
2102470 Học máy

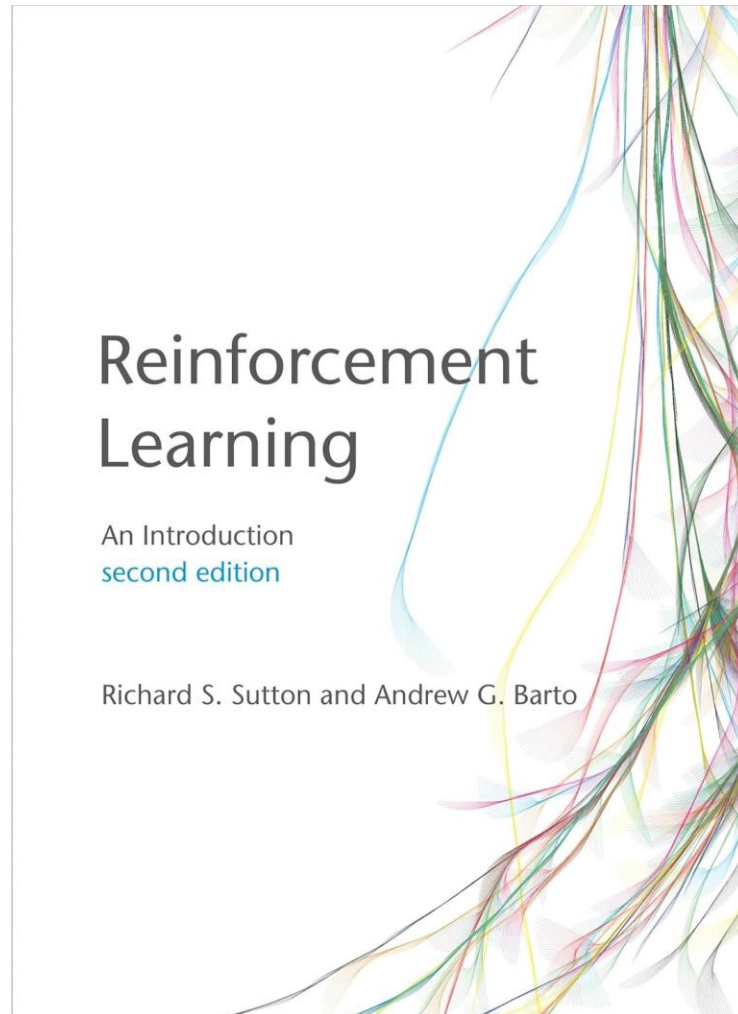
Bài giảng: Các thuật toán học tăng cường

Chương 4: Học tăng cường

Ôn lại bài học trước

- Bạn có nhớ ? % ?

- [B12TLTK1]



<http://incompleteideas.net/book/the-book-2nd.html>

Reinforcement Learning: An Introduction

[Richard S. Sutton](#)
and [Andrew G. Barto](#)

Second Edition (see [here](#) for the first edition)
MIT Press, Cambridge, MA, 2018

[Buy from Amazon](#)

[Errata and Notes](#)

[Full Pdf](#) [Without Margins](#)

[Code](#)

[Solutions](#) -- send in your solutions for a chapter, get the official ones back (currently incor
[Slides and Other Teaching Aids](#)

[Links to pdfs of the literature sources cited in the book](#) (Many thanks to Daniel Plopl)

Latex Notation -- Want to use the book's notation in your own work? Download this [.sty fi](#)

<https://mitpress.mit.edu/9780262039246/reinforcement-learning/>

Nội dung chính

- 4.2 Các thuật toán học tăng cường
 - 4.2.1 Thuật toán lặp giá trị
 - 4.2.2 Thuật toán lặp chiến lược
 - 4.2.3 Thuật toán Q-learning
 - 4.2.4 Các thuật toán khác

