An ϵ-Optimal Grid-Based Algorithm for Partially Observable Markov Decision Processes

Blai Bonet BONET@CS.UCLA.EDU
Cognitive Systems Lab., Dept. of Computer Science, University of California, Los Angeles, CA 90024 USA

Abstract

We present an ϵ-optimal grid based algorithm for POMDPs that is tractable in ϵ^{-1}, the discount factor and the maximum absolute value of the cost function, but exponential in the dimension of the state space. To the best of our knowledge, this is the first optimal grid-based algorithm for POMDPs: all other optimal algorithms that we know are based on Sondik's representation of the Value Function. We also propose a robustness criterion for grid-based algorithms and show that the new algorithm is robust in such sense.

1. Introduction

The theory of Markov Decision Processes (MDPs) is a mathematical framework for modeling sequential decision tasks that had become very popular in AI for three important reasons. First, it provides an clean framework for modeling complex real-life problems that have large state-space (even infinite) and complex dynamics and cost functions. Second, MDPs provide mathematical foundation for independently-developed learning algorithms in Reinforcement Learning (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996). And third, general and efficient algorithms for solving MDPs had been developed, the most important being Value Iteration and Policy Iteration.

The MDP model assumes the existence of a physical system that evolves in discrete time and that is controlled by an agent. The system dynamics are governed by probabilistic transition functions that maps states and controls to states. At every time point, the agent applies a control and incurs a cost that depends in the current state of the system and the control. Thus, the task is to find a control strategy (also known as policy) that minimizes the expected total cost over the infinite horizon time setting.

A Partially Observable Markov Decision Process (POMDP) is an MDP in which the agent does not know the state of the system. This is an important departure from the MDP model since even if the agent knows the optimal strategy for the underlying MDP, it cannot apply such strategy. Thus, the agent needs to estimate the state of the system and then act accordingly. The POMDP problem is to find an optimal control strategy that map estimates to controls. It is known that estimates of the form of *probability distributions* over the set of possible states are sufficient for optimal behavior. These probability distributions, also known as *belief states*, allow the agent to compute the probability of the system being at any given state. The POMDP framework also extends the MDPs by allowing controls to return information about the system; for example, performing a blood test over a patient, or reading a radar sensor. Such information is used to compute new belief states from previous ones. Therefore, a solution for a POMDP is a strategy that maps belief states to controls. Unfortunately, state-of-the-art optimal algorithms for POMDPs are not as advanced as for MDPs in the sense that they only solve very small problems.

In this paper, we present a new optimal algorithm for POMDPs that might solve larger problems than the current best optimal algorithms. The new algorithm belongs to the class of grid-based algorithms that had been used to solve large POMDP problems. This is a relevant contribution since, as far as we know, all known grid-based algorithms do not offer optimality guarantees. We also propose a novel robustness criterion for optimal grid-based algorithms and show that the new algorithm is robust in that sense. The main ideas in the paper are general enough so that they can be applied to "transform" other grid-based algorithms into optimal and robust algorithms. Therefore, part of the contribution is to lay down mathematical foundations for optimal grid-based algorithms for POMDPs.

We organize the paper as follows. In Sect. 2, we give a formal definition for MDPs and POMDPs, give an overview of the current algorithms for POMDPs and present the robustness criterion. In Sect. 3, we show basic mathematical results about POMDPs that are used to derive the algorithm. The new algorithm is presented in Sect. 4 together with its complexity and

optimality guarantees. The paper finishes with a brief discussion in Sect. 5. All proofs are included in the Appendix.

2. Preliminaries

This section contains a brief review of the MDP and POMDP framework. We use notation and presentation style as in (Bertsekas, 1995); the reader is referred there for an excellent exposition of MDPs.

The basic MDP model is characterized by

(M1) A finite state space $S = \{1, \dots, n\}$,

(M2) a finite set of controls $U(i)$ for each state $i \in S$,

(M3) transition probabilities $p_{i,u}(j)$ for all $u \in U(i)$ that are equal to the probability of the next state being j after applying control u in state i, and

(M4) a cost $g(i, u)$ associated to $u \in U(i)$ and $i \in S$.

A strategy or policy π is an infinite sequence $(\mu_0, \mu_1, \dots)$ of functions where μ_k maps states to controls so that the agent applies the control $\mu_k(i)$ in state $x_k = i$ at time k, the only restriction being that $\mu_k(i) \in U(i)$ for all $i \in S$. If $\pi = (\mu, \mu, \dots)$, the policy is called *stationary* (i.e. the control does not depend in time) and is simply denoted by μ. The cost associated to policy π when the system starts at x_0 is:

$$J_\pi(x_0) = \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k))\right\} \quad (1)$$

where the expectation is taken with respect to the probability distribution induced by the transition probabilities, and where the number $\alpha \in [0, 1]$, called the *discount factor*, is used to discount future costs at a geometric rate.

The MDP *problem* is to find an *optimal policy* π^* satisfying $J^*(i) \stackrel{\text{def}}{=} J_{\pi^*}(i) \le J_\pi(i)$ $(i = 1, \dots, n)$, for every other policy π. Although there could be none or more than one optimal policy, the optimal cost vector J^* is always unique. The existence of π^* and how to compute it are non-trivial mathematical problems. However, when $\alpha < 1$ the optimal policy always exists and, more important, there exists a stationary policy that is optimal. In such case, J^* is the unique solution to the *Bellman Optimality* equations:

$$J^*(i) = \min_{u\in U(i)} g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J^*(j). \quad (2)$$

Also, if J^* is a solution for (2) then the *greedy* stationary policy μ^* with respect to J^*:

$$\mu^*(i) = \operatorname*{argmin}_{u\in U(i)} \left\{ g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J^*(j) \right\} \quad (3)$$

is an optimal stationary policy for the MDP. Therefore, solving the MDP problem is equivalent to solving (2).

