Assignment 3

Reinforcement Learning: Multi-Armed Bandits and Q-Learning

Airline Revenue Management

Background. An airline sells tickets for a single flight with 75 seats over a 180-day booking window. One customer arrives per day, and each customer has a willingness-to-pay that generally increases as the departure date approaches. You will compare three pricing strategies. The **Static** policy charges the same price every day. The **Bandit** (Thompson Sampling) policy selects one price *per flight* for all seats from five discrete tiers: $199, $249, $299, $349, and $499, learning which tier maximizes total flight revenue. The **Q-Learning** policy also chooses from these same five price tiers but conditions its decision on a state defined by two factors: remaining seat occupancy (bucketed as >70%, 30–70%, or <30% of seats remaining) and time to departure (bucketed as ≥60 days, 15–59 days, or <15 days), giving it a 3×3 = 9-state space.

Instructions. Build models using the [Week 3 web application] and answer the following questions. Stick to 3 sentences maximum per question. Upload your completed assignment to BruinLearn as a PDF by the end of the day (11:59pm PT).

Question 1. Experiment with the static price slider to find the price that maximizes average revenue over all 100 flights (Note: Increase the speed to 50x to run the simulation faster). Try running around 5 simulations to find the (roughly) optimal price. What price did you choose, and what is the average revenue per flight?

Question 2. The Bandit (Thompson Sampling) strategy considers the following price arms: $199, $249, $299, $349, $499. Hit reset and run the bandit algorithm for 100 flights. What price does it eventually settle on, and how does average revenue over all flights compare to your optimal price from Question 1? What are the advantages of the bandit approach over finding the optimal static price manually?

Question 3. Next, run Q-Learning (under default parameters) for 100 flights. Compare the average revenue over all flights with the revenue under the bandit from Question 2. Note the bandit approach picks one price per flight, whereas Q-Learning tracks seat inventory and days remaining. How does this additional state information change the optimal pricing strategy under Q-Learning?

Question 4. An airline executive wants to deploy Q-Learning pricing across all routes. Give two advantages over fixed pricing and two drawbacks or risks of real-world deployment.