DESPOT: Online POMDP Planning with Regularization

Adhiraj Somani  Nan Ye  David Hsu  Wee Sun Lee
Department of Computer Science
National University of Singapore
adhirajsomani@gmail.com, \{yenan,dyhsu,leews\}@comp.nus.edu.sg

Abstract

POMDPs provide a principled framework for planning under uncertainty, but are computationally intractable, due to the “curse of dimensionality” and the “curse of history”. This paper presents an online search algorithm that alleviates these difficulties by focusing on a set of sampled scenarios. The execution of all policies on the sampled scenarios is captured in a Determinized Sparse Partially Observable Tree (DESPOT). The Regularized DESPOT (R-DESPOT) algorithm searches the DESPOT for a policy, while optimally balancing the size of the policy and the accuracy of its value estimate obtained under the sampled scenarios. We give an output-sensitive performance bound for all policies derived from a DESPOT and show that R-DESPOT works well if a small optimal policy exists. We also give an anytime algorithm that approximates R-DESPOT. Experiments show strong results, compared with two of the fastest online POMDP algorithms. Source code along with experimental settings are available at http://bigbird.comp.nus.edu.sg/pmwiki/farm/appl/.

1 Introduction

Partially observable Markov decision processes (POMDPs) provide a principled general framework for planning in partially observable stochastic environments. However, POMDP planning is computationally intractable in the worst case [1]. The challenges arise from three main sources. First, a POMDP may have a large number of states. Second, as the state is not fully observable, the agent must reason with beliefs, which are probability distributions over the states. Roughly, the size of the belief space grows exponentially with the number of states. Finally, the number of action-observation histories that must be considered for POMDP planning grows exponentially with the planning horizon. The first two difficulties are usually referred to as the “curse of dimensionality”, and the last one, the “curse of history”. To address these difficulties, online POMDP planning (see [2] for a survey) chooses one action at a time and interleaves planning and plan execution. At each time step, the agent performs a $D$-step lookahead search. It plans the immediate next action for the current belief only and reasons in the neighborhood of the current belief, rather than over the entire belief space. Our work adopts this online planning approach.

Recently an online POMDP planning algorithm called POMCP has successfully scaled up to very large POMDPs [3]. POMCP, which is based on Monte Carlo tree search, tries to break the two curses by sampling states from the current belief and sampling histories with a black-box simulator. It uses the UCT algorithm [4] to control the exploration-exploitation trade-off during the lookahead search. However, UCT is sometimes overly greedy and suffers the worst-case performance of $\Omega\left(\exp\left(\exp\left(\ldots\exp(1)\ldots\right)\right)\right)$ samples to find a sufficiently good action [5].

This paper presents a new algorithm for online POMDP planning. It enjoys the same strengths as POMCP – breaking the two curses through sampling – but avoids POMCP’s extremely poor worst-case behavior by evaluating policies on a small number of sampled scenarios [6]. In each planning step, the algorithm searches for a good policy derived from a Determinized Sparse Partially Observable Tree (DESPOT) for the current belief, and executes the policy for one step. A DESPOT summarizes the execution of all policies under $K$ sampled scenarios. It is structurally similar to a standard belief tree, but contains only belief nodes reachable under the $K$ scenarios.

\(^1\)Composition of $D − 1$ exponential functions.
A belief tree of height $D = 2$ (gray) and a corresponding DESPOT obtained with 2 sampled scenarios. Every tree node represents a belief. Every colored dot represents a scenario.

We can view a DESPOT as a sparsely sampled belief tree. While a belief tree of height $D$ contains $O(|A|^D |Z|^D)$ nodes, where $|A|$ and $|Z|$ are the sizes of the action set and the observation set, respectively, a corresponding DESPOT contains only $O(|A|^D K)$ nodes, leading to dramatic improvement in computational efficiency when $K$ is small.

One main result of this work is an output-sensitive bound, showing that a small number of sampled scenarios is sufficient to give a good estimate of the true value of any policy $\pi$, provided that the size of $\pi$ is small (Section 3). Our Regularized DESPOT (R-DESPOT) algorithm interprets this lower bound as a regularized utility function, which it uses to optimally balance the size of a policy and its estimated performance under the sampled scenarios. We show that R-DESPOT computes a near-optimal policy whenever a small optimal policy exists (Section 4). For anytime online planning, we give a heuristic approximation, Anytime Regularized DESPOT (AR-DESPOT), to the R-DESPOT algorithm (Section 5). Experiments show strong results of AR-DESPOT, compared with two of the fastest online POMDP algorithms (Section 6).