The equation (2) can be solved by considering the Dynamic Programming (DP) operators:

$$(T_\mu J)(i) = g(i, \mu(i)) + \alpha \sum_{j=1}^{n} p_{i,\mu(i)}(j) J(j), \quad (4)$$

$$(TJ)(i) = \min_{u\in U(i)} g(i, u) + \alpha \sum_{j=1}^{n} p_{i,u}(j) J(j) \quad (5)$$

that map n-dimensional vectors to n-dimensional vectors. It is not hard to show that when $\alpha < 1$ the operators T_μ and T are contraction mappings with unique fix points J_μ and J^* satisfying:

$$J_\mu = T_\mu J_\mu = \lim_{k\to\infty} T_\mu^k J_0, \quad (6)$$

$$J^* = TJ^* = \lim_{k\to\infty} T^k J_0 \quad (7)$$

where J_0 is the zero n-dimensional vector. The Value Iteration algorithm computes J^* iteratively by using (2) as an update rule. Thus, starting from any vector J, Value Iteration computes a succession of vectors $\langle J_k \rangle_k$ as $J_0 = J$ and $J_{k+1} = TJ_k$. The algorithm stops when $J_{k+1} = J_k$, or when the residual $\max_{i\in S} |J_{k+1}(i) - J_k(i)|$ is sufficiently small. In the latter case, the suboptimality of the resulting policy is bounded by a constant multiplied by the residual.

2.1 Partially Observable MDPs

The POMDP framework, first studied in the Operations Research community, had attracted a lot of interest from AI. A good introduction to POMDPs can be found in (Astrom, 1965; Sondik, 1971; Lovejoy, 1991b; Cassandra et al., 1994). A POMDP is characterized by:

(P1) A finite state space $S = \{1, \dots, n\}$,

(P2) a finite set of controls $U(i)$ for each state $i \in S$,

(P3) transition probabilities $p_{i,u}(j)$ for all $u \in U(i)$ equal to the probability of the next state being j after applying u in i,

(P4) a finite set of observations $O(i, u) \subseteq O$ that may result after applying $u \in U(i)$ in $i \in S$,

(P5) observation probabilities $p_{i,u}(o)$ for all $u \in U(i)$ and $o \in O(i, u)$ equal to the probability of receiving o in i after applying u, and

(P6) a cost $g(i, u)$ associated to $u \in U(i)$ and $i \in S$.

It can be shown that finding an optimal strategy to this problem is equivalent to solving an associated MDP problem in *belief space*, the so-called belief-MDP. Its basic elements are:

(B1) A belief space B of prob. distributions over S,

(B2) a set of controls $U(x) = \cap\{U(i) : x(i) > 0\}$, and

(B3) a cost $g(x,u) = \sum_{i=1}^{n} g(i,u)\, x(i)$ for each $u \in U(x)$ and $x \in B$.

Definition (B2) is motivated from work in the planning community in which controls are defined in terms of *preconditions* and *effects*. Thus, $U(x)$ is the set of all controls whose preconditions are satisfied with probability 1. The use of preconditions allows the engineer to easily specify real-life problems at different levels of granularity. For example, if connecting a 120 Volts device to a 240 Volts outlet has effects that the engineer doesn't want to model, then such situations can be avoided with a simple precondition.

The transition probabilities of the belief-MDP are determined by the abilities of the agent. A full capable and rational agent *should* perform Bayesian updating in order to behave optimally. However, in the most general case, the agent might not be able to do that.[1] Thus, we will assume that after applying a control u in belief state x, the agent next belief is in a set $A \subseteq B$ with probability $\nu_u(x, A)$. Here, $\nu_u(x, \cdot)$ is a probability measure over the space of belief states called the *transition measure* associated with control u and belief x. The DP operators for the belief-MDP are:

$$(T_\mu J)(x) = g(x, \mu(x)) + \alpha \int J(z)\, \nu_{\mu(x)}(x, dz), \quad (8)$$

$$(TJ)(x) = \min_{u \in U(x)} \; g(x,u) + \alpha \int J(z)\, \nu_u(x, dz) \quad (9)$$

where μ is a stationary policy in belief space, $J : B \to \mathbb{R}$ is a real function over B, and $\alpha \in [0,1]$ is the discount factor.[2] As before, when $\alpha < 1$ the DP operators are contraction mappings with unique fix points. This fact guarantees the existence of optimal policies and that there is an optimal policy that is stationary.

From now on, we assume that the agent performs Bayesian updating whenever it applies a control and receives an observation. In this case, the transition measures are discrete measures defined by[3]

$$\nu_u(x, \{z\}) = \sum_{o \in O(x,u)} p_{x,u}(o)\, \mathbf{1}_{\{x_u^o\}}(z) \quad (10)$$

where $O(x,u)$ is the set of possible observations after applying control u in belief state x, $p_{x,u}(o)$ is the probability of receiving observation o after applying u in x, and x_u^o is the Bayesian update of x after u and o; i.e.,

$$x_u^o(i) = \frac{x_u(i)\, p_{i,u}(o)}{p_{x,u}(o)}, \quad (11)$$

$$x_u(i) = \sum_{j=1}^{n} x(j)\, p_{j,u}(i), \quad (12)$$

$$p_{x,u}(o) = \sum_{i=1}^{n} x_u(i)\, p_{i,u}(o), \quad (13)$$

$$O(x,u) = \{o : p_{x,u}(o) > 0\}. \quad (14)$$

The DP operators associated with Eq.(10) are:

$$(T_\mu J)(x) = g(x, \mu(x)) + \alpha \sum_{o \in O(x,\mu(x))} p_{x,\mu(x)}(o) J(x^o_{\mu(x)})$$

$$(TJ)(x) = \min_{u \in U(x)} \; g(x,u) + \alpha \sum_{o \in O(x,u)} p_{x,u}(o) J(x_u^o).$$

Unfortunately, the Value Iteration method is no longer feasible since each DP update has to be over an uncountable number of belief states (similarly for Policy Iteration). Thus, the question of how to compute the optimal stationary policy, or an approximation to it, is a major problem in the field.

