Exercise 3.24 Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places. □

return:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

Sử dụng ký hiệu $G_t^{(A)}$ cho return bắt đầu ở **A**:

$$\begin{aligned} G_t^{(A)} &= 10 + \gamma \cdot 0 + \gamma^2 \cdot 0 + \gamma^3 \cdot 0 + \gamma^4 \cdot 0 + \gamma^5 R_{t+6} + \gamma^6 R_{t+7} + \dots \\ &= 10 + \gamma^5 G_{t+5}^{(A)}. \end{aligned}$$

2 Giá trị trạng thái là kỳ vọng của return

Giá trị tối ưu tại **A** được định nghĩa:

$$v_*(A) = \mathbb{E}\left[G_t \,\middle|\, S_t = A,\ \pi = \pi_* \right].$$

Nói cách khác, **lấy (3.8) rồi lấy kỳ vọng** dưới chính sách tối ưu.

Lấy kỳ vọng hai vế ⇒ phương trình Bellman "rút gọn"

$$\begin{aligned} v_*(A) &= \mathbb{E}[\, G_t^{(A)}] = 10 + \gamma^5 \, \mathbb{E}[\, G_{t+5}^{(A)}] \\ &= 10 + \gamma^5 \, v_*(A). \end{aligned}$$

Đây chính là công thức mình đã viết.

$$\boxed{v_*(A) = 10 + \gamma^5 \, v_*(A)}$$

$$v_*(A) = \frac{10}{1 - \gamma^5}$$

với $\gamma = 0.9$ (thông số của ví dụ):

$$\gamma^5 = 0.9^5 = 0.59049, \qquad 1 - \gamma^5 = 0.40951$$

$$\boxed{v_*(A) = \frac{10}{0.40951} \approx 24.419}$$