4.2.1.Value iteration algorithm

- Thuật toán lặp giá trị
 - Một trong những thuật toán cơ bản
 - Giúp tác tử xác định hành động tốt nhất cần thực hiện trong mỗi trạng thái để tối đa hóa phần thưởng dài hạn
 - Thuật toán dựa trên mô hình
 - Yêu cầu mô hình đầy đủ của môi trường (như xác suất chuyển đổi trạng thái và phần thưởng)
 - => Phù hợp với các môi trường nhỏ, đơn giản

Thuật toán lặp giá trị

- Cập nhật giá trị của từng trạng thái (theo cách lặp lại) bằng cách sử dụng phương trình tối ưu Bellman cho đến khi các giá trị “hội tụ”

Phương trình tối ưu Bellman

$$V^*(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^*(s')]$$

Thuật toán lặp giá trị

- Các bước

- B1: Khởi tại ban đầu cho giá trị của mỗi trạng thái, thường là 0

$$V_0(s) = 0$$

- B2: Cập nhật $V(s)$ cho các trạng thái s

$$V_{k+1}(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$

- B3: Nếu giá trị hội tụ thì gán là giá trị tối ưu $V^*(s)$ và chuyển sang bước 4, nếu không thì lặp lại bước 2

- B4: Xác định chính sách/chiến thuật tối ưu

$$\pi^*(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^*(s')]$$

Thuật toán lặp giá trị

- [B12TLTK1]

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\text{terminal}) = 0$

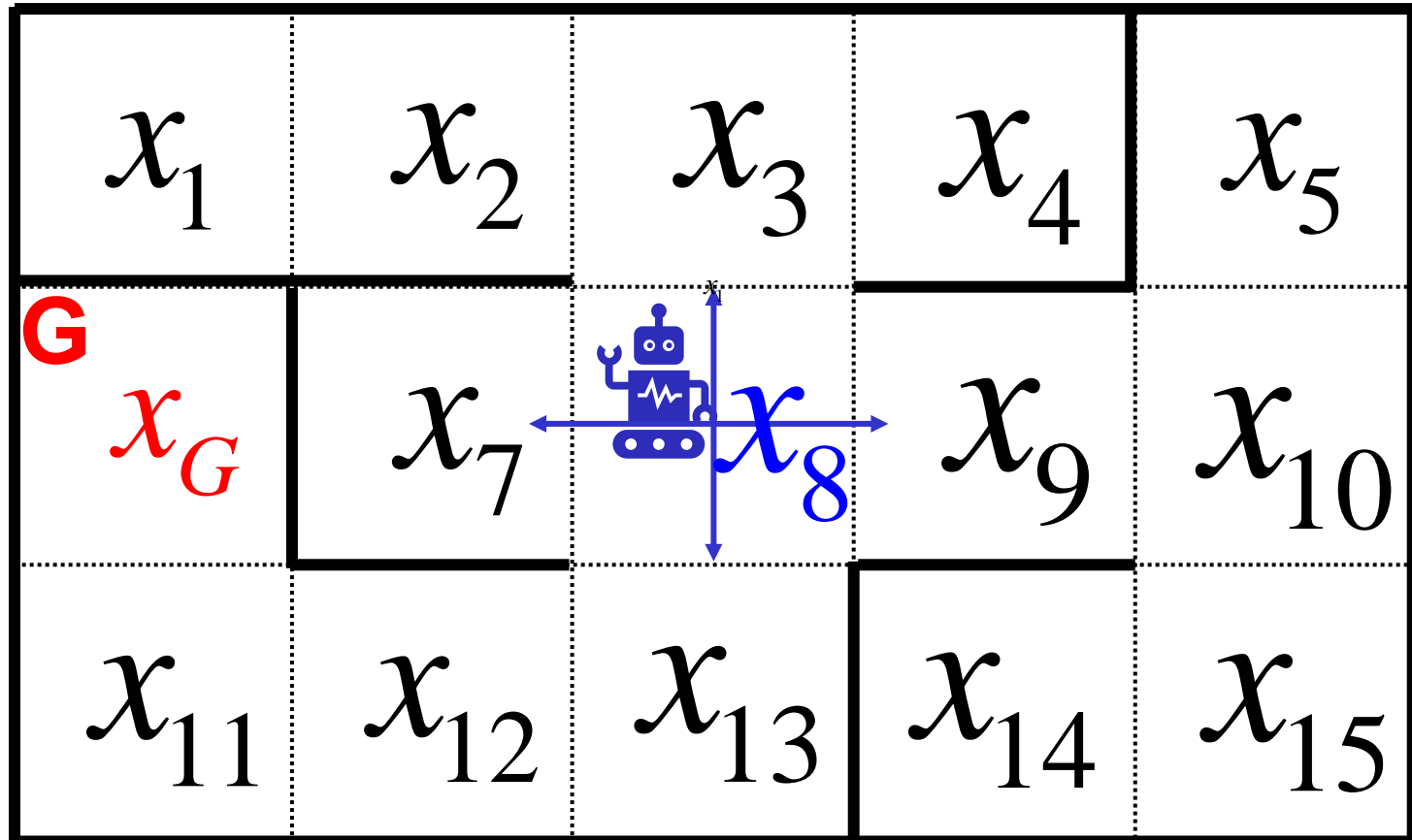
Loop:

```
|  $\Delta \leftarrow 0$   
| Loop for each  $s \in \mathcal{S}$ :  
|    $v \leftarrow V(s)$   
|    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$   
|    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$   
until  $\Delta < \theta$ 
```

