Reinforcement Learning from Scratch

Avani Mawal

January 2024

Involves interactions between an agent and its environment. Markov Decision Processes (MDPs) provide a mathematical framework for decision-making under uncertainty. Components of MDPs:

Comprise states, actions, transition function, reward signal, horizon, discount factor, and initial state distribution. State Space and Observability:

State space can be infinite or finite. States may be fully observable or partially observable in more general cases (POMDPs). Action Space:

Action space is a set of actions varying by state. Actions can have discrete or continuous variables. Temporal Aspects:

Horizon and discount factor introduce a time dimension in MDPs. POMDPs:

In partially observable MDPs, the agent observes a noisy state, not the full system state. Stationarity Assumption:

Transition function and reward signal are assumed to be stationary, with constant probabilities. Agent's Objective:

Objective is to maximize expected return over multiple episodes. Policies:

Universal plans prescribing actions for states. Can be deterministic or stochastic. Value Functions:

State-value functions summarize expected return from a state. Action-value functions summarize return from a state-action pair. Action-advantage functions show improvement over default for a specific action. Policy Evaluation and Improvement:

Policy evaluation estimates value functions from a policy. Policy improvement extracts a greedy policy from value functions. Policy iteration alternates between evaluation and improvement for optimal policies. Value Iteration and Generalized Policy Iteration:

Value iteration truncates policy-evaluation, entering policy-improvement early. Generalized policy iteration refines value estimates and improves policy iteratively. Exploration-Exploitation Trade-off:

Fundamental trade-off where exploration competes with exploitation for maximizing reward. Balancing Exploration and Exploitation:

Agents seek equilibrium considering environment uncertainty for effective decision-making. Exploration Strategies:

Epsilon-greedy, decaying epsilon-greedy, optimistic initialization, UCB, Thompson sampling, softmax. Strategies leverage estimates, uncertainty, or random selection for exploration. Regret as a Measure:

Regret measures how far agent's actions deviate from optimal. Prediction Problem:

Focus on estimating values of agents' behaviors. Introduces methods like Monte Carlo prediction and temporal-difference learning. Control Problem Ahead:

Next chapter addresses improving agents' behaviors in the control problem. Similar to the split between policy evaluation and improvement, understanding prediction and control problems separately enhances methods. Overall Understanding:

RL challenges arise from agents' inability to observe the underlying MDP. Challenges lead to a field where methods balance exploration-exploitation and handle sequential-evaluative feedback.]Reinforcement Learning Problem Overview: Reinforcement Learning (RL) Basics:

Involves interactions between an agent and its environment. Markov Decision Processes (MDPs) provide a mathematical framework for decision-making under uncertainty. Components of MDPs:

Comprise states, actions, transition function, reward signal, horizon, discount factor, and initial state distribution. State Space and Observability:

State space can be infinite or finite. States may be fully observable or partially observable in more general cases (POMDPs). Action Space:

Action space is a set of actions varying by state. Actions can have discrete or continuous variables. Temporal Aspects:

Horizon and discount factor introduce a time dimension in MDPs. POMDPs:

In partially observable MDPs, the agent observes a noisy state, not the full system state. Stationarity Assumption:

Transition function and reward signal are assumed to be stationary, with constant probabilities. Agent's Objective:

Objective is to maximize expected return over multiple episodes. Policies:

Universal plans prescribing actions for states. Can be deterministic or stochastic. Value Functions:

State-value functions summarize expected return from a state. Action-value functions summarize return from a state-action pair. Action-advantage functions show improvement over default for a specific action. Policy Evaluation and Improvement:

Policy evaluation estimates value functions from a policy. Policy improvement extracts a greedy policy from value functions. Policy iteration alternates between evaluation and improvement for optimal policies. Value Iteration and Generalized Policy Iteration:

Value iteration truncates policy-evaluation, entering policy-improvement early. Generalized policy iteration refines value estimates and improves policy iteratively. Exploration-Exploitation Trade-off:

Fundamental trade-off where exploration competes with exploitation for maximizing reward. Balancing Exploration and Exploitation:

Agents seek equilibrium considering environment uncertainty for effective decision-making. Exploration Strategies:

Epsilon-greedy, decaying epsilon-greedy, optimistic initialization, UCB, Thompson sampling, softmax. Strategies leverage estimates, uncertainty, or random selection for exploration. Regret as a Measure:

Regret measures how far agent's actions deviate from optimal. Prediction Problem:

Focus on estimating values of agents' behaviors. Introduces methods like Monte Carlo prediction and temporal-difference learning. Control Problem Ahead:

Next chapter addresses improving agents' behaviors in the control problem. Similar to the split between policy evaluation and improvement, understanding prediction and control problems separately enhances methods. Overall Understanding:

RL challenges arise from agents' inability to observe the underlying MDP. Challenges lead to a field where methods balance exploration-exploitation and handle sequential-evaluative feedback.