Region-Based Value Iteration for Partially Observable Markov Decision Processes

Hui Li
Xuejun Liao
Lawrence Carin

HL1@EE.DUKE.EDU
XJLIAO@EE.DUKE.EDU
LCARIN@EE.DUKE.EDU

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

Abstract

An approximate region-based value iteration (RBVI) algorithm is proposed to find the optimal policy for a partially observable Markov decision process (POMDP). The proposed RBVI approximates the true polyhedral partition of the belief simplex with an ellipsoidal partition, such that the optimal value function is linear in each of the ellipsoidal regions. The position and shape of each region, as well as the gradient (alpha-vector) of the optimal value function in the region, are parameterized explicitly, and are estimated via efficient expectation maximization (EM) and variational Bayesian EM (VBEM), based on a set of selected sample belief points. The RBVI maintains a much smaller number of alpha-vectors than point-based methods and yields a more parsimonious representation that approximates the true value function in the maximum likelihood (ML) sense. The results on benchmark problems show that the proposed RBVI is comparable in performance to state-of-the-art algorithms, despite of the small number of α -vectors that are used.

1. Introduction

The partially observable Markov decision process (POMDP) (Sondik, 1971; Lovejoy, 1991; Kaelbling et al., 1998) provides a rich mathematical framework for planning under uncertainty. A POMDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, T, \mathcal{O}, \Omega, R)$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$ respectively denote a finite set of states, actions, and observations, T are state-transition matrices with $T_{ss'}(a)$ the

probability of transiting to state s' by taking action a in state s , Ω are observation functions with $\Omega_{s'o}(a)$ the probability of observing o after performing action a and transiting to state s' , and R is a reward function with $R(s, a)$ the expected immediate reward received by taking action a in state s .

The true underlying states are not observed directly in a POMDP. Instead the agent maintains a belief state, defined as the probability distribution over the states given past actions and observations. The belief state constitutes a continuous-state Markov process (Smallwood & Sondik, 1973). Given that at time $t-1$ the agent has belief state b and takes action a , and that the agent observes o at time t , the belief state b' at time t is updated according to Bayes rule

$$b'(s') = \frac{\sum_s b(s) T_{ss'}^a \Omega_{s'o}^a}{p(o|b, a)} \quad (1)$$

$$\text{with } p(o|b, a) = \sum_{s'} \sum_s b(s) T_{ss'}^a \Omega_{s'o}^a \quad (2)$$

The goal in POMDP planning is to find a policy that maps any belief state to an optimal action, with the objective of maximizing the expected future reward of each belief state over a specified horizon.

Many POMDP planning algorithms developed to date belong to the family of value iteration, which recursively computes the optimal value function using dynamic programming, starting from a zero-length horizon to a horizon of arbitrary length. Let $V^n(b)$ denote the optimal value function representing the maximum expected future reward of any belief point b for a horizon of length n , and let $\gamma \in [0, 1)$ be a discount factor. Then $V^n(b)$ can be computed from $V^{n-1}(b)$ as

$$V^n(b) = \max_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} b(s) R(s, a) + \gamma \sum_{o \in \mathcal{O}} p(o|b, a) V^{n-1}(b') \right] \quad (3)$$

where b' is given by (1). Smallwood and Sondik (1973) proved that the solution to (3) has a simple form

$$V^n(b) = \max_k \left[\sum_{s \in \mathcal{S}} \alpha_k^n(s) b(s) \right] = \max_k (b^T \alpha_k^n) \quad (4)$$

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

where the superscript T denotes matrix (vector) transpose and $\alpha_k^n = [\alpha_k^n(1), \alpha_k^n(2), \dots, \alpha_k^n(|\mathcal{S}|)]^T$ is called an α -vector. Equation (4) implies that for any finite horizon $V^n(b)$ is piecewise linear and convex in b and the gradient of its k -th hyperplane is α_k^n . In light of (4), the value iteration problem becomes one of computing α^n from α^{n-1} , which is solved by substituting (4) and (1) into (3),

$$\begin{aligned} V^n(b) &= \max_{a \in \mathcal{A}} \left\{ \sum_{s \in \mathcal{S}} b(s) R(s, a) \right. \\ &\quad \left. + \gamma \sum_{o \in \mathcal{O}} \max_k \sum_{s \in \mathcal{S}} b(s) \sum_{s' \in \mathcal{S}} T_{ss'}^a \Omega_{s'o}^a \alpha_k^{n-1}(s') \right\} \\ &= \max_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} b(s) \left[R(s, a) \right. \\ &\quad \left. + \gamma \sum_{o \in \mathcal{O}} \sum_{s' \in \mathcal{S}} T_{ss'}^a \Omega_{s'o}^a \alpha_{l(b,a,o)}^{n-1}(s') \right] \end{aligned} \quad (5)$$

where $l(b, a, o) = \arg \max_k \{ \sum_{s \in \mathcal{S}} b(s) \sum_{s' \in \mathcal{S}} T_{ss'}^a \Omega_{s'o}^a \times \alpha_k^{n-1}(s') \}$. As each bracketed expression in the rightmost side of (5) produces a possible α^n , the number of distinct α^n can be as large as $|\mathcal{A}| \{ |\alpha^{n-1}| \}^{|\mathcal{O}|}$, where $|\mathcal{A}|$ and $|\mathcal{O}|$ denote the size of $\mathcal{A}$ and $\mathcal{O}$, respectively, and $\{ |\alpha^{n-1}| \}$ denotes the number of distinct α^{n-1} . It is clear that the number of distinct α^n will quickly become intractable as the horizon length n increases.

Often many of the α^n produced in the rightmost side of (5) are *inactive*, where α_k^n is inactive if $b^T \alpha_k^n \leq b^T \alpha_j^n \forall j \neq k \forall b$. Exact value iteration algorithms such as the one-pass algorithm (Smallwood & Sondik, 1973), the linear-support algorithm (Cheng, 1988), and the witness algorithm (Kaelbling et al., 1998) seek to identify the complete set of *active* α -vectors, pruning away the inactive ones. However, the pruning requires solving many linear programs and could be computationally very expensive.

