Markov Decision Processes

Chen-Yu Wei

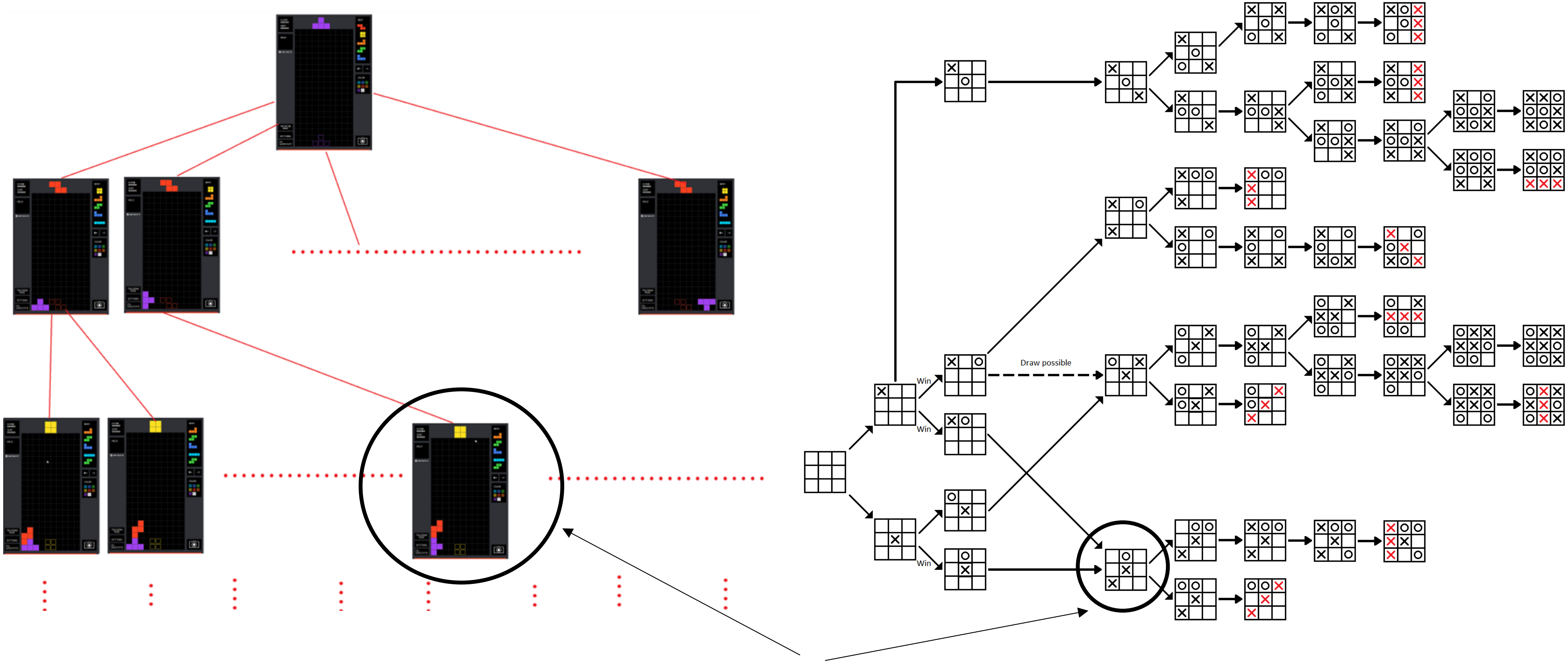
Sequence of Actions

To win the game, the learner has to take a sequence of actions $a_1 \to a_2 \to \cdots \to a_H$.

The effect of a particular action may not be revealed instantaneously.

- Some effect may be revealed instantaneously
- Some may be revealed later

Sequence of Actions



state

(a summary of the current status in a multi-stage game)

Deterministic World

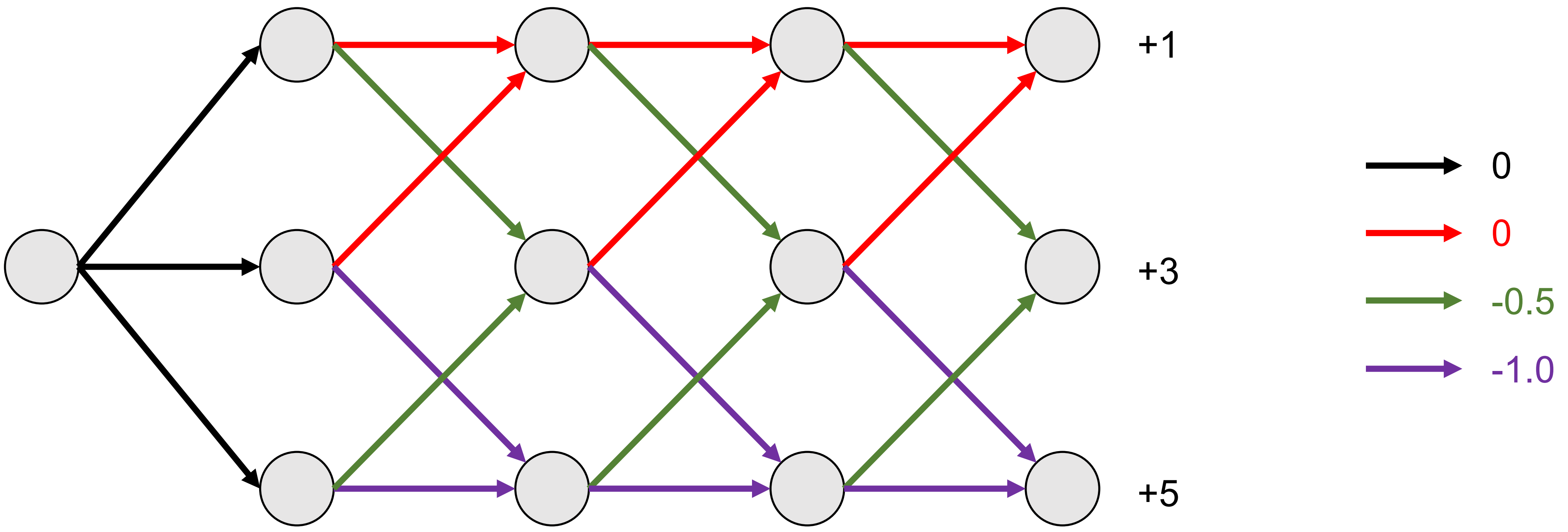
$R(a)$

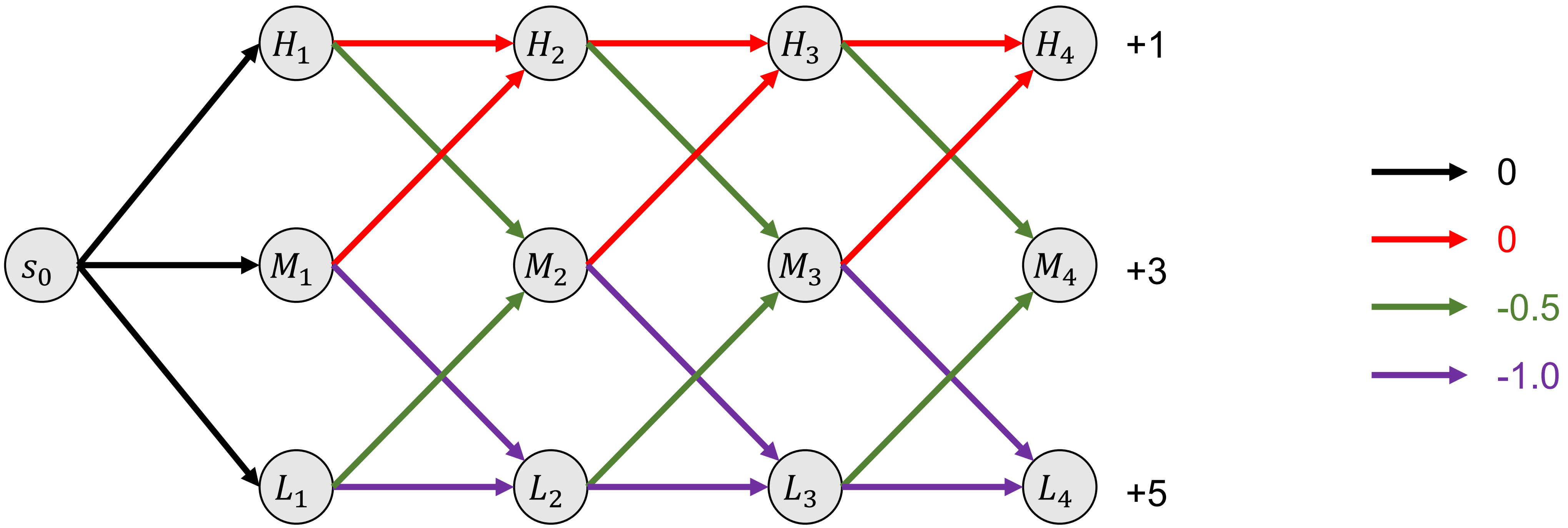
$R(x, a)$

$\pi^* = \arg\max_a R(a)$

$\pi^*(\cancel{x}) = \arg\max_a R(x, a)$

Which path gives us the highest **total reward**?

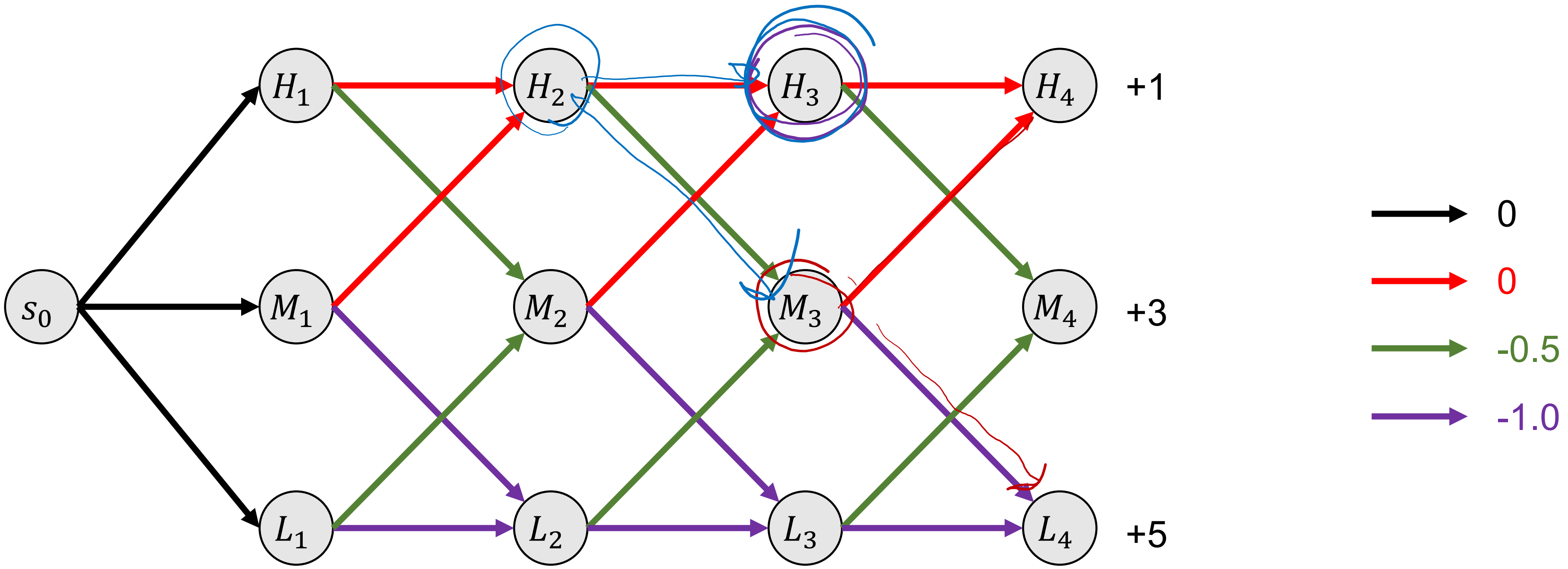




$V^\star(s)$ ≔ maximum total reward starting from state s

$Q^\star(s, a)$ ≔ maximum total reward starting from state s and taking action a **for one step**, and then following the optimal strategy

$\pi^\star(s)$ ≔ optimal decision on state s



$V^\star(H_4) = 1$ $\quad Q^\star(H_3, R) = 0+1=1$ $\quad V^\star(H_3) = 2.5$ $\quad Q^\star(H_2, R) = 0+2.5=2.5$ $\quad V^\star(H_2) = 3.5$

