RL for helicopter control

Note Title 11/3/2005

Many RL algorithms:

- Q-learning } Value iteration
- TD(λ) }
- Pegasus } Policy search

Policy search:

Idea: optimize value function by searching in the space of policies

Traditionally:

$$V(s) = \max_a \; R(s,a) + \gamma \sum_{s'} P(s' \mid s, a) V(s')$$

Alternatively:

$$V(\pi) = R(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) \Big[R(s', \pi(s')) + \gamma \sum_{s''} P(s'' \mid s', \pi(s')) \big[\ldots \big] \Big]$$

optimize $V(\pi)$ by gradient ascent on the parameters: $\pi(s)$

In RL $P(s'|s,a)$ is unknown, but we get to observe trajectories $\langle s, a, s', a', s'', a'', \ldots \rangle$

$$V(\pi) = \frac{1}{n} \sum_{i=1}^{n} R(s_i, \pi(s_i)) + \gamma R(s'_i, \pi(s'_i)) + \gamma^2 R(s''_i, \pi(s''_i)) + \ldots$$

Problem: we can't optimize this expression since it changes for every policy π. In other words the sequence s, s', s'' depends on the actual policy.

Solution: express V in terms of π & a sequence of random numbers. Instead of recording state sequences, record random number sequences.