Class 13

Shikhar Saxena

February 21, 2023

Contents

Reinforcement Learning

$P(s'|s,a)$ and/or $r(s,a,s')$ are not known any s, a, s'.

Model Based RL

- Go out in the real environment and estimate $P(s'|s,a)$.
- Naive Way: Fix a randomized policy that almost surely explores all s, a, s' combinations.
- Once the model's empirical probabilities $\hat{P}(s'|s,a)$ are robust, exploit the model built. Exploit essentially means treating this model as ground-truth, solve the underlying MDP to obtain $\hat{\pi}^*$ and hope that this is same as π^*.

But this requires a lot of computation. Essentially, can be used for game settings (where states are finite) because this approach doesn't make sense in a continuous state space.

Problems with this approach

1. Till we learn $\hat{P}$, we might incurr a lot of regret.
2.

Model Based RL: Certainty Equivalence

Combination of exploration and exploitation. We treat the current estimate $\hat{P}$ as the ground truth and always keep employing the optimal policy as per the current estimate i.e., $\hat{\pi}^*$. Keep refining the model and keep employing the best policy as per the current model.

Problems with this approach

1. It forces you to learn in a direction. because of which all state action pairs cannot be explored because of which accuracy for our estimates might be low.
2. You might converge to a locally optimal policy due to inefficient exploration.

Model Free RL

We don't bother learning the underlying model. Our prime interest is to directly learn V^π instead of the model. These methods try to learn $Q^f(s, a)$ directly for all (s, a) pairs. Such methods are called *value function based direct methods*. Example: Monte-Carlo methods, TD learning, Q-learning, actor-critic methods.

Actor-critic algorithm is based on policy-iteration while Q-learning and SARSA are based on value iteration. Some methods directly search for π^* in policy space. Example: Policy Gradient Method.