## 2 Related Work

There are two main approaches to POMDP planning: offline policy computation and online search. In offline planning, the agent computes beforehand a policy contingent upon all possible future scenarios and executes the computed policy based on the observations received. Although offline planning algorithms have achieved dramatic progress in computing near-optimal policies (e.g., [7, 8, 9, 10]), they are difficult to scale up to very large POMDPs, because of the exponential number of future scenarios that must be considered.

In contrast, online planning interleaves planning and plan execution. The agent searches for a single best action for the current belief only, executes the action, and updates the belief. The process then repeats at the new belief. A recent survey lists three main categories of online planning algorithms: heuristic search, branch-and-bound pruning, and Monte Carlo sampling [2]. AR-DESPOT contains elements of all three, and the idea of constructing DESPOTs through deterministic sampling is related to those in [11, 6]. However, AR-DESPOT balances the size of a policy and its estimated performance during the online search, resulting in improved performance for suitable planning tasks.

During the online search, most algorithms, including those based on Monte Carlo sampling (e.g., [12, 13]), maintain an explicit representation of the belief as a probability distribution over the state space. This, however, limits their scalability for POMDPs with many states, as a single belief update can take time quadratic in the number of states. In contrast, DESPOT algorithms represent the belief as a set of particles, just as POMCP [3] does, and do not perform belief update during the online search.

Online search and offline policy computation are complementary and can be combined, e.g., by using approximate or partial policies computed offline as the default policies at the bottom of the search tree for online planning (e.g., [14, 15]) or as macro-actions to shorten the search horizon [16].

## 3 Determinized Sparse Partially Observable Trees

### 3.1 POMDP Preliminaries

A POMDP is formally a tuple $(S, A, Z, T, O, R)$, where $S$ is a set of states, $A$ is a set of actions, $Z$ is a set of observations, $T(s, a, s′) = p(s′|s, a)$ is the probability of transitioning to state $s′$ when the agent takes action $a$ in state $s$, $O(s, a, z) = p(z|s, a)$ is the probability of observing $z$ if the agent takes action $a$ and ends in state $s$, and $R(s, a)$ is the immediate reward for taking action $a$ in state $s$.

A POMDP agent does not know the true state, but receives observations that provide partial information on the state. The agent maintains a belief, often represented as a probability distribution over $S$. It starts with an initial belief $b_0$. At time $t$, it updates the belief $b_t$ according to Bayes’ rule by incorporating information from the action taken at time $t − 1$ and the resulting observation: $b_t = \pi(b_{t-1}, a_{t-1}, z_t)$. A policy $\pi: \mathcal{B} \mapsto A$ is a mapping that specifies the action $a \in A$ at belief $b \in \mathcal{B}$. The value of a policy $\pi$ at a belief $b$ is the expected total discounted reward obtained by

\[
\mathbb{E}_t \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] = \sum_{s \in S} \sum_{a \in A} \sum_{z \in Z} \sum_{s′ \in S} T(s, a, s′) O(s, a, z) R(s, a) \pi(b)
\]
following $\pi$ with initial belief $b$: $V_\pi(b) = E\left(\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(b_t)) \mid b_0 = b\right)$, for some discount factor $\gamma \in [0, 1)$.

One way of online POMDP planning is to construct a belief tree (Figure 1), with the current belief $b_0$ as the initial belief at the root of the tree, and perform lookahead search on the tree for a policy $\pi$ that maximizes $V_\pi(b_0)$. Each node of the tree represents a belief. A node branches into $|A|$ action edges, and each action edge branches further into $|Z|$ observation edges. If a node and its child represent beliefs $b$ and $b'$, respectively, then $b' = \tau(b, a, z)$ for some $a \in A$ and $z \in Z$. To search a belief tree, we typically truncate it at a maximum depth $D$ and perform a post-order traversal. At each leaf node, we simulate a default policy to obtain a lower bound on its value. At each internal node, we apply Bellman's principle of optimality to choose a best action:

$$V(b) = \max_{a \in A} \left\{ \sum_{s \in S} b(s)R(s, a) + \gamma \sum_{z \in Z} p(z|b, a)V(\tau(b, a, z)) \right\},$$

which recursively computes the maximum value of action branches and the average value of observation branches. The results are an approximately optimal value $V_\pi(b_0)$ and the corresponding policy $\hat{\pi}$, represented as a policy tree. A policy tree retains the chosen action branches, but all observation branches from the belief tree. The size of such a policy is the number of tree nodes.