Output a deterministic policy, $\pi \approx \pi_*$, such that
$$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$$

Ví dụ 1

- Ví dụ 1: Bài toán robot di chuyển trong mê cung
 - Sử dụng thuật toán lặp giá trị tìm đường đi cho robot di chuyển đến đích G



Ví dụ 1

- Tập trạng thái: $\{x_1, x_2, \dots, x_{15}, x_G\}$ x_G : trạng thái đích
- Tín hiệu điều khiển u_k tại một trạng thái là tập con
$$U(x) = \{\text{Up, Down, Left, Right, No move}\}$$
- Hàm chuyển trạng thái $f(x_k, u_k)$ xác định trạng thái tiếp theo nếu thực hiện tín hiệu điều khiển u_k
- Tín hiệu củng cố ~ reward tại bước lặp thứ i :
$$r_i(x_k, u_k) = -1 \quad k = 1, 2, \dots, 15 \quad u_k \in U(x_k)$$
- Tập hiệu củng cố tại trạng thái đích:
$$r_i(x_G, u) = 0 \quad \forall u \in U(x_G)$$

Ví dụ 1

- Chú ý:

- Có thể có các cách biểu diễn khác

$V^{(i)}(x_k)$: Giá trị của trạng thái x_k tại vòng lặp thứ i

- Để đơn giản ta xét các chuyển đổi là xác định, $T = 1$ cho hành động đã chọn
- Thực tế: Nếu các chuyển đổi trạng thái mang tính không xác định
 - Hành động của tác tử có thể không phải lúc nào cũng dẫn đến trạng thái mong muốn
 - Ví dụ, nếu tác tử cố gắng di chuyển lên, có khả năng nó có thể di chuyển sang trái hoặc phải
 - Xác suất: Giả sử tác tử thành công trong việc di chuyển theo hướng mong muốn với xác suất 80% và lệch sang một bên (hướng vuông góc) với xác suất 10% cho mỗi hướng

Ví dụ 1

- Chú ý $\gamma = 1$

$$V_{k+1}(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$



$$V_{i+1}(x_k) = \max_u \left(r(x_k, u, x_j) + V_i(x_j) \right)$$



$$V_{i+1}(x_k) = \max \left(r(x_k, u) + V_i(x_j) \right)$$

Tại vòng lặp thứ i

Ví dụ 1

- B1: Thiết lập ban đầu $V_0(x_k) = 0$

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_1

$$x_2 = f(x_1, \text{Right})$$

$$x_1 = f(x_1, \text{Nomove})$$

$$\begin{aligned} V_1(x_1) &= \max(r_1(x_1, \text{Right}) + V_0(x_2); r_1(x_1, \text{Nomove}) + V_0(x_1)) \\ &= \max(-1 + 0; -1 + 0) = \max(-1; -1) = -1 \end{aligned}$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_2