Algorithms based on Sondik's Representation

These algorithms are based on the fact that $T^k J_0$ can be represented as

$$(T^k J_0)(x) = \min_{\gamma \in \Gamma_k} \sum_{i=1}^{n} x(i)\, \gamma(i) \quad (15)$$

where Γ_k is a *finite* collection of n-dimensional vectors. This result is known as Sondik's piecewise linear and convex representation of the Value Function (Sondik, 1971). Sondik's algorithm works in stages by computing Γ_{k+1} from Γ_k and stopping when k is sufficiently large to guarantee a given bound. Unfortunately, Γ_k grows in size double exponentially in k and, although different techniques had been proposed to remove redundant vectors from Γ_k, the worst-case growth is always exponential. Therefore, even with fixed dimension $|S|$, all known ϵ-optimal algorithms that work with Sondik's representation are exponential. See (Smallwood & Sondik, 1973; Littman, 1996; Cassandra et al., 1997; Zhang & Liu, 1997; Zhang & Lee, 1997; Cassandra, 1998; Kaelbling et al., 1999).

[1] For example, some real-world approaches to robotics do approximate Bayesian updating by means of different sampling techniques as Monte-Carlo and Gibbs sampling (Thrun, 2000).

[2] To be mathematically correct, $\nu_u(\cdot,\cdot)$ must satisfy two conditions:

(*i*) $\nu_u(x, \cdot)$ is a measure for all $x \in B$, and

(*ii*) $\nu_u(\cdot, A)$ is a measurable function for all measurable $A \subseteq B$.

The first condition is required by the definition while the second is a technical one that guarantees all mathematical objects are well-defined (Fristedt & Gray, 1997, Ch.26).

[3] $\mathbf{1}_A$ refers to the indicator function of the set A.

Grid-Based Algorithms

A *grid* G over the belief space is a *finite* collection of points together with a projective map $\eta : B \to G$. A *grid-based approximation* to J^* is a real-valued vector $\tilde{J} : G \to \mathbb{R}$ so that $J^*(x)$ is approximated by $\tilde{J}(\eta(x))$. A *grid-based algorithm* is an algorithm that from input G outputs an approximation $\tilde{J}$.[4]

Several grid-based algorithms had been proposed for finding approximate solutions to POMDPs, e.g. (Hauskrecht, 2000; Bonet & Geffner, 1998; Bonet & Geffner, 2000; Brafman, 1997; Lovejoy, 1991a). Although some grid-based algorithms had shown impressive performance over benchmark problems, none of them guarantees a bound on the quality of the result. That is, a bound on

$$\|J^* - \tilde{J} \circ \eta\| \stackrel{\text{def}}{=} \sup_{x \in B} |J^*(x) - \tilde{J}(\eta(x))|. \qquad (16)$$

Quite often, the map η is defined in terms of a distance function (metric) σ in belief space such that $\eta(x)$ is the nearest grid-point to x (under σ). In such case, we say that G is a *topological* grid and we define the *mesh* of the grid as the maximum separation between a grid-point and its surrogated points, i.e.

$$\sup_{x \in G} \sup \{\sigma(x, y) : y \in \eta^{-1}(x)\}. \qquad (17)$$

We conclude this section with the definition of a robustness criterion for algorithms based on topological grids. We say that a grid-based method is *robust in the strong sense* if, independently of the position of the grid-points, the approximation error goes to zero as the mesh goes to zero:

Definition 1 (Strong Robustness) *Let G be a topological grid on B, $\eta : B \to G$ the projection induced by G, and $\tilde{J} : G \to \mathbb{R}$ a grid-based approximation to J^*. We say that $\tilde{J}$ is* robust in the strong sense *if*

$$\lim_{\epsilon \searrow 0} \sup_{G} \|J^* - \tilde{J} \circ \eta\| = 0 \qquad (18)$$

where the sup is over all grids with mesh at most ϵ.

Our goal is to obtain a strong-robust grid-based algorithm for POMDPs.

3. Basic Mathematical Results

It should be clear that for achieving strong robustness, some notion of continuity for J^* is necessary: if $J^*(x)$ is approximated by $J^*(y)$, then $|J^*(x) - J^*(y)|$ should go to zero as x "approaches" y. To achieve that, we propose to replace the sup metric in belief space $\sigma(x, y) = \max_i |x(i) - y(i)|$ with the following:

Definition 2 *Let $\rho : B \times B \to \mathbb{R}_+$ be the function defined by $\rho(x, y) = |S|$ if x, y have different support (i.e. $\exists i$ such that $x(i) + y(i) > 0$ and $x(i)y(i) = 0$), and by $\rho(x, y) = \sum_{i=1}^{n} |x(i) - y(i)|$ otherwise.*

Note that $\rho(x, y) = |S|$ if and only if x and y have different support. The following result justifies the claim it is a metric and establishes properties about ρ.

