DEPARTMENT OF COMPUTER SCIENCE
ST. FRANCIS XAVIER UNIVERSITY

St. Francis Xavier University
Department of Computer Science

CSCI-531 - Reinforcement Learning
Temporal Difference Learning

Part I: TD Fundamentals

1. **TD Update Rule Derivation**
   (a) Starting from the value function definition $V(s) = \mathbb{E}[G_t|S_t = s]$, show how the TD target $r_{t+1} + \gamma V(s_{t+1})$ is derived.
   (b) Write the complete TD(0) update equation and explain each component.
   (c) Explain why TD learning is called "bootstrapping" and compare it to Monte Carlo.

2. **TD Error Analysis**

   Consider a simple 2-state chain: Start $\to$ Goal, where the agent receives a reward of +5 when reaching the Goal. $\gamma = 0.9$ and $\alpha = 0.1$.

   Current value estimates: $V(\text{Start}) = 3.0$, $V(\text{Goal}) = 0$ (terminal)

   (a) Calculate the TD error when transitioning from Start to Goal.
   (b) What will be the new value estimate for the Start state after this update?
   (c) After many episodes, what should $V(\text{Start})$ converge to and why?

Part II: TD Prediction Algorithms

3. **TD(0) Algorithm Implementation**

   Consider a 3-state Markov chain with states $A \to B \to C$ (terminal).

   - Transition $A \to B$ gives reward $r = 2$
   - Transition $B \to C$ gives reward $r = 8$

- $\gamma = 0.8$, $\alpha = 0.5$
- Initial estimates: $V(A) = 0$, $V(B) = 0$, $V(C) = 0$

(a) Trace through the first episode ($A \to B \to C$) showing all TD updates.

(b) Show the updates for the second episode with the new value estimates.

(c) What are the true values for this chain? Compare with your TD estimates.

4. **Learning Rate Effects**

Using the same 3-state chain from Question 4, compare TD learning with different learning rates.

(a) Calculate $V(A)$ after the first episode with $\alpha = 0.1$ and $\alpha = 0.9$.

(b) Discuss the trade-offs between high and low learning rates in TD learning.

Part III: TD Control - SARSA

5. **SARSA Algorithm Understanding**

(a) Write the SARSA update equation and explain why it's called "SARSA".

(b) Explain why SARSA is considered an "on-policy" method.

(c) What is the role of ϵ-greedy action selection in SARSA?

6. **SARSA Numerical Example**

Consider a simple 2-state MDP with states $\{S_1, S_2\}$ and actions $\{a_1, a_2\}$:

- From S_1: action a_1 goes to S_2 with reward +1, action a_2 stays in S_1 with reward +0
- From S_2: both actions return to S_1 with reward +2
- $\gamma = 0.9$, $\alpha = 0.5$, $\epsilon = 0.1$

Initial Q-values: $Q(s, a) = 0$ for all state-action pairs.

(a) Given the sequence $(S_1, a_1, +1, S_2, a_2)$, calculate the SARSA update for $Q(S_1, a_1)$.

(b) If we're in state S_2 with current Q-values $Q(S_2, a_1) = 1.5$ and $Q(S_2, a_2) = 2.0$, what action would ϵ-greedy select with $\epsilon = 0.1$?

(c) Explain how SARSA would behave differently from Q-learning in a "cliff walking" environment.

Part IV: TD Control - Q-Learning

7. **Q-Learning Algorithm**

(a) Write the Q-learning update equation and identify the key difference from SARSA.

(b) Explain why Q-learning is "off-policy" and what this means for learning.

(c) Under what conditions does Q-learning converge to the optimal action-value function?

8. **Q-Learning vs SARSA Comparison**

(a) For the same 2-state MDP from Question 7, calculate the Q-learning update for the sequence $(S_1, a_1, +1, S_2)$ assuming $Q(S_2, a_1) = 1.5$ and $Q(S_2, a_2) = 2.0$.

(b) Create a small example where SARSA and Q-learning would learn different policies.

(c) When would you choose SARSA over Q-learning and vice versa?

Part V: Advanced TD Concepts

9. **Bias-Variance Trade-off in TD Learning**
   (a) Compare the bias and variance characteristics of Monte Carlo, TD(0), and Dynamic Programming.
   (b) Explain why TD learning often converges faster than Monte Carlo in practice.
10. **TD Learning Applications**
    (a) Design a TD learning approach for a robot navigation problem. Specify states, actions, rewards, and explain your choices.
    (b) Discuss potential challenges and solutions when applying TD learning to this problem.
11. **Theoretical Understanding**
    (a) Prove that the TD error can be written as the sum of changes in value estimates along a trajectory.
    (b) Explain the relationship between TD(0), TD(1), and Monte Carlo methods.