$$x_1 = f(x_2, \text{Left})$$

$$x_3 = f(x_2, \text{Right})$$

$$x_2 = f(x_2, \text{Nomove})$$

$$V_1(x_2) = \max \left(\begin{array}{l} r_1(x_2, \text{Left}) + V_0(x_1); r_1(x_2, \text{Right}) + V_0(x_3); \\ r_1(x_2, \text{Nomove}) + V_0(x_2) \end{array} \right)$$

$$= \max(-1 + 0; -1 + 0; -1 + 0) = \max(-1; -1; -1) = -1$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_3

$$x_2 = f(x_3, \text{Left}) \quad x_3 = f(x_3, \text{Nomove}) \quad x_4 = f(x_3, \text{Right}) \\ x_8 = f(x_3, \text{Down})$$

$$V_1(x_3) = \max \left(\begin{array}{l} r_1(x_3, \text{Down}) + V_0(x_8); r_1(x_3, \text{Left}) + V_0(x_2); \\ r_1(x_3, \text{Right}) + V_0(x_4); r_1(x_3, \text{Nomove}) + V_0(x_3) \end{array} \right) \\ = \max(-1 + 0; -1 + 0; -1 + 0; -1 + 0) \\ = \max(-1; -1; -1; -1) = -1$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_4

$$x_3 = f(x_4, \text{Left})$$

$$x_4 = f(x_4, \text{Nomove})$$

$$\begin{aligned} V_1(x_4) &= \max(r_4(x_4, \text{Left}) + V_0(x_3); r_4(x_4, \text{Nomove}) + V_0(x_4)) \\ &= \max(-1 + 0; -1 + 0) = \max(-1; -1) = -1 \end{aligned}$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_8

$$x_3 = f(x_8, \text{Up})$$

$$x_7 = f(x_8, \text{Left}) \quad x_8 = f(x_8, \text{Nomove}) \quad x_9 = f(x_8, \text{Right})$$
$$x_{13} = f(x_8, \text{Down})$$

$$V_1(x_8) = \max \left(\begin{array}{l} r_1(x_8, \text{Up}) + V_0(x_3); r_1(x_8, \text{Down}) + V_0(x_{13}); \\ r_1(x_8, \text{Left}) + V_0(x_7); r_1(x_8, \text{Right}) + V_0(x_9); \\ r_1(x_8, \text{Nomove}) + V_0(x_8) \end{array} \right)$$

$$= \max(-1 + 0; -1 + 0; -1 + 0; -1 + 0; -1 + 0)$$

$$= \max(-1; -1; -1; -1; -1) = -1$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_{11}

$$x_G = f(x_{11}, \text{Up})$$

$$x_4 = f(x_4, \text{Nomove}) \quad x_{12} = f(x_{11}, \text{Right})$$

$$\begin{aligned} V_1(x_{11}) &= \max \left(\begin{array}{l} r_1(x_{11}, \text{Up}) + V_0(x_G); r_1(x_{11}, \text{Right}) + V_0(x_{12}); \\ r_1(x_{11}, \text{Nomove}) + V_0(x_{11}) \end{array} \right) \\ &= \max(-1 + 0; -1 + 0; -1 + 0) = \max(-1; -1; -1) = -1 \end{aligned}$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1

Trạng thái x_G

Có cần phải tính toán cụ thể ?

$$x_G = f(x_G, \text{Nomove})$$

$$x_{11} = f(x_G, \text{Down})$$

$$V_1(x_G) = \max \left(\begin{array}{l} r_1(x_G, \text{Down}) + V_0(x_{11}); \\ r_1(x_G, \text{Nomove}) + V_0(x_G) \end{array} \right)$$

$$= \max(-1 + 0; 0 + 0) = \max(-1; 0) = 0$$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 1
 - Tính toán tương tự cho tất cả các trạng thái



