POLICY GRADIENT METHODS, CURVATURE, AND DISTRIBUTION SHIFT

SHAM KAKADE

These are notes from a talk given by Dr. Sham Kakade. The talk can be accessed **here**.

1. What is reinforcement learning?

Reinforcement learning is an area of machine learning that is concerned with the behaviour of agents who are concerned with maximizing utility. In a sense, it is paradigmatically divergent from **supervised** & **unsupervised** learning, which focus upon recognizing patterns within a given domain.

Reinforcement learning is studied by scholars schooled in a multitude of different fields. Game theorists, control theorists, researchers in operations research, and statisticians, among others, study reinforcement learning.

Basic reinforcement is modeled as a Markov decision process (MDP), consisting of four elements:

- A set of environment and agent states, $\mathcal{S}$. Can be either finite or infinite. We assume it's finite or countably infinite.
- A set of actions $\mathcal{A}$ of an agent. Can be either finite or finite, but we typically assume it's finite.
- A transition function $P : \mathcal{S} \times \mathcal{A} \to \delta(\mathcal{S})$, where $\delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$. $P\left(s' \,\middle|\, s, a \right)$. In other words, it's the *probability simplex.*
- $P_a(s, s') = \Pr\left(\mathcal{S}_{t+1} = s' \,\middle|\, \mathcal{S}_t = s, a_t = a\right)$
- A feward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. This is the agent's reward for taking action a at state s.
- A discount factor $\gamma \in [0, 1)$, which defines a horizon for the problem. Intuitively, γ considers the tiem-scale of the pay-off (are we rewarding delayed gratification? In which case, γ will be large. Are we rewarding immediate gratification? In which case, γ will be small.)
- An initial state distribution $\mu \in \delta(\mathcal{S})$, which specifies how the initial state s_0 is generated.

A Markov decision process occurs over discrete steps of time t with can be characterized with a state s_t and an agent's action a_t. The goal of the agent is to learn a policy $\pi : A \times S \to [0, 1]$, $\pi(a, s) = \Pr\left(a_t = a \,\middle|\, s_t = s \right)$

Date: November 2, 2020.

2. What are policy methods?

In a given MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, the agent interacts with the environment according to the following protocol: the agent starts at s_0 μ & at each time step $t = 0, 1, 2, \ldots$, the agent takes an action $a_t \in \mathcal{A}$, obtains the immediate reward $r_t(s_t, a_t)$, and observes the next state s_{t+1} sampled according to s_{t+1} $P\left(\cdot \,\middle|\, s_t, a_t \right)$

The interaction record τ at time t,

$$\tau_t = (s_0, a_0, r_1 \ldots s_t)5$$

is called the trajectory at time t and includes the state at time t, s_t.

A policy specifies a decision-making strategy in which the agent is informed by the history of their observations (in this sense, a Markov decision process should be considered conceptually different from a Markov chain in terms of agnosticity to history).

A policy is a (possibly randomized) mapping from a trajectory to an action. Thus,

$$\pi : \mathcal{H} \to \delta(\mathcal{A})$$

3. What was Dr. Kakade talking about?

References

[1] A. Agarwal, N. Jiang, S.M. Kakade, W. Sun; Reinforcement Learning: Theory and Algorithms; accessed October 27, 2020; available at `https://rltheorybook.github.io/`