Reinforcement Learning Training 2025

Bellman's Equation

- Foundation to solving MDP and RL problems.

Recall (1)

- Markov decision process has transition probabilities

$$\mathtt{P}\Big[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\Big]$$

which transitions the agen to state S_{t+1} and a reward of R_{t+1}

- Cumulative reward at time t

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

Recall (2)

- A value function is an expected cumulative reward

$$v_\pi(s) = \mathtt{E}_\pi[G_t|S_t = s]$$

- A action-value value function is an expected cumulative reward from taking action a

$$q_\pi(s, a) = E_\pi[G_t|S_t = s, A_t = a]$$

> Note that v and q depend on the policy π.

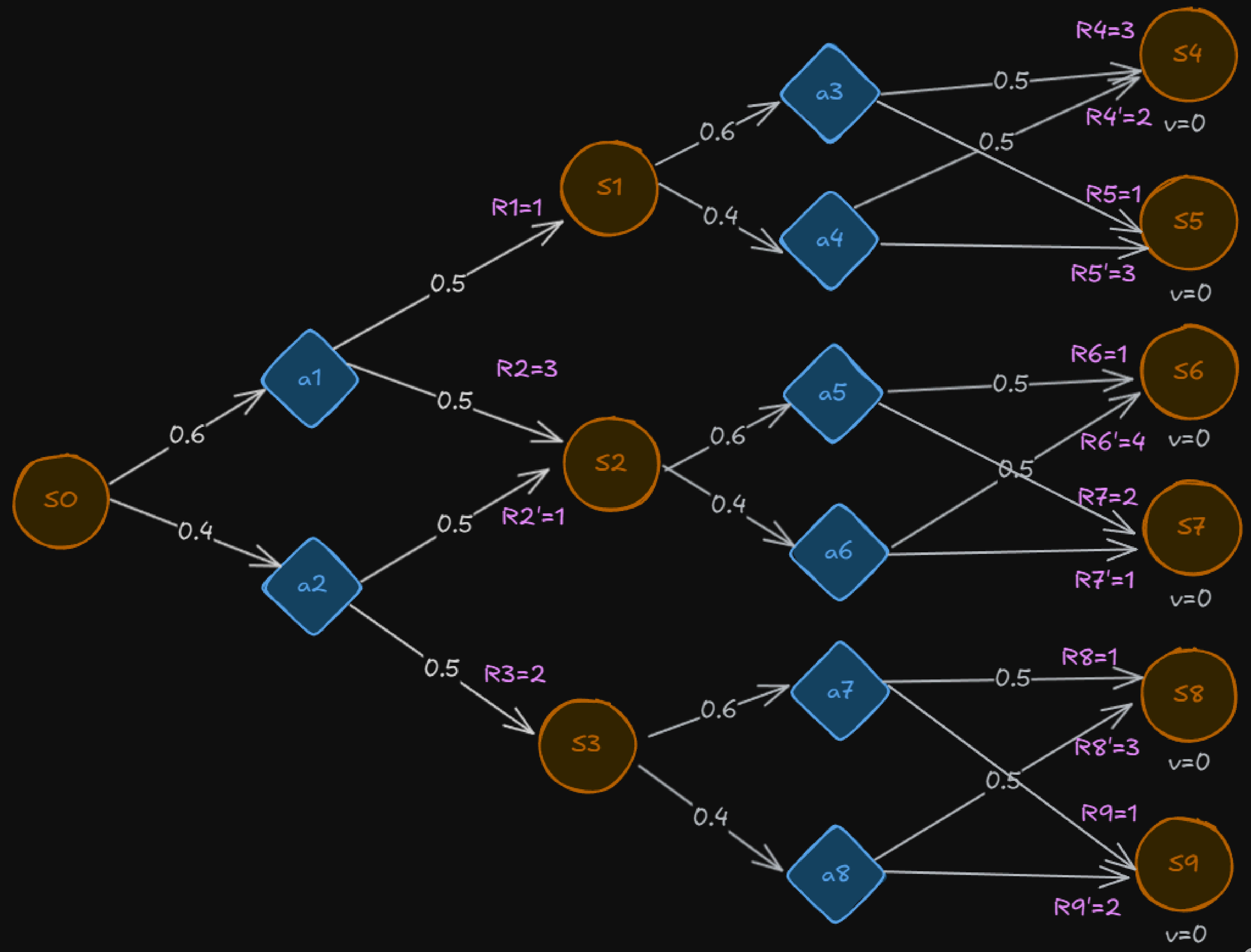
Bellman Equation

- Allows relationships among v and q.