Ví dụ 1

- B2: Cập nhật giá trị, vòng 1 $V_1(x_k)$

-1	-1	-1	-1	-1
0	-1	-1	-1	-1
-1	-1	-1	-1	-1

Ví dụ 1

- B2: Cập nhật giá trị, vòng 2 $V_2(x_k)$

-2	-2	-2	-2	-2
0	-2	-2	-2	-2
-1	-2	-2	-2	-2

Ví dụ 1

- Sinh viên tự thực hiện một số vòng lặp để hiểu rõ về thuật toán
 - B2: Cập nhật giá trị, vòng j $V_j(x_k)$

Ví dụ 1

- B2: Cập nhật giá trị, vòng 3 $V_3(x_k)$

-3	-3	-3	-3	-3
0	-3	-3	-3	-3
-1	-2	-3	-3	-3

Ví dụ 1

- B2: Cập nhật giá trị, vòng 4 $V_4(x_k)$

-4	-4	-4	-4	-4
0	-4	-4	-4	-4
-1	-2	-3	-4	-4

Ví dụ 1

- B2: Cập nhật giá trị, vòng 5 $V_5(x_k)$

-5	-5	-5	-5	-5
0	-5	-4	-5	-5
-1	-2	-3	-5	-5

Ví dụ 1

- B2: Cập nhật giá trị, vòng 6 $V_6(x_k)$

-6	-6	-5	-6	-6
0	-5	-4	-5	-6
-1	-2	-3	-6	-6

Ví dụ 1

- B2: Cập nhật giá trị, vòng 7 $V_7(x_k)$

-7	-6	-5	-6	-7
0	-5	-4	-5	-6
-1	-2	-3	-7	-7

Ví dụ 1

- B2: Cập nhật giá trị, vòng 8 $V_8(x_k)$

-7	-6	-5	-6	-7
0	-5	-4	-5	-6
-1	-2	-3	-8	-7

Ví dụ 1


- Cập nhật giá trị, vòng 9
 $V^*(x_k)$

$$V_9(x_k) = V_8(x_k)$$

-7	-6	-5	-6	-7
0	-5	-4	-5	-6
-1	-2	-3	-8	-7

Ví dụ 1

- B4: Tín hiệu điều khiển

-7	-6	-5	-6	-7
0	-5	 -4	-5	-6
-1	-2	-3	-8	-7

Ví dụ 1

- B4: Tín hiệu điều khiển

Trạng thái x_8

$$x_3 = f(x_8, \text{Up})$$

$$x_7 = f(x_8, \text{Left}) \quad x_8 = f(x_8, \text{Nomove}) \quad x_9 = f(x_8, \text{Right})$$

$$x_{13} = f(x_8, \text{Down})$$

$$\pi^*(x_8) = \max \left(\begin{array}{l} r(x_8, \text{Up}) + V^*(x_3); \boxed{r(x_8, \text{Down}) + V^*(x_{13})}; \\ r(x_8, \text{Left}) + V^*(x_7); r(x_8, \text{Right}) + V^*(x_9); \\ r(x_8, \text{Nomove}) + V^*(x_8) \end{array} \right)$$

