Anytime Error Minimization Search

Stéphane Ross

School of Computer Science

November 1\textsuperscript{st}, 2006
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
5. Results
6. Future Work
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
Recall on POMDPs

Model

\[ < S, A, \Omega, T, R, O, \gamma, b_0 > \]

- \( S \): Finite set of states
- \( A \): Finite set of actions
- \( \Omega \): Finite set of observations
- \( T : S \times A \times S \rightarrow [0, 1] \), the transition function ( \( P(s'|s, a) \) )
- \( R : S \times A \rightarrow \mathbb{R} \), the reward function
- \( O : S \times A \times \Omega \rightarrow [0, 1] \), the observation function ( \( P(o|s', a) \) )
- \( \gamma \in [0, 1] \): Discount factor
- \( b_0 \): Initial belief state
Recall on POMDPs

**Belief state**
- Probability distribution over $S$
- Sufficient statistic of the complete history:
  \[ b_t(s) = P(s_t = s | b_0, a_0, o_0, a_1, o_1, \ldots, a_{t-1}, o_{t-1}) \]
- It is maintained through the belief update function
  \[ \tau : \beta \times A \times \Omega \rightarrow \beta \]
  \[ b_{t+1}(s') = O(s', a, o) \sum_{s \in S} T(s, a, s')b_t(s) \]

**Policy**
- Mapping from belief states to actions
- Optimality criterion for the infinite-horizon case:
  \[ \pi^* = \arg \max_{\pi \in \Pi} E(\sum_{t=0}^{\infty} \gamma^t \sum_{s \in S} b_t(s)R(s, \pi(b_t)) | b_0) \]
Recall on POMDPs

Value function

- Defines the expected sum of discounted reward obtained by following a policy $\pi$ from a certain belief state $b$:
  \[ V^\pi(b) = R(b, \pi(b)) + \gamma \sum_{o \in \Omega} P(o|b, a) V^\pi(\tau(b, a, o)) \]
- The value function of the optimal policy satisfies:
  \[ V^*(b) = \max_{a \in A} [R(b, a) + \gamma \sum_{o \in \Omega} P(o|b, a) V^*(\tau(b, a, o))] \]
- Therefore, the optimal policy can be defined as:
  \[ \pi^*(b) = \arg \max_{a \in A} [R(b, a) + \gamma \sum_{o \in \Omega} P(o|b, a) V^*(\tau(b, a, o))] \]
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
Offline vs. Online algorithms

Offline algorithms

- Compute a policy $\pi$ defined for all belief states prior to execution.
- Done via value iteration or policy iteration.
- Advantages:
  - ✓ Almost no computation required during execution.
  - ✓ Our patience is the only time constraint!
- Disadvantages:
  - ✗ All possible execution paths must be considered.
  - ✗ Complexity grows exponentially in the planning horizon.
  - ✗ If the environment changes slightly, we must recompute everything.
 Offline vs. Online algorithms

**Online algorithms**

- Computes the best action to do in a belief state only when it encounters it during execution.
- This is generally done via a lookahead search.
- **Advantages:**
  - ✓ Only the possible path from the current belief state must be considered.
  - ✓ No additional computations required if the environment changes.
  - ✓ Immediately executable in any environment.
- **Disadvantages:**
  - ✗ Require more computations during execution.
  - ✗ Planning time available during execution is generally very short.
Offline vs. Online algorithms

In practice, the most efficient way is to combine both approach:

- Compute an approximate value function via some offline algorithm.
- Improve the policy for the current belief state via an online lookahead search.

Advantages:

- ✓ Trade-off between online/offline computations can be customized according to real-time constraints.
- ✓ Immediatly executable in different environments with few offline computations.
- ✓ Shorter online planning horizon required for near-optimal policy.
- ✓ Improvement of the policy computed offline.
- ✓ Improvement only done for belief states encountered during execution.

