Compact, Convex Upper Bound Iteration for Approximate POMDP Planning

Tao Wang
University of Alberta

Joint work with Pascal Poupart Michael Bowling Dale Schuurmans
U.Waterloo U.Alberta U.Alberta
New approach to approximate POMDP planning

- Quadratic value function approximator
- Upper bound on true value
- Compact (bounded size) representation
- Competitive approximation quality
New approach
to approximate POMDP planning

- Quadratic value function approximator
- Upper bound on true value
- Compact (bounded size) representation
- Competitive approximation quality
New approach to approximate POMDP planning

- Quadratic value function approximator
- Upper bound on true value
- Compact (bounded size) representation
- Competitive approximation quality
New approach to approximate POMDP planning

- Quadratic value function approximator
- Upper bound on true value
- Compact (bounded size) representation
- Competitive approximation quality
Partially Observable Markov Decision Process (POMDP)

- General framework for optimal decision making under uncertainty
  - How to act based on past experience
  - Maximize long term reward

- Wide range of applications
  - Robotics and autonomous agent design
    - Helicopter control, Robot navigation and mapping
    - Nursing, Elderly assistance
  - Others:
    - Preference elicitation, stochastic resource allocation, spoken dialogue systems, active gesture recognition
POMDP model

Observation function
\[ p(o' | a, s') \]

Transition function
\[ p(s' | s, a) \]

Reward function
\[ R(s, a) \]

Goal: choose actions to maximize long term reward
Belief state

- Probability distribution over underlying states

\[ b = \begin{bmatrix}
  p(s^1) \\
p(s^2) \\
\vdots \\
p(s^{|s|})
\end{bmatrix} \]

- Sufficient summary of history for decision making

\[ b(\overline{s}_t) = p(\overline{s}_t | b_0 a_1 o_1 a_2 o_2 \cdots a_t o_t) \]
Belief state

$|\mathcal{S}| = 2$

\[
b = \begin{bmatrix}
p(s^1) \\
1 - p(s^1)
\end{bmatrix}
\]
Belief state

$|\mathcal{S}| = 3$

$$b = \begin{bmatrix} p(s^1) \\ p(s^2) \\ 1 - p(s^1) - p(s^2) \end{bmatrix}$$
Belief state

\[ |\mathcal{S}| = 4 \]

\[ b = \begin{bmatrix} p(s^1) \\ p(s^2) \\ p(s^3) \\ 1 - p(s^1) - p(s^2) - p(s^3) \end{bmatrix} \]
Updating belief state

\[ b'_{(b,a,o')} (s') = p(o'|a, s') \sum_s p(s'|s, a) b(s) / Z \quad \text{Bayes rule} \]

\[ Z = p(o'|b, a) = \sum_{s'} p(o'|a, s') \sum_s p(s'|s, a) b(s) \]
POMDP solving

- Policy maps belief states to actions $\pi : B \rightarrow A$

- Value function is expected total discounted future reward starting from each belief state

$$V^\pi(b) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b \right]$$

- Planning problem: find policy that maximizes value function

- **Provably hard (even to approximate)**
POMDP approximation approaches

- **Value function approximation**
  - Hauskrecht 2000
  - Spaan&Vlassis 2005
  - Pineau et al. 2003
  - Parr&Russell 1995

- **Policy based optimization**
  - Ng&Jordan 00; Poupart & Boutilier 03,04; Amato et al. 06

- **Stochastic sampling**
  - Kearns et al. 02; Thrun 00
Value function based approaches

- Optimal value function
  (satisfies Bellman equation)

\[
V^*(b) = \max_a r(b, a) + \gamma \sum_{b'} p(b' | b, a) V^*(b')
\]

\[
= \max_a r(b, a) + \gamma \sum_{o'} p(o' | b, a) V^*(b'_{b,a,o'})
\]

- Difficulty: belief space is continuous & high dimensional
Optimal 1-step decision

\[ V_1(b) = \max_a b \cdot r_a \]

\[ \Gamma_1 = \{ r_{a_1}, r_{a_2}, r_{a_3} \} = \{ \alpha_1, \alpha_2, \alpha_3 \} \]

\[ V_1(b) = \max_{\alpha \in \Gamma_1} b \cdot \alpha \]