Point-based methods (Lovejoy, 1991; Brafman, 1997; Poon, 2001; Pineau et al., 2003) alleviate the computational load by focusing only on a finite set of belief points $\mathcal{B}$. In particular, Pineau et al. (2003) suggested constructing $\mathcal{B}$ as a set of points most uniformly sampled from the belief points reachable by the POMDP model, and backing up an α -vector for each point in $\mathcal{B}$. The resulting algorithm, point-based value iteration (PBVI), was proven to be a practical POMDP solution scaling up to large problems. The idea has subsequently been pursued in a number of papers (Spaan & Vlassis, 2004; Smith & Simmons, 2004; Smith & Simmons, 2005), where various heuristics were proposed to further improve the algorithmic efficiency.

It is known that PBVI has a time complexity of $O(|\mathcal{B}| |\mathcal{S}| |\Gamma_{\text{PBVI}}| |\mathcal{A}| |\mathcal{O}|)$ (Pineau et al., 2003) for a single iteration, where Γ_{PBVI} denotes the set of α -vectors backed up by PBVI in the previous iteration. As PBVI

backs up an α -vector for each point in $\mathcal{B}$, the complexity is equal to $O(|\mathcal{B}|^2 |\mathcal{S}| |\mathcal{A}| |\mathcal{O}|)$. Clearly the size of $\mathcal{B}$ greatly affects the efficiency of PBVI. In this paper, we introduce a *region-based value iteration* (RBVI) algorithm, which backs up a single α -vector for each of the convex regions over which the optimal value function is linear. The RBVI is a two-stage processing: first it computes $V^n(b) \forall b \in \mathcal{B}$ from α^{n-1} using (5); second, it uses an estimator Υ to estimate the region-based α^n from the samples $\{b, V^n(b)\}_{b \in \mathcal{B}}$. Computation of $\{V^n(b)\}_{b \in \mathcal{B}}$ has a time complexity of $O(|\mathcal{B}| |\mathcal{S}| |\Gamma_{\text{RBVI}}| |\mathcal{A}| |\mathcal{O}|)$, where Γ_{RBVI} denotes the set of α -vectors backed up by RBVI in the previous iteration; Υ has a complexity denoted by $O(\Upsilon)$. Clearly, $|\Gamma_{\text{RBVI}}| < |\Gamma_{\text{PBVI}}|$, hence the first stage of RBVI is always more efficient than PBVI. Therefore any Υ satisfying $O(\Upsilon) < O(|\mathcal{B}| |\mathcal{S}| |\Gamma_{\text{PBVI}}| |\mathcal{A}| |\mathcal{O}|)$ will yield a RBVI that computes more efficiently than PBVI.

As implied by (4), the belief simplex is partitioned into a finite number of polyhedral regions, such that $V^n(b)$ is linear in b in each region. In the proposed RBVI algorithm we approximate this polyhedral partition with an ellipsoidal partition, in which each polyhedral region is approximated by one or more ellipsoidal regions and a single α -vector is estimated and backed up for each ellipsoidal region. We derive efficient estimators Υ to jointly estimate the α -vectors as well as the position and shape parameters of each ellipsoidal region, using the samples $\{b, V^n(b)\}_{b \in \mathcal{B}}$. For completeness and comparison, we give two versions of Υ , based on expectation-maximization (EM) and variational Bayesian EM (VBEM). We show in Section 8 that the time complexity of either EM or VBEM is no worse than $O(|\mathcal{B}| [\dim(U^T b)]^2 |\Gamma_{\text{RBVI}}| N_\Upsilon)$, where N_Υ is the number of EM or VBEM iterations, and $\dim(U^T b) \leq |\mathcal{S}|$, with $\dim$ denoting the dimensionality and U a matrix of full column rank.

2. Ellipsoidal Approximation of the Polyhedral Piecewise Linearity

For a horizon of length n , the proposed RBVI algorithm is designed to probabilistically partition the belief simplex into a number of ellipsoidal region such that in each of the regions the optimal value function $V^n(b)$ is linear in b . See Figure 1 for an illustration and (6) for the mathematical formulation. This ellipsoidal partition is an approximation of the true polyhedral partition (Smallwood & Sondik, 1973). The position and shape of each ellipsoidal region, along with the α -vector (gradient) of $V^n(b)$ in each region, are estimated from the samples $\{b, V^n(b)\}_{b \in \mathcal{B}}$.

For notational simplicity, we write the optimal value

function as $V(b)$ in Sections 3 and 4, dropping the dependence on the horizon length n . The discussions apply to any finite n , unless otherwise indicated.

Figure 1. Partition of the belief simplex into K ellipsoidal regions, such that the optimal value function $V(b)$ is linear in b in each region. Here it is assumed that $|\mathcal{S}| = 3$ and the true piecewise linear convex partition are 3 polygons. $K = 3$ ellipses are used to cover (probabilistically) the belief simplex and preserve the piecewise linear convexity of $V(b)$.

We now formally describe the proposed ellipsoidal partition of the belief simplex. Let B_k represent the k -th ellipsoidal region in which the optimal value function $V(b)$ is linear in b . Assume $\{B_k : k = 1, \dots, K\}$ form an approximate partition of the belief simplex. Then $V(b)$ can be modeled as

$$V(b) = \begin{cases} \alpha_1^T b + e_1 & \text{if } b \in B_1 \\ \alpha_2^T b + e_2 & \text{if } b \in B_2 \\ \vdots & \vdots \\ \alpha_K^T b + e_K & \text{if } b \in B_K \end{cases}, e_k \sim \mathcal{N}(0, \mu_k^2) \quad (6)$$

where e_k is introduced to reflect any noise or error due to the approximation, and it is assumed to be a zero-mean Gaussian distribution with variance μ_k . The Gaussian noise assumption produces a least-square (LS) solution of α_k , $k = 1, \dots, K$, using the data from region k (probabilistically defined) as will be shown in (18), which is reasonable since all belief samples are equally important (the importance has been accounted for in the sampling — a more important belief point will be sampled more frequently, see Section 5).