For online POMDP planning, we represent a belief as a set of particles, i.e., sampled states. We start with an initial belief. At each time step, we search for a policy $\hat{\pi}$, as described above. The agent executes $a$, the first action of $\hat{\pi}$, and receives an observation $z$. We then apply particle filtering to incorporate information from $a$ and $z$ into an updated new belief. The process then repeats.

### 3.2 DESPOT

While a standard belief tree captures the execution of all policies under all possible scenarios, a DESPOT captures the execution of all policies under a sampled set of scenarios (Figure 1). It contains all the action branches, but only the observation branches under the sampled scenarios.

To construct a DESPOT, we apply a deterministic simulative model to $K$ scenarios for an initial belief $b_0$ and all possible action sequences. A scenario is an abstract simulation trajectory starting with some state $s_0$. Formally, a scenario for a belief $b$ is a random sequence $\phi = (s_0, \phi_1, \phi_2, \ldots)$, in which the start state $s_0$ is sampled according to $b$ and each $\phi_i$ is a real number sampled independently and uniformly from the range $[0, 1]$. The deterministic simulative model is function $g: S \times A \times R \mapsto S \times Z$, such that if a random number $\phi$ is distributed uniformly over $[0, 1]$, then $(s', z') = g(s, a, \phi)$ is distributed according to $p(s', z'|s, a)$. When we simulate this model for an action sequence $(a_1, a_2, a_3, \ldots)$ under a scenario $(s_0, \phi_1, \phi_2, \ldots)$, the simulation generates a trajectory $(s_0, a_1, s_1, z_1, a_2, s_2, z_2, \ldots)$, where $(s_t, z_t) = g(s_{t-1}, a_t, \phi_t)$ for $t = 1, 2, \ldots$. The simulation trajectory traces out a path $(a_1, z_1, a_2, z_2, \ldots)$ from the root of the standard belief tree. We add all the nodes and edges on this path to the DESPOT. Each DESPOT node $b$ contains a set $\Phi_b$, consisting of all scenarios that it encounters. The start states of the scenarios in $\Phi_b$ form a particle set that represents $b$ approximately. We insert the scenario $(s_0, \phi_0, \phi_1)$ into the set $\Phi_b$ and insert $(s_1, \phi_{t+1}, \phi_{t+2}, \ldots)$ into the set $\Phi_b$ for the belief node $b_t$ reached at the end of the subpath $(a_1, z_1, a_2, z_2, \ldots, a_t, z_t)$, for $t = 1, 2, \ldots$. Repeating this process for every scenario and every action sequence completes the construction of the DESPOT.

A DESPOT is determined completely by the $K$ scenarios, which are sampled randomly a priori. Intuitively, a DESPOT is a standard belief tree with some observation branches removed. While a belief tree of height $D$ has $O(|A|^D |Z|^D)$ nodes, a corresponding DESPOT has only $O(|A|^D K)$ nodes, because of reduced observation branching under the sampled scenarios. Hence the name Determinized Sparse Partially Observable Tree (DESPOT).

To evaluate a policy $\pi$ under sampled scenarios, define $V_{\pi, \phi}$ as the total discounted reward of the trajectory obtained by simulating $\pi$ under a scenario $\phi$. Then $\hat{V}_\pi(b) = \sum_{\phi \in \Phi_b} V_{\pi, \phi} / |\Phi_b|$ is an estimate of $V_\pi(b)$, the value of $\pi$ at $b$, under a set of scenarios, $\Phi_b$. We then apply the usual belief tree search from the previous subsection to a DESPOT to find a policy having good performance under the sampled scenarios. We call this algorithm Basic DESPOT (B-DESPOT).