Theorem 1 *ρ is a distance metric over the set of belief states. For all $x, y \in B$, $u \in U(x)$ and $o \in O(x, u)$, if $\rho(x, y) < |S|$, then*

(i) $U(x) = U(y)$ *and* $O(x, u) = O(y, u)$,
(ii) $\rho(x_u, y_u) \leq \rho(x, y)$,
(iii) $|p_{x,u}(o) - p_{y,u}(o)| \leq \sum_{i=1}^{n} p_{i,u}(o)|x_u(i) - y_u(i)|$,
(iv) $\rho(x_u^o, y_u^o) \leq \frac{2}{p \wedge q} \sum_{i=1}^{n} p_{i,u}(o)|x_u(i) - y_u(i)|$ *where* $p = p_{x,u}(o)$, $q = p_{y,u}(o)$ *and* $a \wedge b = \min\{a, b\}$.

An interesting fact about the metric ρ is that its associated topology,[5] denoted by ρ-topology, has isolated points. Indeed, the "deterministic" belief states are the isolated points. The following result justifies the use of ρ and permits us to obtain the algorithm.

Theorem 2 (Key Theorem) *Suppose $\alpha \in [0, 1)$ and $\|g\| < \infty$. Then, the optimal cost function J^* is a* uniform continuous *map from B to $\mathbb{R}$. Indeed, for integer $m \geq 1$ (given below) define $F(x) = x^{(1-\alpha)^m}$. Then, for small $\rho(x, y)$,*

$$|J^*(x) - J^*(y)| < \frac{\|g\|}{1 - \alpha} F(\rho(x, y)) \qquad (19)$$

where $\|g\| \stackrel{def}{=} \max_{i \in S} \max_{u \in U(i)} |g(i, u)|$.[6]

The integer m is chosen as the minimum positive integer so that $|O|^{(1-\alpha)^m} < \tau$ where $\tau > 1$ satisfies

$$x + \alpha\tau(2x)^{(1-\alpha)} - x^{(1-\alpha)} \leq 0 \qquad (20)$$

for all x in an interval $[0, \xi]$ (where ξ, that depends in τ, increases as τ decreases). This choice of m guarantees

$$x + \alpha\beta F(2x) \leq F(x) \qquad (21)$$

[4] Other more general definitions for grid-based algorithm had been given, see (Hauskrecht, 2000).

[5] The collection of open sets with respect to ρ.

[6] That J^* is a continuous function (with respect to ρ) is trivial since it is the limit of a uniform converging sequence of continuous functions. The fact that is uniform continuous is more interesting since the metric space (B, ρ) is not compact. The theorem goes one step further by giving the modulus of continuity.

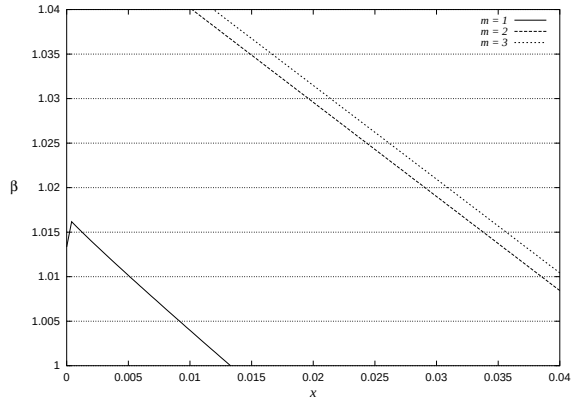

Figure 1. Contours plot of $x + \alpha\beta F(2x) - F(x)$ at level set 0 for $\alpha = 0.95$ and $m = 1, 2, 3$.

for all $x \in [0, \xi]$ and all $\beta \leq \tau$ (a fact needed in the proof of the theorem). Fig. 1 shows a contour plot of $x + \alpha\beta F(2x) - F(x)$ for different x's, β's and m's, e.g. $\tau = 1.01$ is good enough for $\alpha = 0.95$ so that (21) holds for $x \in [0, 0.005]$. Thus, we say that $\rho(x, y)$ is "small" when it is $< \xi$. This bound is loose and can be improved with a more careful analysis.

It is interesting to note that a result as the Key Theorem is not possible for the sup metric. In fact, the following example shows a POMDP such that J^* is not continuous in the σ-topology.

Example 1: Let $S = \{1, 2\}$ and consider two controls u_1, u_2 such that $U(1) = \{u_1\}$, $U(2) = \{u_1, u_2\}$, $g(1, u_1) = 1$, $g(2, u_1) = g(2, u_2) = 0$, $p_{2,u_2}(2) = 1$ and the transition probabilities for u_1 are given by two parameters p_1, p_2 as shown in Fig. 2. Also, there is just one observation that is always received with probability 1. Each belief state in this problem is of the form $(p, 1-p)$ so it can be represented by $p \in [0, 1]$. When $\alpha < 1$, the corresponding POMDP is guaranteed to have a solution, and it is easy to check that J^* becomes

$$J^*(p) = \begin{cases} p + \alpha J^*(pp_1 + (1-p)(1-p_2)) & \text{if } p > 0, \\ 0 & \text{if } p = 0. \end{cases}$$

Note that $pp_1 + (1-p)(1-p_2) = 1 + p(p_1 + p_2 - 1) - p_2$. Thus, if $p_1 + p_2 = 1$, then $J^*(1-p_2) = 1 - p_2 + \alpha J^*(1-p_2)$. Hence, if $p_1 + p_2 = 1$,

$$J^*(p) = \begin{cases} p + \alpha(1-p_2)/(1-\alpha) & \text{if } p > 0, \\ 0 & \text{if } p = 0. \end{cases}$$

