CS294-40 Learning for Robotics and Control Lecture 1 - 8/28/2008

Introduction

Lecturer: Pieter Abbeel *Scribe: Pieter Abbeel*

1 Lecture outline

- Class logistics.
- Slideshow and movies on current autonomous robotics, on algorithms they use, and on future directions.
- Markov decision processes.

2 Markov decision processes (MDPs)

2.1 Definition

A (discounted infinite horizon) Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{D}, R)$.
Here

1. $\mathcal{S}$ is the set of possible states for the system;
2. $\mathcal{A}$ is the set of possible actions;
3. $\mathcal{T}$ represents the (typically stochastic) system dynamics;
4. $\mathcal{D}$ is the initial-state distribution, from which the start state s_0 is drawn;
5. $R : \mathcal{S} \mapsto \Re$ is the reward function.

Acting in a Markov decision process results in a sequence of states and actions $s_0, a_0, s_1, a_1, s_2, \ldots$.

A policy π is a sequence of mappings $(\mu_0, \mu_1, \mu_2, \ldots)$, where, at time t the mapping $\mu_t(\cdot)$ determines the action $a_t = \mu_t(s_t)$ to take when in state s_t.

The objective is to find policies that maximize the expected sum of rewards accumulated over time. In particular, a policy π is good if its utility

$$U(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi]$$

is high.

To represent the system dynamics, we can use the state-transition distribution notation

$$s_{t+1} \sim P_{sa}(\cdot|s_t, a_t).$$

We will also often use the following notation:

$$s_{t+1} = F(s_t, a_t, w_t).$$

Here F is a deterministic function, and w_t is a random disturbance.

2.2 Examples

2.2.1 Car

One (approximate) way to model the state of a car is to use the following six state variables: northing (n), easting (e), north velocity ($\dot{n}$), east velocity ($\dot{e}$), heading (θ), angular rate ($\dot{\theta}$). Hence the state space $\mathcal{S} = \Re^6$.

The actions (or control inputs) are (i) steering angle, (ii) throttle, (iii) brake.

The perturbances capture both environmental perturbations as well as unmodeled aspects of the car dynamics.

We could have the following dynamics model $s_{t+1} = F(s_t, a_t, w_t)$:

$$\begin{aligned}
n_{t+1} &= n_t + \dot{n}_t \Delta t, \\
e_{t+1} &= e_t + \dot{e}_t \Delta t, \\
\theta_{t+1} &= \theta_t + \dot{\theta}_t \Delta t, \\
\dot{n}_{t+1} &= f_n(\dot{n}_t, \dot{e}_t, \dot{\theta}_t, a_t, w_t) \\
\dot{e}_{t+1} &= f_e(\dot{n}_t, \dot{e}_t, \dot{\theta}_t, a_t, w_t) \\
\dot{\theta}_{t+1} &= f_\theta(\dot{n}_t, \dot{e}_t, \dot{\theta}_t, a_t, w_t)
\end{aligned}$$

The reward function could be $R(s_t) = \mathbf{1}\{\text{in goal region}\} - 100 * \mathbf{1}\{\text{in collision}\}$. Here $\mathbf{1}\{\cdot\}$ is an indicator function, taking the value "1" when its argument is true, and "0" otherwise. The functions f_n, f_e, f_θ are deterministic functions modeling the car's dynamics.