The idea of using sampled scenarios for planning is exploited in hindsight optimization (HO) as well [17, 18]. HO plans for each scenario independently and builds $K$ separate trees, each with $O(|A|^D)$ nodes. In contrast, DESPOT captures all $K$ scenarios in a single tree with $O(|A|^D K)$ nodes and allows us to reason with all scenarios simultaneously. For this reason, DESPOT can provide stronger performance guarantees than HO.
4 Regularized DESPOT

To search a DESPOT for a near-optimal policy, B-DESPOT chooses a best action at every internal node of the DESPOT, according to the scenarios it encounters. This, however, may cause overfitting: the chosen policy optimizes for the sampled scenarios, but does not perform well in general, as many scenarios are not sampled. To reduce overfitting, our R-DESPOT algorithm leverages the idea of regularization, which balances the estimated performance of a policy under the sampled scenarios and the policy size. If the subtree at a DESPOT node is too large, then the performance of a policy for this subtree may not be estimated reliably with $K$ scenarios. Instead of searching the subtree for a policy, R-DESPOT terminates the search and uses a simple default policy from this node onwards.

To derive R-DESPOT, we start with two theoretical results. The first one provides an output-sensitive lower bound on the performance of any arbitrary policy derived from a DESPOT. It implies that despite its sparsity, a DESPOT contains sufficient information for approximate policy evaluation, and the accuracy depends on the size of the policy. The second result shows that by optimizing a lower bound on the performance of any arbitrary policy derived from a DESPOT, we can find a policy with small size and high value. For convenience, we assume that $R(s, a) \in [0, R_{\text{max}}]$ for all $s \in S$ and $a \in A$, but the results can be easily extended to accommodate negative rewards. The proofs of both results are available in the supplementary material.

Formally, a policy tree derived from a DESPOT is a subtree having the same root as the DESPOT, but only one action branch at each belief node. More commonly, it is represented more compactly by labeling a node by the action edge that follows and then removing the action edge. Let $T$ denote the class of all policy trees derived from DESPOT and $K$ from this bound, we can find a policy with small size and high value. For convenience, we assume that $R(s, a) \in [0, R_{\text{max}}]$ for all $s \in S$ and $a \in A$, but the results can be easily extended to accommodate negative rewards. The proofs of both results are available in the supplementary material.

Theorem 1 For any $\tau, \alpha \in (0, 1)$, every policy tree $\pi \in \Pi_{b_0, D, K}$ satisfies

$$V_\pi(b_0) \geq 1 - \frac{\alpha}{1 + \alpha} \hat{V}_\pi(b_0) - \frac{R_{\text{max}}}{(1 + \alpha)(1 - \gamma)} \cdot \frac{\ln(4/\tau) + |\pi| \ln(KD|A||Z|)}{\alpha K},$$

(2)

with probability at least $1 - \tau$, where $\hat{V}_\pi(b_0)$ is the estimated value of $\pi$ under any set of $K$ randomly sampled scenarios for belief $b_0$.

The second term on the right hand side (RHS) of (2) captures the additive error in estimating the value of policy tree $\pi$, and depends on the size of $\pi$. We can make this error arbitrarily small by choosing a suitable $K$, the number of sampled scenarios. Furthermore, this error grows logarithmically with $|A|$ and $|Z|$, indicating that the approximation scales well with large action and observation sets. The constant $\alpha$ can be tuned to tighten the bound. A smaller $\alpha$ value allows the first term on the RHS of (2) to approximate $\hat{V}_\pi$ better, but increases the additive error in the second term. We have specifically constructed the bound in this multiplicative-additive form, due to Haussler [19], in order to apply efficient dynamic programming techniques in R-DESPOT.

Now a natural idea is to search for a near-optimal policy $\pi$ by maximizing the RHS of (2), which guarantees the performance of $\pi$ by accounting for both the estimated performance and the size of $\pi$.

Theorem 2 Let $\pi^*$ be an optimal policy at a belief $b_0$. Let $\pi$ be a policy derived from a DESPOT that has height $D$ and is constructed from $K$ randomly sampled scenarios for belief $b_0$. For any $\tau, \alpha \in (0, 1)$, if $\pi$ maximizes

$$\frac{1 - \alpha}{1 + \alpha} \hat{V}_\pi(b_0) - \frac{R_{\text{max}}}{(1 + \alpha)(1 - \gamma)} \cdot \frac{|\pi| \ln(KD|A||Z|)}{\alpha K},$$

(3)

among all policies derived from the DESPOT, then

$$V_\pi(b_0) \geq \frac{1 - \alpha}{1 + \alpha} V_{\pi^*}(b_0) - \frac{R_{\text{max}}}{(1 + \alpha)(1 - \gamma)} \left( \frac{\ln(4/\tau) + |\pi^*| \ln(KD|A||Z|)}{\alpha K} + (1 - \alpha) \left( \frac{\gamma D}{K} + (1 - \alpha) \right) \right),$$

with probability at least $1 - \tau$.