Letting $z(b) = k$ indicate that $b \in B_k$, we write (6) equivalently as

$$p(V(b)|b, z(b) = k, \alpha_k) = (2\pi\mu_k^2)^{-1/2} \exp\left\{-\frac{(V(b) - \alpha_k^T b)^2}{2\mu_k^2}\right\} \quad (7)$$

The k -th ellipsoidal region in the belief simplex is represented probabilistically as

$$p(b|z(b) = k, c_k, D_k) = ((2\pi)^d |D_k|)^{-1/2} \exp\left\{-\frac{1}{2}(b - c_k)^T D_k^{-1} (b - c_k)\right\} \quad (8)$$

where $d = |\mathcal{S}|$ is the number of states in the POMDP model; c_k is a d -dimensional column and D_k is a $d \times d$ matrix¹; c_k and D_k represent the position and shape parameters, respectively, of the k -th ellipsoidal region.

For any b , the prior distribution of $z(b)$ over the K ellipsoidal regions is assumed to be $p(z(b) = k) = \omega_k$. The joint probability of $\{b, V(b), z(b)\}$ is given by

$$p(V(b), b, z(b) = k | \omega_k, c_k, D_k, \alpha_k) = \omega_k p(b|z(b) = k; c_k, D_k) p(V(b)|b, z(b) = k; \alpha_k) \quad (9)$$

Integrating out z yields the joint probability of $(b, V(b))$,

$$p(V(b), b | \omega_k, c_k, D_k, \alpha_k) = \sum_{k=1}^K \omega_k p(b|z(b) = k, c_k, D_k) p(V(b)|b, z(b) = k, \alpha_k) \quad (10)$$

The proposed RBVI aims to find the α -vectors based on a finite set of samples $\{b, V(b)\}_{b \in \mathcal{B}}$, construction of which is discussed in Section 5. In the following two sections, we focus on estimation of $\{\omega_k, c_k, D_k, \alpha_k\}_{k=1}^K$, assuming that this sample set is given.

3. Parameter Estimation Using Expectation-Maximization (EM)

We consider a given set of belief points $\mathcal{B} = \{b_1, b_2, \dots, b_{|\mathcal{B}|}\}$ and the set of associated optimal values $\mathcal{V} = \{V_1, V_2, \dots, V_{|\mathcal{B}|}\}$, where $V_i = V(b_i)$, $i = 1, 2, \dots, |\mathcal{B}|$. Assuming the pairs (b_i, V_i) , $i = 1, 2, \dots, |\mathcal{B}|$, are independent², and denoting $z_i = z(b_i)$, we have the likelihood function for $(\mathcal{V}, \mathcal{B})$,

$$p(\mathcal{V}, \mathcal{B} | \omega, c, D, \alpha) = \prod_{i=1}^{|\mathcal{B}|} \sum_{k=1}^K \omega_k p(b_i|z_i = k, c_k, D_k) p(V_i|b_i, z_i = k, \alpha_k) \quad (11)$$

where $\omega = \{\omega_k\}_{k=1}^K$, $c = \{c_k\}_{k=1}^K$, $D = \{D_k\}_{k=1}^K$, and $\alpha = \{\alpha_k\}_{k=1}^K$.

We seek the parameters ω, c, D, α that maximize the logarithmic likelihood function $\ln p(\mathcal{V}, \mathcal{B} | \omega, c, D, \alpha)$. Since $\omega_1, \dots, \omega_K$ represent a discrete probability distribution, they must be nonnegative and sum to one.

¹As the elements of b sum to one, b can be uniquely represented by its first $d-1$ elements. In this representation, c and D will be $d-1$ dimensional; α will still be d dimensional but its last element becomes an intercept. The equations in this paper are expressed in the original d dimensional belief simplex. However, they are easily converted to expressions in the $d-1$ dimensional representation.

²Once the $\mathcal{B}$ and $\mathcal{V}$ are computed, everything is restricted to the given horizon n , hence the belief points are independent of each other as are their values.

Therefore we have the constrained maximization problem,

$$\begin{aligned} \max_{\omega, c, D, \alpha} \sum_{i=1}^{|\mathcal{B}|} \ln \sum_{k=1}^K \omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k) \\ \text{subject to } \sum_{k=1}^K \omega_k = 1, \omega_k \geq 0, k = 1, \dots, K \end{aligned} \quad (12)$$

By Jensen's inequality, for any δ_i satisfying $\sum_{k=1}^K \delta_k^i = 1$, it holds

$$\begin{aligned} \ln \sum_{k=1}^K \omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k) \\ = \ln \sum_{k=1}^K \delta_k^i \frac{\omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k)}{\delta_k^i} \\ \geq \sum_{k=1}^K \delta_k^i \ln \omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k) \\ - \sum_{k=1}^K \delta_k^i \ln \delta_k^i \end{aligned} \quad (13)$$

where the inequality becomes an equality when

$$\begin{aligned} \delta_k^i &= \delta_k^i(\omega, c, D, \alpha) \\ &= \frac{\omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k)}{\sum_{k=1}^K \omega_k p(b_i | z_i = k, c_k, D_k) p(V_i | b_i, z_i = k, \alpha_k)} \end{aligned} \quad (14)$$

Let

$$\begin{aligned} Q(\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha} | \omega, c, D, \alpha) &= \sum_{i=1}^{|\mathcal{B}|} \sum_{k=1}^K \delta_k^i(\omega, c, D, \alpha) \\ &\times \ln \left[\omega_k p(b_i | z_i = k, \hat{c}_k, \hat{D}_k) p(V_i | b_i, z_i = k, \hat{\alpha}_k) \right] \end{aligned} \quad (15)$$

Iteratively finding $Q(\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha} | \omega, c, D, \alpha)$ (**E-Step**) and maximizing $Q(\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha} | \omega, c, D, \alpha)$ with respect to $\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha}$ (**M-Step**) constitute the Expectation-Maximization (EM) algorithm.