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma\, v_\pi(s')\right]$$

$$q_\pi(s, a) = \sum_{s', r} p(s', r | s, a) \left[r + \gamma \sum_{a'} \pi(a'|s')\, q(s', a') \right]$$

Example



s0
0.6
0.4
a1
a2
0.5
0.5
0.5
0.5
R1=1
R2=3
R2'=1
R3=2
s1
s2
s3
0.6
0.4
0.6
0.4
0.6
0.4
a3
a4
a5
a6
a7
a8
0.5
0.5
0.5
0.5
0.5
0.5
R4=3
R4'=2
R5=1
R5'=3
R6=1
R6'=4
R7=2
R7'=1
R8=1
R8'=3
R9=1
R9'=2
s4
s5
s6
s7
s8
s9
v=0
v=0
v=0
v=0
v=0
v=0

Find $v(s_1)$

$$V(s_1) = 0.6\left[0.5 \times \left(R_4 + \gamma \overset{0}{\cancel{V(s_4)}}\right) + 0.5 \times \left(R_5 + \gamma \overset{0}{\cancel{V(s_5)}}\right)\right]$$

$$+ 0.4\left[0.5 \times \left(R_4' + \gamma \overset{0}{\cancel{V(s_4)}}\right) + 0.5 \times \left(R_5' + \gamma \overset{0}{\cancel{V(s_5)}}\right)\right]$$