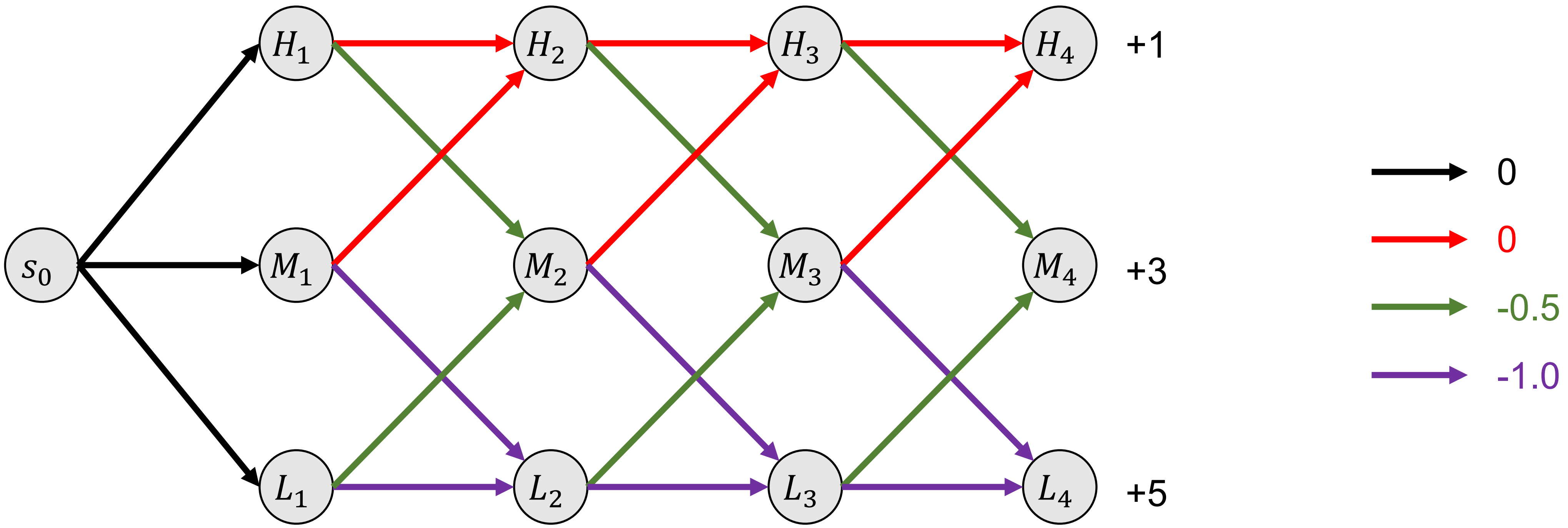
$\boxed{Q^\star(H_3, G)} = -0.5+3=2.5$ $\quad \boxed{Q^\star(H_2, G)} = -0.5+4=3.5$

$V^\star(M_4) = 3$ $\quad Q^\star(M_3, R) = 0+1=1$ $\quad V^\star(M_3) = 4$ $\quad Q^\star(M_2, R) =$ $\quad V^\star(M_2) = ?$

$\boxed{Q^\star(M_3, P)} = -1+5=4$ $\quad Q^\star(M_2, P) =$

$V^\star(L_4) = 5$ $\quad Q^\star(L_3, G) = -0.5+3=2.5$ $\quad V^\star(L_3) = 4$ $\quad Q^\star(L_2, G) =$ $\quad V^\star(L_2) = ?$

$\boxed{Q^\star(L_3, P)} = -1+5=4$ $\quad Q^\star(L_2, P) =$

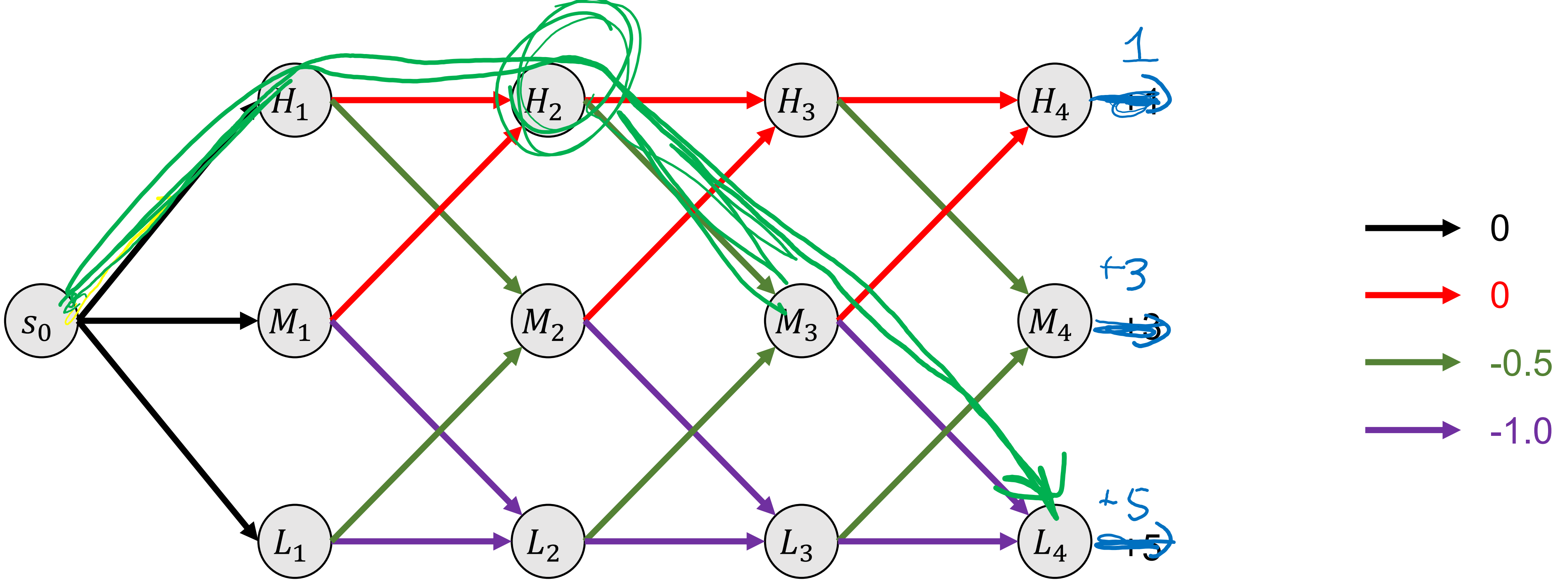


Relation between $Q^\star, V^\star, \pi^\star$:

$$Q^\star(s,a) = R(s,a) + V^\star(\text{next\_state}(s,a))$$

$$V^\star(s) = \max_a Q^\star(s,a)$$

$$\pi^\star(s) = \operatorname*{argmax}_a Q^\star(s,a)$$



General algorithm (Value Iteration) --- see next slide for a more detailed version

Repeat until Q, V no longer changes:

$$Q(s,a) \leftarrow R(s,a) + V(\text{next\_state}(s,a)) \quad \text{for all } (s,a)$$

$$V(s) = \max_a Q(s,a) \quad \text{for all } s$$

$$\pi^\star(s) = \operatorname*{argmax}_a Q(s,a)$$

(for the last layer states. $Q(s,a) \leftarrow R(s,a)$)

Value Iteration:

Initialize $Q_0(s,a) \leftarrow 0,\ V_0(s) \leftarrow 0$ for all (s,a)

For $i = 1, 2, \dots$

$Q_i(s,a) \leftarrow R(s,a) + V_{i-1}(\mathrm{next\_state}(s,a))$ for all (s,a)

$V_i(s) = \max_a Q_i(s,a)$ for all s

If $Q_i(s,a) = Q_{i-1}(s,a)$ for all (s,a): **break**

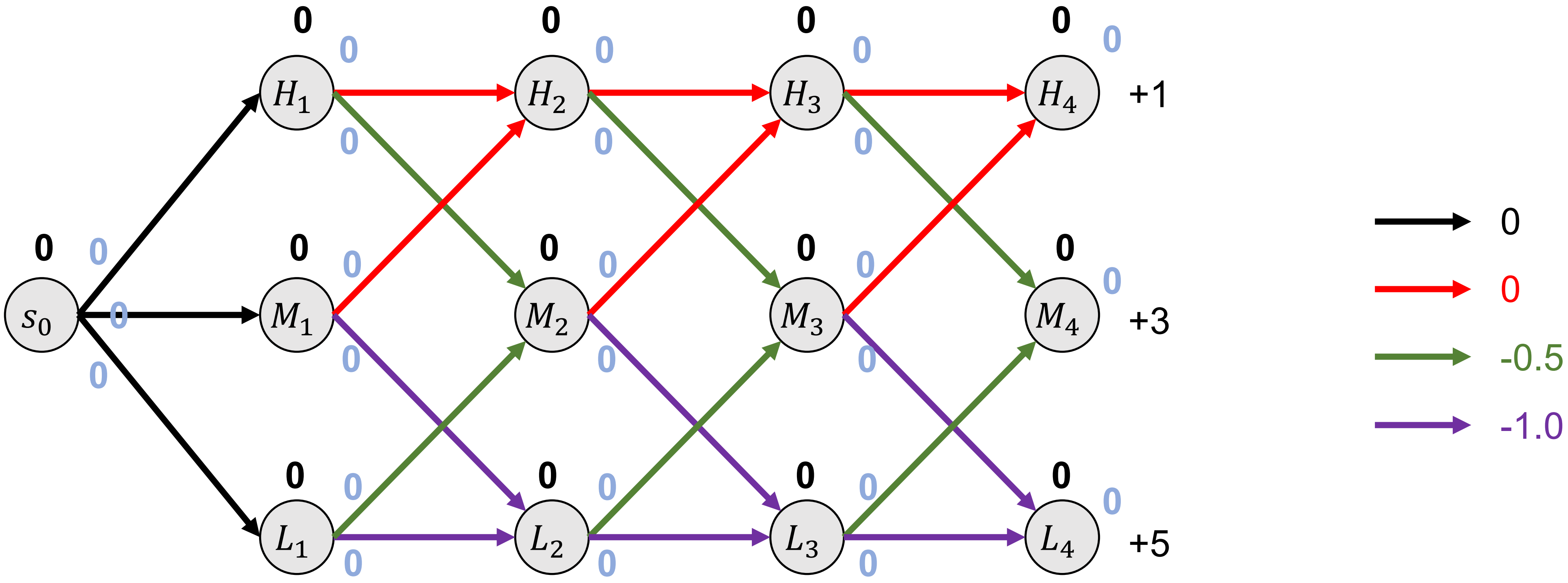
$\pi^\star(s) = \operatorname{argmax}_a Q_{\text{final}}(s,a)$

If s is a terminal state, this line is simply $Q_i(s,a) \leftarrow R(s,a)$

If every path in the graph has length $\leq K$, then Value Iteration will terminate in $\leq K + 1$ iterations.

$Q_0(s, a) = 0$

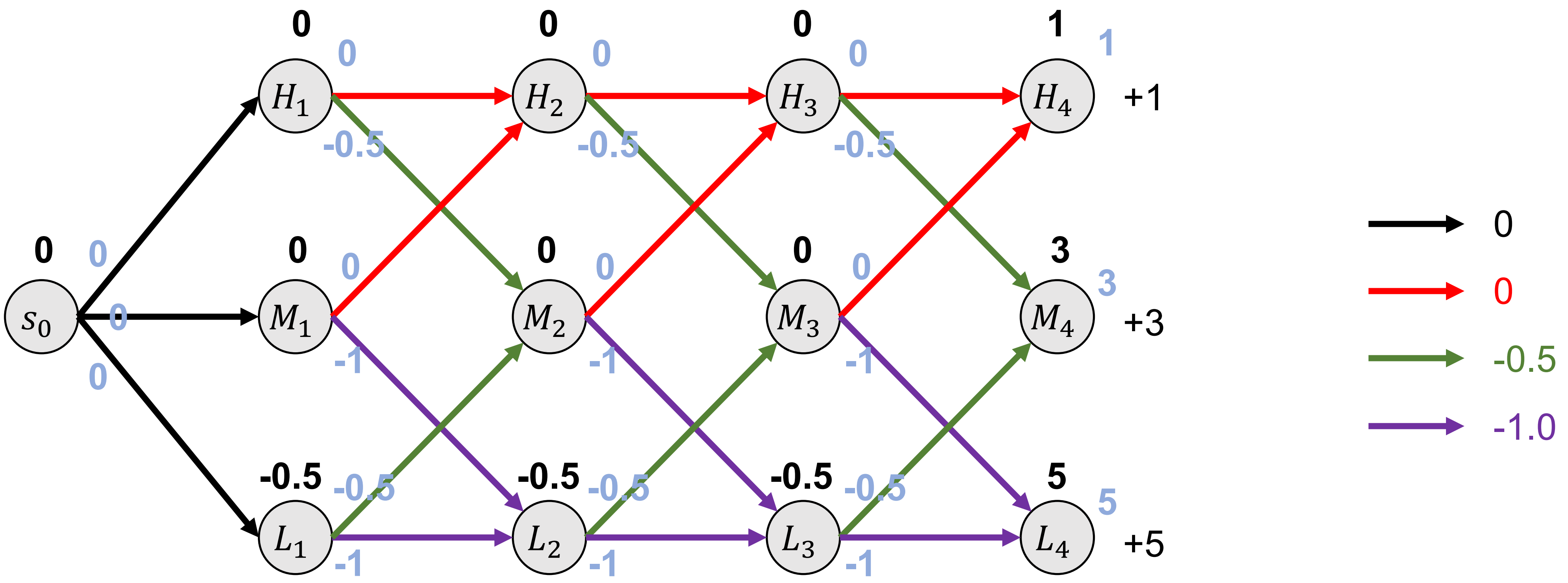
$V_0(s) = 0$



0
0
0
s_0
0
0
0
0
0
H_1
0
M_1
0
0
0
L_1
0
0
0
0
H_2
0
0
M_2
0
0
0
L_2
0
0
0
0
H_3
0
0
M_3
0
0
0
L_3
0
0
0
H_4
+1
0
0
M_4
+3
0
0
L_4
+5
0
0
-0.5
-1.0