Optimal value function is \textbf{piecewise linear convex}
**Optimal n+1-step decision**

Value function representation

\[
V_n(b) = \max_{\alpha_{\pi'} \in \Gamma_n} b \cdot \alpha_{\pi'}
\]

Value function iteration

\[
V_{n+1}(b) = \max_a r(b, a) + \gamma \sum_{o'} p(o' | b, a) V_n(b'_{(b, a, o')})
\]

\[
= \max_{a, \{o' \to \pi'\}} b \cdot \alpha_{a, \{o' \to \pi'\}}
\]
POMDP solution

- A POMDP solution represented as a set of $\alpha$-vectors
  \[ \Gamma_n = \{ \alpha_{\pi'} : \pi' \in \Pi_n \} \]

- The value of any belief can be extracted from
  \[ V_n(b) = \max_{\alpha_{\pi'} \in \Gamma_n} b \cdot \alpha_{\pi'} \]

- Any exact finite-horizon solution $V_n$ can be represented by a finite set of vectors $\Gamma_n$
# policies grows exponentially in the number of observations at one step

\[ |\Gamma_{n+1}| \leq |A| |\Gamma_n|^{O} \]

where $\Gamma = \alpha$-vectors

n = Planning horizon

A = Actions

O = Observations
Current approximation strategies

- Grid based approach
  - Gorden 95
  - Hauskrecht 00
  - Zhou & Hansen 01
  - Bonet 02
- Belief point approach
  - Pineau et al. 03
  - Smith & Simmons 05
  - Spaan & Vlassis 05

value function representation

$\alpha$ -vectors
Our idea

Approximate $V^*(b)$ with a **convex quadratic** upper bound

- Maintain compact (bounded size) representation of value approximation
- Can still model multiple $\alpha$-vectors
- Can be optimized easily
Quadratic approximation

Value function representation

$$\hat{V}(b) = b^\top W b + w^\top b + \omega$$

Would like to enforce

$$\hat{V}_{n+1}(b) \geq \max_a \hat{q}_a(b)$$

Need action-value backup for each action

$$\hat{q}_a(b) = r(b, a) + \gamma \sum_{o'} p(o' | b, a) \hat{V}_n(b'_{(b, a, o')})$$
Quadratic approximation

Combine with belief update

\[ b'_{(b,a,o')} = \frac{M_{a,o'}b}{e^\top M_{a,o'}b} \]

\[ M_{a,o'}(s', s) = p(o'|a, s')p(s'|s, a) \]

Get action-value

\[ q_a(b) = r(b, a) + \gamma \sum_{o'} \frac{b^\top M_{a,o'}^\top WM_{a,o'}b}{e^\top M_{a,o'}b} + (w + \omega e)^\top M_{a,o'}b \]

Theorem 1 \( q_a(b) \) is convex in \( b \).

Corollary 1 \( \max_a q_a(b) \) is convex in \( b \).
Algorithm

\[ \hat{V}_n \quad \hat{q}_a(b) \quad \max_a \hat{q}_a(b) \quad \hat{V}_{n+1} \]

Maintain tight upper bound of the maximum of the action-values
Mathematically

Optimization problem

\[ \min_{W, w, \omega} \int_b (b^\top W b + w^\top b + \omega) \, \mu(b) \, db \]

subject to

\[ b^\top W b + w^\top b + \omega \geq q_a(b), \quad \forall a, b \]

\[ W \succeq 0 \quad \text{(positive semi-definite)} \]

a measure over space of possible beliefs

ensure upper bound

ensure convexity
Two difficulties

- Integral in objective
- Infinite number of linear constraints

But it is a convex optimization problem (SDP plus infinitely many linear constraints)
Integral

Objective \[ \int_b \left( b^\top W b + w^\top b + \omega \right) \mu(b) \, db \]

is equal to \[ \langle W, \mathbb{E}[bb^\top] \rangle + w^\top \mathbb{E}[b] + \omega \] (linear)