$$V(s_1) = 0.6\left[0.5(3) + 0.5(1)\right] + 0.4\left[0.5(2) + 0.5(3)\right]$$

$$= 2.2$$

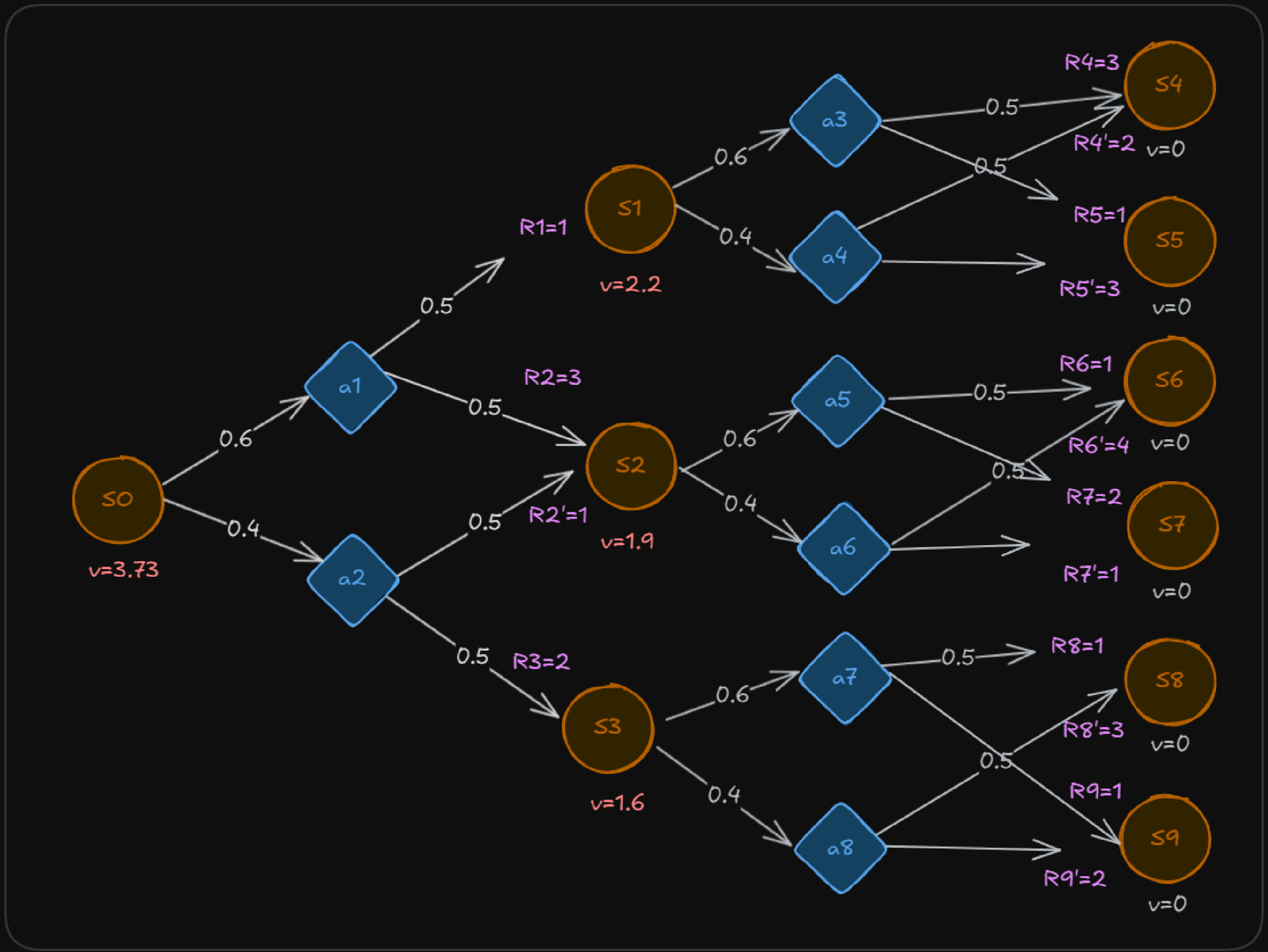
Find $v(s_2)$ and $v(s_3)$

$$V(S_2) = 0.6\,[\,0.5(1) + 0.5(2)\,] + 0.4\,[\,0.5(4) + 0.5(1)\,] = 1.9$$
$$V(S_3) = 0.6\,[\,0.5(1) + 0.5(1)\,] + 0.4\,[\,0.5(3) + 0.5(2)\,] = 1.6$$

Find $v(s_0)$

$$V(S_0) = 0.6\left[0.5(1+2.2) + 0.5(3+1.9)\right]$$
$$+0.4\left[0.5(1+1.9) + 0.5(2+1.6)\right]$$

$$= 0.6(4.25) + 0.4(3.25)$$
$$= 3.73$$


s0
v=3.73
0.6
0.4
a1
a2
0.5
0.5
0.5
0.5
R1=1
R2=3
R2'=1
R3=2
s1
v=2.2
s2
v=1.9
s3
v=1.6
0.6
0.4
0.6
0.4
0.6
0.4
a3
a4
a5
a6
a7
a8
0.5
0.5
0.5
0.5
0.5
0.5
R4=3
R4'=2
R5=1
R5'=3
R6=1
R6'=4
R7=2
R7'=1
R8=1
R8'=3
R9=1
R9'=2
s4
v=0
s5
v=0
s6
v=0
s7
v=0
s8
v=0
s9
v=0

q(a_3)

$$q(a_3) = 0.5\left[R_4 + \gamma \sum_{a'} \overset{0\ \text{(No action)}}{\cancel{(\ldots)}}\right] + 0.5\left[R_5 + \gamma \sum_{a'} \overset{0}{\cancel{(\ldots)}}\right]$$

$$= 0.5(3) + (0.5)(1)$$

$$= 2$$

q(a_4) - q(a_8)