$$= \max(-1-5; -1-3; -1-5; -1-5; -1-4)$$

$$= \max(-6; -4; -6; -6; -5) = -4$$

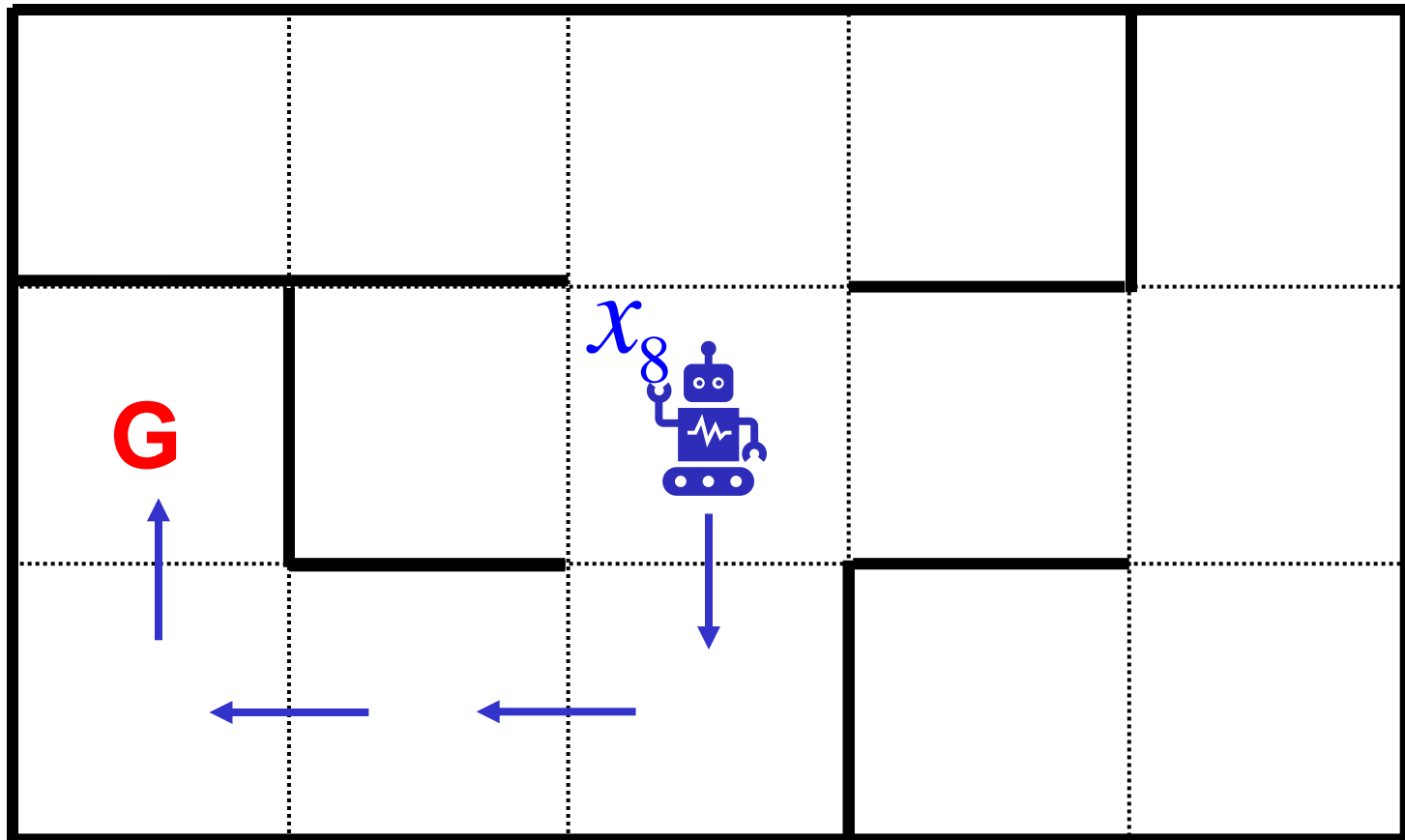
Ví dụ 1

- B4: Tín hiệu điều khiển $x_{13} = f(x_8, \text{Down})$

-7	-6	-5	-6	-7
0	-5	x_8 -4	-5	-6
-1	-2	-3	-8	-7

Ví dụ 1

- Di chuyển robot từ một trạng thái cụ thể x_8 về đích G



4.2.2 Thuật toán lập chiến lược

Policy Iteration (PI) Algorithm

- Thuật toán lặp chiến lược/chính sách
 - Một thuật toán cơ bản khác để giải quyết MDPs, thường được dùng cùng thuật toán lặp giá trị (VI)
 - Đánh giá và cải thiện chính sách theo từng bước cho đến khi nó hội tụ thành một chính sách tối ưu
 - Cách tiếp cận này xen kẽ giữa các bước đánh giá chính sách và cải thiện chính sách để giúp tác tử tìm ra chính sách tốt nhất có thể, nhằm tối đa hóa phần thưởng trong dài hạn

