Reinforcement Learning
MDP và bài toán Frozen Lake

Thành viên nhóm:

Phùng Anh Hùng - 20173150

Nguyễn Đức Vượng - 20173603

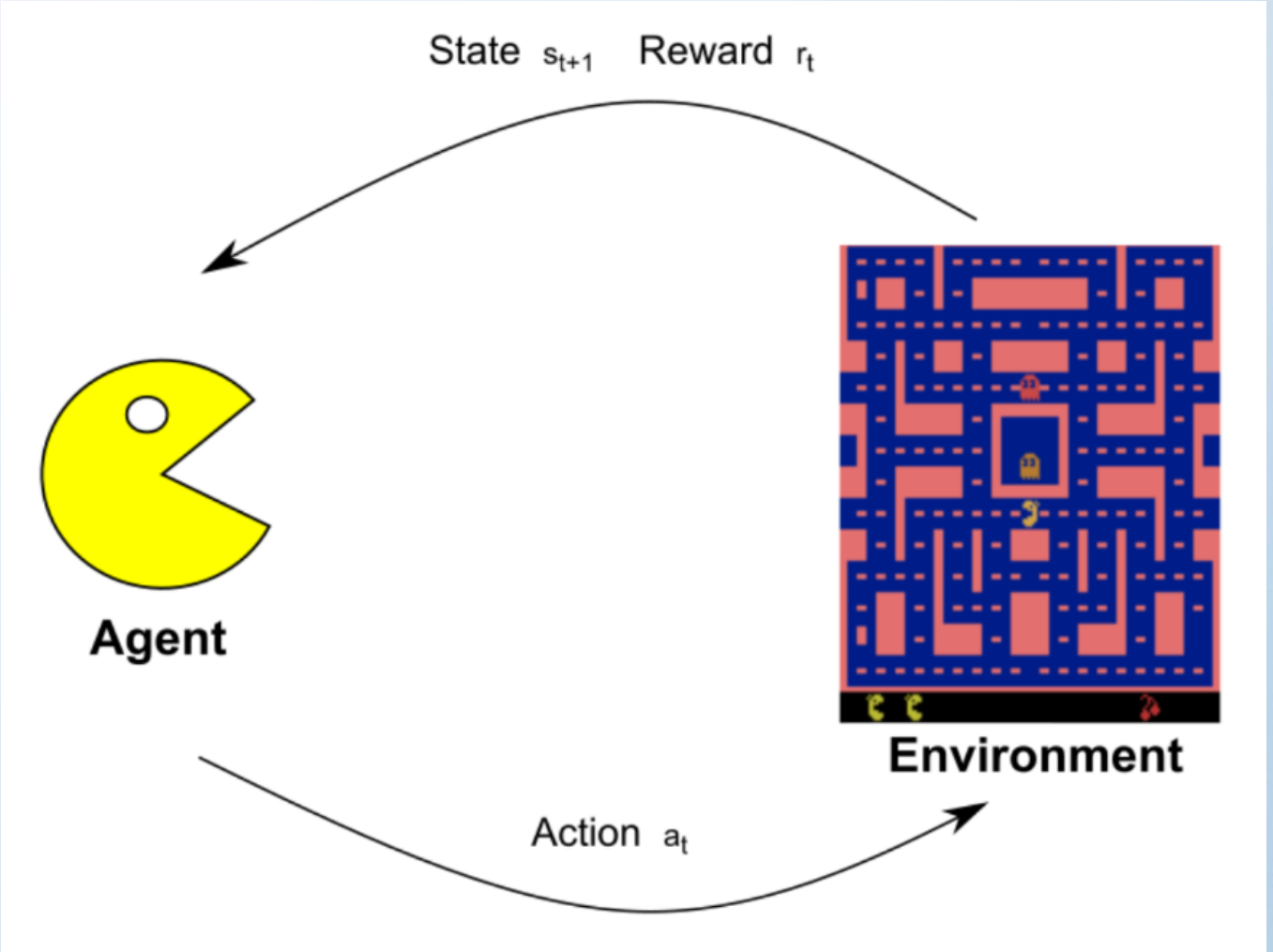
Ngô Việt Trung - 20173415

Nội dung

- Reinforcement Learning và ứng dụng
- Quá trình quyết định Markov
- Bài toán Frozen Lake

Reinforcement Learning là gì?