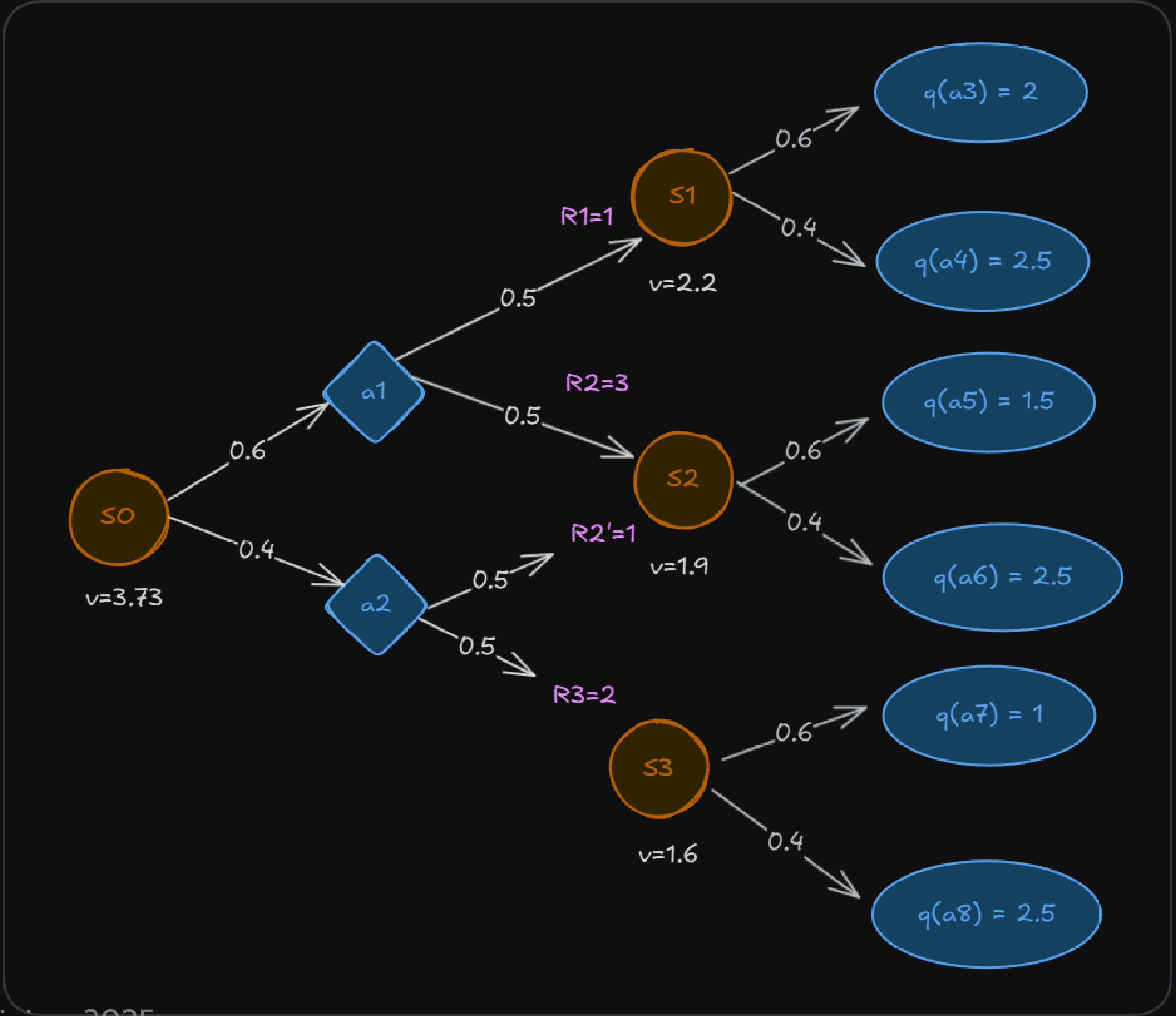
$$q(a_4) = 0.5(2) + 0.5(3) = 2.5$$

$$q(a_5) = 0.5(1) + 0.5(2) = 1.5$$

$$q(a_6) = 0.5(4) + 0.5(1) = 2.5$$

$$q(a_7) = 0.5(1) + 0.5(1) = 1$$

$$q(a_8) = 0.5(3) + 0.5(2) = 2.5$$


s0
v=3.73
0.6
0.4
a1
a2
0.5
0.5
0.5
0.5
R1=1
R2=3
R2'=1
R3=2
s1
v=2.2
s2
v=1.9
s3
v=1.6
0.6
0.4
0.6
0.4
0.6
0.4
q(a3) = 2
q(a4) = 2.5
q(a5) = 1.5
q(a6) = 2.5
q(a7) = 1
q(a8) = 2.5

