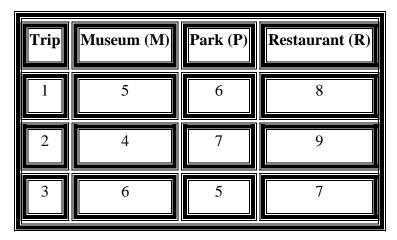
# **Temporal Difference (TD) Learning (Tourist Example)**

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:

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



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

# **Step 1: Temporal Difference Learning Formula**

$$U(s) \leftarrow U(s) + lpha \left[ r + \gamma U(s') - U(s) 
ight]$$

## Where:

- U(s) = Utility of current place.
- r = Immediate reward at current place.
- $\gamma$ \gamma = Discount factor (0.9).
- U(s') = Utility of the next place.
- $\alpha \cdot \text{alpha} = \text{Learning rate } (0.5).$

# **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  $\rightarrow$  Park):

$$U(M) = 5.0 + 0.5 imes [6 + (0.9 imes 6.0) - 5.0]$$
  $U(M) = 5.0 + 0.5 imes [6 + 5.4 - 5.0] = 5.0 + 0.5 imes 6.4$   $U(M) = 5.0 + 3.2 = 8.2$ 

2. Update U(P) (Park  $\rightarrow$  Restaurant):

$$U(P) = 6.0 + 0.5 imes [8 + (0.9 imes 8.0) - 6.0]$$
  $U(P) = 6.0 + 0.5 imes [8 + 7.2 - 6.0] = 6.0 + 0.5 imes 9.2$   $U(P) = 6.0 + 4.6 = 10.6$ 

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

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

$$U(R) = 8.0$$

#### **Trip 2 Updates**

Using the new utilities from Trip 1:

1. Update U(M)

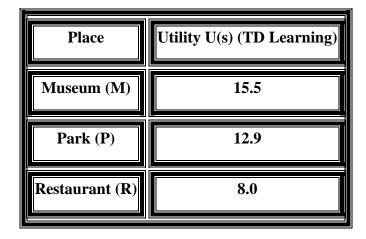
$$U(M) = 8.2 + 0.5 imes [7 + (0.9 imes 10.6) - 8.2]$$
  $U(M) = 8.2 + 0.5 imes [7 + 9.54 - 8.2] = 8.2 + 0.5 imes 8.34$   $U(M) = 8.2 + 4.17 = 12.37$ 

# 2. Update U(P)

$$U(P) = 10.6 + 0.5 imes [9 + (0.9 imes 8.0) - 10.6]$$
  $U(P) = 10.6 + 0.5 imes [9 + 7.2 - 10.6] = 10.6 + 0.5 imes 5.6$   $U(P) = 10.6 + 2.8 = 13.4$ 

# **Final Estimated Utilities After Convergence**

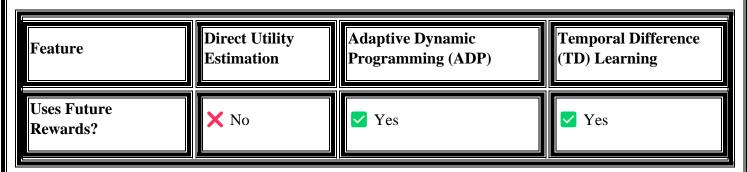
After several iterations, the utilities stabilize around:

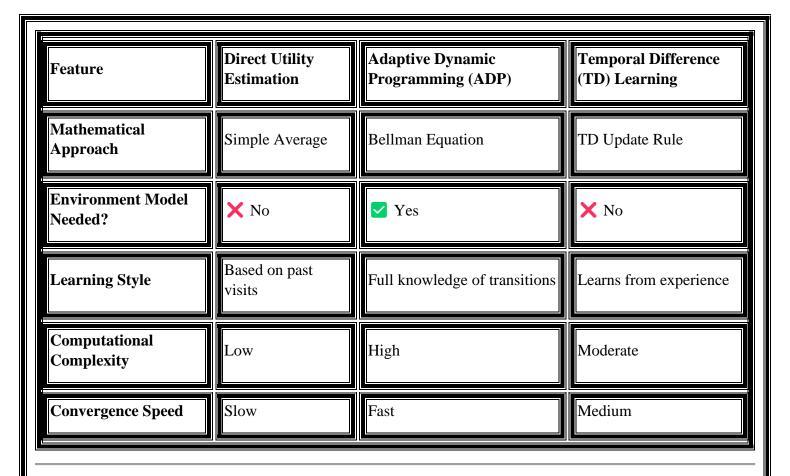


## **Step 4: Interpret the Results**

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





#### **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! ★**