\footnote{Here, we are using the term in the graph-theoretic sense, where a tree $T'$ is said to be subtree of another tree $T$ if it is a subgraph of $T$. At other places in this paper, we use the sense commonly used in data structure literature: $T'$ must include all the descendants of its root in $T$.}
Theorem 2 says that we can find a near-optimal policy with high probability by maximizing (3), if a small optimal policy tree \( \pi^* \) exists. Note that \( \pi^* \) is a globally optimal policy at \( b_0 \). It may or may not lie in \( \Pi_{b_0, D, K} \). The expression in (3) can be rewritten in the form \( \hat{V}_\pi(b_0) - \lambda |\pi| \), similar to that of regularized utility functions in many machine learning algorithms.

We now describe R-DESPOT, which consists of two main steps. In the first step, R-DESPOT constructs a DESPOT \( T \) of height \( D \) using \( K \) scenarios. To improve online planning performance, we may use offline learning to optimize the values for \( D \) and \( K \). In the second step, we perform bottom-up dynamic programming on the DESPOT and derive a policy tree that maximizes (3), as follows. For a policy \( \pi \) derivable from the DESPOT \( T \), let \( \pi_0 \) be the subtree of \( \pi \) rooted at \( b \), define the \textit{regularized discounted weighted utility} of \( \pi_b \) as

\[
\nu(\pi_b) = \gamma^{\text{depth}(b)} \frac{|\Phi_b|}{K} \hat{V}_{\pi_b}(b) - \lambda |\pi_b|,
\]

The regularized utility \( \hat{V}_\pi(b_0) - \lambda |\pi| \) is just \( \nu(\pi_{b_0}) \), or \( \nu(\pi) \). We can compute \( \nu(\pi_b) \) recursively using

\[
\nu(\pi_b) = \hat{R}(b, \pi(b)) + \sum_{b' \in \text{Children}_b} (\nu)(\pi_{b'}),
\]

where \( \hat{R}(b, \pi(b)) = \gamma^{\text{depth}(b)} \frac{1}{|\Phi_b|} \sum_{\phi \in \Phi_b} R(s_\phi, \pi(b)) - \lambda \) and \( s_\phi \) is the state associated with \( \phi \) at \( b \), and \( \text{Children}_b(b) \) is the set of belief nodes reachable from \( b \) in \( \pi \). Let \( \pi_0 \) be the default policy. For any belief node \( b \) in \( T \), let \( \nu^*(b) \) be the maximum possible \( \nu(\pi_b) \) value for a policy \( \pi \) derivable from \( T \) and having \( b \) as a node, and \( \text{Children}(b, a) \) be the set of child belief nodes of \( b \) action branch \( a \) in the full DESPOT. Define \( \nu^*(b) = \gamma^{\text{depth}(b)} \hat{V}_{\pi_0}(b) \) for any \( b \) not in \( T \), then for any \( b \) in \( T \),

\[
\nu^*(b) = \max \{ \gamma^{\text{depth}(b)} \hat{V}_{\pi_0}(b) - \lambda, \max_{\alpha} [\hat{R}(b, \alpha) + \sum_{b' \in \text{Children}(b, \alpha)} \nu^*(b')] \}.
\]

Let \( \pi_0^* \) be the policy maximizing the regularized utility \( \hat{V}_\pi(b_0) - \lambda |\pi| \), then \( \nu^*(b_0) \) is the regularized value of \( \pi_0^* \), and the above equation gives a bottom up dynamic programming algorithm to find the policy in time linear in the size of the DESPOT.

5 \textit{Anytime Regularized DESPOT}

To further improve online planning performance for large-scale POMDPs, we introduce AR-DESPOT, an anytime approximation of R-DESPOT. AR-DESPOT applies heuristic search and branch-and-bound pruning to uncover the more promising parts of a DESPOT and then searches the partially constructed DESPOT for a policy that maximizes the regularized utility in Theorem 2. A brief summary of AR-DESPOT is given in Algorithm 1. Below we provides some details on how AR-DESPOT performs the heuristic search (Section 5.1) and constructs the upper and lower bounds for branch-and-bound pruning (Sections 5.2 and 5.3).