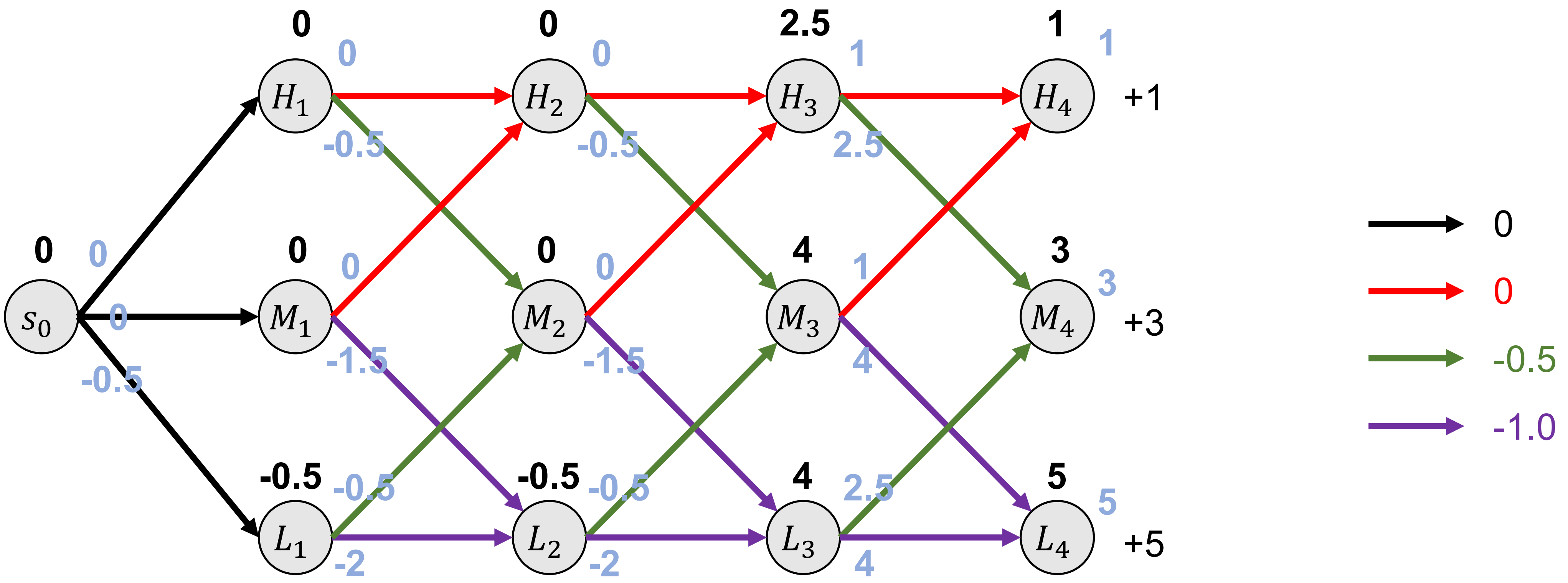
$$Q_1(s,a) = R(s,a) + V_0(\text{next}(s,a))$$

$$V_1(s) = \max_a Q_1(s,a)$$



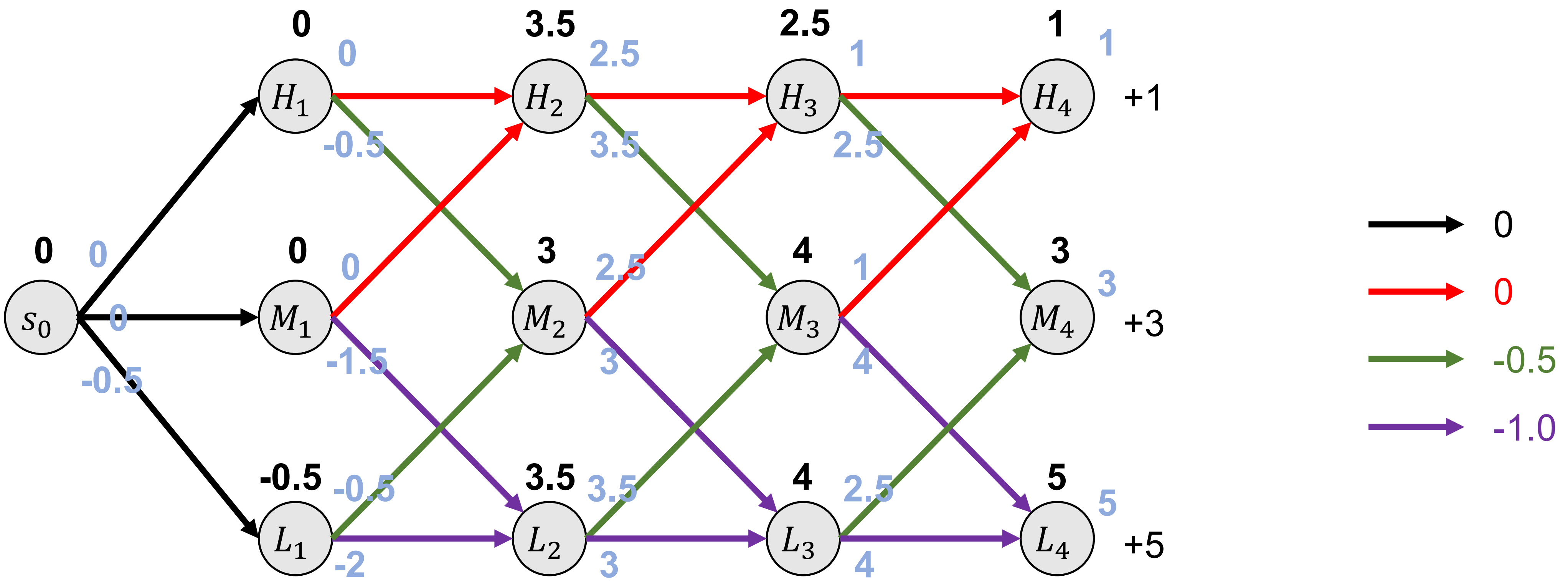
$$Q_2(s,a) = R(s,a) + V_1(\text{next}(s,a))$$

$$V_2(s) = \max_a Q_2(s,a)$$



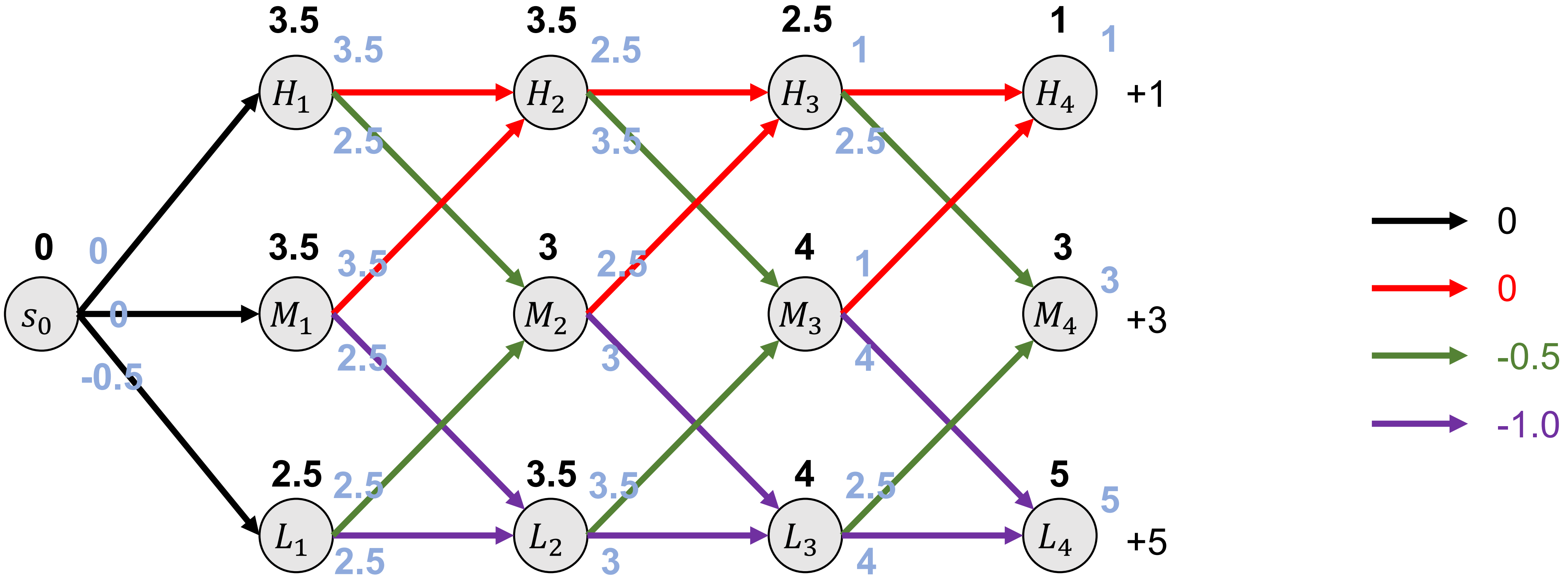
$$Q_3(s,a) = R(s,a) + V_2(\text{next}(s,a))$$

$$V_3(s) = \max_a Q_3(s,a)$$



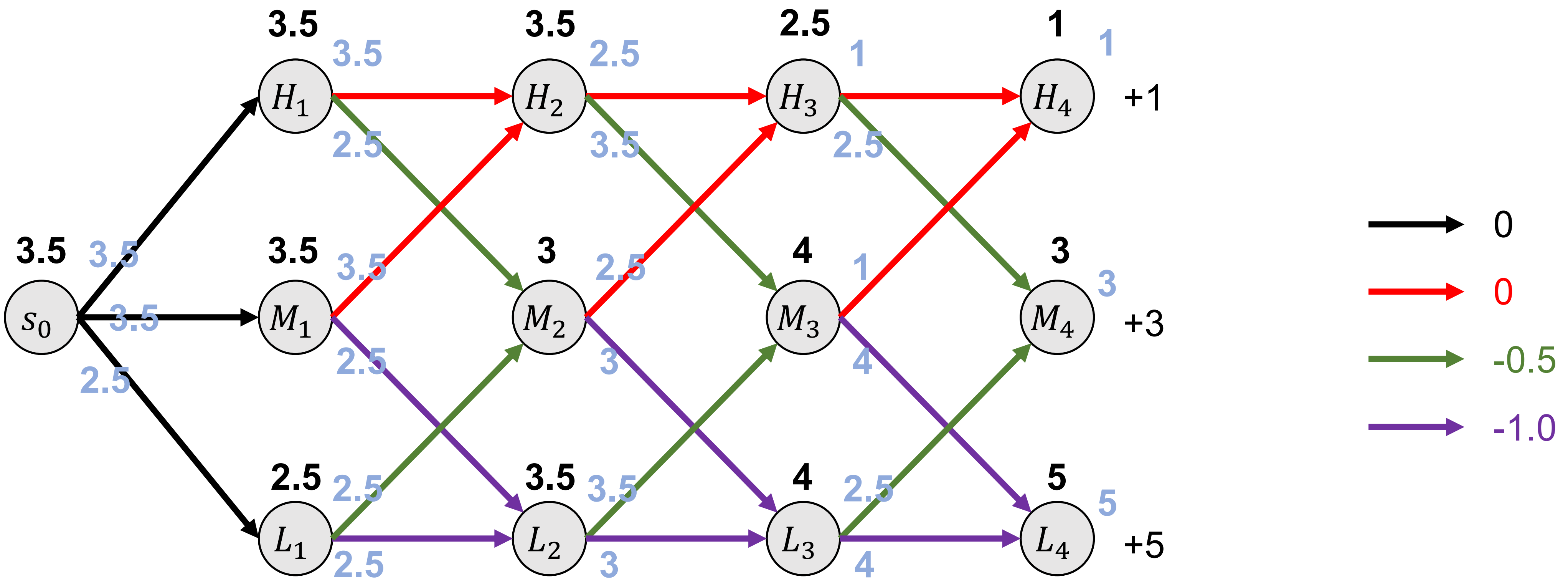
$$Q_4(s,a) = R(s,a) + V_3(\text{next}(s,a))$$

$$V_4(s) = \max_a Q_4(s,a)$$



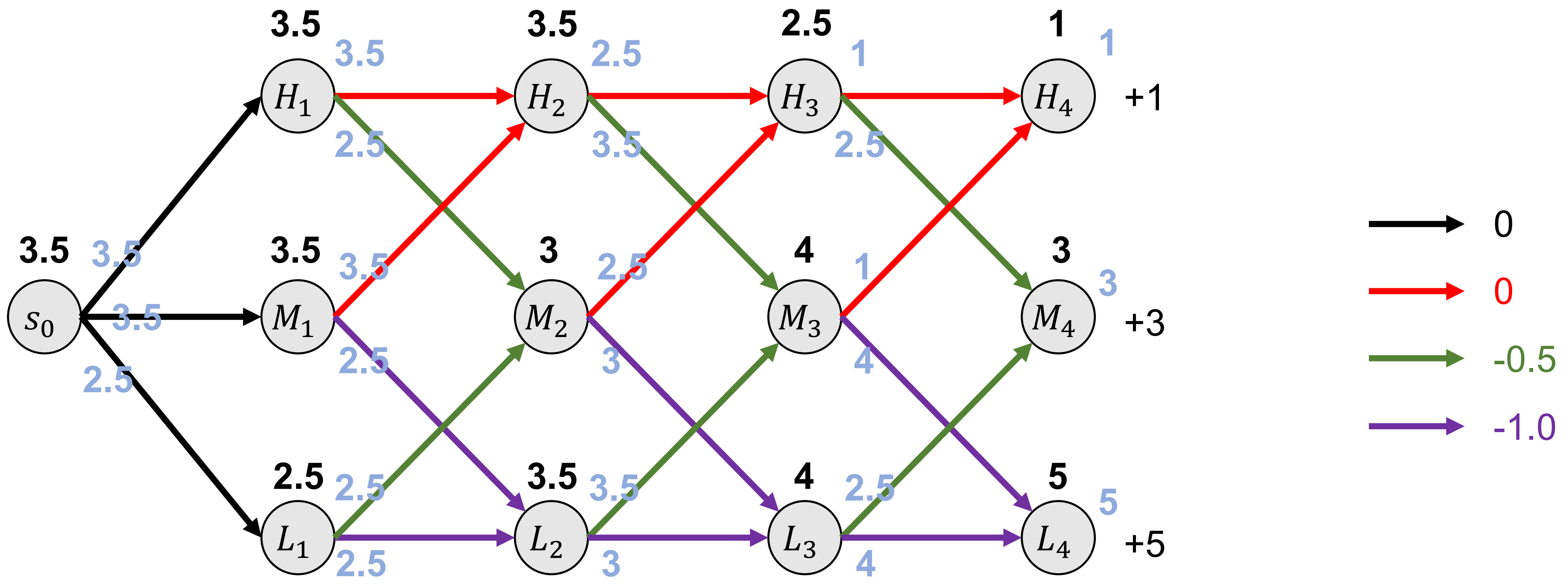
$$Q_5(s,a) = R(s,a) + V_4(\text{next}(s,a))$$