Disadvantages:

- ✗ None !?
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
Quick Survey of Online Algorithms

AND/OR graph

Search algorithms proceed by constructing an AND/OR graph
Values of fringe nodes

- **Approximate value functions used at the fringe**
  - **Lower Bounds**:
    - Blind policy
    - PBVI style algorithms
  - **Upper Bounds**:
    - MDP
    - QMDP
    - FIB
    - Grid based algorithms
Quick Survey of Online Algorithms

Values of parent nodes
- Values of parent nodes are obtained from their children values
  - Lower Bounds:
    - $Low(b) = \max_{a \in A} Low(b, a)$
    - $Low(b, a) = R(b, a) + \gamma \sum_{o \in \Omega} P(o|b, a) Low(\tau(b, a, o))$
  - Upper Bounds:
    - $Upp(b) = \max_{a \in A} Upp(b, a)$
    - $Upp(b, a) = R(b, a) + \gamma \sum_{o \in \Omega} P(o|b, a) Upp(\tau(b, a, o))$

Branch & Bound pruning
- Action $a$ can be pruned in belief $b$ when $Upp(b, a) < Low(b)$
Quick Survey of Online Algorithms

Some algorithms...

- **Satia & Lave (1973)**
  - Best-First-Search with lower and upper bounds
  - Branch & Bound pruning

- **BI-POMDP (Washington 1997)**
  - Uses the AO* algorithm with lower and upper bounds
  - Explores the node with highest width between the bounds

- **Rollout (Bertsekas 1998)**
  - Assumes a given approximate policy
  - Performs a policy improvement step via trajectory sampling
Quick Survey of Online Algorithms

Some algorithms...

- McAllester (1999)
  - Depth-Limited search with observation sampling
  - Factored POMDP Representation

- RTBSS (Paquet 2005)
  - Depth-First-Search with lower and upper bounds
  - Branch & Bound pruning
  - Factored POMDP Representation

- SOVI (Shani & Brafman 2005)
  - Online HSVI
Quick Survey of Online Algorithms

In summary...

- Lookahead search in an AND/OR graph
- Approximate value functions used at the fringe
- Various techniques used to reduce complexity
  - Branch & Bound Pruning
  - Factored POMDP Representation
  - Sampling
  - Search Heuristics
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
Motivation

A k-step lookahead online search reduces the error by a factor $\gamma^k$.

It has a complexity in $O((|A||\Omega|)^k(|S|^2 + C_\gamma))$.

Can we do better than this?
- k-step lookahead might explore useless paths (e.g. probability near 0, optimal value already known, etc.)
- Hence, variable depth search might be better to get more precision where needed.

How should we explore the tree to reduce the error as quickly as possible?
- Explore the fringe node that contributes the most to the error in $b_0$.
How can we measure the error contribution of a fringe node to the error in $b_0$?

- From the optimal value function equation, we can derive:
  
  $$e(b_0) \leq \sum_{b \in \text{fringe}(T)} \gamma^{d(b)} P(b|\pi^*) e(b)$$

  - $d(b)$ is the depth of $b$ in $T$
  - $P(b|\pi^*) = \prod_{0}^{d(b)-1} P(o_i|b_i, a_i)\pi^*(b_i, a_i)$
  - $e(b) = V^*(b) - \text{Low}(b)$

  Contribution of a fringe node: $\gamma^{d(b)} P(b|\pi^*) e(b)$

  However, we do not know $V^*$ and $\pi^*$ to evaluate $e(b)$ and $P(b|\pi^*)$

  - $\hat{e}(b) = \text{Upp}(b) - \text{Low}(b) \geq e(b)$
  - $\hat{\pi}(b) = \arg \max_{a \in A} \text{Upp}(b, a)$
Heuristic

- \( b^*(T) = \arg \max_{b \in \text{fringe}(T)} \gamma^d(b) P(b|\hat{\pi}(b))\hat{e}(b) \)

- Is this a good heuristic?
  - Yes! It guides the search toward nodes that:
    - have a high uncertainty on their values
    - have a high probability to be reached in the future
    - are reached by promising actions
Function $AEMS(t, L, U, \epsilon)$