5.1 DESPOT Approximation

We approximate the DESPOT by adapting the HSVI heuristic [20] to uncover it greedily. Initially, the approximation \( T \) we are constructing contains only the root node \( b_0 \), with \( \Phi_{b_0} \) being initialized to contain \( K \) scenarios either sampled from the initial belief, or updated from the previous belief using sequential importance resampling (SIR) on the particles of the previous belief. Then a series of “trials” is made, each of which augments \( T \) by tracing a path from the root to a leaf and adding new nodes to the end of the path.

HSVI maintains an upper bound \( U(b) \) and a lower bound \( L(b) \) on \( \hat{V}^*(b) = \arg \max_{\pi} \hat{V}_\pi(b) \) for each belief node \( b \) in current approximation. \( \hat{V}^*(b) \) is the maximum empirical value achievable by an arbitrary policy for \( b \) on the set of scenarios \( \Phi_b \). In addition, it also maintains an upper bound \( U(b, a) \) and a lower bound \( L(b, a) \) on \( Q_{\pi^*}(b, a) = \sum_{\phi \in \Phi_b} R(s_\phi, a) + \gamma \sum_{b' \in \text{Children}(b, a)} (\nu)(\pi_{b'}) \).

The Trial function traces a path from the root node \( b_0 \) to a leaf node, choosing successive nodes of the path as follows: it first selects the action \( a^* \) maximizing \( U(b, a) \) for the current node \( b \), and then selects the observation \( z^* \) maximizing the weighted excess uncertainty \( WEU_{b, \nu^*}(z) = \frac{\Phi_{b_0}(\phi^*, a^*)}{\Phi_{b_0}(\phi^*, a^*)} \text{excess}(\tau(b, a^*, z^*)) \), where \( \text{excess}(b') = U(b') - L(b') - \epsilon \gamma^{-\text{depth}(b')} \) and \( \epsilon \) is the desired accuracy at the root. If the excess uncertainty of the new node \( \tau(b, a^*, z^*) \) is negative, the trial
Algorithm 1 Anytime Regularized DESPOT

1: $b_0 \leftarrow$ Initial belief
2: loop
3: $T \leftarrow$ CONSTRUCTTREE($b_0$)
4: Compute $\pi_0^*$ for $T$ using Eq 6
5: Execute action $\pi_0^*(b_0)$
6: Receive real observation $z$
7: $b_0 \leftarrow$ Update($b_0$, $a$, $z$)

CONSTRUCTTREE($b_0$)
1: $\Phi_{b_0} \leftarrow K$ random scenarios for $b_0$
2: $T \leftarrow$ a tree having a root node $b_0$ only
3: while time remaining do
4: $b \leftarrow$ TRIAL($b_0$)
5: Backup lower and upper bounds for nodes on the path from $b$ to $b_0$

TRIAL($b$)
1: if depth($b$) > $D$ then
2: return $b$
3: if $b$ is a leaf node then
4: Expand($b$)
5: $a^* \leftarrow \arg \max_{a \in A} \hat{U}(b, a)$
6: $z^* \leftarrow \arg \max_{z \in Z^a_b} \text{WEU}_{b, a^*}(z)$
7: $b \leftarrow \tau(b, a^*, z^*)$
8: if excess($b$) > 0 then
9: return TRIAL($b$)
10: else
11: return $b$

returns $b$, otherwise it continues from the new node. At the end of the trial, backup operations are performed on both the upper and lower bounds from $b$ back to the root. The lower bound backup operation is as follows:

$$L(b) = \max_{a \in A} \left\{ \sum_{\phi \in \Phi_b} R(s_{\phi}, a) + \gamma \sum_{z \in Z^a_b} \left[ \frac{\Phi_{\tau(b, a, z)}}{|\Phi_b|} \right] L(\tau(b, a, z)) \right\}.$$  \hspace{1cm} (7)

where $Z^a_b$ is the set of observations generated when action $a$ is taken on all particles in $\Phi_b$. The upper bound backup is similar except that the upper bound $U$ is used in placed of $L$. Trials are repeated as long as there is time remaining and the uncertainty at the root is larger than the desired value.

The upper and lower bounds of leaf nodes (initial bounds) are determined as described in the following subsections. When a leaf node is chosen in the path of a trial, it is expanded one level deep before the best action and observation are selected. The expansion is done for every action, by adding belief nodes for every observation seen on executing the action for all particles in the current belief.

