Reinforcement Learning Notes

Kiran Kannar

July 17, 2025

Contents

1 Introduction to Reinforcement Learning

Computational approach to **goal-directed** learning by an agent from interactions with an environment.

- Trial-and-error search
- Delayed reward
- Exploration vs exploitation
- Model-based vs model-free

Broadly, maximize reward signa l, despite uncertainty about the environment -¿ Optimization problem.

1.1 Elements of RL

A Markov Decision Process (MDP) involving:

- Policy : Mapping from states to actions
- Reward Signal : Scalar signal that indicates the **desirability** of the state i.e. immediate value
- Value Function : Expected cumulative reward from a state i.e. far-sighted judgment of value of starting from a particular state.
- Model (of the environment): Predicts the next state and reward given the current state and action; used for planning.

Value function is computed without **explicit search** over possible sequences of future states and actions. Focus is on highest value, and not on highest reward, even though value is computed from rewards.

2 Multi-Armed Bandit Problems

Setting: A single state (aka situation) with multiple actions, one of which must be selected repeatedly. Each arm is slot machine with a stationary probability distribution of rewards, say, $\mathcal{N}(\mu_i, \Sigma_i)$. The goal is to maximize the expected reward over some time horizon.

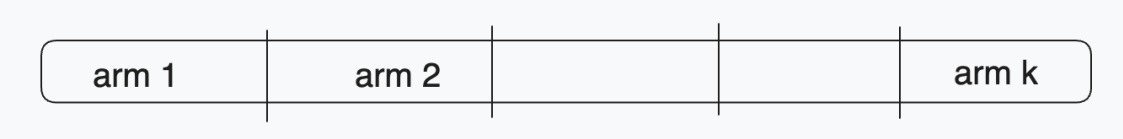

Figure 1: A single state with multiple actions.

The expected reward of taking an action is given by:

$$q_*(a) = \mathbb{E}[R_t | A_t = a] \tag{1}$$

This is the value of taking that action. We do not however know the value of each action, but we can estimate it by sampling the reward. Thus, we approximate the value of an action as $Q_t(a) \approx q_*(a)$.

2.1 Action-Value Methods

Methods to estimate the value of an action and then use that estimate to select actions.

2.1.1 Sample-Average Method:

Each estimate is the average of all the sample of rewards collected up to that point.
Suppose $Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{N_t(a)} R_i$, where $N_t(a)$ is the number of times action a has been selected up to time t.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot 1(A_i = a)}{\sum_{i=1}^{t-1} 1(A_i = a)} \tag{2}$$

Action Selection:

- Greedy: $A_t = \text{argmax}_a Q_t(a)$
- ϵ-greedy: $A_t = \begin{cases} \text{argmax}_a Q_t(a) & \text{with probability } 1 - \epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases}$

With ϵ-greedy, as $t \to \infty$, each action is sampled infinite times, and therefore the estimate $Q_t(a)$ converges to $q_*(a)$. The greedy selection, however, does not have strong convergence properties.

With 'k' arms in ϵ-greedy method, the probability of selecting the greedy action is,

$$\begin{aligned} P(\text{greedy action}) &= (1-\epsilon).P(\text{action is greedy}) + \epsilon.P(\text{action is random}) \\ &= (1-\epsilon).1 + \epsilon.\frac{1}{k} \\ &= 1 - \epsilon + \frac{\epsilon}{k} \end{aligned} \tag{3}$$

The expected reward of the greedy action is TODO