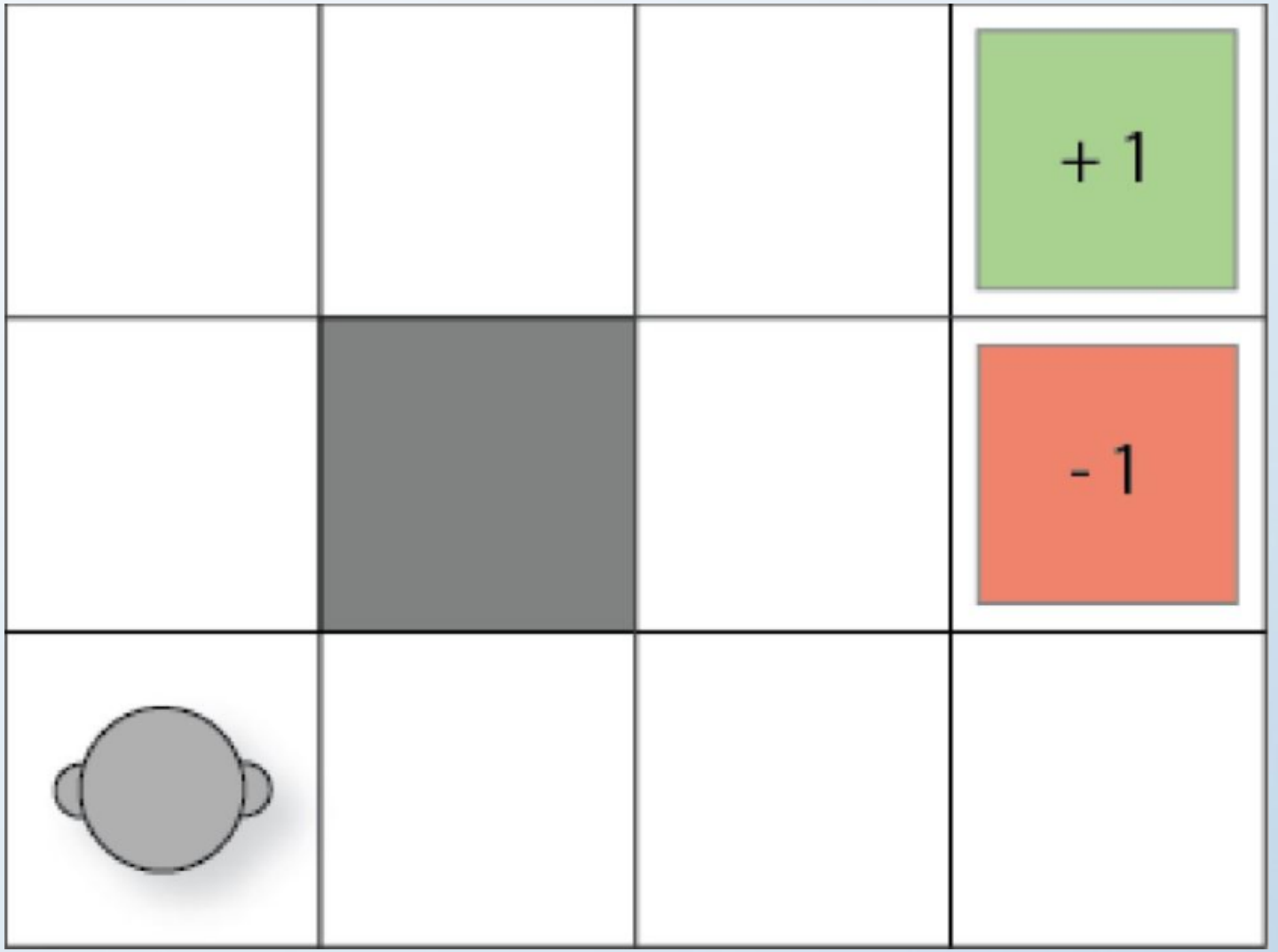
- Lĩnh vực máy học
- Quyết định tuần tự
- Đạt mục tiêu mong muốn