5.2 Initial Upper Bounds
AR-DESPOT requires initial upper and lower bounds. It is always possible to use the non-informative bound of $R_{\max}/(1 - \gamma)$ as the initial upper bound, but better solutions may sometimes be obtained by exploiting domain knowledge to compute good initial upper bounds. We give a domain independent upper bound, which is the average of the upper bound for each particle at a belief. The upper bound for a particle is the maximum return that can be obtained by following any policy with respect to its random number sequence. All the possible ways a particle may evolve under its random-number sequence depend only on the action taken at each step, and therefore form a trellis with actions as edges and the different states possible at each step as nodes. The trellis terminates when either a terminal state is encountered or the maximum search depth is reached. The path with highest value in the trellis gives the upper bound for the particle. This provides a deterministic upper bound which can be computed in $O(K|S||A|D)$ time.

5.3 Initial Lower Bounds and Default Policies
Any history-based policy can be used to generate the initial lower bound by using it to run simulations using the same random sequences that are used to construct the DESPOT. Computing the lower bound by $l$-step simulation at $b$ takes $O(|\Phi_b|l)$ time.

A history-based policy is often large, and in practice, we can use a particle-based approximations to history-based policies as default policies. For example, consider the history-based policy $\pi$ such that $\pi(b) = f(\text{mode}(b))$, where $\text{mode}(b)$ is the mode of $b$, and $f : S \rightarrow A$ is some mapping which is used to generate a good action for a given state. The particle-based approximation determines the
mode using $\Phi_b$. We use such particle-based approximations in the experiments. However, note that unlike history-based policies, this type of policies do not satisfy the requirement for Theorem 1 to hold, although the difference is unlikely to be important in practice.

6 Experiments

To evaluate AB-DESPOT experimentally, we compared it with four other algorithms. Anytime Basic DESPOT (AB-DESPOT), which is AR-DESPOT without the dynamic programming step, helps to understand the benefit of regularization. AEMS2 is an early successful online POMDP algorithm [21, 2]. POMCP has scaled up to very large POMDPs [3]. SARSOP is a state-of-the-art offline POMDP algorithm [10]. It helps to calibrate the best performance achievable for POMDPs of moderate size. In our online planning tests, each algorithm was given exactly 1 second per step to choose an action. For AR-DESPOT and AB-DESPOT, $K = 500$ and $D = 90$. The regularization parameter $\lambda$ for AR-DESPOT was selected offline by running the algorithm with a training set distinct from the online test set. The discount factor is $\gamma = 0.95$. For POMCP, we used the implementation from the original authors, but had to modify it in order to support very large number of observations and strictly follow the online planning time limit. We evaluated the algorithms on four different domains, including a very large one with about $10^{50}$ states.

The results are reported in Table 1. In summary, compared with AEMS2, AR-DESPOT is competitive on smaller POMDPs, but scales up much better on large POMDPs. Compared with POMCP, AR-DESPOT performs better than POMCP on the smaller POMDPs and scales up just as well.

We first tested the algorithms on Tag [7], a standard benchmark problem. In Tag, the agent’s goal is to find and tag a target that intentionally moves away. Both the agent and target operate in a grid with 29 possible positions. The agent knows its own position but can observe the target’s position only if they are in the same location. The agent can either stay in the same position or move to the four adjacent positions, paying a cost for each move. It can also perform the tag action and is rewarded if it successfully tags the target, but is penalized if it fails. For POMCP, we used the Tag implementation that comes with the package, but modified it slightly to improve its default rollout policy. The modified policy always tags when the agent is in the same position as the robot, providing better performance. For AR-DESPOT, we use a simple particle set default policy, which moves the agent towards the mode of the target in the particle set. For the upper bound, we average the upper bound for each particle as described in Section 5.2. The results (Table 1) show that AR-DESPOT gives comparable performance to AEMS2.

Theorem 1 suggests that AR-DESPOT may still perform well when the observation space is large, if a good small policy exists. To examine the performance of AR-DESPOT on large observation spaces, we experimented with an augmented version of Tag called LaserTag. In LaserTag, the agent moves in a $7 \times 11$ rectangular grid with obstacles placed in 8 random cells. The behavior of the agent and opponent are identical to that in Tag, except that in LaserTag the agent knows it location before the game starts, whereas in Tag this happens only after the first observation is seen. The agent is equipped with a laser that gives distance estimates in 8 directions. The distance between 2 adjacent cells is considered one unit, and the laser reading in each direction is generated from a normal distribution centered at the true distance of the agent from the nearest obstacle in that direction, with a standard deviation of 2.5 units. The readings are discretized into whole units, so an observation comprises a set of 8 integers. For a map of size $7 \times 11$, $|Z|$ is of the order of $10^6$. The environment for LaserTag is shown in Figure 2. As can be seen from Table 1, AR-DESPOT outperforms POMCP on this problem. We can also see the effect of regularization by comparing AR-DESPOT with AB-DESPOT. It is not feasible to run AEMS2 or SARSOP on this problem in reasonable time because of the very large observation space.