$b_0 \leftarrow$ initial belief state

while $b_0$ is not terminal do

$t_0 \leftarrow time()$

while $time() < t_0 + t$ and not $SOLVED(b_0, \epsilon)$ do

$b^* \leftarrow \arg \max_{b \in fringe(T)} \gamma^{d(b)} P(b|\pi(b))b(b)$

EXPAND($b^*$)

UPDATEANCESTORS($b^*$)

end while

$\hat{a} \leftarrow \arg \max_{a \in A} Low(b_0, a)$

DOACTION( $\hat{a}$ )

$o \leftarrow GETNEWOBSERVATION()$

$b_0 \leftarrow child(b_0, \hat{a}, o)$

$T \leftarrow SUBTREE(b_0)$

end while
Choice 1\textsuperscript{st} iteration

Expand 1\textsuperscript{st} iteration

Choice 2\textsuperscript{nd} iteration
Example

Expand 2\textsuperscript{nd} iteration

Update 2\textsuperscript{nd} iteration
Example

Update 2\textsuperscript{nd} iteration

Choice 3\textsuperscript{rd} iteration
Example

Expand 3rd iteration

Update 3rd iteration

Motivation
Error
Heuristic
Algo

Recall
Offline vs. Online
Survey
AEMS
Results
Future Work
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
### Results

#### Tag

\[ S = 870, |A| = 5, |\Omega| = 30 \]

<table>
<thead>
<tr>
<th>Method ( \text{\textit{MDP}} )</th>
<th>Reward</th>
<th>Offline Time (s)</th>
<th>Online Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMDP</td>
<td>-16.77</td>
<td>0.875</td>
<td>-</td>
</tr>
<tr>
<td>RTBSS(5)( \text{\textit{MDP}} ) ( \text{\textit{Blind}} )</td>
<td>-10.3</td>
<td>( \sim 1 )</td>
<td>0.387</td>
</tr>
<tr>
<td>PBVI</td>
<td>-9.18</td>
<td>180880</td>
<td>-</td>
</tr>
<tr>
<td>Satia( \text{\textit{FIB}} ) ( \text{\textit{Blind}} )</td>
<td>-8.35</td>
<td>( \sim 1 )</td>
<td>0.911</td>
</tr>
<tr>
<td>BBSLS</td>
<td>( \sim -8.3 )</td>
<td>( \sim 100000 )</td>
<td>-</td>
</tr>
<tr>
<td>BPI</td>
<td>-6.65</td>
<td>250</td>
<td>-</td>
</tr>
<tr>
<td>HSVI1</td>
<td>-6.37</td>
<td>10113</td>
<td>-</td>
</tr>
<tr>
<td>HSVI2</td>
<td>-6.36</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>RTBSS(6)( \text{\textit{MDP}} )</td>
<td>-6.32</td>
<td>( \sim 1 )</td>
<td>1.070</td>
</tr>
<tr>
<td>BI-POMDP( \text{\textit{FIB}} ) ( \text{\textit{Blind}} )</td>
<td>-6.22</td>
<td>( \sim 1 )</td>
<td>0.882</td>
</tr>
<tr>
<td>AEMS( \text{\textit{FIB}} ) ( \text{\textit{Blind}} )</td>
<td>-6.19</td>
<td>( \sim 1 )</td>
<td>0.841</td>
</tr>
<tr>
<td>PERSEUS</td>
<td>-6.17</td>
<td>1670</td>
<td>-</td>
</tr>
<tr>
<td>PBVI</td>
<td>( \sim -6.12 )</td>
<td>( \sim 900000 )</td>
<td>-</td>
</tr>
</tbody>
</table>
## Results

### RockSample[7,8]

\[ |S| = 12545, |A| = 13, |\Omega| = 2 \]