$$V_5(s) = \max_a Q_5(s,a)$$



$$Q_6(s,a) = R(s,a) + V_5(\text{next}(s,a))$$

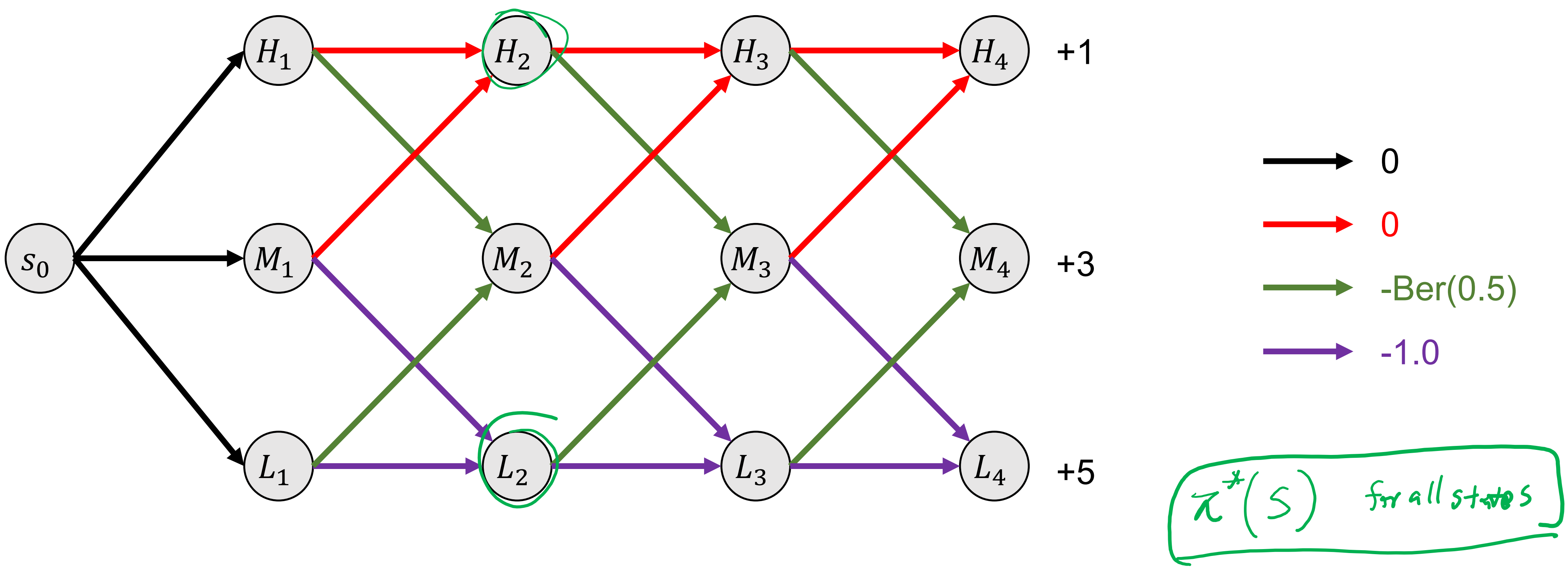
$$V_6(s) = \max_a Q_6(s,a)$$

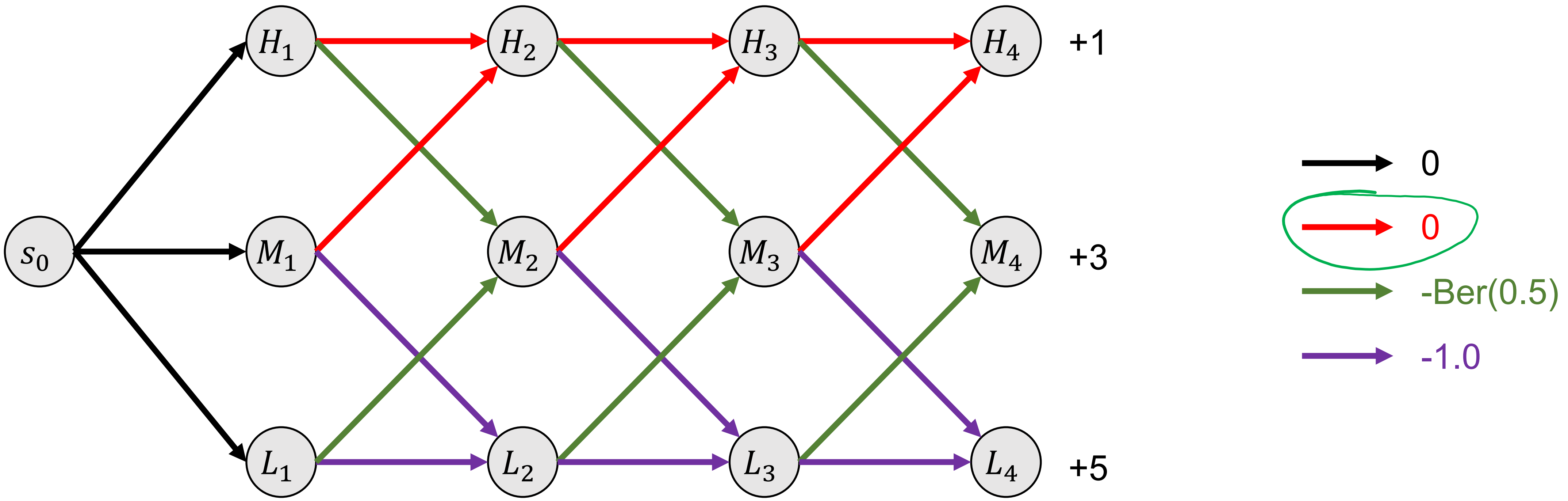


Stochastic Worlds

Now, suppose taking an action does **not** lead to the desired state deterministically.

Instead, with probability 0.8, it goes to the state as specified in the figure;
with probability 0.1 each, it goes to the other two states.





$\boxed{V^\star(H_4)} = 1$

$Q^\star(H_3, R) =$

$Q^\star(H_3, G) =$

$V^\star(H_3) =$

$V^\star(M_4) = 3$

$Q^\star(M_3, R) =$

$Q^\star(M_3, P) =$

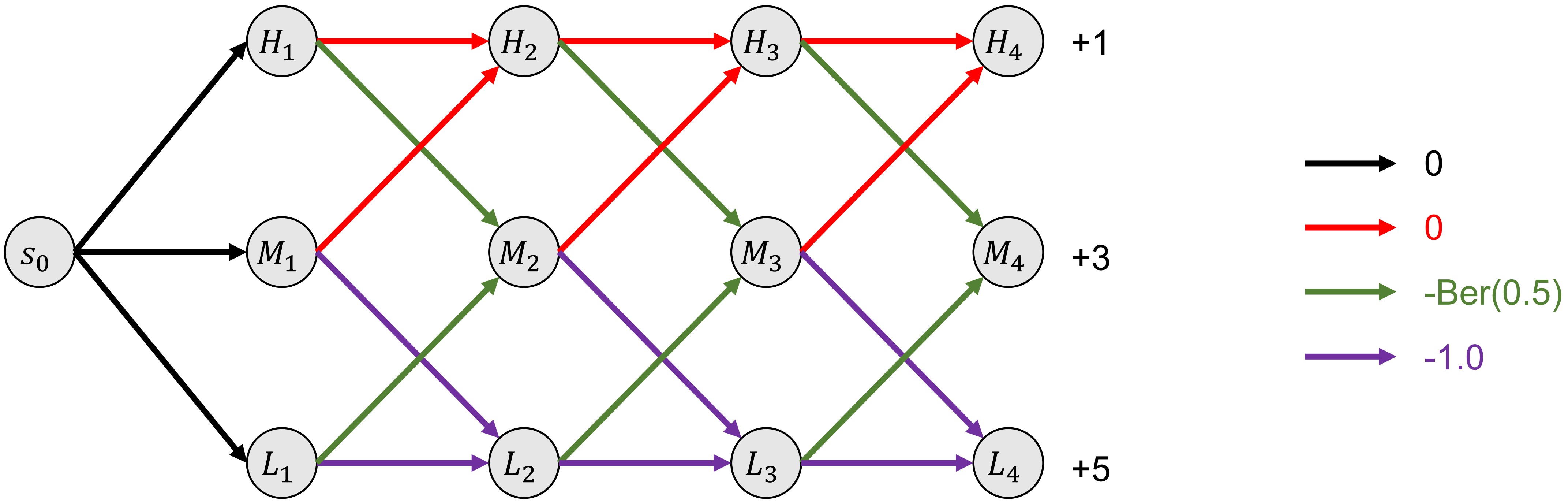
$V^\star(M_3) =$

$V^\star(L_4) = 5$

$Q^\star(L_3, G) =$

$Q^\star(L_3, P) =$

$V^\star(L_3) =$



$V^\star(H_3) =$

$Q^\star(H_2, R) =$

$Q^\star(H_2, G) =$

$V^\star(H_2) =$

$V^\star(M_3) =$

$Q^\star(M_2, R) =$

$Q^\star(M_2, P) =$

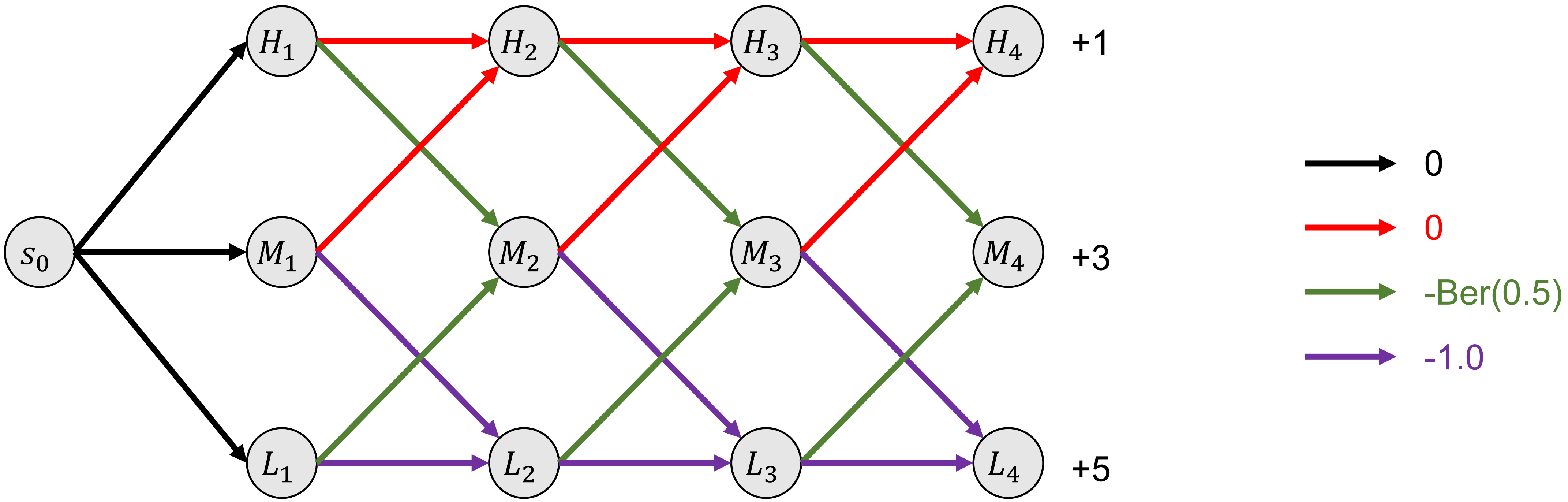
$V^\star(M_2) =$

$V^\star(L_3) =$

$Q^\star(L_2, G) =$

$Q^\star(L_2, P) =$

$V^\star(L_2) =$



Relation between $Q^\star, V^\star, \pi^\star$:

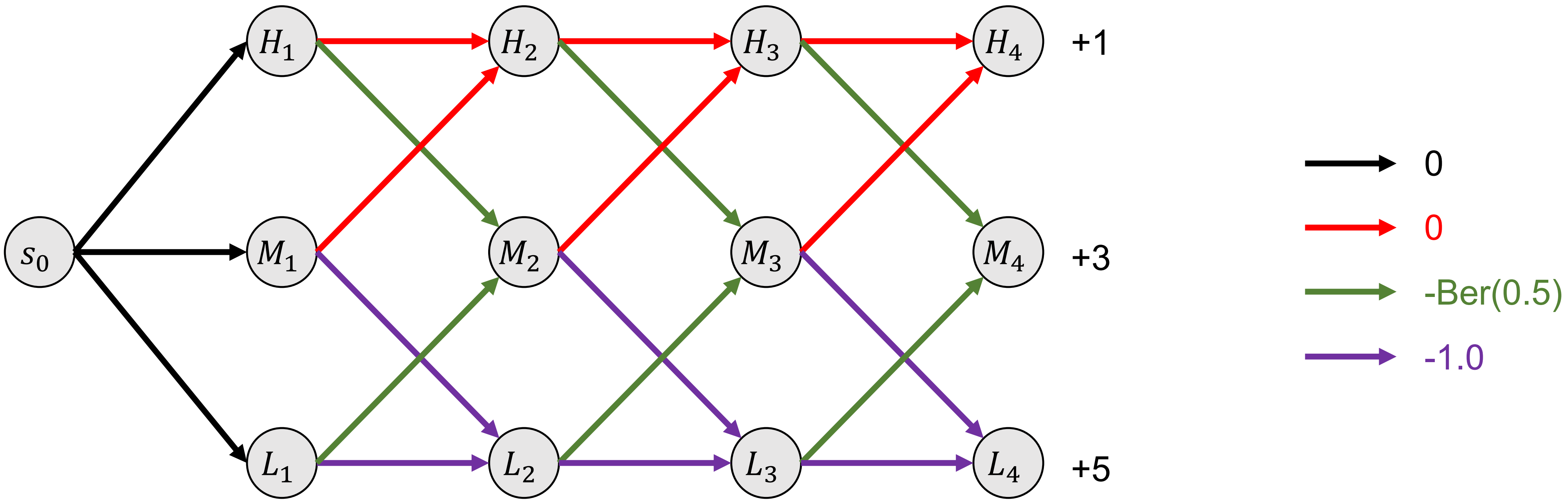
$$Q^\star(s,a) = R(s,a) + \sum_{s'} P(s'|s,a)\, V^\star(s')$$

$$V^\star(s) = \max_a Q^\star(s,a)$$

$$\pi^\star(s) = \operatorname*{argmax}_a Q^\star(s,a)$$

Transition probability $P(s'|s,a)$:

The probability of going to state s' if we take action a on state s



Value Iteration:

Repeat until Q, V no longer changes:

$$Q(s,a) \leftarrow R(s,a) + \sum_{s'} P(s'|s,a)\,V(s') \qquad \text{for all } (s,a)$$

$$V(s) = \max_a Q(s,a) \qquad \text{for all } s$$

$$\pi^\star(s) = \operatorname*{argmax}_a Q(s,a)$$

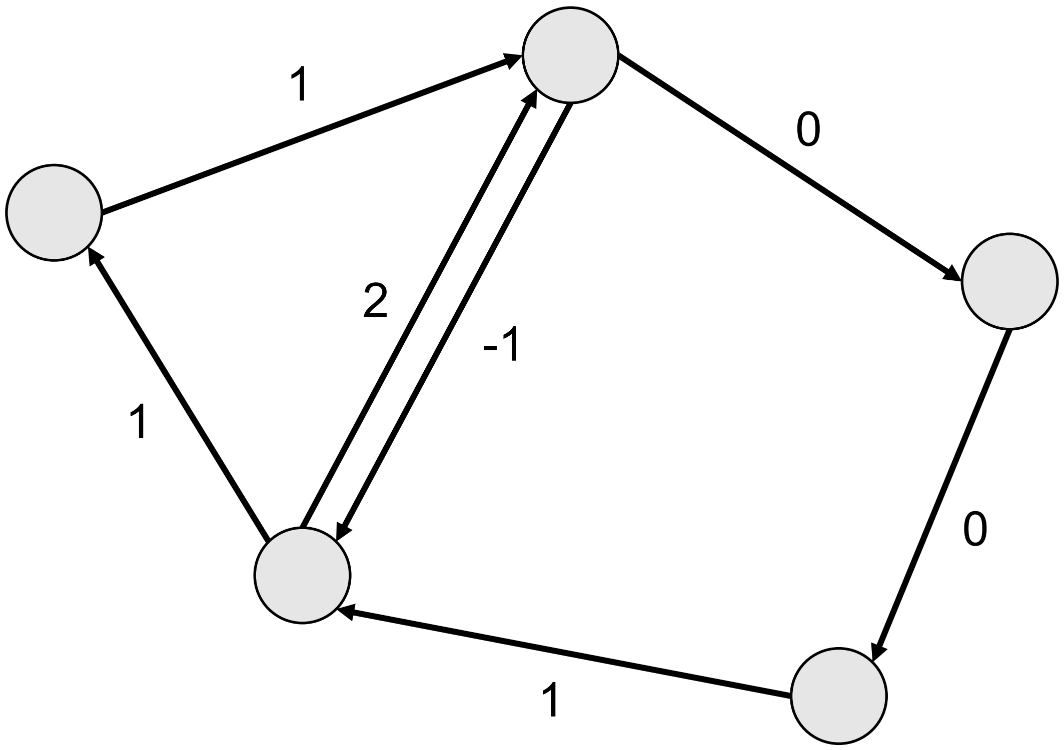
Infinite / Long-Horizon Case

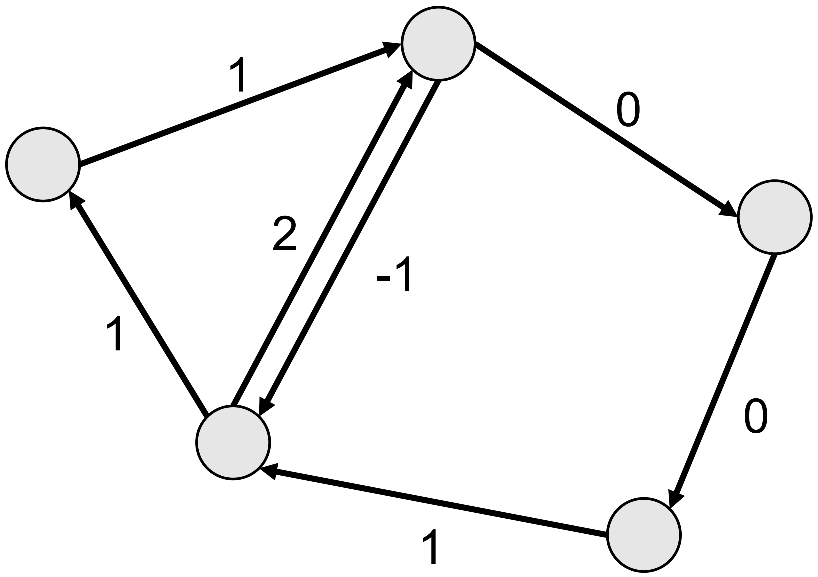
What if there is no “terminal state”?

- Tasks with terminal states
  - Tic-Tac-Toe
  - Tetris
  - Robot doing housework

- Tasks without terminal states (or tasks with very **long horizon** or **loops**)
  - Driving
  - Inventory control

How to gain most reward **in the long run**?   (or, in 1000 rounds)





Repeat until Q, V no longer changes? (Won’t terminate)

$$Q(s,a) \leftarrow R(s,a) + \sum_{s'} P(s'|s,a)\,V(s') \qquad \text{for all } (s,a)$$

$$V(s) = \max_a Q(s,a) \qquad \text{for all } s$$

$$\pi^\star(s) = \operatorname*{argmax}_a Q(s,a)$$

Discounting

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\star(s')$$

$$V^\star(s) = \max_a Q^\star(s,a)$$

$$\pi^\star(s) = \operatorname{argmax}_a Q^\star(s,a)$$

To calculate $Q^\star, V^\star, \pi^\star$ --- see later slide for a more detailed version

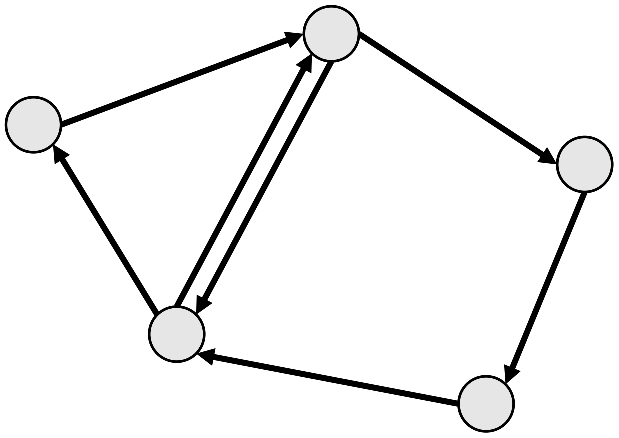
Repeat until Q, V **becomes stable:**

$$Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V(s') \quad \text{for all } (s,a)$$

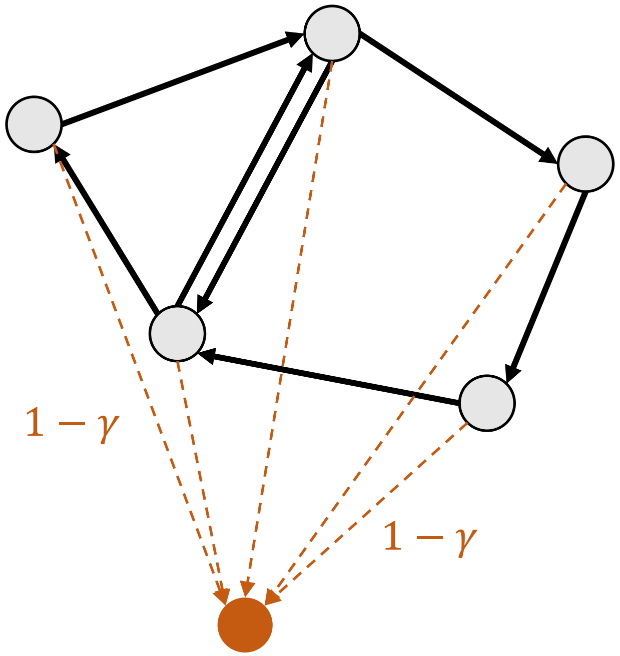
$$V(s) = \max_a Q(s,a) \quad \text{for all } s$$

$$\pi^\star(s) = \operatorname{argmax}_a Q(s,a)$$

Discounting: Equivalent View



Original Transition



Consider the $Q^\star, V^\star$ in this modified graph.

Modified Transition: on any state, taking any action,

$$\text{Next state} = \begin{cases} \text{terminal state} & \text{with probability } 1-\gamma \\ \text{following } P(s'|s,a) & \text{with probability } \gamma \end{cases}$$

Discounting

Effective horizon length (the expected length before going to the dummy terminal state) = $\frac{1}{1-\gamma}$

- Goal of discounting: Make $V^\star, Q^\star$ finite even when there are positive loops
  - γ should be strictly smaller than 1
- The modification should still approximate the learner's goal well enough
  - γ should be sufficiently close to 1

Usually, γ is set to some value like 0.99, but it depends on the application

Discounting

Value Iteration with discounting

Initialize $Q_0(s,a) \leftarrow 0,\ V_0(s) \leftarrow 0$ for all (s,a)

For $i = 1, 2, \ldots$

$Q_i(s,a) \leftarrow R(s,a) + \gamma V_{i-1}(\text{next\_state}(s,a))$ for all (s,a)