In the E-Step, we use the current estimates to compute $\delta_k^i(\omega, c, D, \alpha)$ and $Q(\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha} | \omega, c, D, \alpha)$. In the M-Step, we maximize $Q(\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha} | \omega, c, D, \alpha)$ with respect to $\{\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha}\}$, subject to the constraints in (12), the solutions of which are obtained analytically and are given in the following:

$$\omega_k = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \delta_k^i, \quad \hat{c}_k = \frac{\sum_{i=1}^{|\mathcal{B}|} \delta_k^i b_i}{\sum_{i=1}^{|\mathcal{B}|} \delta_k^i} \quad (16)$$

$$\hat{D}_k = \frac{\sum_{i=1}^{|\mathcal{B}|} \delta_k^i (b_i - \hat{c}_k)(b_i - \hat{c}_k)^T}{\sum_{i=1}^{|\mathcal{B}|} \delta_k^i} \quad (17)$$

$$\hat{\alpha}_k = \left[\sum_{i=1}^{|\mathcal{B}|} \delta_k^i b_i b_i^T \right]^{-1} \sum_{i=1}^{|\mathcal{B}|} \delta_k^i V_i b_i \quad (18)$$

4. Parameter Estimation Using Variational Bayesian EM (VB-EM)

As an alternative to standard EM, we consider the variational Bayesian EM (VBEM), which is known to be less sensitive to local maxima and singularities of the likelihood function (Corduneanu & Bishop,

2001). In contrast to EM, which maximizes the likelihood function $p(\mathcal{V}, \mathcal{B} | \omega, c, D, \alpha)$, VBEM maximizes the marginal likelihood $p(\mathcal{V}, \mathcal{B} | \omega)$, where ω is treated as model hyper-parameters. For a given set of belief points $\mathcal{B} = \{b_1, \dots, b_{|\mathcal{B}|}\}$ and the set of associated optimal values $\mathcal{V} = \{V_1, \dots, V_{|\mathcal{B}|}\}$, the marginal likelihood is obtained by integrating out the model parameters $\Theta = \{c, D, \alpha\}$ and the hidden variables $z = \{z_1, \dots, z_{|\mathcal{B}|}\}$, and is expressed in logarithm as

$$\begin{aligned} \ln p(\mathcal{V}, \mathcal{B} | \omega) &= \ln \int_{\Theta} \sum_{z_1, \dots, z_{|\mathcal{B}|}=1}^K p(\mathcal{V}, \mathcal{B}, z, \Theta) \\ &= \ln \int_{\Theta} \sum_{z_1, \dots, z_{|\mathcal{B}|}=1}^K q(z, \Theta) \frac{p(\mathcal{V}, \mathcal{B}, z, \Theta)}{q(z, \Theta)} \\ &\geq \int_{\Theta} \sum_{z_1, \dots, z_{|\mathcal{B}|}=1}^K q(z, \Theta) \ln \frac{p(\mathcal{V}, \mathcal{B}, z, \Theta)}{q(z, \Theta)} \\ &= \mathcal{L}(\mathcal{V}, \mathcal{B} | \omega) \end{aligned} \quad (19)$$

where the inequality results from Jensen's inequality. Assuming factorization $q(z, \Theta) = q(z)q(\Theta)$, we have

$$\mathcal{L}(\mathcal{V}, \mathcal{B} | \omega) = \int_{\Theta} \sum_{z_1, \dots, z_{|\mathcal{B}|}=1}^K q(z)q(\Theta) \ln \frac{p(\mathcal{V}, \mathcal{B}, z, \Theta)}{q(z)q(\Theta)} \quad (20)$$

Instead of directly maximizing the log marginal likelihood, which is difficult to evaluate, VBEM maximizes its lower bound as given by (20), alternatively solving for $q(z)$ and $q(\Theta)$, keeping one fixed while solving for the other. Taking into consideration that $\sum_{z_1, \dots, z_{|\mathcal{B}|}=1}^K q(z) = 1$ and $\int_{\Theta} q(\Theta) = 1$, these solutions can be easily obtained as

$$q(z) \propto \exp \{ \langle \ln p(\mathcal{V}, \mathcal{B}, z | \Theta) \rangle_{\Theta} \} \quad (21)$$

$$q(\Theta) \propto p(\Theta) \exp \{ \langle \ln p(\mathcal{V}, \mathcal{B}, z | \Theta) \rangle_z \} \quad (22)$$

where $\langle \cdot \rangle_{\Theta}$ denotes expectation with respect to $q(\Theta)$. Iteratively finding $q(z)$ (**VB-E Step**) and finding $q(\Theta)$ (**VB-M Step**) constitutes the VBEM algorithm.

As our likelihood function in (11) has an exponential form, we choose conjugate-exponential priors for the model parameters $\Theta = \{\omega, c, D\}$. Moreover, we assume the model parameters are independent of each other *a priori*. The variational posteriors $q(\Theta)$ can therefore be solved analytically and they all have the same form as their corresponding prior distributions. Specifically, we choose independent Normal-Wishart priors for (c_k, D_k^{-1}) , $k = 1, \dots, K$, and independent Normal priors for α_k , $k = 1, \dots, K$, i.e.,

$$p(c, D, \alpha) = \prod_{k=1}^K N(c_k) W(D_k^{-1}) N(\alpha_k) \quad (23)$$

where

$$N(c_k | D_k) = Z_{c_k} e^{-\frac{1}{2}(c_k - m_k^0)^T \beta_k^{-1,0} D_k^{-1} (c_k - m_k^0)} \quad (24)$$

$$W(D_k^{-1}) = Z_{D_k} e^{-\frac{1}{2} \text{tr}(S_k^{-1,0} D_k^{-1})} |D_k^{-1}|^{(n_k^0 - d - 1)/2} \quad (25)$$

$$N(\alpha_k) = Z_{\alpha_k} e^{-\frac{1}{2}(\alpha_k - \eta_k^0)^T \Sigma_k^{-1,0} (\alpha_k - \eta_k^0)} \quad (26)$$

and Z_{D_k} , Z_{c_k} , and Z_{α_k} are normalization constants.

