1. Your TD algorithm generates the following episode using policy \( \pi \) when interacting with its environment. This is the first episode that has been generated.

<table>
<thead>
<tr>
<th>Timestep</th>
<th>Reward</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>S1</td>
<td>A1</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>S1</td>
<td>A2</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>S1</td>
<td>A1</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>S1</td>
<td>A1</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Assume the discount factor, \( \gamma \), is \( \frac{1}{2} \), the step size, \( \alpha \), is 0.1, and that all \( q_\pi \) values are currently 0.

What are the estimates of: \( q_\pi(S_1, A_1) \) and \( q_\pi(S_1, A_2) \) assuming the use of 2-step Sarsa? Show your work.

**Solution:** Let’s look at the rewards:

After the second timestep,

\[
q_\pi(S_0, A_0) = q_\pi(S_0, A_0) + \alpha[R_1 + \gamma R_2 + \gamma^2 q_\pi(S_2, A_2) - q_\pi(S_0, A_0)] \\
q_\pi(S_1, A_1) = q_\pi(S_1, A_1) + \alpha[R_1 + \gamma R_2 + \gamma^2 q_\pi(S_1, A_1) - q_\pi(S_1, A_1)] \\
= 0 + .1[16 + .5 \times 12 + .5^2 \times 0 - 0] \\
= 0 + .1[16 + 6] \\
= 2.2
\]

After the third timestep:

\[
q_\pi(S_1, A_1) = q_\pi(S_1, A_1) + \alpha[R_2 + \gamma R_3 + \gamma^2 q_\pi(S_3, A_3) - q_\pi(S_1, A_1)] \\
q_\pi(S_1, A_2) = q_\pi(S_1, A_2) + \alpha[R_2 + \gamma R_3 + \gamma^2 q_\pi(S_1, A_1) - q_\pi(S_1, A_2)] \\
= 0 + .1[12 + .5 \times 24 + .5^2 \times 2.2 - 0] \\
= 0 + .1[12 + 12 + 0.55] \\
= 2.455
\]

After the fourth timestep:

\[
q_\pi(S_2, A_2) = q_\pi(S_2, A_2) + \alpha[R_3 + \gamma R_4 + \gamma^2 q_\pi(S_4, A_4) - q_\pi(S_2, A_2)] \\
q_\pi(S_1, A_1) = q_\pi(S_1, A_1) + \alpha[R_3 + \gamma R_4 + \gamma^2 q_\pi(T, \circ) - q_\pi(S_1, A_1)] \\
= 2.2 + .1[24 + .5 \times 16 + .5^2 \times 0 - 2.2] \\
= 2.2 + .1[24 + 8 + 0 - 2.2] \\
= 2.2 + 2.98 \\
= 5.18
\]
After the fifth timestep:

\[
q_\pi(S_3, A_3) = q_\pi(S_3, A_3) + \alpha[R_4 + \gamma q_\pi(S_4, A_4) - q_\pi(S_3, A_3)]
\]

\[
q_\pi(S_1, A_1) = q_\pi(S_1, A_1) + \alpha[16 + \gamma q_\pi(T, \odot) - q_\pi(S_1, A_1)]
\]

\[
= 5.18 + 0.1[16 + 0.5 \times 0 - 5.18]
\]

\[
= 5.18 + 0.1[16 - 5.18]
\]

\[
= 5.18 + 1.082
\]

\[
= 6.262
\]

So, our estimates are: \(q_\pi(S_1, A_1) = 6.262, q_\pi(S_1, A_2) = 2.455\).