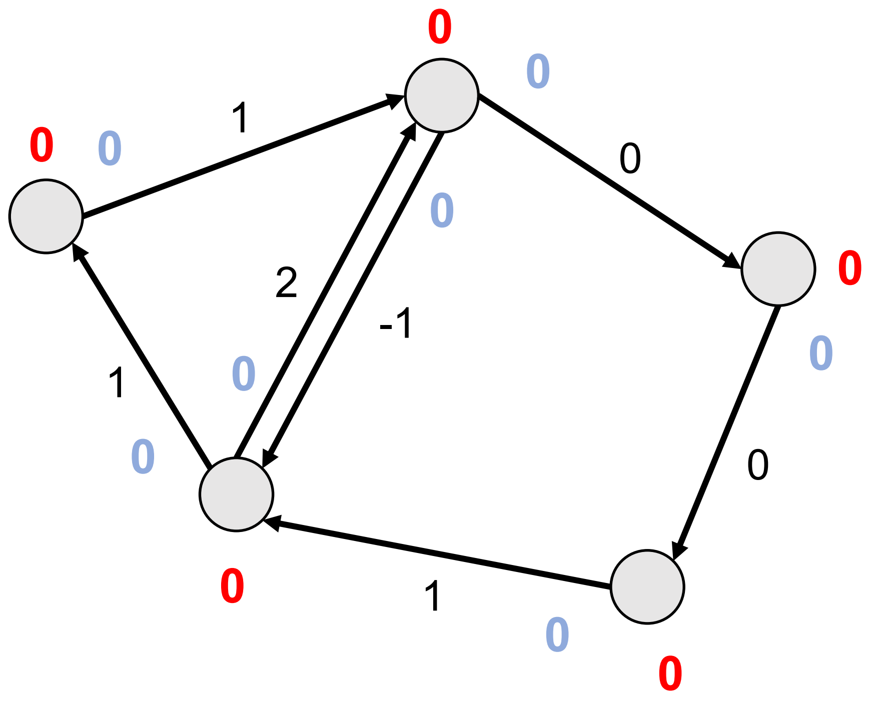
$V_i(s) = \max_a Q_i(s,a)$ for all s

If $|Q_i(s,a) - Q_{i-1}(s,a)| \leq \epsilon$ for all (s,a): **break**

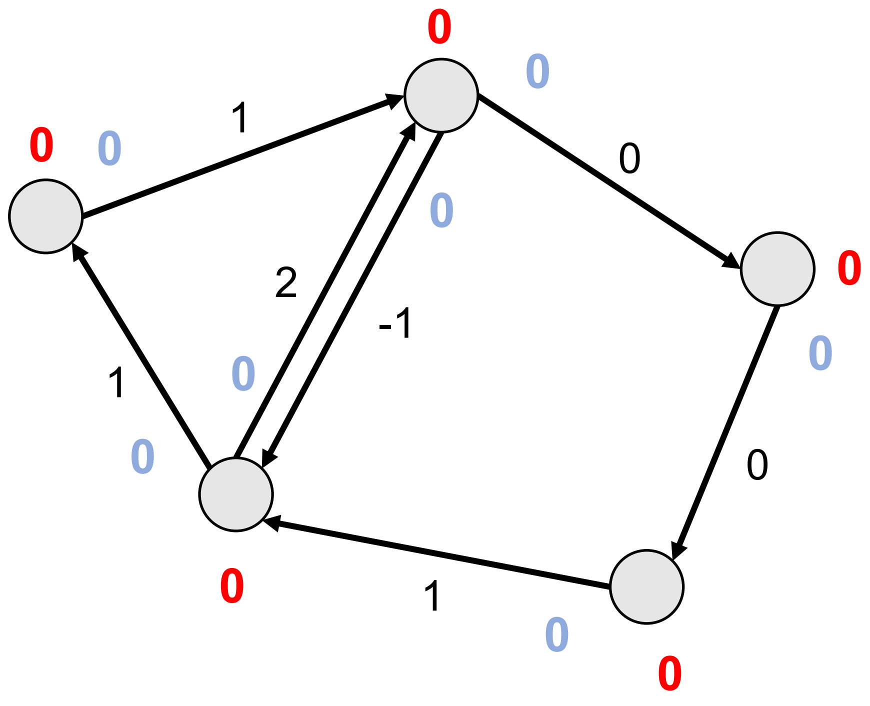
$$\pi^\star(s) = \operatorname*{argmax}_a Q_{\text{final}}(s,a)$$

$Q_0(s,a) = 0$

$V_0(s) = 0$



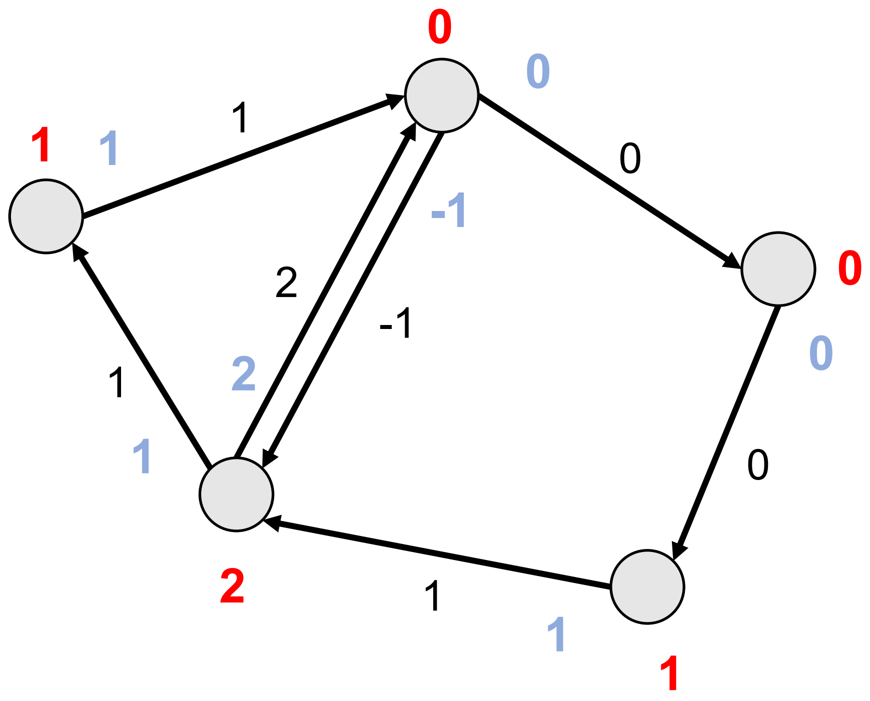
$\gamma = 0.95$



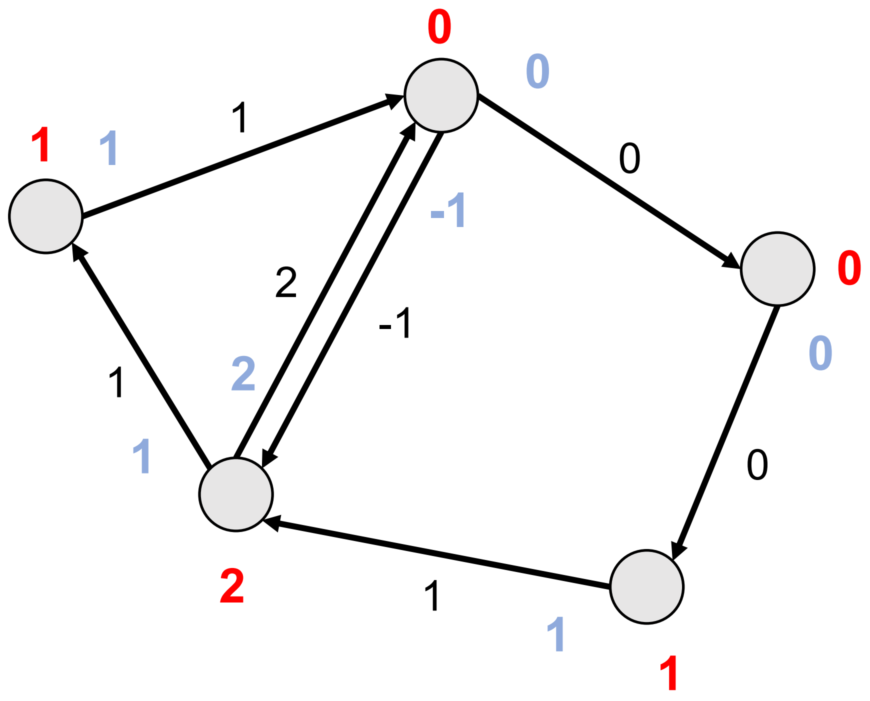
$\gamma = 1.0$

$$\boldsymbol{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$\boldsymbol{V_i(s)} \leftarrow \max_a Q_i(s,a)$$



$\gamma = 0.95$

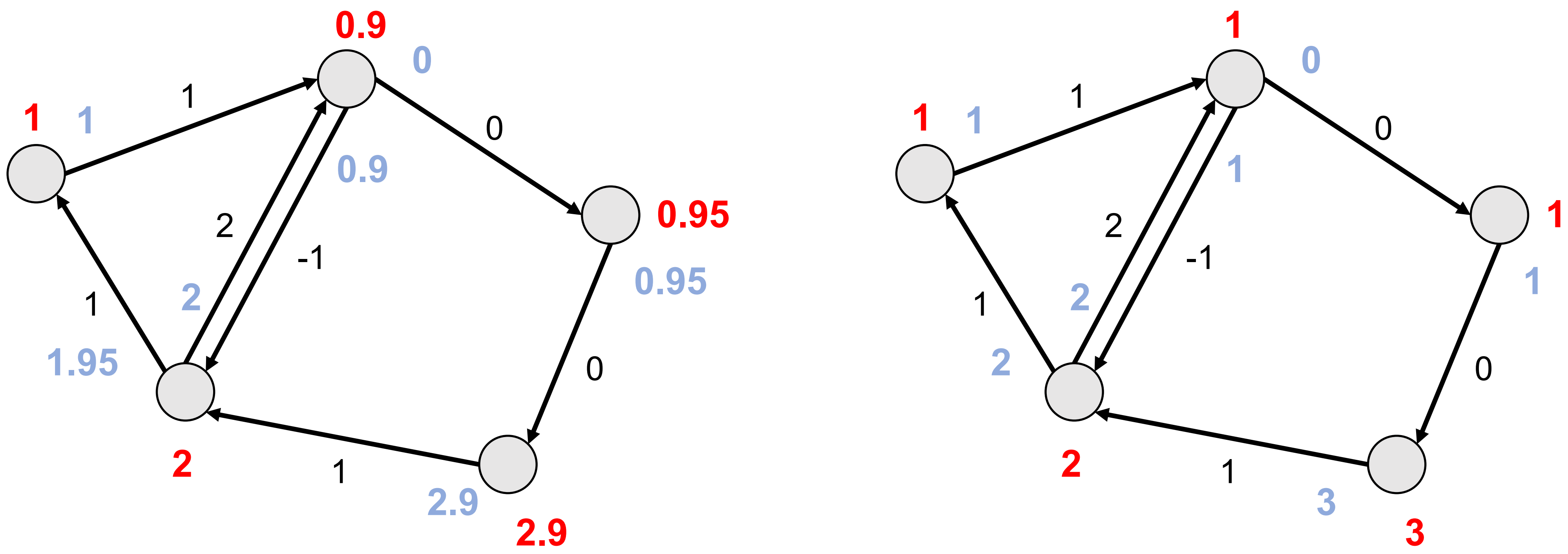


$\gamma = 1.0$

Iteration $i = 1$

$$\boldsymbol{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$\boldsymbol{V_i(s)} \leftarrow \max_a Q_i(s,a)$$

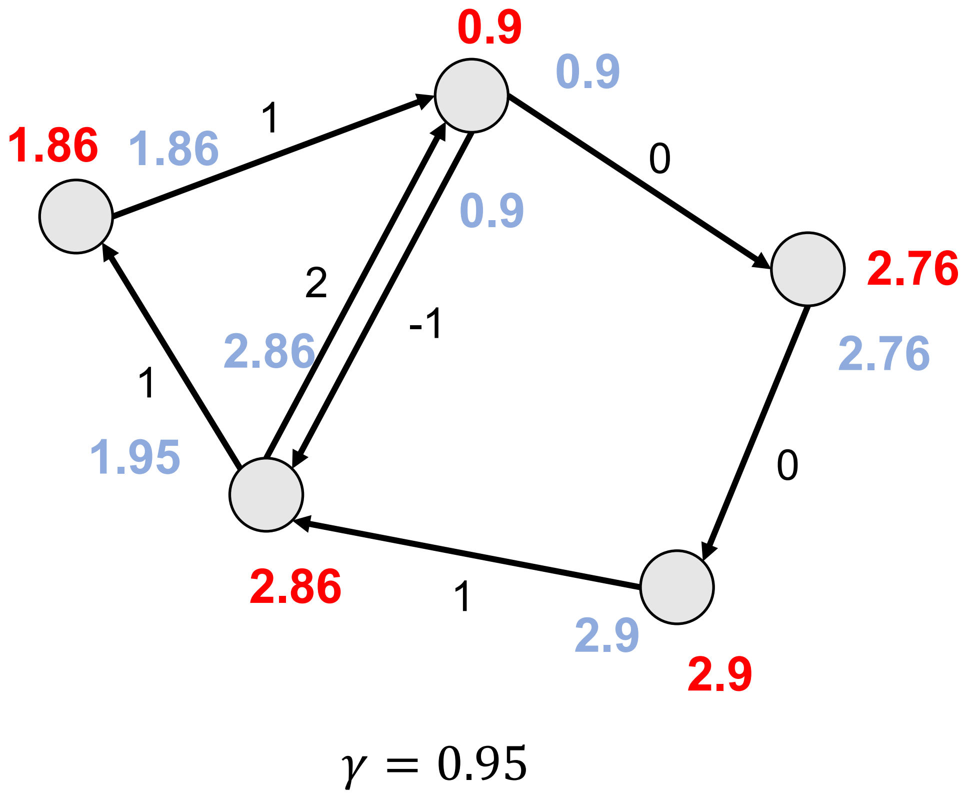


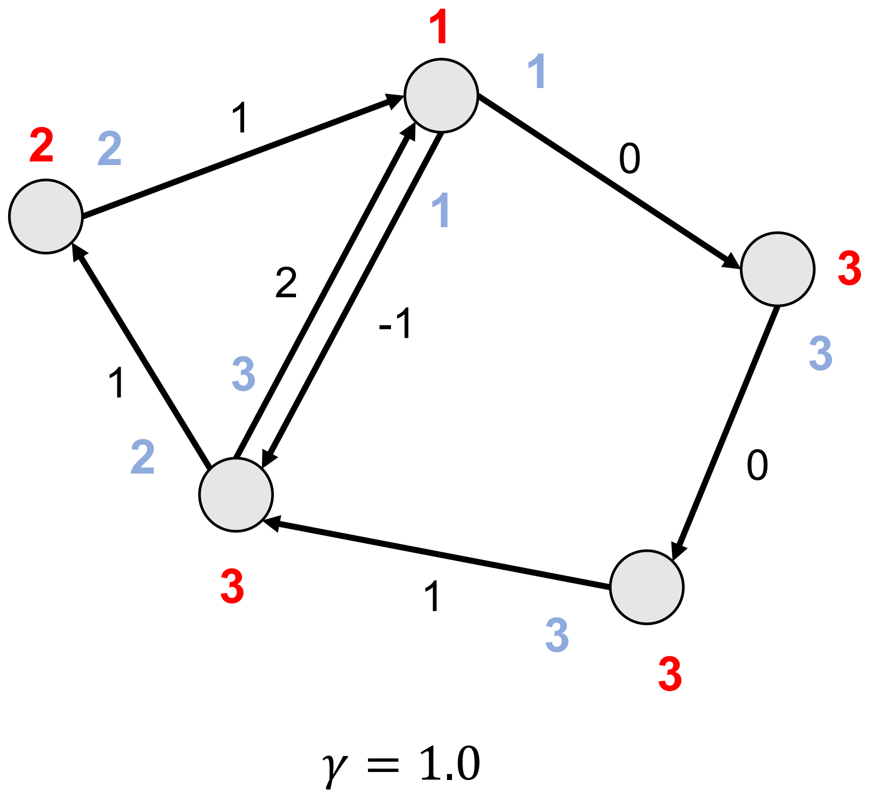
$\gamma = 0.95$ $\gamma = 1.0$

Iteration $i = 2$

$$Q_i(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$V_i(s) \leftarrow \max_a Q_i(s,a)$$