VB-E Step Substituting (9) into (21) and assuming the pairs $(V_i, b_i), i = 1 \dots, |\mathcal{B}|$ are independent, we obtain $q(z) = \prod_{i=1}^{|\mathcal{B}|} q(z_i)$, where, for $k = 1, \dots, K$,

$$\begin{aligned} q(z_i = k) &\propto e^{(\ln p(V_i, b_i, z_i=k|\omega, c, D, \alpha))_{q(c, D, \alpha)}} \\ &= e^{\ln \omega_k - \frac{d}{2} \ln 2\pi + \frac{1}{2} \langle \ln |D_k^{-1}| \rangle_{D_k}} \\ &\quad \times e^{-\frac{1}{2} \text{tr} \left(\langle D_k^{-1} ((b_i - c_k)^T (b_i - c_k)) \rangle_{c_k | D_k} \rangle_{D_k} \right)} \\ &\quad \times e^{-\frac{1}{2} \ln 2\pi \mu_k^2 + \langle -\frac{1}{2\mu_k^2} (V_i - \alpha_k^T b_i)^2 \rangle_{\alpha_k}} \\ &\propto \exp \left\{ \frac{1}{2} \left[\sum_{j=1}^d \psi \left(\frac{n_k + 1 - j}{2} \right) + \ln |S_k| \right] \right. \\ &\quad \left. - \frac{1}{2\mu_k^2} \left[(V_i - \eta_k^T b_i)^2 + b_i^T \Sigma_k b_i \right] - \frac{1}{2} \ln 2\pi \mu_k^2 \right. \\ &\quad \left. - \frac{d}{2} \left[(b_i - m_k)^T n_k S_k (b_i - m_k) + \beta_k \right] \right\} \quad (27) \end{aligned}$$

As $\sum_{k=1}^K q(z_i = k) = 1$, we must normalize the right-most side of (27) to get the magnitudes of $q(z_i = k)$. For notational simplicity, we denote $\delta_k^i = q(z_i = k)$ in the remainder of the paper.

VB-M Step Substituting (9), (24), (25), (26) into (22) and assuming the pairs $(V_i, b_i), i = 1 \dots, |\mathcal{B}|$ are independent, we have

$$\begin{aligned} q(c, D, \alpha) &\propto p(c, D, \alpha) \prod_{i=1}^{|\mathcal{B}|} e^{(\ln p(V_i, b_i, z_i=k|\omega, c, D, \alpha))_{z_i=k}} \\ &\propto \prod_{k=1}^K |D_k^{-1}|^{\frac{1}{2} (n_k^0 + \sum_{i=1}^{|\mathcal{B}|} \delta_k^i - d - 1)} e^{-\frac{1}{2} \text{tr} (S_k^{-1,0} D_k^{-1})} \\ &\quad \times e^{-\frac{1}{2} \sum_{i=1}^{|\mathcal{B}|} \delta_k^i (b_i - c_k)^T D_k^{-1} (b_i - c_k)} \\ &\quad \times e^{-\frac{1}{2} \text{tr} [(c_k - m_k^0)(c_k - m_k^0)^T \beta_k^{-1,0} D_k^{-1}]} \\ &\quad \times e^{-\frac{1}{2} \sum_{i=1}^{|\mathcal{B}|} \frac{\delta_k^i}{\mu_k^2} (\alpha_k^T b_i b_i^T \alpha_k - 2\alpha_k^T b_i V_i)} \\ &\quad \times e^{-\frac{1}{2} (\alpha_k - \eta_k^0)^T \Sigma_k^{-1,0} (\alpha_k - \eta_k^0)} \quad (28) \end{aligned}$$

It is seen from (28) that the variational posterior $q(c, D, \alpha)$ has the same form as the prior $p(c, D, \alpha)$, with the hyper-parameters given by

$$\beta_k^{-1, new} = \beta_k^{-1,0} + \sum_{i=1}^{|\mathcal{B}|} \delta_k^i \quad (29)$$

$$m_k^{new} = \frac{\beta_k^{-1,0} m_k^0 + \sum_{i=1}^{|\mathcal{B}|} \delta_k^i b_i}{\beta_k^{-1, new}}, \quad n_k^{new} = n_k^0 + \sum_{i=1}^{|\mathcal{B}|} \delta_k^i \quad (30)$$

$$S_k^{-1, new} = S_k^{-1,0} + \sum_{i=1}^{|\mathcal{B}|} \delta_k^i b_i b_i^T + \beta_k^{-1,0} m_k^0 m_k^{0,T} - m_k^{new} m_k^{new,T} \beta_k^{-1, new} \quad (31)$$

$$\Sigma_k^{-1, new} = \Sigma_k^{-1,0} + \frac{\sum_{i=1}^{|\mathcal{B}|} \delta_k^i b_i b_i^T}{\mu_k^2} \quad (32)$$

$$\eta_k = (\Sigma_k^{-1, new})^{-1} (\Sigma_k^{-1,0} \eta_k^0 + \frac{\sum_{i=1}^{|\mathcal{B}|} b_i V_i \delta_k^i}{\mu_k^2}) \quad (33)$$

At the end of the VBEM algorithm, we let $\alpha_k = \eta_k$ for $k = 1 \dots, K$, and use them as the α -vectors in the POMDP optimal value function for the horizon length considered currently.

Estimation of Hyper-parameters ω Following the type-II maximum likelihood approach (Corduneanu & Bishop, 2001), we estimate the model hyper-parameters ω by maximizing the the lower bound as given by (20), which is evaluated as

$$\begin{aligned} \mathcal{L}(\mathcal{V}, \mathcal{B}|\omega) &= \int_{\Theta} \sum_{z_1 \dots z_{|\mathcal{B}|}=1}^K q(z) q(\Theta) \ln \frac{p(\mathcal{V}, \mathcal{B}, z, \Theta)}{q(z) q(\Theta)} \\ &= \sum_{z_1 \dots z_{|\mathcal{B}|}=1}^K \int_{\Theta} q(z) q(\Theta) \ln \frac{p(\Theta) \prod_{i=1}^{|\mathcal{B}|} \omega_{z_i} p(V_i, b_i | z_i, \Theta)}{q(z) q(\Theta)} \\ &= \sum_{i=1}^{|\mathcal{B}|} \sum_{z_i=1}^K \int_{c, D, \alpha} q(c, D, \alpha) q(z_i) \ln p(V_i, b_i | c, D, \alpha, z_i) \\ &\quad - KL(q(c, D, \alpha) || p(c, D, \alpha)) - \sum_{i=1}^{|\mathcal{B}|} \sum_{z_i=1}^K q(z_i) \ln q(z_i) \\ &\quad + \sum_{i=1}^{|\mathcal{B}|} \sum_{k=1}^K q(z_i = k) \ln \omega_k \quad (34) \end{aligned}$$

where the first term can be obtained in a similar way as in (27) and (28), and the KL distance are easily obtained by substituting (23) and (28) into the definitions of the KL distance. To conserve space, the details are omitted.