To demonstrate the performance of AR-DESPOT on large state spaces, we experimented with the RockSample problem [20]. The RockSample$(n, k)$ problem mimics a robot moving in an $n \times n$ grid

---

1http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Applications.html
Finally, we implemented Pocman sensing action. Does not appear beneficial to this problem, possibly because it is mostly deterministic, except for the behavior of the algorithms in very large state spaces. For POMCP, we used the implementation method described in Section 5.2. As in [3], we use a particle filter to represent the belief to examine the same horizon length as the maximum depth of the search tree. For the initial upper bound, we use the particle set. The optimal policy for the resulting state is used as the default policy. The optimal and each rock is labeled as good or bad depending on whichever condition is more prevalent in the environment containing $k$ rocks, each of which may be good or bad. At each step, the robot either moves to an adjacent cell, samples a rock, or senses a rock. Sampling gives a reward of +10 if the rock is good and -10 otherwise. Both moving and sampling produce a null observation. Sensing produces an observation in $\{\text{good}, \text{bad}\}$, with the probability of producing the correct observation decreasing exponentially with the agent’s distance from the rock. A terminal state is reached when the agent moves past the east edge of the map. For AR-DESPOT, we use a default policy derived from the particle set as follows: a new state is created with the positions of the robot and the rocks unchanged, and each rock is labeled as good or bad depending on whichever condition is more prevalent in the particle set. The optimal policy for the resulting state is used as the default policy. The optimal policy for all states is computed before the algorithm begins, using dynamic programming with the same horizon length as the maximum depth of the search tree. For the initial upper bound, we use the method described in Section 5.2. As in [3], we use a particle filter to represent the belief to examine the behavior of the algorithms in very large state spaces. For POMCP, we used the implementation in [3] but ran it on the same platform as AR-DESPOT. As the results for our runs of POMCP are not available from [3], we also reproduce their reported results in Table 1. The results in Table 1 indicate that AR-DESPOT is able to scale up to very large state spaces. Regularization does not appear beneficial to this problem, possibly because it is mostly deterministic, except for the sensing action.

Finally, we implemented Pocman, the partially observable version of the video game Pacman, as described in [3]. Pacman has an extremely large state space of approximately 10^{50}. We compute an approximate upper bound for a belief by summing the following quantities for each particle in it, and taking the average over all particles: reward for eating each pellet discounted by its distance from pocman; reward for clearing the level discounted by the maximum distance to a pellet; default per-step reward of −1 for a number of steps equal to the maximum distance to a pellet; penalty for eating a ghost discounted by the distance to the closest ghost being chased (if any); penalty for dying discounted by the average distance to the ghosts; and half the penalty for hitting a wall if pocman tries to double back along its direction of movement. This need not always be an upper bound, but AR-DESPOT can be modified to run even when this is the case. For the lower bound, we use a history-based policy that searches a random ghost, if visible, when pocman is under the effect of a powerpill, and avoids ghosts and doubling-back when it is not. This example shows that AR-DESPOT can be used successfully even in cases of extremely large state space.

7 Conclusion

This paper presents DESPOT, a new approach to online POMDP planning. Our R-DESPOT algorithm and its anytime approximation, AR-DESPOT, search a DESPOT for an approximately optimal policy, while balancing the size of the policy and the accuracy on its value estimate. Theoretical analysis and experiments show that the new approach outperforms two of the fastest online POMDP planning algorithms. It scales up better than AEMS2, and it does not suffer the extremely poor worst-case behavior of POMCP. The performance of AR-DESPOT depends on the upper and lower bounds supplied. Effective methods to construct such bounds, possibly automatically, will be an interesting topic for further investigation.

Acknowledgments. This work is supported in part by MoE A*STAR grant 2010-T2-2-071, National Research Foundation Singapore through the SMART IRG program, and US Air Force Research Laboratory under agreement FA2386-12-1-4031.
References