Clearly, J^* is discontinuous at $p = 0$ with respect to the σ-topology. On the other hand, $p = 0$ is an isolated point in the ρ-topology, so J^* is continuous with respect to the latter. Also note that the jump can be made arbitrary large. □

Thus, in the most general case, any grid-based method based on the sup metric is not robust in the strong sense. In the case when all $U(i)$ are identical, it is easy to show that J^* is uniform continuous with respect to the σ-topology (see footnote 6), though its modulus of continuity is yet to be computed. In any case,

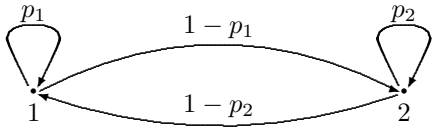

Figure 2. Transition probabilities for u_1 in Example 1.

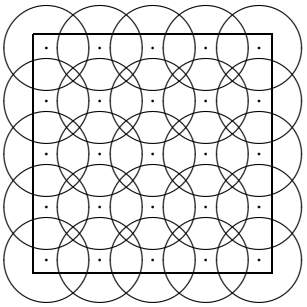

Figure 3. An ϵ-cover of the unit square.

Theorem 3 *Let σ' be a metric in belief space such that the σ'-topology is identical to the σ-topology, e.g. σ' = Euclidean distance. Then, any grid-based method such that η maps belief states into their nearest grid-points (under σ') is not robust in the strong sense for general* POMDPs.

An immediate consequence of the Key Theorem is that any "reasonable" grid-based algorithm is robust in the strong sense. In the next section, we present a very simple grid-based algorithm.

4. An Optimal Grid-Based Algorithm

An *ϵ-cover* for B is a finite collection of balls[7] $\{B_1, \ldots, B_m\}$, open or closed with respect to ρ, such that each B_i is of radius ϵ centered at some x_i, $x_i \notin B_j$ for $i \neq j$, and $B \subseteq \cup_{i=1}^m B_i$. Fig. 3 shows an example of an ϵ-cover for the unit square. It is not hard to see that an ϵ-cover always exists for every $\epsilon > 0$ and that the x_i can be chosen so that all $x_i(j)$ are rational. If, in addition, ϵ is rational, then we call the cover a *rational* ϵ-cover. Each such cover induces a projection $\eta : B \to \{x_1, \ldots, x_m\}$ by $\eta(x) = x_i$ if and only if i is

[7]The open ball with center x and radius ϵ is $\{y : \rho(x, y) < \epsilon\}$, for the closed ball replace $<$ with $\leq$.

Input: A number $\epsilon > 0$, and a finite collection $\{x_1, \ldots, x_m\}$ such that the open/closed balls of radius ϵ centered at x_i form a rational cover.

Output: A m-dimensional real vector $\tilde{J}$ such that

$$\|J^* - \tilde{J} \circ \eta\| \ < \ \tfrac{2\|g\|}{(1-\alpha)^2} F(\epsilon)$$

$$\|J^* - J_{\tilde{\mu}}\| \ < \ \tfrac{2\alpha(1+\alpha)\|g\|}{(1-\alpha)^3} F(\epsilon)$$

where $\tilde{\mu}$ is the greedy policy with respect to $\tilde{J} \circ \eta$.

Procedure:

1. **Let** $\tilde{J}_0(x_i) = 0$ for $i = 1, \ldots, m$, and set $k = 0$.
2. **Update** $\tilde{J}_{k+1} = T(\tilde{J}_k \circ \eta)$.
3. **Increase** k and **Goto** 2 if

$$k \leq (\log F(\epsilon) - 2\log(1-\alpha))/\log \alpha,$$

4. **Return** $\tilde{J}_{k+1}$ otherwise.

Figure 4. An ϵ-optimal grid-based algorithm for POMDPs.

minimum such that $x \in B_i$. By the Key Theorem,

$$|J^*(x) - J^*(\eta(x))| < \frac{\|g\|}{1-\alpha} F(\epsilon) \tag{22}$$

for sufficiently small ϵ. Therefore, our goal is to construct a "good" approximation to $J^* \circ \eta$. Consider the sequence of vectors $\tilde{J}_k : \{x_1, \ldots, x_m\} \to \mathbb{R}$:

$$\tilde{J}_0(x_i) = \ 0, \tag{23}$$

$$\tilde{J}_{k+1}(x_i) = \min_{u \in U(x_i)} g(x_i, u) + \alpha \sum_{o \in O(x_i, u)} p_{x_i,u}(o) \tilde{J}_k(\eta((x_i)_u^o)).$$

This is the Value Iteration algorithm applied to an MDP with state space $\{x_1, \ldots, x_m\}$ and transition probabilities

$$p_{x_i,u}(x_j) \ = \sum_{o : \eta((x_i)_u^o) = x_j} p_{x_i,u}(o). \tag{24}$$

It is easy to see that the associated DP operators are identical to the restriction of $T_\mu(J \circ \eta)$ and $T(J \circ \eta)$ to the set $\{x_1, \ldots, x_m\}$.

Since η^{-1} partitions the belief space into a finite number of pieces, Eq. (23) defines a grid-based algorithm for POMDPs. The following results bound the approximation error and the loss incurred by the resulting greedy policy when the Value Iteration method $\langle \tilde{J}_k \circ \eta \rangle_{k \geq 0}$ is stopped.