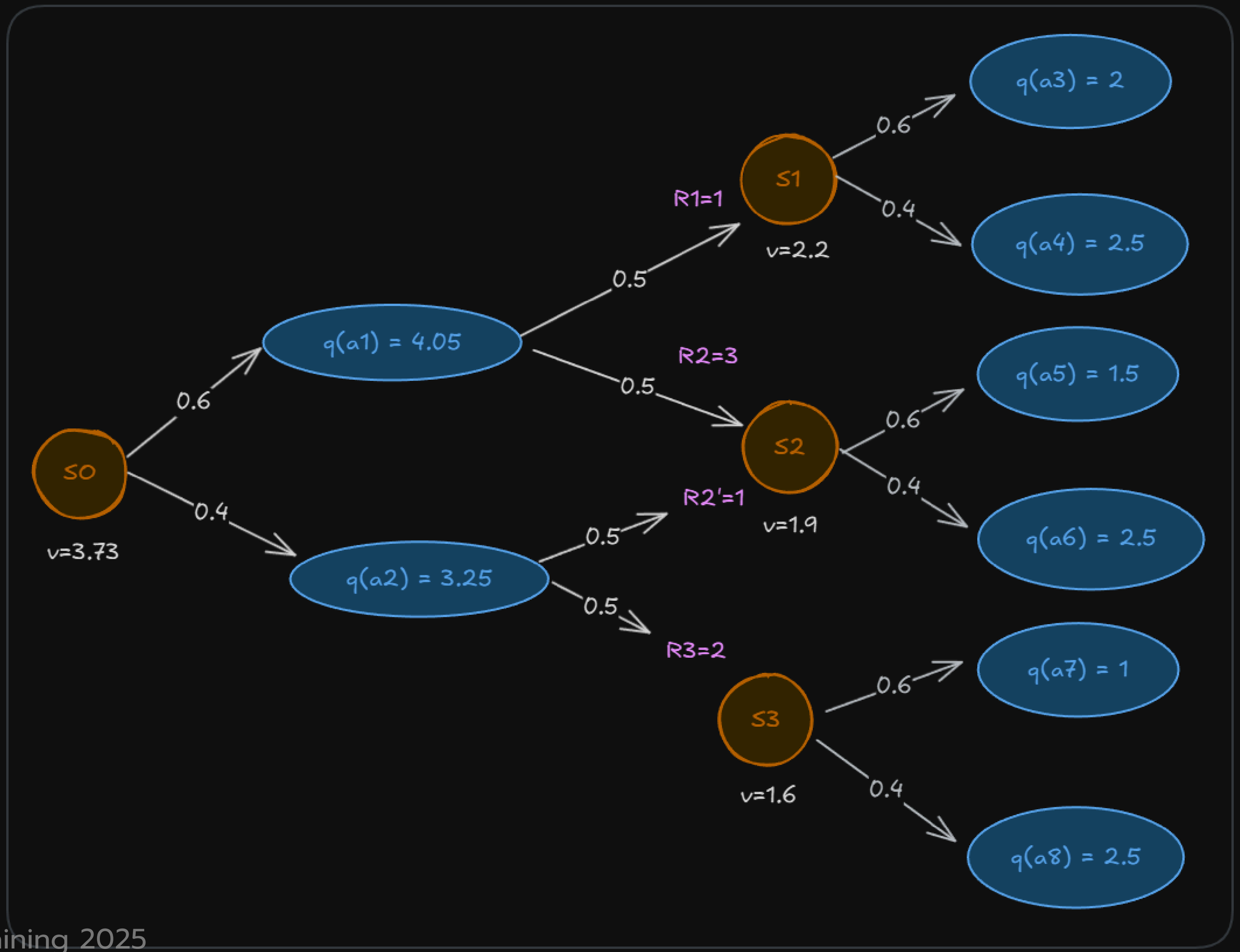
q(a_1) - q(a_2)

$$q(a_1) = 0.5\left[1 + \cancelto{1}{\gamma}\left(0.6(2) + 0.4(2.5)\right)\right] + 0.5\left[3 + \cancelto{1}{\gamma}\left(0.6(1.5) + 0.4(2.5)\right)\right]$$

