Simple RL Algorithm Comparisons

1. Q-Learning (Tabular)

Scenario: A robot exploring a 3-room house.

States: Kitchen, Living Room, Bathroom

Actions: Go Left, Go Right

Rewards: +1 in Kitchen, 0 elsewhere

You learn a table of expected rewards: Q[state][action] -> expected reward

Example: Q[Living Room][Go Left] = 0.8

Update Rule:

Q(s, a) <- Q(s, a) + alpha * [r + gamma * max_a' Q(s', a') - Q(s, a)]

Intuition: I thought this action was worth 0.2, but I got a reward of 1. I should increase my estimate!

2. DQN (Deep Q-Network)

Scenario: You're playing Atari (e.g., Breakout) with pixel inputs.

State: Image frame

Action: Discrete actions (left, right, no-op)

Model: A neural network predicting Q(s, a)

You learn a function that maps state to Q-values for all actions:

Q(s, a) -> predicted reward

Update Rule:

L = (Q(s, a) - [r + gamma * max_a' Q_target(s', a')])^2

Intuition: My model said this action gives 0.4 reward, but it's really 0.8. Let's update the weights to fix that.

3. PPO (Proximal Policy Optimization)

Scenario: A robot balancing a pole, learning which direction to push.

State: Position, velocity, angle, etc.

Simple RL Algorithm Comparisons

Action: Push left or right

Model: Policy network pi(a | s) + Value network

You learn a probability distribution over actions:

pi(a | s) -> action probabilities

Example: pi(push-left | state) = 0.6

Update Rule:

L = E_t[min(r_t(theta) * A_t, clip(r_t(theta), 1 - epsilon, 1 + epsilon) * A_t)]

Where r_t = pi_theta(a_t | s_t) / pi_theta_old(a_t | s_t)

Intuition: This action was better than expected - increase its probability, but not too much!

Summary Table

Q-Learning:

- Learns: Q-table (expected reward per action)
- Analogy: Robot exploring rooms
- Update: Adjust estimate if actual reward is different from expectation

DQN:

- Learns: Neural network for Q(s, a)
- Analogy: Atari agent
- Update: Predict future reward using a neural net

PPO:

- Learns: A policy (probabilities over actions)
- Analogy: Pole-balancing robot
- Update: Slightly adjust the policy based on action advantage