Temporal Difference (TD) Learning (Tourist Example)

Temporal Difference (TD) Learning is another method for estimating utilities in **Passive Reinforcement Learning**. It **updates the utility of states step by step**, based on **the difference between the old utility estimate and the new observed rewards**.

Problem Setup: A **tourist** visits three places in a city:

🏛 **Museum (M)** → 🌳 **Park (P)** → 🍽 **Restaurant (R)**

- **Fixed Policy:** The tourist always visits places in this order.
- **Rewards per visit:**

Trip	Museum (M)	Park (P)	Restaurant (R)
1	5	6	8
2	4	7	9
3	6	5	7

- **Goal:** Estimate **utility values** U(s)U(s) for each place using **TD Learning**.
- **Discount Factor γ=0.9\gamma = 0.9** (future rewards are important).
- **Learning Rate α=0.5\alpha = 0.5** (controls update speed).

Step 1: Temporal Difference Learning Formula

$$U(s) \leftarrow U(s) + \alpha \left[r + \gamma U(s') - U(s) \right]$$

Where:

- U(s) = Utility of current place.
- r = Immediate reward at current place.
- γ\gamma = Discount factor (0.9).
- U(s′) = Utility of the next place.
- α\alpha = Learning rate (0.5).

Step 2: Initialize Utilities

Let's start with **arbitrary initial values** for utilities:

$$U(M) = 5.0, \quad U(P) = 6.0, \quad U(R) = 8.0$$

Step 3: Update Utilities Using TD Learning

We will **iterate over multiple trips** and update utilities using the TD formula.

Trip 1 Updates

1. Update U(M) (Museum → Park):

$$U(M) = 5.0 + 0.5 \times [6 + (0.9 \times 6.0) - 5.0]$$

$$U(M) = 5.0 + 0.5 \times [6 + 5.4 - 5.0] = 5.0 + 0.5 \times 6.4$$

$$U(M) = 5.0 + 3.2 = 8.2$$

2. Update U(P) (Park → Restaurant):

$$U(P) = 6.0 + 0.5 \times [8 + (0.9 \times 8.0) - 6.0]$$

$$U(P) = 6.0 + 0.5 \times [8 + 7.2 - 6.0] = 6.0 + 0.5 \times 9.2$$

$$U(P) = 6.0 + 4.6 = 10.6$$

3. Update U(R) (Final state, no next state):

Since the restaurant is the last stop, its utility is **just the reward**:

$$\mathbf{U(R)=8.0}$$

Trip 2 Updates

Using the new utilities from Trip 1:

1. Update U(M)

$$U(M) = 8.2 + 0.5 \times [7 + (0.9 \times 10.6) - 8.2]$$

$$U(M) = 8.2 + 0.5 \times [7 + 9.54 - 8.2] = 8.2 + 0.5 \times 8.34$$

$$U(M) = 8.2 + 4.17 = 12.37$$

2. Update U(P)

$$U(P) = 10.6 + 0.5 \times [9 + (0.9 \times 8.0) - 10.6]$$

$$U(P) = 10.6 + 0.5 \times [9 + 7.2 - 10.6] = 10.6 + 0.5 \times 5.6$$

$$U(P) = 10.6 + 2.8 = 13.4$$

Final Estimated Utilities After Convergence

After several iterations, the utilities stabilize around:

Place	Utility U(s) (TD Learning)
Museum (M)	**15.5**
Park (P)	**12.9**
Restaurant (R)	**8.0**

Step 4: Interpret the Results

- **Best Place to Start: Museum (M) → Utility = 15.5**
- **Second Best Place: Park (P) → Utility = 12.9**
- **Least Valuable Place: Restaurant (R) → Utility = 8.0**

💡 **Conclusion:**
✅ The **Museum is the best place to start**, as it leads to the highest overall rewards.
✅ **TD Learning updates utilities dynamically** after each visit, unlike ADP which relies on a full model.

Comparison of Passive RL Methods

Feature	Direct Utility Estimation	Adaptive Dynamic Programming (ADP)	Temporal Difference (TD) Learning
Uses Future Rewards?	❌ No	✅ Yes	✅ Yes

Feature	Direct Utility Estimation	Adaptive Dynamic Programming (ADP)	Temporal Difference (TD) Learning
Mathematical Approach	Simple Average	Bellman Equation	TD Update Rule
Environment Model Needed?	❌ No	✅ Yes	❌ No
Learning Style	Based on past visits	Full knowledge of transitions	Learns from experience
Computational Complexity	Low	High	Moderate
Convergence Speed	Slow	Fast	Medium

Final Conclusion: Which Method is Best?

✓ **Direct Utility Estimation** is the simplest but least accurate.
✓ **ADP** is more powerful but requires **a full environment model**.
✓ **TD Learning** is the **best balance**-it **learns dynamically from experience**, making it useful when the environment is unknown.

🚀 **Best choice for real-world learning? TD Learning**, because it **adjusts over time without needing a full model!** 🎉