RL Cheat Sheet

Definitions

- **State(S):** Current Condition
- **Reward(R):** Instant Return from environment to appraise the last action
- **Value(V):** Expected Long-term reward with discount, as opposed to short-term reward R
- **Action-Value(Q):** This is similar to Value, except, it takes extra parameter, action A
- **Policy(π):** Approach of agent to determine next action based on current state
- **Exploitation** is about using already known info to maximize rewards
- **Exploration** is about exploring and capturing more information

Value Function $V(s)$

- Long term value of state S
- State value function $V(s)$ of a MRP is expected reward from state s
- $V(s) = E(G_t|S_t = s)$

Action Value Function $q(s, a)$

- $q_\pi(s,a) = E_\pi(G_t|S_t = s, A_t = a)$

Bellmen Equation

- $V(s) = R(s) + \gamma E_{s' \in s}[V(s')]$
- $V(s) = R(s) + \gamma \sum_{s' \in S} P_{ss'}(V(s'))$
- $V = R + \gamma PV \rightarrow V = (I - \gamma P)^{-1} R$

Discount Factor (γ)

- Varies between 0 to 1
- Closer to 0 → Agent tend to consider immediate reward
- Closer to 1→ Agent tend to consider future reward with greater

Q-Learning

4. Create reward matrix R where R_{sa} = reward for taking action a in state s and set γ parameter.
5. Initialize Q matrix to 0
6. Set initial random state and assign this to current state
7. Select One among all possible actions of current state
   a. Use this action to get new state
   b. Get maximum Q value for this state based on all previous actions
   c. Compute Q matrix using
   $$Q_{sa} = R_{sa} + \gamma \times Max[Q_{s'a'}]$$
   $$\forall\, s' accessible\ from\ s$$
   $$\forall a' available\ in\ s'$$
8. Repeat 4 until current state=goal state

Monte Carlo Policy Evaluation

1. To evaluate Value(s) $V_\pi(S)$
2. At any time step t when state s is visited in an episode
   a. Increment Counter N(s)<- N(s)+1
   b. Increment total return S(s) <- S(s)+G_t
3. Value estimated is mean $V(s) = S(s)/N(s)$

Policy Gradients

- $p_\theta(s_1, a_1, s_2, a_2 \dots) = p(s_1) \times \Pi_{t=1}^{T} p(s_{t+1}|s_t, a_t)\pi_\theta(s_t, a_t)$
- Goal is to $\theta^* = \arg\max_\theta E_{\tau \sim P_\theta(\tau)} [\sum_t r(s_t, a_t)] = \arg\max J(\theta)$
- $J(\theta) = \frac{1}{N} \sum_i \sum_t r(s_i, a_i)$
- $\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \times \nabla_\theta \log p_\theta(\tau)$
- $\nabla_\theta J_\theta = \frac{1}{N} \sum_{i=1}^{N} \{\sum_{t=1} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \sum_{t=1} r(s_{i,t}, a_{i,t})$
- $\theta \leftarrow \theta + \nabla_\theta J(\theta)$
- $\log \pi_\theta(a_{i,t}|s_{i,t})$ is the log probability of action, defines how likely are we going to see $a_{i,t}$ as action

Actor Critic

- Q-V = Advantage
- $Q^\pi(s_t, a_t) = \sum_{t'=t}^{T} E_{\pi_0}[r(s_t, a_t)|s_t, a_t]$ Reward of action a_t in s_t
- $V^\pi(s_t) = E_{a_t \sim \pi_\theta(a_t|s_t)}[Q^\pi(s_t, a_t)]$ total reward from st
- $A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$ how much better A_t is
- $\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\, A^\pi(s_t, a_t)$
- $Q^\pi(s_t, a_t) = r(s_t, a_t) + \sum_{t'=t+1}^{T} E_{\pi_\theta}[r(s'_t, a'_t)|s_t, a_t]$
  $= r(s_t, a_t) + E_{\pi_{t+1} \sim p(s_{t+1}|s_t, a_t)}[V^\pi(s_{t+1})]$
- $A^\pi(s_t, a_t) = r(s_t, a_t) + V^\pi(s_{t+1}) - V^\pi(s_t)$

Markov Process

- Consists of $< s, p >$ tuple where s are states and p is state transition matrix
- $P_{ss'} = P(s_{t+1} = s'|s_t = s)$
- $\mu_{t+1} = p^T \mu_t$ where $\mu_t = [\mu_{t,1} \dots \mu_{t,n}]^T$

Markov Reward Process

- Consists of $< s, p, R, \gamma >$ tuple where R is reward γ is discount
- $R = E[R_{t+1}|S_t = S] = R(s)$
- $G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0} \gamma^k R_{t+k+1}$ is Total discounted reward

Discounted Reward cons: Uncertainty may not be fully represented. Immediate rewards values> delayed, Avoid ∞ rewards in cycle

Markov Decision Process

- Consists of $< s, A, p, R, \gamma >$ tuple where A is action
- $P_{ss'} = P(S_{t+1} = S'|S_t = S, A_t = a)$

Discounted Reward cons: Uncertainty may not be fully represented. Immediate rewards values> delayed, Avoid ∞ rewards in cycle

Policy

- $\pi(a|s) = P(a_t = a|s_t = s)$
- Either deterministic or stochastics. In deterministic P=1 for one a_t
- $P_\pi(s'|s) = \sum \pi(a|s) \times P(s'|s, a)$ for stochastic process.
- One step expected reward $r_\pi = \sum_a \pi(a|s) r(s, a)$
- For rewards as function of transition states

$$r_\pi = \sum_a \pi(a|s) \sum_s \pi(a|s) \times \sum_{s'} P(s'|s, a) \times r(s, a, s')$$

Relation Between V_π and q_π

- $V_\pi(s) = \sum_{a \in A} \pi(a|s_t = s) q_\pi(s, a)$
- $V_\pi(s) = \sum_{a \in A} \pi(a|s) \times \{r(s, a) + \gamma \times \sum_{s' \in S} P(s'|s, a) V_\pi(s)\}$
- $V_\pi(s) = r(s) + \gamma \sum_{a \in A} \pi(a|s) \sum p(s'|s, a) \times V_\pi(s')$
- $q_\pi(s, a) = r(s, a) + \gamma \times \sum_{s' \in S} P(s'|s, a) V_\pi(s)$
- $q_\pi(s, a) = r(s, a) + \gamma \times \sum P(s'|s, a)\{\sum_{a' \in A} \pi(a'|s') q_\pi(s', a')\}$
- $q_\pi(s, a) = r(s, a) + \sum_{s \in S'} p(s'|s, a) \sum_{a' \in A} q_\pi(s', a') \times P(a'|s')$

Optimality Condition

- $V_\pi^*(s) = \max_\pi V_\pi(s)\ \forall s \in S$ similarly for $q_\pi^*(s, a)$