances in Neural Information Processing Systems 14*, 1579–1586.

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.