Assume measure \( \mu(b) \) is Dirichlet distribution on \( b \)

then \( \mathbb{E}[bb^\top] \) and \( \mathbb{E}[b] \) have closed form
Infinite constraints

Have \( b^\top W b + w^\top b + \omega \geq q_a(b), \ \forall a, b \)

infinitely many linear constraints on \( Ww\omega \)

Optimal constraint generation: most violated constraint

\[
\min_b b^\top W b + w^\top b + \omega - q_a(b)
\]

subject to \( b \geq 0, \ \sum_s b(s) = 1 \)

Unfortunately, not necessarily a convex minimization problem in \( b \)
Strategies for constraint generation

1. Sample belief space + “corners”

Or: optimal constraint generation

2. Non-convex minimization

3. Tighten constraints to linear upper bound, solve convex approximation

   Proposition 1 The tightest linear upper bound on $q_a(b)$ is given by $q_a(b) \leq u_a^\top b$ for a vector $u_a$ such that $u_a^\top 1_s = q_a(1_s)$ for each corner belief state $1_s$.

4. Relax the upper bound, but after belief generated, use real action-value bounds
Experimental results

- Benchmark problems
- Mean discounted reward & Run time
  - 10 runs
  - 1000 trajectories
- Competitors
  - Perseus (Spaan & Vlassis 05)
  - PBVI (Pineau et al. 03)
## Problem characteristics

| Problems      | $|S|$ | $|A|$ | $|O|$ |
|----------------|-----|-----|-----|
| Maze          | 20  | 6   | 8   |
| Tiger-grid    | 33  | 5   | 17  |
| Hallway       | 57  | 5   | 21  |
| Hallway2      | 89  | 5   | 17  |
| Aircraft      | 100 | 10  | 31  |
## Mean discounted reward

<table>
<thead>
<tr>
<th>Avg. Reward</th>
<th>CQUB</th>
<th>Perseus</th>
<th>PBVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hallway</td>
<td>0.58 ±0.14</td>
<td>0.51 ±0.06</td>
<td>0.53 ±0.03</td>
</tr>
<tr>
<td>Hallway2</td>
<td>0.43 ±0.25</td>
<td>0.34 ±0.16</td>
<td>0.35 ±0.03</td>
</tr>
<tr>
<td>Tiger-grid</td>
<td>2.16 ±0.02</td>
<td>2.34 ±0.02</td>
<td>2.25 ±0.06</td>
</tr>
<tr>
<td>Maze</td>
<td>45.35 ±3.28</td>
<td>30.49 ±0.75</td>
<td>46.70 ±2.00</td>
</tr>
<tr>
<td>Aircraft</td>
<td>16.70 ±0.58</td>
<td>12.73 ±4.63</td>
<td>16.37 ±0.42</td>
</tr>
</tbody>
</table>
## Compact representation

<table>
<thead>
<tr>
<th>Size</th>
<th>CQUB</th>
<th>Perseus</th>
<th>PBVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maze</td>
<td>231</td>
<td>460</td>
<td>1160</td>
</tr>
<tr>
<td>Tiger-grid</td>
<td>595</td>
<td>4422</td>
<td>15510</td>
</tr>
<tr>
<td>Hallway</td>
<td>1711</td>
<td>3135</td>
<td>4902</td>
</tr>
<tr>
<td>Hallway2</td>
<td>4095</td>
<td>4984</td>
<td>8455</td>
</tr>
<tr>
<td>Aircraft</td>
<td>5151</td>
<td>10665</td>
<td>47000</td>
</tr>
</tbody>
</table>
Conclusions

New approximation algorithm
- Quadratic value function approximator
- Compact representation
- Competitive approximation quality
- Provable upper bound on the optimal values
- Computational cost independent of iteration number
Future work

- Set of quadratics
- Use belief state compression & factored models
- Combine with sampling
- Interpretation: 2\textsuperscript{nd} order Taylor expansion
Take home message

A new perspective to value function approximation for POMDP planning

- **Approximate** $V^*(b)$ with a **convex quadratic** upper bound
- Compact representation
- Computational cost independent of iteration number