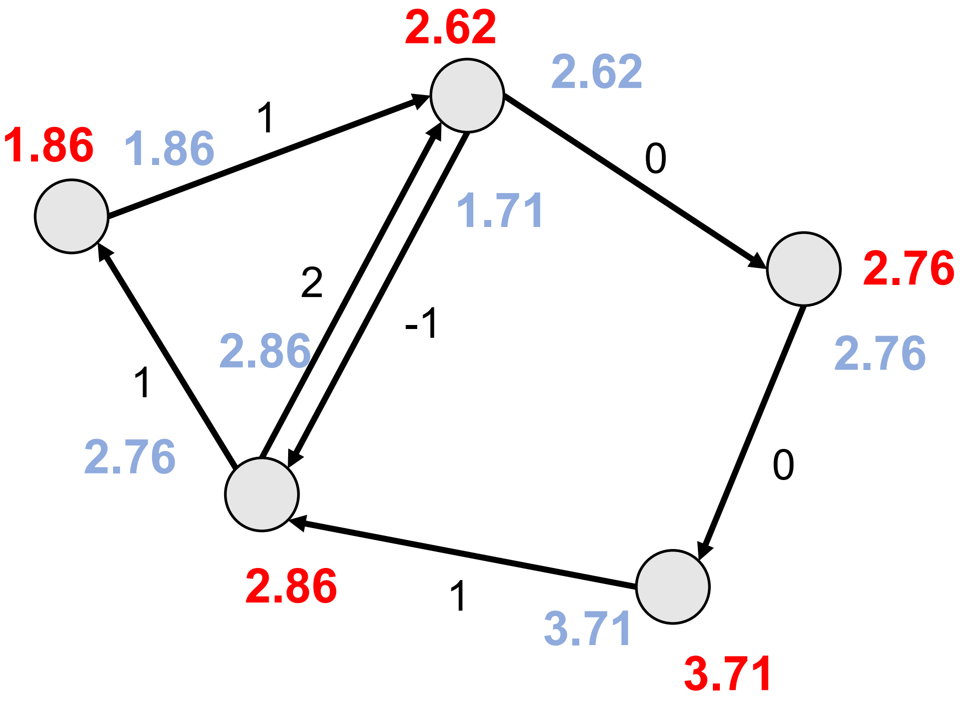




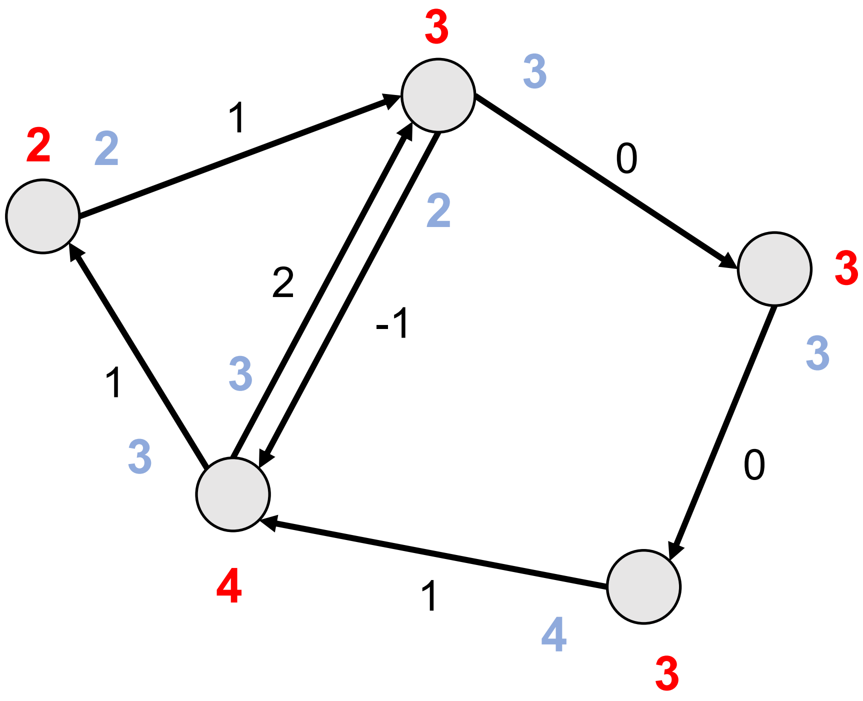
Iteration $i = 3$

$$Q_i(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$V_i(s) \leftarrow \max_a Q_i(s,a)$$



$\gamma = 0.95$

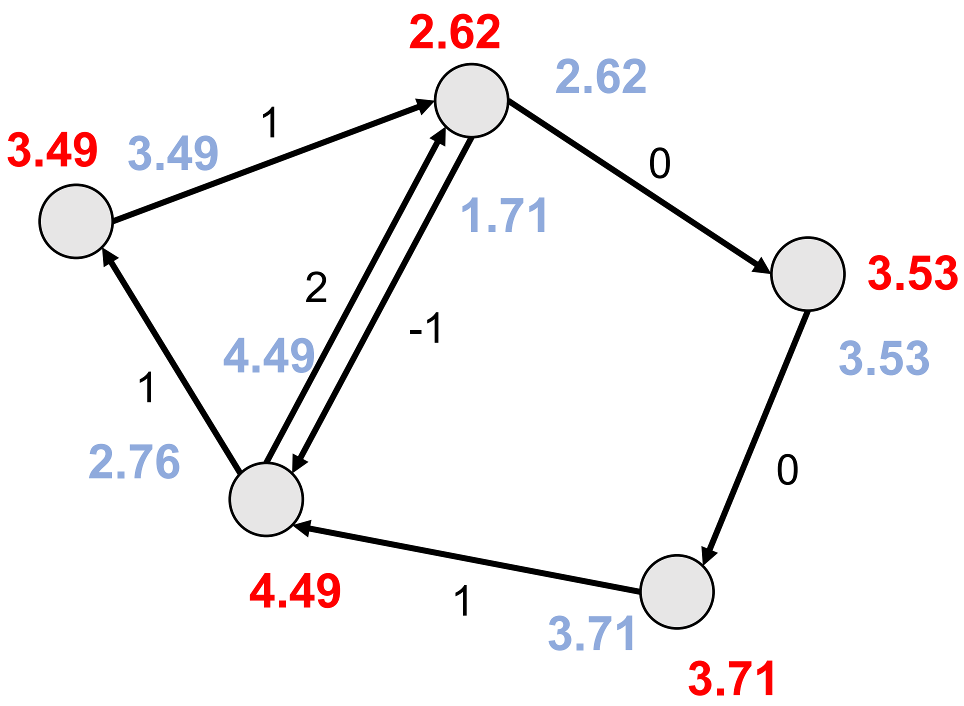


$\gamma = 1.0$

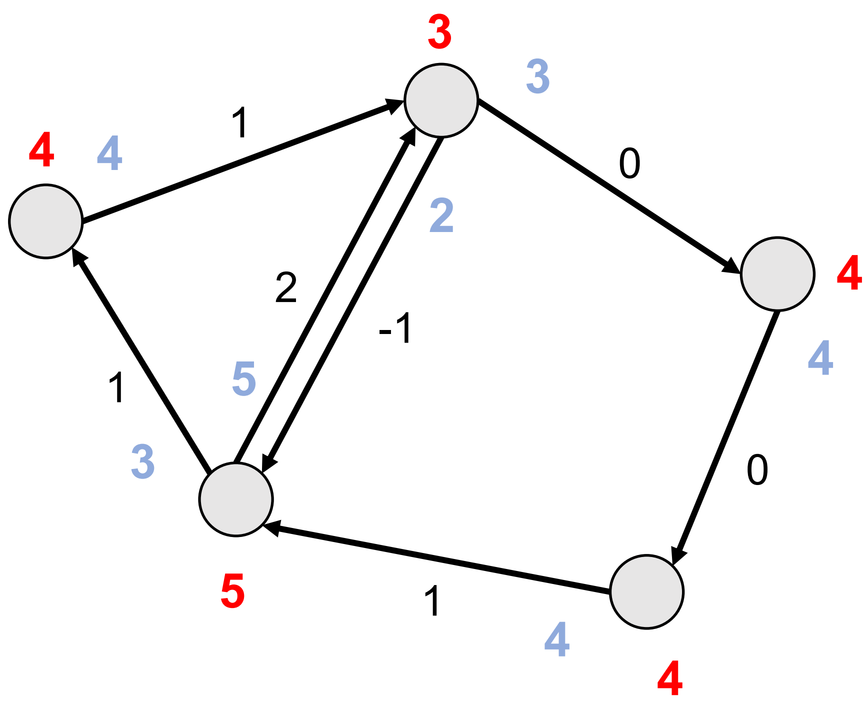
Iteration $i = 4$

$$Q_i(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$V_i(s) \leftarrow \max_a Q_i(s,a)$$



$\gamma = 0.95$

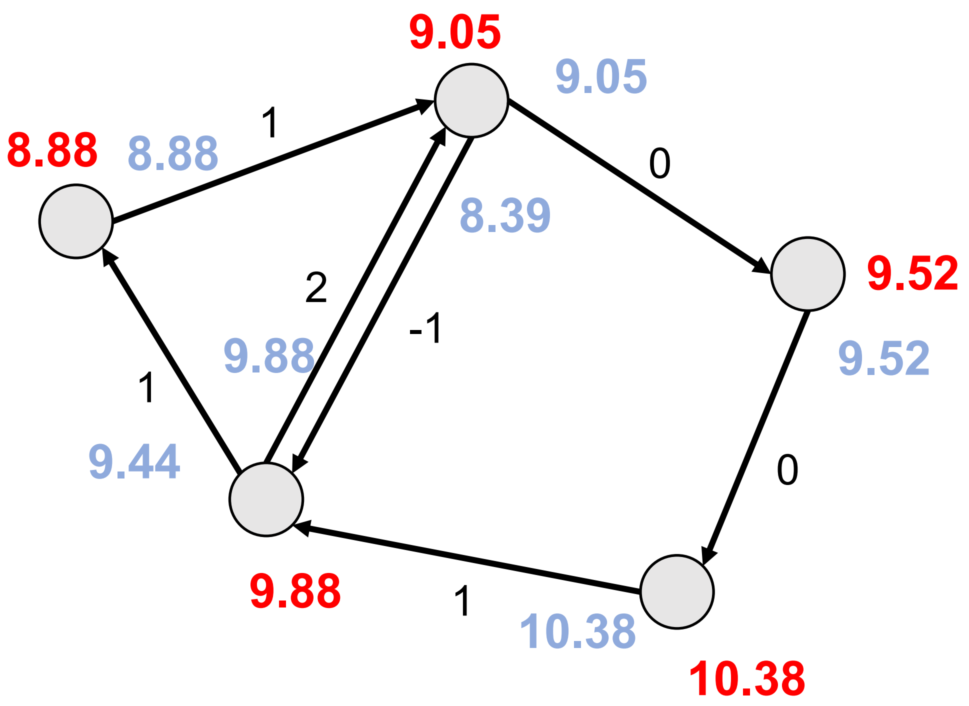


$\gamma = 1.0$

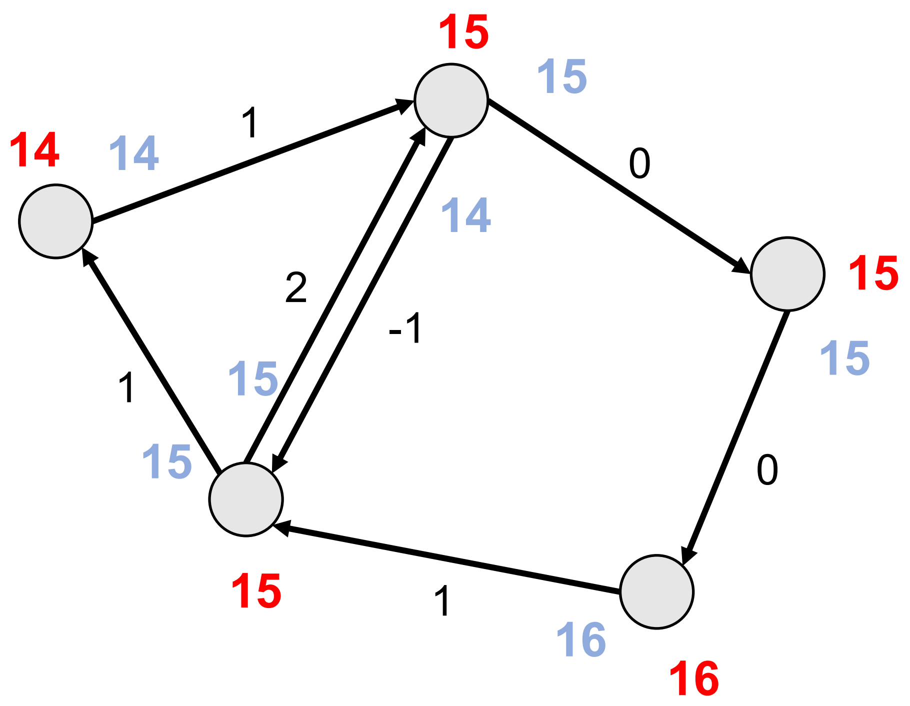
Iteration $i = 5$

$$\boldsymbol{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$\boldsymbol{V_i(s)} \leftarrow \max_a Q_i(s,a)$$