Given $q(c, D, \alpha)$ and $q(z)$, we maximize $\mathcal{L}(\mathcal{V}, \mathcal{B}|\omega)$ with respect to ω . Considering $\sum_{k=1}^K \omega_k = 1$, we obtain the solution as $\omega_k = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} q(z_i = k)$. As the VBEM directly maximizes the lower bound $\mathcal{L}(\mathcal{V}, \mathcal{B}|\omega)$, it is guaranteed to increase monotonically and serves as a convergence criterion.

5. Selection of Sample Belief Points $\mathcal{B}$

Let $\mathcal{B}_0$ be a set of initial belief points at time $t = 0$. For time $t = 1, 2, \dots$, let $\mathcal{B}_t = \{b' : p(o|b, a) > 0, b \in \mathcal{B}_{t-1}, a \in \mathcal{A}, o \in \mathcal{O}\}$, where b' is given by (1). Then $\cup_{t=0}^{\infty} \mathcal{B}_t$ is the set of belief points reachable by the POMDP by starting from $\mathcal{B}_0$. It is known that for most practical POMDP models, $\cup_{t=0}^{\infty} \mathcal{B}_t$ covers only a subset of the belief simplex (Pineau et al., 2003). It is therefore sufficient to only find optimal actions for the belief points in $\cup_{t=0}^{\infty} \mathcal{B}_t$, instead of all points in the belief simplex.

However, the belief points in $\cup_{t=0}^{\infty} \mathcal{B}_t$ may still be infinite. To overcome this, we follow the idea in (Pineau et al., 2003) to perform pruning for $\mathcal{B}_t$ at every t . For any given $b \in \mathcal{B}_{t-1}$, let $\mathcal{B}_{t,b} = \{b' : a \in \mathcal{A}, o \sim p(o|b, a)\}$, where b' is given by (1). Define $\rho_t(b')$ as the L_1 distance of b' from $\cup_{\tau=0}^{t-1} \mathcal{B}_\tau$, i.e., $\rho_t(b') = \min_{b'' \in \cup_{\tau=0}^{t-1} \mathcal{B}_\tau} \|b' - b''\|_1$. Let $b'_{t,b,\max} = \arg \max_{b' \in \mathcal{B}_{t,b}} \rho_t(b')$. We construct $\mathcal{B}_t = \{b'_{t,b,\max} : \rho(b'_{t,b,\max}) > \zeta, b \in \mathcal{B}_{t-1}\}$, where ζ is a threshold and is usually set to zero. That is, we prune down the points in $\mathcal{B}_t$ to retain only those that are

no closer than ζ to $\cup_{\tau=0}^{t-1} \mathcal{B}_\tau$ and that are farther away from $\cup_{\tau=0}^{t-1} \mathcal{B}_\tau$ than all other points expanded from the same point in $\mathcal{B}_{t-1}$. Since for every point in $\mathcal{B}_{t-1}$ we generate at most one point in $\mathcal{B}_t$, this pruning process produces a $\mathcal{B}_t$ that contains at most the same number of belief points as $\mathcal{B}_{t-1}$, for $t = 1, 2, \dots$. In fact, the number of points in $\mathcal{B}_t$ will begin to diminish at some t until finally at t^* we have zero points generated in $\mathcal{B}_{t^*}$, which indicates termination of the belief expansion. Clearly this can always be made to happen by using a large enough threshold ζ .

Upon termination of belief expansion at time t^* , the set of sample belief points is obtained as $\mathcal{B} = \cup_{t=0}^{t^*} \mathcal{B}_t$. From the belief pruning, it is clear that the belief points in $\mathcal{B}$ spread most uniformly among the belief points reachable by the POMDP and are therefore most representative. Since in the value iteration we will never encounter belief points unreachable by the POMDP, the $\mathcal{B}$ provides a representative sample set to the RBVI, for *any* horizon length.

Given any $b \in \mathcal{B}$, the associated optimal-value $V^n(b)$ for the length- n horizon is computed from (5), using the α^{n-1} estimated for the length- $(n-1)$ horizon. The initial α -vectors α^0 represent the terminal rewards and are determined by the reward structure of the POMDP.

6. The Complete RBVI Algorithm

We now give the complete procedures of the proposed RBVI algorithm. Since we have two methods to estimate the α -vectors and reselective model parameters, the algorithm has two versions: RBVI-EM, RBVI-VBEM. The procedures of the RBVI-EM are summarized in Table 1, where $std(\mathcal{V})$ denotes standard deviation of the elements in $\mathcal{V}$. For the RBVI-VBEM, we simply modify Steps 2.2 and 2.3, replacing the E-Step with VB-E-Step and M-Step with VB-M-Step. To execute the policy, we use the estimated α -vectors in (5) to find the optimal action.

7. Experimental Results

We demonstrate the performance of the proposed RBVI on four benchmark problems. The first three, namely, Tiger-grid, Hallway, and Hallway2, were introduced in (Littman et al., 1995) and have since been widely used to test scalable POMDP solutions. The fourth problem is the Tag problem introduced in (Pineau et al., 2003), which is relatively new and has a larger problem size. Detailed descriptions of these problems have been given by the authors who introduced them. Here we only summarize the numerical

Table 1. The Proposed RBVI Algorithm with EM as a Parameter Estimator

Input: POMDP model $(\mathcal{S}, \mathcal{A}, T, \mathcal{O}, \Omega, R)$, K , initial α -vectors α^0 , convergence criterion for EM/VBEM, convergence criterion for the value function or maximum length of horizon T ; Output: the α -vectors α^n	
1. Obtain a set of belief points $\mathcal{B} = \{b_1, b_2, \dots, b_{ \mathcal{B} }\}$, as described in Section 5; let $n = 1$.	
2. Repeat until $n = T$ or $\frac{1}{ \mathcal{B} } \sum_{i=1}^{ \mathcal{B} } V^n(b_i)$ converges in n :	
2.1. Compute $V^n(b)$, $\forall b \in \mathcal{B}$, using (5); record the values $\mathcal{V} = \{V_1, V_2, \dots, V_{ \mathcal{B} }\}$, where $V_i = V^n(b_i)$, $i = 1, 2, \dots, \mathcal{B} $.	
2.2. Initialize $\{\omega, c, D, \alpha\}$; let $\mu_k = std(\mathcal{V})$, $k = 1, 2, \dots, K$.	
2.3. Repeat until logarithm of the likelihood function in (11) converges:	
2.3.1. E-Step: compute δ_k^i using (14) along with (7) and (8).	
2.3.2. M-Step: compute the updated parameters $\{\hat{\omega}, \hat{c}, \hat{D}, \hat{\alpha}\}$ using (16), (17), (18); let $c = \hat{c}$, $D = \hat{D}$, $\omega = \hat{\omega}$, and $\alpha = \hat{\alpha}$.	
2.4. Record $\alpha^n = \alpha$; let $n = n + 1$.	

values of $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{O}|$ of these problems in Table 2.

