National University
of computer and emerging sciences

FAST School of Computing

Department of Artificial Intelligence and Data Science

Course: RL

Semester: Fall 2025

Task # I

Course Instructors: Dr. Ahmad Din
Course TAs:

Max Marks= 15
Due Date: As mentioned on Google Classroom

Instructions:

- Plagiarism is strictly prohibited and may lead to academic penalties.
- Only the coordinator is required to submit the assignment.

Group Coordinator: Haider Niaz	**Reg.** # 22i-0481	**Section:** AI-A
Group Member 1: Abdullah Mazhar	**Reg.** # 22i-0622	**Section:**AI-B
Group Member 2: Abdullah Kaif Sheikh	**Reg.** # 22i-2142	**Section:** AI-A
Group Member 3: Katrina Bodani	**Reg.** # 22i-0545	**Section:** AI-B

1. **RL Topic Name:** Dynamic Retail Pricing via Q-Learning - A Reinforcement Learning Framework for Enhanced Revenue Management.
2. **Short Title:** Dynamic Retail Pricing via Q-Learning.
3. **Brief Overview of the RL Algorithm:**

Q-Learning is a model-free reinforcement learning algorithm that enables agents to learn optimal decision-making policies through trial-and-error interactions with an environment. Unlike supervised learning approaches that require labeled training data, Q-Learning discovers effective behaviors by exploring different actions and observing their outcomes without needing a predefined model of the environment's dynamics.

The algorithm operates by maintaining a Q-table that stores expected cumulative rewards for each state-action pair. These Q-values represent the quality of taking a particular action in a given state and are updated iteratively using the Bellman equation:

$$Q(\text{state, action}) = (1 - \alpha) \times Q(\text{state, action}) + \alpha \times (\text{reward} + \gamma \times \max Q(\text{next\_state}, a')).$$

The learning rate (α) determines how much weight to assign to new experiences versus existing knowledge, while the discount factor (γ) balances the importance of immediate rewards against future rewards.

Q-Learning employs an epsilon-greedy exploration strategy to balance exploration and exploitation. During exploration, the agent randomly selects actions to discover potentially better strategies, while during exploitation, it chooses actions with the highest known Q-values. This balance is crucial for avoiding local optima and finding globally optimal policies. As the agent interacts with the environment over many episodes, the Q-values gradually converge toward the optimal policy that maximizes cumulative rewards.

The algorithm's model-free nature makes it particularly valuable for complex, dynamic environments where building accurate predictive models is difficult or impossible. Q-Learning has been successfully applied across diverse domains including robotics, game playing, resource allocation, and autonomous systems, demonstrating its versatility as a reinforcement learning approach for sequential decision-making problems.

4. Brief Overview of the RL Algorithm + Application:

This paper applies Q-Learning to optimize dynamic pricing strategies in retail environments, addressing limitations of traditional operations research methods that rely on static demand models and require frequent manual updates. The authors model retail pricing as a Markov Decision Process where the agent (pricing system) must learn optimal price points to maximize profit.

The environment is defined by states representing product type and day of the week (weekday/weekend) to capture different demand patterns, actions consisting of discrete price points the retailer can set for products, and rewards calculated as profit:

$$(\text{Price} \times \text{Demand}) - (\text{Cost} \times \text{Demand}).$$

The demand function incorporates price elasticity using the formula:
Demand = Base Demand + (Base Demand × Elasticity × (Price - Base Price) / Base Price), capturing consumer sensitivity to price changes.

The Q-Learning agent begins with zero-initialized Q-values for all state-action pairs. During training, it runs multiple episodes where it selects prices using an epsilon-greedy strategy, observes the resulting demand and profit, and updates its Q-table using:

$$Q(\text{state, action}) = (1 - \alpha) \times Q(\text{state, action}) + \alpha \times (\text{reward} + \gamma \times \max Q(\text{next\_state}, a')).$$

The learning rate (α) controls how quickly the agent adapts to new market information, while the discount factor (γ) determines the weight given to future profits. Through repeated interactions, the agent learns which prices maximize profit under different market conditions.