Theorem 4 (Goodness) *Suppose* $\alpha \in [0, 1)$ *and* $\|g\| < \infty$. *Let* $\tilde{J}_k$ *be defined by* (23), *then*

$$\|J^* - \tilde{J}_k \circ \eta\| \ < \ \frac{\|g\|}{(1-\alpha)^2} F(\epsilon) + \frac{\alpha^k \|g\|}{1-\alpha}. \tag{25}$$

Corollary 5 (Policy Loss) *Suppose* $\alpha \in [0, 1)$ *and* $\|g\| < \infty$. *Let* μ^k *be the greedy policy, in belief space, with respect to* $\tilde{J}_k \circ \eta$, *i.e.* $T_{\mu^k}(\tilde{J}_k \circ \eta) = T(\tilde{J}_k \circ \eta)$. *Then,*

$$\|J^* - J_{\mu^k}\| \ < \ \frac{2\alpha(1+\alpha)\|g\|}{(1-\alpha)^3} F(\epsilon) \tag{26}$$

for every $k \geq k_0$ *where* k_0 *is such that* $\alpha^{k_0} \leq \frac{F(\epsilon)}{1-\alpha}$.

These results show the correctness of the grid-based algorithm in Fig. 4. The algorithm is robust in the strong sense, tractable in the parameters ϵ^{-1}, α, $|O|$ and $\|g\|$, and only exponential in the dimension $|S|$. The exponentiality in $|S|$ cannot be removed since it is known that solving general POMDPs, either optimally or ϵ-optimally, is NP-HARD (Lusena et al., 2001).

5. Discussion

Since the number of grid-points grows exponentially with the dimension, the problem of how to solve the grid problem is still a research topic. In stochastic shortest-path problems, we think that the grid could be solved by using algorithms like RTDP and LAO* with "good" heuristic functions (Barto et al., 1995; Hansen & Zilberstein, 2001; Bonet & Geffner, 2000). In summary, we have presented a new grid-based optimal algorithm for general POMDPs and proposed a robustness criterion for such algorithms. Our results and methodology are general enough so that they can be applied to other algorithms, e.g. some of the recent adaptive and multi-resolution grid methods. Thus, the paper lays down mathematical foundations for new and better POMDPs algorithms. In the near future, we plan to make empirical comparisons between our algorithm and those based on Sondik's representation.

Acknowledgements

I like to thank the anonymous ICML reviewers for their comments that helped me to improve the presentation. Also, thanks to Thomas Phan that read a previous version of the paper. This work was partly supported by a scholarship from USB/CONICIT from Venezuela.

References

Astrom, K. (1965). Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Appl.*, *10*, 174–205.

Barto, A., Bradtke, S., & Singh, S. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, *72*, 81–138.

Bertsekas, D. (1995). *Dynamic programming and optimal control, (2 vols)*. Athena Scientific.

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bonet, B., & Geffner, H. (1998). Learning sorting and decision trees with POMDPs. *Proceedings of ICML-98*.

Bonet, B., & Geffner, H. (2000). Planning with incomplete information as heuristic search in belief space. *Proceedings of AIPS-2000* (pp. 52–61). AAAI Press.

Brafman, R. (1997). A heuristic variable grid solution for POMDP's. *Proceedings AAAI-97* (pp. 727–733).

Cassandra, A., Kaebling, L., & Littman, M. (1994). Acting optimally in partially observable stochastic domains. *Proceedings AAAI94* (pp. 1023–1028).

Cassandra, A., Littman, M., & Zhang, N. (1997). Incremental pruning: A simple, fast, exact algorithm for Partially Observable Markov Decision Processes. *Proceedings UAI-97* (pp. 54–61). Morgan Kaufmann.

Cassandra, A. R. (1998). *Exact and approximate algorithms for partially observable Markov decision processes*. Doctoral dissertation, Brown University.

Fristedt, B., & Gray, L. (1997). *A modern approach to probability theory*. Birkhauser.

Hansen, E., & Zilberstein, S. (2001). LAO*: A heuristic search algorithm that finds solutions with loops. *Artificial Intelligence*, *129*, 35–62.

Hauskrecht, M. (2000). Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, *13*, 33–94.

Kaelbling, L., Littman, M., & Cassandra, A. (1999). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Littman, M. L. (1996). *Algorithms for sequential decision making*. Doctoral dissertation, Brown University.

Lovejoy, W. (1991a). Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, *39*.

Lovejoy, W. (1991b). A survey of algorithmic techniques for partially observed Markov decision processes. *Annals of Operations Research*, *28*, 47–66.

Lusena, C., Mundhenk, M., & Goldsmith, J. (2001). Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, *14*, 83–103.

Smallwood, R., & Sondik, E. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, *21*, 1071–1088.

Sondik, E. (1971). *The optimal control of partially observable Markov decision processes*. Doctoral dissertation, Stanford University.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. MIT Press.

Thrun, S. (2000). Probabilistic algorithms in robotics. *AI Magazine*, *21*.

Zhang, N. L., & Lee, S. S. (1997). Planning with partially observable Markov decision processes: Advances in exact solution method. *Proceedings UAI-97* (pp. 523–530). Morgan Kaufmann.

Zhang, N. L., & Liu, W. (1997). A model approximation scheme for planning in partially observable stochastic domains. *Journal of Artificial Intelligence Research*, *7*, 199–230.