$\gamma = 0.95$

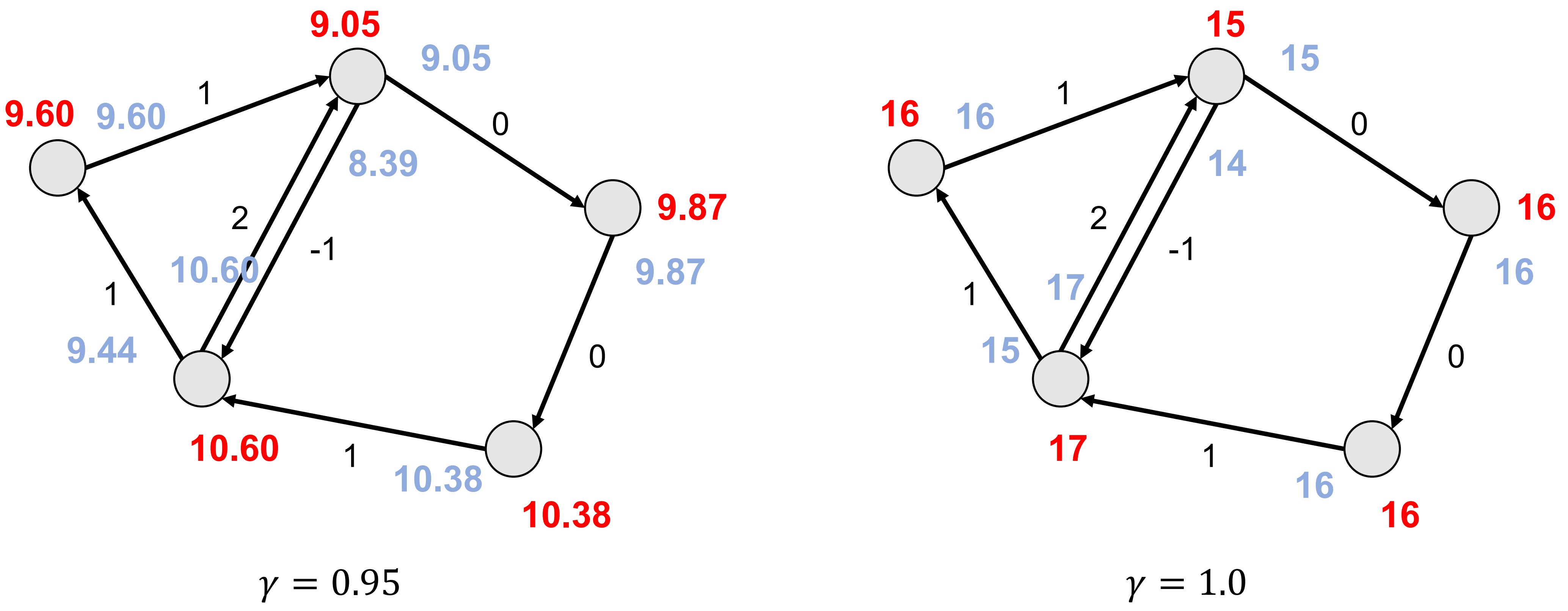


$\gamma = 1.0$

Iteration $i = 20$

$$\boldsymbol{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

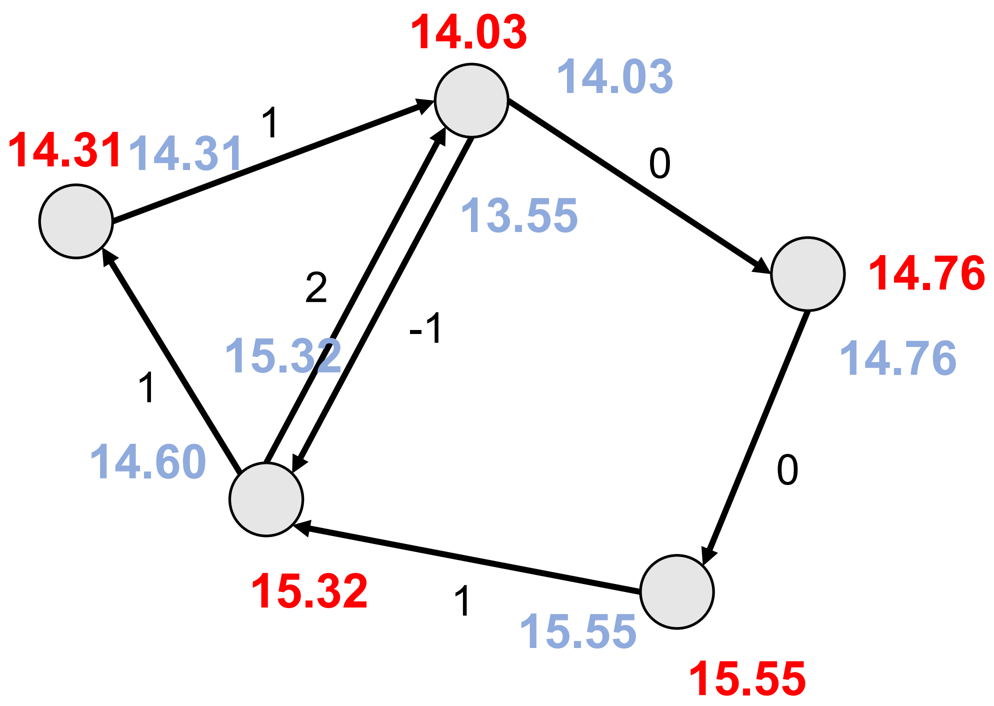
$$\boldsymbol{V_i(s)} \leftarrow \max_a Q_i(s,a)$$



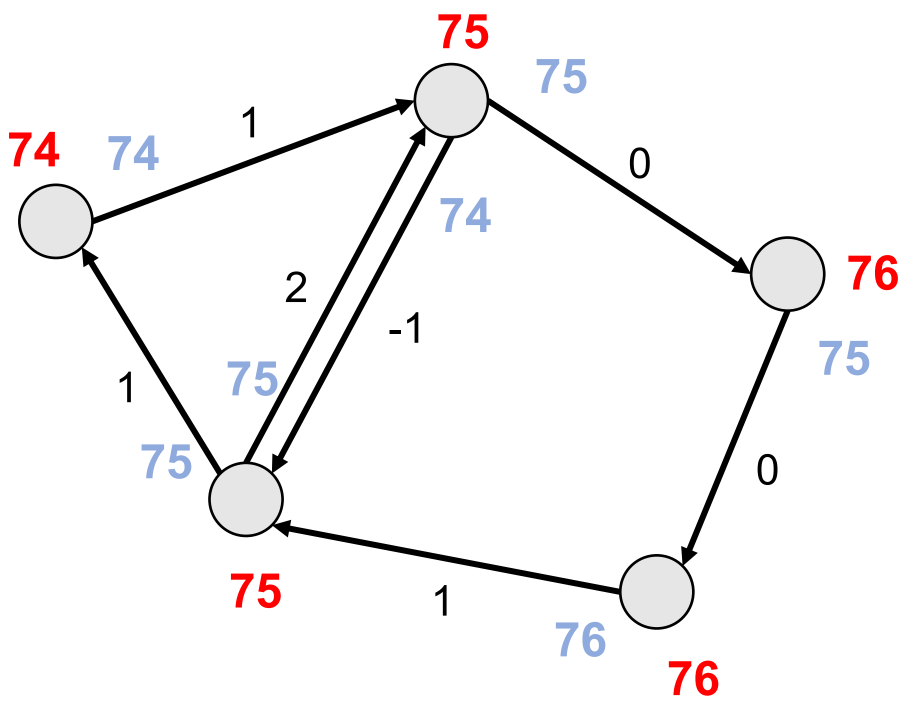
Iteration $i = 21$

$$\boldsymbol{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$\boldsymbol{V_i(s)} \leftarrow \max_a Q_i(s,a)$$



$\gamma = 0.95$

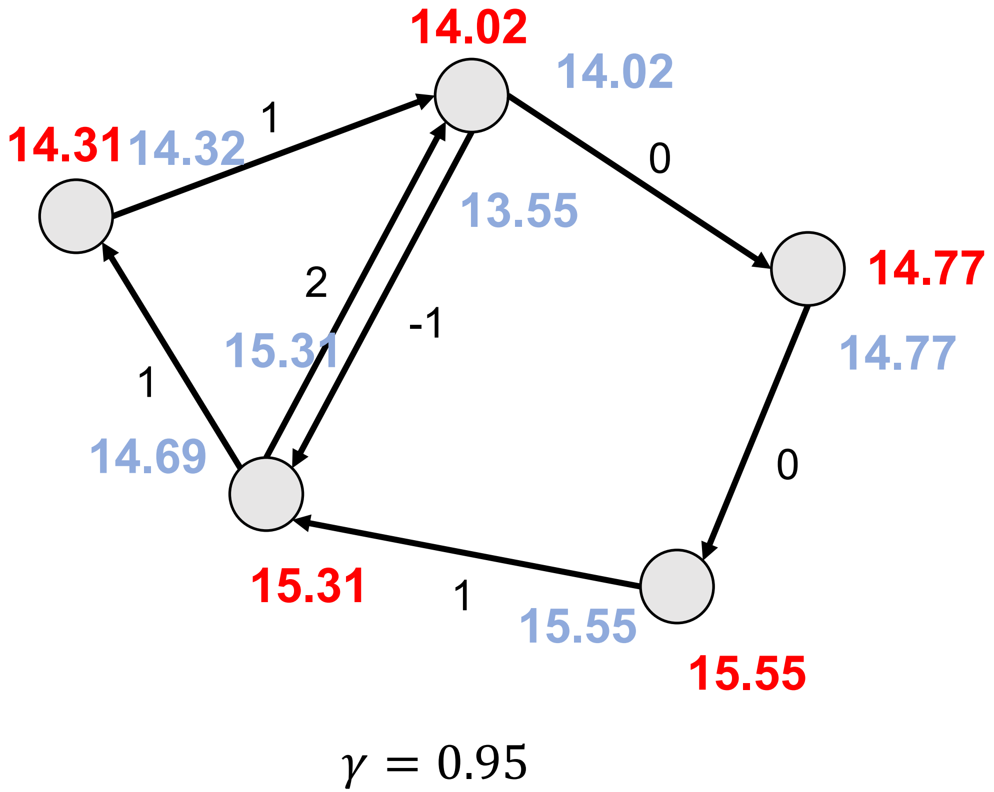


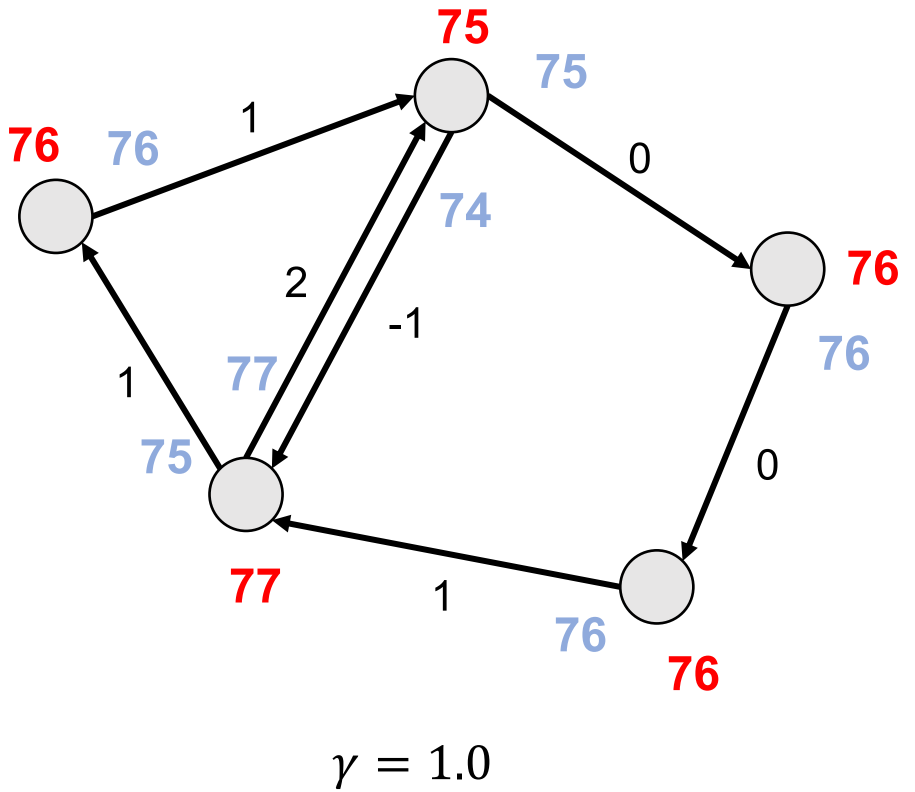
$\gamma = 1.0$

Iteration $i = 100$

$$\color{blue}{Q_i(s,a)} \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_{i-1}(s')$$

$$\color{red}{V_i(s)} \leftarrow \max_a Q_i(s,a)$$





Iteration $i = 101$