Testing on over 15,000 electronic products from Datafiniti's database demonstrated Q-Learning's superior adaptability compared to traditional scipy.optimize methods. For the Samsung 49" 4K Q6F, Q-Learning achieved optimal demand of 101.5 units at $820.3 versus traditional optimization's 260.1 units at $509.5. The Samsung 65" 4K Q7F showed 285.0 units at $1253.6 with Q-Learning compared to 264.3 units at $1360.2 using conventional approaches.

Key advantages of the RL approach include automatic adaptation to changing market conditions without human intervention, no dependency on extensive historical datasets, ability to discover complex price-demand relationships through exploration, and improved revenue generation through flexible pricing strategies. Unlike traditional methods that optimize based on fixed assumptions, Q-Learning continuously learns from market feedback, making it robust to unexpected shifts in consumer behavior or competitive dynamics.

5. Overview of Paper # 1 which has the RL topic + Application

Title of the paper 1[1]: Dynamic Retail Pricing via Q-Learning - A Reinforcement Learning Framework for Enhanced Revenue Management

Review of the paper

This paper presents a reinforcement learning approach using Q-Learning to optimize dynamic pricing in retail, comparing it against traditional operations research methods. The authors argue that while conventional pricing strategies rely on static demand models and require frequent manual updates, Q-Learning offers a more adaptive solution that continuously learns from market interactions.

Methodology and Key Findings

The authors implement Q-Learning in a simulated retail environment incorporating base demand, price elasticity, and operational costs. The algorithm learns optimal pricing policies by maintaining a Q-table that maps state-action pairs to expected rewards, updated through the Bellman equation. States represent product types and days (weekday/weekend), actions are discrete price points, and rewards equal profit. Using an epsilon-greedy exploration strategy, the agent balances trying new prices against exploiting known optimal ones across multiple learning episodes.

The study evaluates performance using a dataset of over 15,000 electronic products from Datafiniti's database, comparing Q-Learning results against traditional scipy.optimize methods. Results show Q-Learning frequently achieves better demand optimization—for example, the Samsung 65" 4K Q7F reached 285.0 units demand at $1253.6 compared to traditional methods' 264.3 units at $1360.2. The authors conclude that reinforcement learning surpasses traditional approaches in revenue generation and pricing flexibility.

Strengths

The paper addresses a practical problem with clear business relevance, demonstrating RL's applicability beyond theoretical settings. The comparison with traditional optimization methods provides useful context for understanding Q-Learning's advantages. The use of real product data lends credibility to the findings, and the methodology is clearly explained with appropriate mathematical formulations. The results effectively illustrate RL's ability to adapt to market dynamics without manual intervention.

Weaknesses and Limitations

Several significant limitations undermine the paper's contributions. First, the evaluation relies entirely on a simulated environment rather than real-world deployment, raising questions about practical applicability. The demand function used (Equation 1) appears overly simplistic, assuming linear price elasticity without accounting for competitor pricing, seasonality, inventory constraints, or other market complexities that retailers face.

The comparison methodology is problematic—the paper doesn't clearly explain whether traditional methods and Q-Learning use identical assumptions or if Q-Learning benefits from additional information. Some results seem counterintuitive (e.g., Sony 43" 4K UHD showing 203.8 vs. 506.9 units demand), suggesting potential issues with the experimental setup or demand model.

The paper lacks critical implementation details: training duration, convergence criteria, hyperparameter selection process, and computational costs. No sensitivity analysis examines how results vary with different elasticity values or market conditions. The state space (only product type and day) seems insufficient for capturing real pricing complexities. Additionally, there's no discussion of exploration-exploitation tradeoffs during deployment or how the system would handle new products without historical data.

Overall Contribution

While the paper demonstrates Q-Learning's potential for dynamic pricing, the contribution is primarily pedagogical rather than advancing the state-of-the-art. The simplified simulation and lack of real-world validation limit practical insights. Future work should address deployment challenges, incorporate richer market dynamics, and validate findings in actual retail environments.

6. References

[1] M. Apte, K. Kale, P. Datar, and P. R. Deshmukh, "*Dynamic Retail Pricing via Q-Learning - A Reinforcement Learning Framework for Enhanced Revenue Management*," arXiv preprint arXiv:2411.18261v1, Nov. 2024.