Appendix: Proofs

Proof of Theorem 1: That ρ is a metric is trivial so it is left to the reader. Let $x, y \in B$ be such that $\rho(x,y) < |S|$. Since x and y have identical support, it is obvious that $U(x) = U(y)$ and $O(x,u) = O(y,u)$. For (ii) and (iii),

$$\begin{aligned}
&\rho(x_u, y_u) \\
&\quad = \sum_{i=1}^{n} |x_u(i) - y_u(i)| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} p_{j,u}(i)\,(x(j) - y(j)) \right| \\
&\quad \leq \sum_{1 \leq i,j \leq n} p_{j,u}(i)\,|x(j) - y(j)| = \rho(x,y),
\end{aligned}$$

$$\begin{aligned}
&|p_{x,u}(o) - p_{y,u}(o)| \\
&\quad = \left| \sum_{i=1}^{n} x_u(i) p_{i,u}(o) - y_u(i) p_{i,u}(o) \right| \\
&\quad \leq \sum_{i=1}^{n} p_{i,u}(o)\,|x_u(i) - y_u(i)|.
\end{aligned}$$

For the last one, let $p = p_{x,u}(o)$ and $q = p_{y,u}(o)$ and assume, without loss of generality, $p > q$. Then,

$$\begin{aligned}
&\rho(x_u^o, y_u^o) \\
&\quad = \frac{1}{pq} \sum_{i=1}^{n} p_{i_u}(o)\,\left| q x_u(i) - p y_u(i) \right| \\
&\quad \leq \frac{1}{pq} \sum_{i=1}^{n} p_{i_u}(o)\,\left(p\,|x_u(i) - y_u(i)| + |p-q| x_u(i) \right) \\
&\quad = \frac{1}{pq} \left(p \sum_{i=1}^{n} p_{i_u}(o)\,|x_u(i) - y_u(i)| + |p-q| p \right) \\
&\quad \leq \frac{2}{p \wedge q} \sum_{i=1}^{n} p_{i,u}(o)\,|x_u(i) - y_u(i)|.
\end{aligned}$$

□

Proof of Key Theorem: Let $x, y \in B$ be such that $\rho(x,y) < |S|$ is small. Thus, $U(x) = U(y)$ and $O(x,u) = O(y,u)$. First, we show using induction that $|(T^k J_0)(x) - (T^k J_0)(y)| < \|g\| F(\rho(x,y))/(1-\alpha)$. For the base case,

$$\begin{aligned}
&|(TJ_0)(x) - (TJ_0)(y)| \\
&\qquad = \left| \left(\min_{u \in U(x)} g(x,u) + \alpha \int J_0(z) \nu_u(x, dz) \right) - \right. \\
&\qquad\qquad \left. \left(\min_{u \in U(y)} g(y,u) + \alpha \int J_0(z) \nu_u(y, dz) \right) \right| \\
&\qquad \leq |g(x,u) - g(y,u)| \;\leq\; \|g\| \rho(x,y) \\
&\qquad < \frac{\|g\|}{1-\alpha} F(\rho(x,y)) \quad (\text{since } F(x) > x \text{ for small } x)
\end{aligned}$$

where the $u \in U(x)$ in the first inequality is the control that minimizes the second term, and we have assumed, without loss of generality, that the first term is larger than the second. The inductive step is

$$
\begin{aligned}
&|(T^{k+1}J_0)(x) - (T^{k+1}J_0)(y)| \\
&\le |g(x,u) - g(y,u)| + \\
&\quad \alpha \left| \sum p_{x,u}(o)(T^k J_0)(x_u^o) - p_{y,u}(o)(T^k J_0)(y_u^o) \right| \\
&\le \|g\| \rho(x,y) + \\
&\quad \alpha \sum_o \Big[(p \wedge q) \big| (T^k J_0)(x_u^o) - (T^k J_0)(y_u^o) \big| + \|T^k J_0\| |p - q| \Big] \\
&< \|g\| \rho(x,y) \;+\; \frac{\alpha \|g\| \rho(x,y)}{1-\alpha} \;+\; \frac{\alpha \|g\|}{1-\alpha} \sum_o (p \wedge q) F(\rho(x_u^o, y_u^o)) \\
&= \frac{\|g\| \rho(x,y)}{1-\alpha} \;+\; \frac{\alpha \|g\|}{1-\alpha} \sum_o (p \wedge q)\, F(\rho(x_u^o, y_u^o)) \\
&\le \frac{\|g\| \rho(x,y)}{1-\alpha} \;+\; \frac{\alpha \|g\|}{1-\alpha} F\left(2\rho(x,y)\right) \sum_o (p \wedge q)^{1-(1-\alpha)^m} \\
&\le \frac{\|g\| \rho(x,y)}{1-\alpha} \;+\; \frac{\alpha \|g\|}{1-\alpha} F\left(2\rho(x,y)\right) |O| \left(\frac{1}{|O|} \right)^{1-(1-\alpha)^m} \\
&\le \frac{\|g\|}{1-\alpha} F(\rho(x,y)) \quad \text{(by the choice of } m\text{)}.
\end{aligned}
$$

Where $p = p_{x,u}(o)$ and $q = p_{y,u}(o)$ for some $u \in U(x)$ as in the base case. The bound on $\sum_o (p \wedge q)^{1-(1-\alpha)^m}$ comes from the fact that the expression is maximized when all $p = q = 1/|O|$. In the third inequality, we used

$$
\begin{aligned}
\sum_o \|T^k J_0\| \, |p - q| &\le \frac{\|g\|}{1-\alpha} \sum_o \sum_{i=1}^n p_{i,u}(o) |x_u(i) - y_u(i)| \\
&= \frac{\|g\|}{1-\alpha} \rho(x_u, y_u) \le \frac{\|g\|}{1-\alpha} \rho(x,y).
\end{aligned}
$$