The proposed RBVI is compared to six state-of-the-art POMDP algorithms: Grid (Brafman, 1997), PBUA (Poon, 2001), PBVI (Pineau et al., 2003), BPI (Poupart & Boutilier, 2003), Perseus (Spaan & Vlassis, 2004), HSVI (Smith & Simmons, 2004; Smith & Simmons, 2005), where BPI is based on policy iteration while the others are point-based methods for value iteration. The comparison is made in terms of performance, policy computation time, and $|\Gamma|$ the number of α -vectors upon convergence of the optimal value function. To replicate the experiments of previous authors, we measure the algorithms' performance by the discounted-accumulative-reward averaged over N_{test} independent tests. For Tiger-Grid, $N_{\text{test}} = 151$ and each test terminates after the agent takes 500 actions; during a test the agent is reset each time it reaches the goal. For Hallway and Hallway2, $N_{\text{test}} = 251$ and each test terminates when the agent reaches the goal or has taken a maximum of 251 actions. For Tag, $N_{\text{test}} = 1000$ and each test terminates when the agent successfully tags its component or has taken a maximum of 100 actions.

For the RBVI, the EM is initialized as follows: $D = \{D_k = I_d\}_{k=1}^K$ where I_d is a $d \times d$ identity matrix, $c = \{c_k\}_{k=1}^K \subset \mathcal{B}$ are K belief samples that are most separated from each other, and $\alpha = \{\alpha_k = \alpha_{\text{LS}}\}$ where α_{LS} is the least-squares (LS) estimate using $\{b, V(b)\}_{b \in \mathcal{B}}$. The prior distribution in VBEM is constructed such that the resulting means of the param-

Table 2. Results on the benchmark problems, where T denotes time in seconds, $|\Gamma|$ =number of α -vectors, n.a.=not applicable, and n.v.=not available. The results marked with * are those we obtained by coding the respective algorithms in Matlab; other results may have been coded in languages other than Matlab and executed on computer platforms different from ours.

Method	Reward	$T(s)$	$ \Gamma $
Tiger-Grid $ \mathcal{S} = 33, \mathcal{A} = 5, \mathcal{O} = 17$			
Grid(Brafman, 1997)	0.94	n.v.	174
PBUA(Poon, 2001)	2.30	12116	660
PBVI(Pineau et al., 2003)	2.25	3448	470
PBVI (*)	2.23	2239	970
BPI(Poupart & Boutilier, 2003)	2.22	1000	120
Perseus(Spaan & Vlassis, 2004)	2.34	104	134
HSVI1(Smith & Simmons, 2004)	2.35	10341	4860
HSVI2(Smith & Simmons, 2005)	2.30	52	1003
RBVI-EM (*)	1.95	135	10
RBVI-VBEM (*)	2.05	64	10
Hallway $ \mathcal{S} = 57, \mathcal{A} = 5, \mathcal{O} = 21$			
PBUA(Poon, 2001)	0.53	450	300
PBVI(Pineau et al., 2003)	0.53	288	86
PBVI (*)	0.54	1166	408
BPI(Poupart & Boutilier, 2003)	0.51	185	43
Perseus(Spaan & Vlassis, 2004)	0.51	35	55
HSVI1(Smith & Simmons, 2004)	0.52	10836	1341
HSVI2(Smith & Simmons, 2005)	0.52	2.4	147
RBVI-EM (*)	0.54	88	10
RBVI-VBEM (*)	0.54	84	10
Hallway2 $ \mathcal{S} = 89, \mathcal{A} = 5, \mathcal{O} = 17$			
PBUA(Poon, 2001)	0.35	27898	1840
PBVI(Pineau et al., 2003)	0.34	360	95
PBVI (*)	0.35	2345	572
BPI(Poupart & Boutilier, 2003)	0.32	790	60
Perseus(Spaan & Vlassis, 2004)	0.35	10	56
HSVI1(Smith & Simmons, 2004)	0.35	10010	1571
HSVI2(Smith & Simmons, 2005)	0.35	1.5	114
RBVI-EM (*)	0.30	90	15
RBVI-VBEM (*)	0.31	103	15
Tag $ \mathcal{S} = 870, \mathcal{A} = 5, \mathcal{O} = 30$			
PBVI(Pineau et al., 2003)	-9.180	180880	1334
BPI(Poupart & Boutilier, 2003)	-6.65	250	17
Perseus(Spaan & Vlassis, 2004)	-6.17	1670	280
HSVI1(Smith & Simmons, 2004)	-6.37	10113	1657
HSVI2(Smith & Simmons, 2005)	-6.36	24	415
RBVI-EM (*)	-6.56	2481	15
RBVI-VBEM (*)	-6.34	2430	15

eters equal to the corresponding initial parameters in EM.

The experimental results of the proposed RBVI algorithm are summarized in Table 2, in comparison to the other six algorithms. It is demonstrated that for each of the four problems the RBVI achieves performance competitive to the other algorithms and yet uses a significantly smaller number of α -vectors. Moreover the RBVI is also one of the best algorithms in terms of the running time spent in policy computation. Since the running time is heavily dependent on computer platforms, we will also perform a more accurate time comparison in Section 8 using the Big O notation.

