Module 15
POMDP Bounds

CS 886 Sequential Decision Making and Reinforcement Learning

University of Waterloo

Bounds

- POMDP algorithms typically find approximations to optimal value function or optimal policy
  - Need some performance guarantees

- Lower bounds on V^*
  - V^π for any policy π
  - Point-based value iteration
- Upper bounds on V^*
  - QMDP
  - Fast-informed bound
  - Finite Belief-State MDP

CS886 (c) 2013 Pascal Poupart

Lower Bounds

- Lower bounds are easy to obtain
- For any policy π, V^{π} is a lower bound since $V^{\pi}(b) \leq V^{*}(b)\ \forall \pi, b$
- The main issue is to evaluate a policy π

CS886 (c) 2013 Pascal Poupart

Point-based Value Iteration

- Theorem: If V_0 is a lower bound, then the value functions V_n produced by point-based value iteration at each iteration n are lower bounds.

- Proof by induction
  - Base case: pick V_0 to be a lower bound
  - Inductive assumption: $V_n(b) \leq V^*(b)\ \forall b$
  - Induction:
    - Let T_{n+1} be the set of α-vectors for some set B of beliefs
    - Let T^*_{n+1} be the set of α-vectors for **all** beliefs
    - Hence $V_{n+1}(b) = \max_{\alpha \in \mathrm{T}_{n+1}} \alpha(b) \leq \max_{\alpha \in \mathrm{T}^*_{n+1}} \alpha(b) \leq V^*(b)$

CS886 (c) 2013 Pascal Poupart

Upper Bounds

- Idea: make decision based on more information than normally available to obtain higher value than optimal.

- POMDP: states are hidden
- MDP: states are observable
- Hence $V_{MDP} \geq V_{POMDP}$

CS886 (c) 2013 Pascal Poupart

QMDP Algorithm

- Derive upper bound from MDP Q-function by allowing the state to be observable
- Policy: $s_t \rightarrow a_t$

QMDP(POMDP)

Solve MDP to find Q_{MDP}

$Q_{MDP}(s,a) = R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \max_{a'} Q_{MDP}(s',a')$

Let $\bar{V}(b) = \max_{a} \sum_{s} b(s) Q_{MDP}(s,a)$

Return $\bar{V}$

CS886 (c) 2013 Pascal Poupart

Fast Informed Bound

- QMDP upper bound is too loose
  - Actions depend on current state (too informative)

- Tighter upper bound: fast Informed bound (FIB)
  - Actions depend on previous state (less informative)

$$V_{MDP} \geq V_{FIB} \geq V^*$$

FIB Algorithm

- Derive upper bound by allowing the previous state to be observable
- Policy: $s_{t-1}, a_{t-1}, o_t \rightarrow a_t$

FIB(POMDP)

Find Q_{FIB} by value iteration

$Q_{FIB}(s,a) = R(s,a) + \gamma \sum_{o'} \max_{a'} \sum_{s'} \Pr(s'|s,a) \Pr(o'|s',a)\, Q_{FIB}(s',a')$

Let $\bar{V}(b) = \max_{a} \sum_{s} b(s) Q_{FIB}(s,a)$

Return $\bar{V}$

CS886 (c) 2013 Pascal Poupart

FIB Analysis

- Theorem: $V_{MDP} \geq V_{FIB} \geq V^*$
- Proof:

1) $Q_{MDP}(s,a) = R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \max_{a'} Q(s',a')$

$= R(s,a) + \gamma \sum_{s'o'} \Pr(s'|s,a) \Pr(o'|s',a) \max_{a'} Q(s',a')$

$\geq R(s,a) + \gamma \sum_{o'} \max_{a'} \sum_{s'} \Pr(s'|s,a) \Pr(o'|s',a)\, Q(s',a')$

$= Q_{FIB}(s,a)$

2) $V_{FIB} \geq V^*$ since V_{FIB} is based on observing the previous state (too informative)

CS886 (c) 2013 Pascal Poupart

Finite Belief-State MDP

- Belief state MDP: all beliefs are treated as states
$$V^*(b) = \max_a Q^*(b,a)$$
- QMDP and FIB: value of each interior belief is interpolated: i.e., $\bar{V}(b) = \max_a \sum_s b(s) Q_{FIB}(s,a)$
- Idea: retain subset of beliefs
  - Interpolate value of remaining beliefs

CS886 (c) 2013 Pascal Poupart

Finite Belief-State MDP

- Belief state MDP

$$Q(b,a) = R(b,a) + \gamma \sum_{o'} \Pr(o'|b,a) \max_{a'} Q(b^{a,o}, a')$$

- Let B be a subset of representative beliefs
- Approximate $Q(b^{a,o}, a')$ with lowest interpolation
  - Linear program

$$Q(b^{a,o}, a') = \min_{c} \textstyle\sum_{b \in B} c_b Q(b, a')$$

such that $\sum_b c_b = 1$ and $c_b \geq 0 \ \forall b$

CS886 (c) 2013 Pascal Poupart

Finite Belief-State MDP Algorithm

- Derive upper bound by interpolating values based on a finite subset of values

FiniteBeliefStateMDP(POMDP)

Find Q_B by value iteration

$Q_B(b,a) = R(b,a) + \gamma \sum_{o'} \Pr(o'|b,a) \max_{a'} Q_B(b^{ao'},a') \;\; \forall b \in B, a$

where $Q_B\left(b^{ao'},a'\right) = \min_c \sum_{b \in B} c_b Q_B(b,a')$

such that $\sum_{b \in B} c_b = 1$ and $c_b \geq 0 \; \forall b \in B$

Let $\bar{V}(b) = \max_a \sum_s b(s) Q_B(s,a)$

Return $\bar{V}$

CS886 (c) 2013 Pascal Poupart