Class 1

Shikhar Saxena

January 03, 2023

Contents

Resources

1. Sutton & Barto
2. Bertsekas; Reinforcement learning and optimal control
3. Applied porbability models with optimization applications by Sheldon Ross (for MDP's)
4. Other recent books by Warren Powell, Sean Meyn, Sham Kakde, Abhijith Gosavi, Ashwin Rao.

- Some course notes (links from slides). First and Second Link will be followed by Sir mostly. Third Link: (david silver lectures).

Grading

- Quiz 1: 10%
- Midsem: 20%
- Quiz 2: 10%
- Endsem: 20%
- Assignment: 20%
- Project: 20%

Probably two-three assignments and group-based project.

Course Outline

- Module 1 (3-4 lecs)
  - ⋆ Probability & markov chain
- Module 2 (5-6 lecs)
  - ⋆ MDPs
- Module 3 (3-4 lecs)
  - ⋆ Intro to RL
- Module 4 & 5 (12-14 lecs)
  - ⋆ Adv RL (Prof. Harikumar)

HW for today

Watch

- AlphaGo (2017 documentary)
- David Silver lec 1

What is RL? Mathematical Viewpoint

Essentially an MDP where Markovian transitions are unknown.

- MDP: Markov Chain that you control with actions for maximizing your accumulated reward.

Viewpoint 2: Sequential decision problem

Interaction between Agent and Environment. Agent performs `action` and environment `reward`s the agent and next `state`. Objective is to select the best `action` over time.

SARSA: State action reward (next) state (an rl algorithm).

Select sequence of actions to maximize total reward under environment uncertainity.

- Model based rl
  - ⋆ You learn the mdp and use the optimal policy for that
- another is where we try to pick policies (and then converge to the best policy).
- Immediate gains vs long term gains
  - ⋆ balance exploration and exploitation

Key ingredients of MDP/RL

- Model for the environment
  - ⋆ transition model
    - you move from one state to another
    - represents ***dynamics*** of the environment

- $S_{t+1} = f(H_t, W_t)$
  - $H_t = \{S_1, A_1, \ldots, S_t, A_t\}$: History (state-action pairs)
  - W_t: possible source for randomness (noise)
- Markovian Model: $S_{t+1} = f(S_t, A_t, W_t)$ where W_t is i.i.d noise. In this case, the Markov Property is true.

⋆ reward model

- how are the rewards coming
- $R_{t+1} = g(S_t, A_t)$
- Another model for reward $R_{t+1} = g(S_t, A_t, S_{t+1})$.
- **Reward Hypothesis:** Optimize expected total reward.
- Other metrics finite time expected total reward, time average reward and discounted total expected reward.

◦ policy of the agent

⋆ $\pi = (\pi_1, \pi_2, \ldots)$

- sequence of actions that the agent selects at each time
- policies could be history-based, markovian, deterministic, randomized, stationary,etc.

⋆ optimal policy π^*: highest expected total reward.

⋆ when model is knoen, the optimal policies often turn out to be markovian, deterministic and even stationary (more later).

◦ value function for the policy and/or states

⋆ $V^\pi(s)$ quantifies the expected total reward from policy π when starting in state s.

⋆ $Q^\pi(s, a)$: state action value function for policy π.

RL: minimize the regret (how different you are from the optimal; in an RL scenario you don't really know how much regret you're making).

My objective:

$$V(s) := \max_{\pi \in \Pi} V^\pi(s)$$

and

$$\pi^* = \arg\max_{\pi \in \Pi} V^\pi(s)$$

Classification of RL Problems

- Under uncertainity, obj of RL to learn $Q^*(s, a)$ and/or π^*.
- focus on learning Q^*, value function based algo eg value iteration
- focus on learning π^*, policy based algo eg policy iteration
- (these are model-free algorithms).