6.825 Techniques in Artificial Intelligence

Markov Decision Processes

- Framework
- Markov chains
- MDPs
- Value iteration
- Extensions

MDP Framework

- $S$: states
- $A$: actions
- $\Pr(s_{t+1} \mid s_t, a_t)$: transition probabilities
- $R(s_t)$: real-valued reward

Find a policy: $\pi: S \rightarrow A$
Maximize
- Myopic: $E[s_t | \pi]$ for all $s$
- Finite horizon: $E[\sum_{t=0}^{T} \gamma^t r_t | \pi, s_0]$
- Non-stationary policy: depends on time
- Infinite horizon: $E[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s_0]$
- $0 < \gamma < 1$ is discount factor
- Optimal policy is stationary

Markov Chain

- Markov Chain
- States
- Transitions
- Rewards
- No actions
- Value of a state, using infinite discounted horizon
  $V(s) = R(s) + \sum_{s'} P(s' \mid s) V(s')$
- Assume $\gamma = 0.9$
  $V(1) = 0.1 \cdot (0.5 \cdot V(1) + 0.5 \cdot V(2))$
  $V(2) = 0.1 \cdot (0.2 \cdot V(1) + 1.0 \cdot V(2) + 0.7 \cdot V(3))$
  $V(3) = 0.1 \cdot (0.9 \cdot V(2) + 0.1 \cdot V(3))$

Finding the Best Policy

- MDP + Policy = Markov Chain
- MDP = the way the world works
- Policy = the way the agent works

- $V'(s) = R(s) + \max_{s'} [\gamma \sum_{s'} P(s' \mid s, a) V'(s')]$
- Theorem: There is a unique $V'$ satisfying these equations
- $[\Gamma(s) = \arg \max_{s'} \sum_{s'} P(s' \mid s, a) V'(s')]$

Computing $V'$

- Approaches
  - Value iteration
  - Policy iteration
  - Linear programming

Value Iteration

Initialize $V^0(s) = 0$, for all $s$
Loop for a while until $|V^t - V^{t+1}| < \epsilon(1-\gamma)/\gamma$
Loop for all $s$

- $V^{t+1}(s) = R(s) + \max_{\pi} \sum_{s'} P(s' \mid s, a) V^t(s')$
- Converges to $V'$
- No need to keep $V^t$ vs $V^{t+1}$
- Asynchronous (can do random state updates)
- Assume we want
- Gets to optimal policy in time polynomial in $|A|$, $|S|$, $1/(1-\gamma)$
Big state spaces
- Function approximation for V
- neural nets
- regression trees
- factored representations (represent Pr(s’|s, a) using Bayes net)

Partial Observability
- MDPs assume complete observability (can always tell what state you’re in)
- POMDP (Partially Observable MDP)
- Observation: Pr(O|s, a) [O is observation]
- o, a, o, a, o, a

Worrying too much
- Assumption that every possible eventuality should be taken into account
- sample-based planning: with short horizon in large state space, planning should be independent of state-space size

Leading to Learning
MDPs and value iteration are an important foundation of reinforcement learning, or learning to behave