Therefore, for small $\rho(x,y)$,

$$
\begin{aligned}
|J^*(x) - J^*(y)| &= \lim_{k \to \infty} \left| (T^k J_0)(x) - (T^k J_0)(y) \right| \\
&< \frac{\|g\|}{1-\alpha} F(\rho(x,y)).
\end{aligned}
$$

□

Proof of Theorem 4: First note that

$$
\begin{aligned}
&|J^*(x) - \tilde{J}_k(\eta(x))| \\
&\quad = |J^*(x) - J^*(\eta(x)) + J^*(\eta(x)) - \tilde{J}_k(\eta(x))| \\
&\quad < \frac{\|g\|}{1-\alpha} F(\epsilon) + |J^*(\eta(x)) - \tilde{J}_k(\eta(x))|.
\end{aligned}
$$

Now, use induction on k to find a bound on $|J^*(x) - \tilde{J}_k(x)|$ for $x \in \{x_1, \ldots, x_m\}$ as follows:

$$
\begin{aligned}
|J^*(x) - \tilde{J}_1(x)| &\le \alpha \sum_o p_{x,u}(o) \left| J^*(x_u^o) - \tilde{J}_0(\eta(x_u^o)) \right| \\
&\le \alpha \|J^*\|, \\
|J^*(x) - \tilde{J}_2(x)| &\le \alpha \sum_o p_{x,u}(o) \left| J^*(x_u^o) - \tilde{J}_1(\eta(x_u^o)) \right| \\
&< \alpha \sum_o p_{x,u}(o) \left(\frac{\|g\| F(\epsilon)}{1-\alpha} + \alpha \|J^*\| \right) \\
&= \frac{\|g\| F(\epsilon)}{1-\alpha} \alpha + \alpha^2 \|J^*\|
\end{aligned}
$$

where u is some control belonging to $U(x)$ (check the proof of Key Theorem to see where u comes from). Then,

$$
\begin{aligned}
|J^*(x) - \tilde{J}_{k+1}(x)| &\le \alpha \sum_o p_{x,u}(o) \left| J^*(x_u^o) - \tilde{J}_k(\eta(x_u^o)) \right| \\
&< \alpha \left(\frac{\|g\| F(\epsilon)}{1-\alpha} + \frac{\|g\| F(\epsilon)}{1-\alpha} \sum_{j=1}^{k-1} \alpha^j + \alpha^k \|J^*\| \right) \\
&= \frac{\|g\| F(\epsilon)}{1-\alpha} \sum_{j=1}^{k} \alpha^j + \alpha^{k+1} \|J^*\|.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sup_{x \in B} |J^*(x) - \tilde{J}_k(\eta(x))| \\
&\qquad < \frac{\|g\| F(\epsilon)}{1-\alpha} + \frac{\|g\| F(\epsilon)}{1-\alpha} \sum_{j=1}^{k-1} \alpha^j + \alpha^k \|J^*\| \\
&\qquad \le \frac{\|g\| F(\epsilon)}{(1-\alpha)^2} + \frac{\alpha^k \|g\|}{1-\alpha}.
\end{aligned}
$$

□

To bound the loss incurred by the greedy policy with respect to $\tilde{J}_k \circ \eta$, we use a known result for MDPs. Consider an MDP (with finite or infinite state space) and its DP operators T and T_μ. Let J^* be the unique fix point of T, J a real cost function over the state space, and μ a *greedy* policy with respect to J (i.e., $T_\mu J = TJ$), then the following bound on the suboptimality of μ holds (see (Bertsekas, 1995, Vol 2, pp. 19–24))

$$
\begin{aligned}
&\sup_x J_\mu(x) - J^*(x) \le \\
&\quad \frac{\alpha}{1-\alpha} \left(\sup_x |(TJ)(x) - J(x)| - \inf_x |(TJ)(x) - J(x)| \right)
\end{aligned} \tag{27}
$$

where J_μ is the unique fix point of T_μ, and the sup and inf are taken over the corresponding state space.

Proof of Corollary 5:

$$
\begin{aligned}
&|(T(\tilde{J}_k \circ \eta))(x) - \tilde{J}_k(\eta(x))| \\
&\quad \le |(T(\tilde{J}_k \circ \eta))(x) - (TJ^*)(x)| + |J^*(x) - \tilde{J}_k(\eta(x))| \\
&\quad \le \alpha \sum_o p_{x,u}(o) \left| \tilde{J}_k(\eta(x_u^o)) - J^*(x_u^o) \right| + |J^*(x) - \tilde{J}_k(\eta(x))| \\
&\quad \le (1+\alpha) \left\| J^* - \tilde{J}_k \circ \eta \right\| \\
&\quad < (1+\alpha) \left(\frac{\|g\| F(\epsilon)}{(1-\alpha)^2} + \frac{\alpha^k \|g\|}{1-\alpha} \right) \le (1+\alpha) \frac{2\|g\| F(\epsilon)}{(1-\alpha)^2}
\end{aligned}
$$

for all $k \ge k_0$. Now, use the bound (27) on the suboptimality of μ^k:

$$
\begin{aligned}
J_{\mu^k}(x) - J^*(x) &\le \frac{\alpha}{1-\alpha} \left(\sup_{x \in B} |(T(\tilde{J}_k \circ \eta))(x) - \tilde{J}_k(\eta(x))| \right) \\
&< \frac{\alpha(1+\alpha)}{1-\alpha} \frac{2\|g\| F(\epsilon)}{(1-\alpha)^2}
\end{aligned}
$$

for all $k \ge k_0$. □