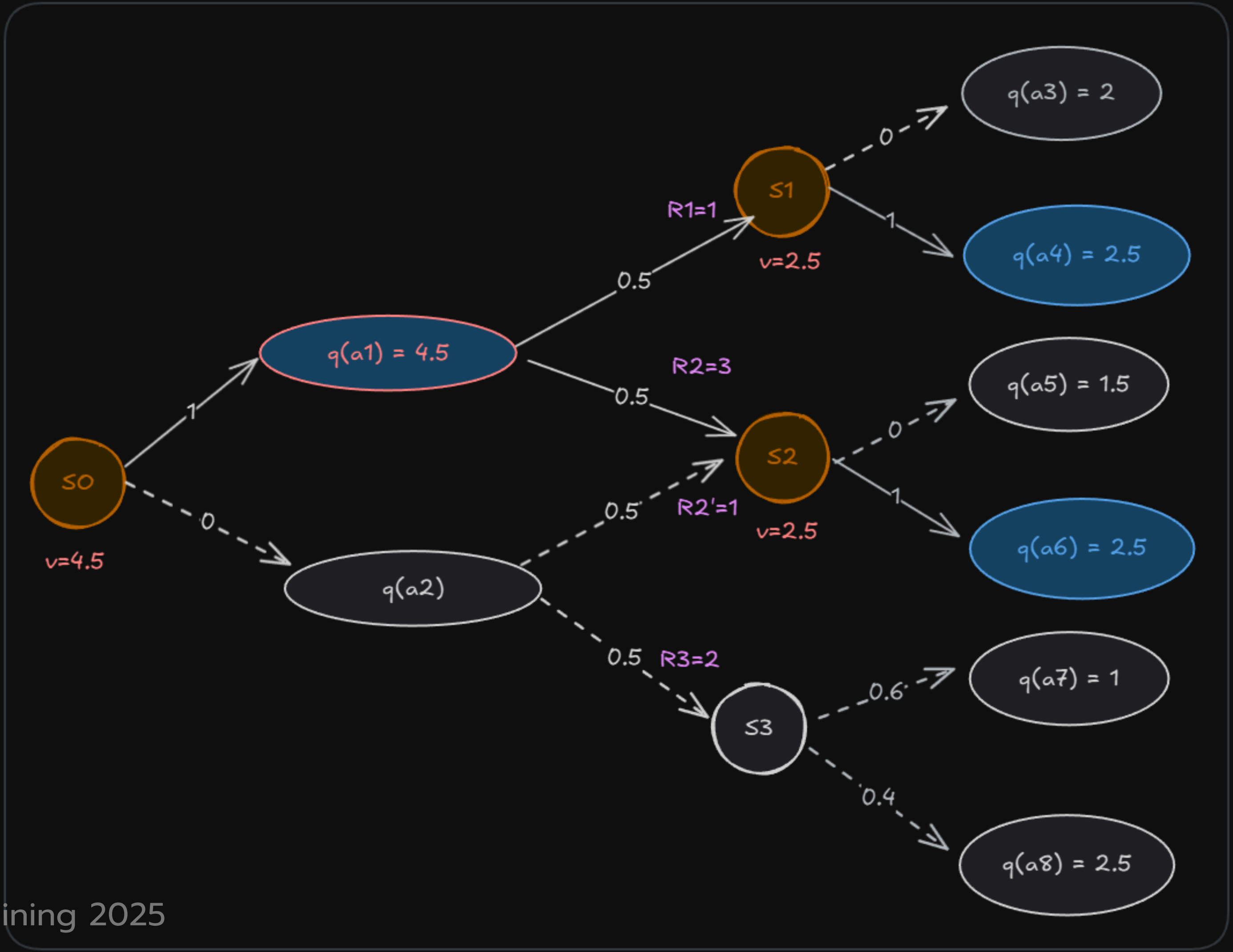
$$= 4.05$$

$$q(a_2) = 0.5\left[1 + \gamma\left(0.6(1.5) + 0.4(2.5)\right)\right] + 0.5\left[2 + \gamma\left(0.6(1) + 0.4(2.5)\right)\right]$$

$$= 3.25$$


S0
v=3.73
0.6
0.4
q(a1) = 4.05
q(a2) = 3.25
0.5
0.5
0.5
0.5
R1=1
R2=3
R2'=1
R3=2
S1
v=2.2
S2
v=1.9
S3
v=1.6
0.6
0.4
0.6
0.4
0.6
0.4
q(a3) = 2
q(a4) = 2.5
q(a5) = 1.5
q(a6) = 2.5
q(a7) = 1
q(a8) = 2.5

Optimal Policy



Optimality Condition

- If agent is following the optimal policy π^* (something we want to find), then the value function will also be optimal.

$$v^*(s) = \max_{\pi} v_{\pi}(s)$$

- It follows that

$$v^*(s) = \max_{a} \left\{ \sum_{s',r} p(s', r|s, a)[r + \gamma v^*(s')] \right\}$$

$$q^*(s, a) = \sum_{s',r} p(s', r|s, a) \cdot \left[r + \gamma \max_{a'} q^*(s', a') \right]$$