<table>
<thead>
<tr>
<th>Method</th>
<th>Reward</th>
<th>Offline Time (s)</th>
<th>Online Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind</td>
<td>7.4</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Satia\textsuperscript{QMDP, Blind}</td>
<td>7.4</td>
<td>25</td>
<td>0.889</td>
</tr>
<tr>
<td>PBVI</td>
<td>7.7</td>
<td>2418</td>
<td>-</td>
</tr>
<tr>
<td>Perseus</td>
<td>8.3</td>
<td>36000</td>
<td>-</td>
</tr>
<tr>
<td>RTDP-BEL</td>
<td>8.7</td>
<td>8362</td>
<td>-</td>
</tr>
<tr>
<td>RTBSS(2)\textsuperscript{QMDP, Blind}</td>
<td>10.3</td>
<td>25</td>
<td>0.896</td>
</tr>
<tr>
<td>HSVI</td>
<td>15.1</td>
<td>10266</td>
<td>-</td>
</tr>
<tr>
<td>QMDP</td>
<td>15.5</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>BI-POMDP\textsuperscript{QMDP, Blind}</td>
<td>18.4</td>
<td>25</td>
<td>0.955</td>
</tr>
<tr>
<td>RTBSS(2)\textsuperscript{QMDP}</td>
<td>20.3</td>
<td>25</td>
<td>0.320</td>
</tr>
<tr>
<td>HSVI2</td>
<td>20.6</td>
<td>1003</td>
<td>-</td>
</tr>
<tr>
<td>AEMS\textsuperscript{QMDP, Blind}</td>
<td>20.7</td>
<td>25</td>
<td>0.884</td>
</tr>
</tbody>
</table>
### Results

#### FieldVisionRockSample[5,7]

\[ |S| = 3201, |A| = 5, |\Omega| = 128 \]

<table>
<thead>
<tr>
<th>Method</th>
<th>Reward</th>
<th>Offline Time (s)</th>
<th>Online Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind</td>
<td>8.1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>RTBSS(1)\textsuperscript{QMDP}_{\text{Blind}}</td>
<td>20.6</td>
<td>4</td>
<td>0.254</td>
</tr>
<tr>
<td>BI-POMDP\textsuperscript{QMDP}_{\text{Blind}}</td>
<td>22.8</td>
<td>4</td>
<td>0.923</td>
</tr>
<tr>
<td>Satia\textsuperscript{QMDP}_{\text{Blind}}</td>
<td>22.8</td>
<td>4</td>
<td>0.949</td>
</tr>
<tr>
<td>AEMS\textsuperscript{QMDP}_{\text{Blind}}</td>
<td>23.4</td>
<td>4</td>
<td>0.943</td>
</tr>
<tr>
<td>PBVI</td>
<td>14.1</td>
<td>7802</td>
<td>-</td>
</tr>
<tr>
<td>RTBSS(1)\textsuperscript{QMDP}_{PBVI}</td>
<td>26.4</td>
<td>7806</td>
<td>0.183</td>
</tr>
<tr>
<td>BI-POMDP\textsuperscript{QMDP}_{PBVI}</td>
<td>26.7</td>
<td>7806</td>
<td>1.015</td>
</tr>
<tr>
<td>Satia\textsuperscript{QMDP}_{PBVI}</td>
<td>26.8</td>
<td>7806</td>
<td>1.015</td>
</tr>
<tr>
<td>AEMS\textsuperscript{QMDP}_{PBVI}</td>
<td>27.2</td>
<td>7806</td>
<td>0.966</td>
</tr>
</tbody>
</table>
Plan

1. Recall on POMDPs
2. Offline vs. Online algorithms
3. Quick Survey of Online Algorithms
4. AEMS
   - Motivation
   - Error Contribution
   - Heuristic
   - Algorithm
5. Results
6. Future Work
Future Work

- Explore different variants of $\hat{\pi}$
  - $\hat{\pi}(b, a) = P(a = \arg \max_{a' \in A} Q^*(b, a'))$
  - Similar to exploration/exploitation trade-off in RL?
  - Goal similar to Recognizers?
- Is handling cycle more efficient?
Questions