Thành phần của Reinforcement Learning

- Agent
- Environment
- Action
- State
- Reward
- Policy
- Value function

Ví dụ



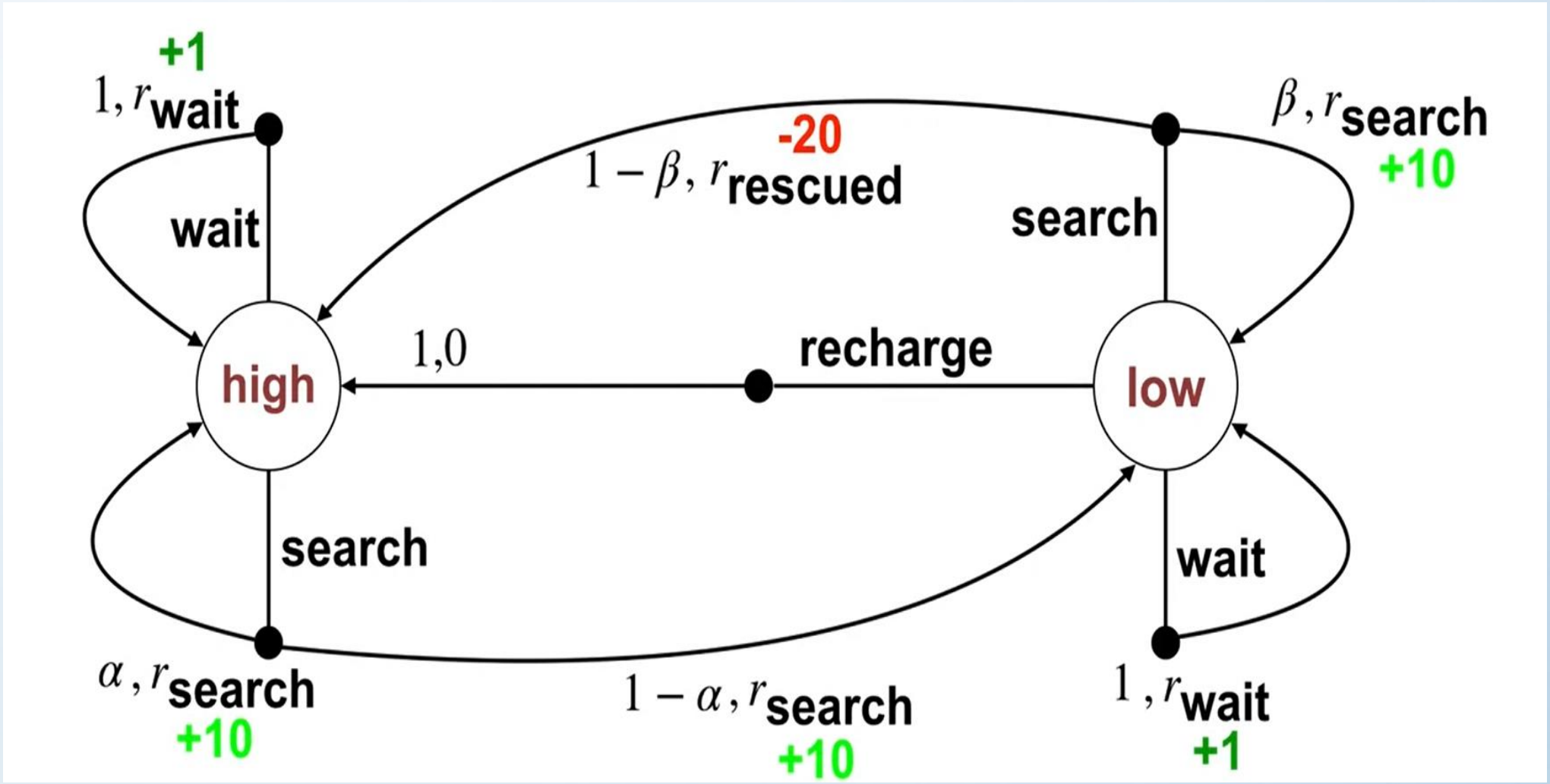
Ứng dụng của RL

- Một số ứng dụng của Reinforcement Learning:
  - Rô-bốt trong công nghiệp tự động hóa và hàng không
  - Xây dựng chiến lược kinh doanh
  - Máy học và xử lí dữ liệu
  - Q-learning và Deep Q-learning
  - Thuật toán giúp “máy chơi game”

