Running Lecture Outline: Understanding RL

[Chirayu Salgarkar]

Fall 2024

Contents

1 26-AUG-24

1.1 Order, Linear, and PDE vs ODE

2 28-AUG-24

2.1 Motvation

Decision making is hard. How we tractably reason over a sequence of decisions is a subject for much research. One potential mechanism for modeling sequential decision making is a *Markov Decision Process*.

2.2 MDP

MDPS consist of a state S, action A, cost C, and transition$\mathscr{T}$.

2.2.1 What is a state?

A state refers to the sufficient statistic of the system to predict the future disregarding the past. This definition is not really precise. More generally, the state is the status of the world. That's a definition. Not the definition. We show state as $s \in S$.

2.2.2 Action

Action refers to the decisions or, more basically, the act of doing something, or the control action.

2.2.3 Cost

The cost, or rewardd, is the instantaneous cost of an individual action within a state. This is denoted $c(s, a)$. Sometimes, we see $c(s, a, t)$ (time-dependence). Sometime, we even have $c(s, a, s', t)$ indicating previous state matters too. Dr. Baheri uses cost and reward interchangaby here.

2.2.4 Transition

Insert Anthony Fantano joke here. The transition refers to the next state given the state and action. In a deterministic world, $s' = \mathscr{T}(s, a)$, but stochastic worlds are more of $s' \cong \mathscr{T}(s, a)$.

Quiz: Given a system, whate are the components of the MDP and how do you formulate them?

Let's identify the MDP components of Tetris.

State: Board configuration. Action: $4 * 10$ Cost: Userdefined. Transition: rule of game. Update of board game + random selection of next piece.

For a self-driving car, what are the MDP components?

We now move to a Markov Decision Problem. This includes the things to define an optimization problem.

2.3 MDP, continued

We first describe the Horizon, and discount.

2.3.1 Horizon

Simply when to make decision.

2.3.2 Discount Factor

Reward is more valuable at the current moment as opposed to the future! (Costs are more valuable when they happen soon.) They are represented as a $r \in \mathbb{R}$, $0 \leq r \leq 1$. Think of it this like

$$c_0 + rc_1 + ... + r^{t-1}c_{t-1}$$

The final goal of RL is to find a *policy*, essentially given state, what action do we have. We sek to find a policy that minimizes the sum of discounted future costs.

3 06-SEP-2024

3.1 Solving Markov Decision processes

Solving an MDP, in general means to find a *policy.* Recall the definition of policy. Then a MDP is a function map mapping decision to reward.

But a better question is finding *optimal policy.* What makes an optimal policy? The ultimate goal is to search over a family of policies in order to find the sequence of actions that maximizes (or sometimes minimize) the notion of the reward. That is, we seek to search over policies. Our goal in the slide is to minimize the Expected value. We then discuss discount factor. The sooner that we get the reward is generally a better choice. Discount value implies that future rewards mean that costs tend to matter less, as seen in previous slides. Therefore, in a case where we find optimal policies, we tend to work in the discounted version of the equation - that is we search over the poliies that maxes or mins the total sum of the cost.

So, how do we solve? First, how NOT to solve: brute force. It simply takes far too long. We aren't doing bruteforce, because it takes far too long.

We use the *Bellman equation.*

Definition 1. *The Bellman equation is defined as:*

$$V^{\pi}(s_t) = \min_{a_t}(c(s_t, \pi(s_t)) + \gamma \mathbb{E}_{s_{t+1}} V^{\pi}(s_{t+1}))$$

Dr. Baheri wants us to recognize how gorgeous this equation is. Next, we discuss value iteraion and policy iteration.

4 Policy-Guided diffusion

Oxford paper has some components of offline RL and diffusion modeling

Understanding Deep Learning Chapter 8 VAE - how the diffusion model works

Understanding Deep Learning , Reinforcement learning, offline RL, decision transformers GNN, transformers,

4.1 SAFE POLICY GUIDED DIFFUSION

from constraint optimization perspective - safety in many terms from convex optimization

5 09-SEP-24

Quiz on Friday: value iteration, policy iteration, temporal difference, computational questions.