To examine why such a small number of α -vectors is sufficient to represent the value function, we show in Figure 2 the model hyper-parameters $\omega = \{\omega_k\}_{k=1}^K$ estimated by VBEM. It is seen that the ω_5 in Hallway and the ω_{11} in Hallway2 are approximately zero, each indicating a spurious region and the associated α -vector. This implies that $K = 10$ for Hallway and $K = 15$ for Hallway2 is more than sufficient to represent the value function on the belief samples. The ω for Tiger-Grid does not seem to show any spurious α -vectors. This may partly explain why the performance of RBVI degrades a little on Tiger-Grid.

8. Time Comparison in Big O Notation

At a given iteration of the optimal value function, the proposed RBVI has a two-stage processing: first, compute $V(b) \forall b \in \mathcal{B}$ from (5) using α -vectors estimated in the previous iteration; second, estimate the new α -vectors for the present iteration. The time complexity of the RBVI therefore consists of two terms, corresponding to the two stages, as summarized in Table 3, where the PBVI time complexity (Pineau et al., 2003) is also shown as a comparison. In Table 3, we have shown the general case when we perform a linear transform $\tilde{b} = U^T b$ (U is computed only once) and obtain $\tilde{b}$ as a projection of b on $\text{span}(U)$, the subspace spanned by the orthonormal columns of U . Usually U is an identity matrix. However, it is possible $\dim(\tilde{b}) = \dim(U^T b) \ll \dim(b) = |\mathcal{S}|$, and replacing b with $\tilde{b}$ can result in great reduction in the EM (VBEM) computations. This happens, for example, when the belief samples in $\mathcal{B}$ are sparse (which is the case in the Tag problem),

The most intensive EM (VBEM) computations are performed in equations (8), (17), and (18) for EM, and (27), (31), and (32) for VBEM. These equations all involve the belief samples in $\mathcal{B}$ which are fixed throughout the RBVI iterations, hence many computations can be computed beforehand only once. By avoiding redundant computations in these equations, it is possible to further bring down the EM/VBEM computational complexity. The time complexity presented in Table 3 represents the case when we perform straightforward computation of these equations.

Since PBVI backs up an α -vector $\forall b \in \mathcal{B}$, $|\Gamma_{\text{PBVI}}| \approx |\mathcal{B}|$, whereas for RBVI $|\Gamma_{\text{RBVI}}| \ll |\mathcal{B}|$. Therefore, the first stage of the RBVI is always computed more efficiently than PBVI, regardless of $|\mathcal{S}|$, $|\mathcal{A}|$, and $|\mathcal{O}|$. The time complexity of the second stage in RBVI has a squared dependence on $\dim(\tilde{b}) = \dim(U^T b)$ and it is independent of $|\mathcal{A}|$ and $|\mathcal{O}|$. The EM and VBEM estimators in RBVI typically converges very fast, resulting

 Figure 2. The $\omega_k = p(z=k)$ estimated by VBEM.

 Table 3. Time Comparison in Big O Notation, where $\dim(\tilde{b})$ denotes the dimensionality of $\tilde{b} = U^T b$ and U is a matrix whose orthonormal columns span the space of $\tilde{b}$.

	Computation involving Eqn. (5)	EM or VBEM estimator (Υ_{EM} or Υ_{VBEM})
PBVI	$O(\mathcal{B} \mathcal{S} \Gamma_{PBVI} \mathcal{A} \mathcal{O})$	Not applicable
Proposed RBVI	$O(\mathcal{B} \mathcal{S} \Gamma_{RBVI} \mathcal{A} \mathcal{O})$	No worse than $O(\mathcal{B} [\dim(\tilde{b})]^2 \Gamma_{RBVI} N_T)$

in a small N_T . Therefore, the second stage of RBVI can be computed more efficiently than PBVI when: either (i) $\frac{|\mathcal{S}|}{|\mathcal{A}||\mathcal{O}|}$ is not too large, as in Tiger-Grid, Hallway, and Hallway2; or (ii) $\dim(\tilde{b}) = \dim(U^T b) \ll \dim(b) = |\mathcal{S}|$ as in Tag where $\dim(\tilde{b}) = 176$.

9. Conclusions

We have proposed an approximate region-based value iteration (RBVI) algorithm for solving the POMDP policy. By approximating the true polyhedral piecewise linearity of the optimal value function with an ellipsoidal piecewise linearity, we are able to find a small number of region-based α -vectors efficiently. The compact representation of the value function brings significant computational savings, particularly when the number of states is comparable to the product of the number of actions and the number of observations, or when the belief samples exhibit some special structures like sparseness. The results on benchmark problems show the the proposed RBVI can produce performance that is competitive to state-of-the-art POMDP algorithms, by using a significantly smaller number of α -vectors. The RBVI time complexity shown in the paper is based on straightforward computation of the related equations. We are currently investigating fast algorithms to avoid redundant computations, in hope of further reducing the complexity order of the RBVI.

References

- Brafman, R. I. (1997). A heuristic variable grid solution method for pomdps. *AAAI* (pp. 727–733).
- Cheng, H.-T. (1988). *Algorithms for partially observable Markov decision processes*. Doctoral dissertation, University of British Columbia, Vancouver, BC.
- Corduneanu, A., & Bishop, C. (2001). Variational Bayesian model selection for mixture distributions. *AI and Statistics* (pp. 27–34).
- Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 99–134.
- Littman, M. L., Cassandra, A. R., & Kaelbling, L. P. (1995). Learning policies for partially observable environments: scaling up. *ICML* (pp. 362–370).
- Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, 39, 162–175.
- Pineau, J., Gordon, G., & Thrun, S. (2003). Point-based value iteration: An anytime algorithm for pomdps. *IJ-CAI* (pp. 1025 – 1032).
- Poon, K.-M. (2001). A fast heuristic algorithm for decision-theoretic planning. Master’s thesis, The Hong Kong University of Science and Technology.
- Poupart, P., & Boutilier, C. (2003). Bounded finite state controllers. *NIPS*.
- Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operational Research*, 21, 1071–1088.
- Smith, T., & Simmons, R. (2004). Heuristic search value iteration for pomdps. .
- Smith, T., & Simmons, R. (2005). Point-based POMDP algorithms: Improved analysis and implementation. .
- Sondik, E. J. (1971). *The optimal control of partially observable Markov processes*. Doctoral dissertation, Stanford University.
- Spaan, M., & Vlassis, N. (2004). A point-based POMDP algorithm for robot planning. *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (pp. 2399–2404).