Quá trình quyết định Markov

- Giới thiệu MDP
- Định nghĩa MDP
- Hàm trả về
- Hàm giá trị
- Phương trình Bellman
- Đánh giá và cải thiện chính sách
- Thuật toán lặp chính sách
- Thuật toán lặp giá trị

Giới thiệu MDP



Định nghĩa MDP

- MDP thể hiện việc đưa ra các quyết định theo thứ tự, trong đó các hành động ảnh hưởng đến trạng thái và kết quả
- Biểu diễn bằng bộ 4 dữ liệu (S, A, P, R):
  - S: Không gian trạng thái
  - A: Tập hữu hạn hành động
  - P: Hàm chuyển tiếp $P(s', s, a) = p(s'|s, a)$ xác đinh xác suất đạt trạng thái s' từ trạng thái s thông qua hành động a
  - R: Hàm phần thưởng

Hàm trả về

- Quỹ đạo τ có dạng: $(S_0, A_0, S_1, A_1, \dots)$
- Mỗi quỹ đạo đem lại một chuỗi phần thưởng
- Hàm trả về là hàm có dạng:

$$G(\tau) = r_0 + r_1 + \cdots = \sum_{t=0}^{\infty} r_t$$

Hàm trả về

- Tăng giá trị của phần thưởng ngắn hạn = giảm giá trị của phần thưởng trong tương lai xa
- Sử dụng hệ số chiết khấu γ
- Công thức hàm trả về:

$$G(\tau) = r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots = \sum_{t=0}^{\infty} \gamma^t r_t$$

Hàm trả về

- Đặt $G_t = G_t(\tau)$ là tổng trả về tính từ bước thứ t, ta có công thức:

$$G_t(\tau) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \dots$$

- Từ đó suy ra công thức:

$$G_t = r_t + \gamma G_{t+1}$$

Hàm giá trị

- Hàm $G(\tau)$ chỉ cho ta cái nhìn về phần thưởng nhận được qua cả quỹ đạo
- Hàm giá trị tính kì vọng của một trạng thái theo một chính sách π

$$V_\pi(s) = E_\pi[G|s_0 = s] = E_\pi[\sum_{t=0}^{k} \gamma^t r_t|s_0 = s]$$

Hàm giá trị hành động

- Khi xét một hành động nhất định ở một trạng thái, ta được hàm giá trị hành động:

$$Q_\pi(s,a) = E_\pi[G|s_0 = s, a_0 = a]$$

$$= E_\pi[\sum_{t=0}^{k} \gamma^t r_t | s_0 = s\,, a_0 = a]$$

Phương trình Bellman

$$V_\pi(s) = E_\pi[G_t|s_0 = s] = E_\pi[r_t + \gamma G_{t+1}|s_0 = s]$$
$$= E_\pi[r_t + \gamma V_\pi(s_{t+1})|s_t = s, a_t \sim \pi(s_t)]$$

$$Q_\pi(s,a) = E_\pi[G_t|s_t = s, a_t = a]$$
$$= E_\pi[r_t + \gamma G_{t+1}|s_t = s, a_t = a]$$
$$= E_\pi[r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})|s_t = s, a_t = a]$$

Đánh giá chính sách

- Áp dụng phương trình Bellman, ta cải thiện giá trị có thể nhận được từ một chính sách π tại trạng thái s theo công thức:

$$V_{k+1}(s) = E_\pi[r_t + \gamma V_k(s_{t+1})|s_t = s]$$
$$= \sum_a \pi(s,a) \sum_{s',r} p(s'|s,a)[r + \gamma V_k(s')]$$

Cải thiện chính sách

- Sử dụng hàm giá trị đã được cải thiện để tính chính sách π' tốt hơn theo công thức:

$$\pi' = argmax_a Q_\pi(s,a)$$
$$= argmax_a \sum_{s',r} p(s'|s,a)[r + \gamma V_\pi(s')]$$

Thuật toán lặp chính sách

- Khởi tạo $V_\pi(s)$ và $\pi(s)$ cho mỗi trạng thái s
- While π không ổn định:

    While V_π không ổn định:

        For mỗi trạng thái s:

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

    For mỗi trạng thái s:

$$\pi = argmax_a \sum_{s',r} p(s'|s, a)[r + \gamma V_\pi(s')]$$

Thuật toán lặp giá trị

- Khởi tạo $V(s)$ cho mỗi trạng thái s
- While $V(s)$ không ổn định:

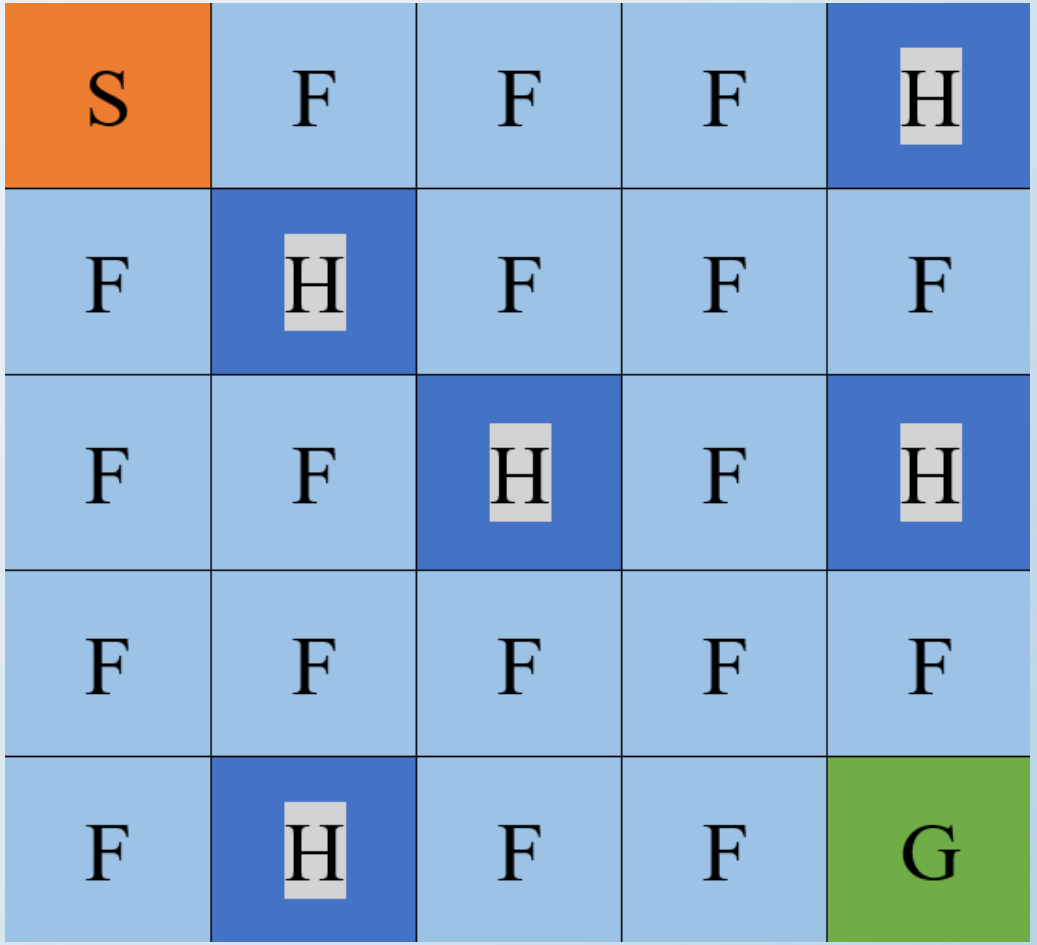
  For mỗi trạng thái s:

$$V(s) = \max_a \sum_{s',r} p(s'|s,a)[r + \gamma V(s')]$$

$$\pi = argmax_a \sum_{s',r} p(s'|s,a)[r + \gamma V(s)]$$

Bài toán Frozen Lake (Hồ băng)

- S: Ô khởi đầu
- G: Ô đích
- F: Ô an toàn
- H: Ô hố



Câu hỏi và nhận xét

Cảm ơn cô và các bạn đã lắng nghe