Thuật toán lập chiến lược

- Policy, chiến lược/chính sách π
 - $\pi(s)$ xác định hành động mà tác tử thực hiện trong mỗi trạng thái
 - Lập chiến lược: bắt đầu bằng một chiến lược tùy ý và cải thiện nó theo từng bước để tìm ra một chiến lược tối ưu
 - Đánh giá chiến lược: Tính $V(s)$ cho mỗi trạng thái theo chiến lược hiện tại
 - Cải thiện chiến lược: Cập nhật chiến lược để tối đa hóa hàm giá trị dựa trên các giá trị gần đây nhất của $V(s)$

Thuật toán lặp chiến lược – Các bước

- B1: Khởi tại ban đầu cho chiến lược, ví dụ chọn ngẫu nhiên hành động đối với mỗi trạng thái

- B2: Đánh giá chiến lược

Tính hàm giá trị $V^\pi(s)$ cho tất cả các trạng thái s thông qua

$$V(s) = \sum_{s'} T(s, \pi(s), s') [R(s, \pi(s), s') + \gamma V(s')]$$

- B3: Cải thiện chiến lược

Đối với mỗi trạng thái s , cập nhật chính sách để chọn hành động tối đa hóa phần thưởng mong đợi dựa trên hàm giá trị hiện tại $V(s)$

$$\pi(s) = \arg \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V(s')]$$

- B4: Kiểm tra sự hội tụ

+ Nếu chiến lược không thay đổi trong bước cải thiện, thì chiến lược đã hội tụ thành chiến lược tối ưu và thuật toán kết thúc

+ Nếu chiến lược thay đổi, lặp lại các bước đánh giá chiến lược và cải tiến chiến lược

Thuật toán lặp chiến lược

- [B12TLTK1]

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π , the policy to be evaluated

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation

Initialize $V(s)$ arbitrarily, for $s \in \mathcal{S}$, and $V(\text{terminal})$ to 0

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

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

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$

Thuật toán lặp chiến lược

- [B12TLTK1]

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization

$V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$; $V(\text{terminal}) \doteq 0$

2. Policy Evaluation

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

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

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement

$\text{policy-stable} \leftarrow \text{true}$

For each $s \in \mathcal{S}$:

$\text{old-action} \leftarrow \pi(s)$

$\pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s',r|s,a) [r + \gamma V(s')]$

If $\text{old-action} \neq \pi(s)$, then $\text{policy-stable} \leftarrow \text{false}$

If policy-stable , then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

4.2.3 Thuật toán Q-learning

Thuật toán Q-learning

- Q-values (quality values) $Q(s, a)$

Q-value tối ưu $Q^*(s, a)$



Chiến lược tối ưu

$$\pi^*(s) = \arg \max_a Q^*(s, a)$$

Thuật toán Q-learning – Các bước

- B1: Khởi tại ban đầu cho Q-value, thường là 0

$$Q_0(s, a) = 0$$

- B2: Cập nhật Q-value

$$Q_{k+1}(s, a) = (1 - \alpha) Q_k(s, a) + \alpha \left(r + \gamma \max_a Q_k(s', a') \right)$$

- B3: Lặp lại quá trình này cho tới khi các Q-value hội tụ hoặc tác tử đã học đủ một chiến lược tối ưu cho việc lựa chọn hành động
- B4: Xác định chính sách/chiến thuật tối ưu

Thuật toán Q-learning

- Chú ý:

α : tốc độ học

γ : hệ số chiết khấu

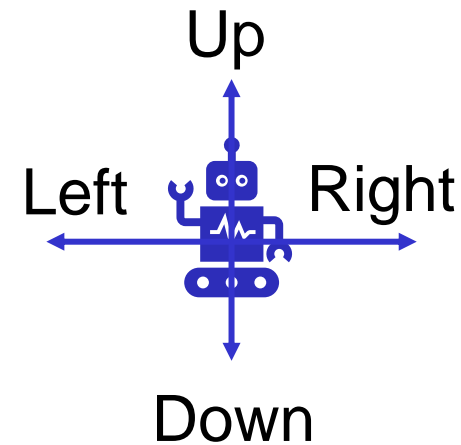
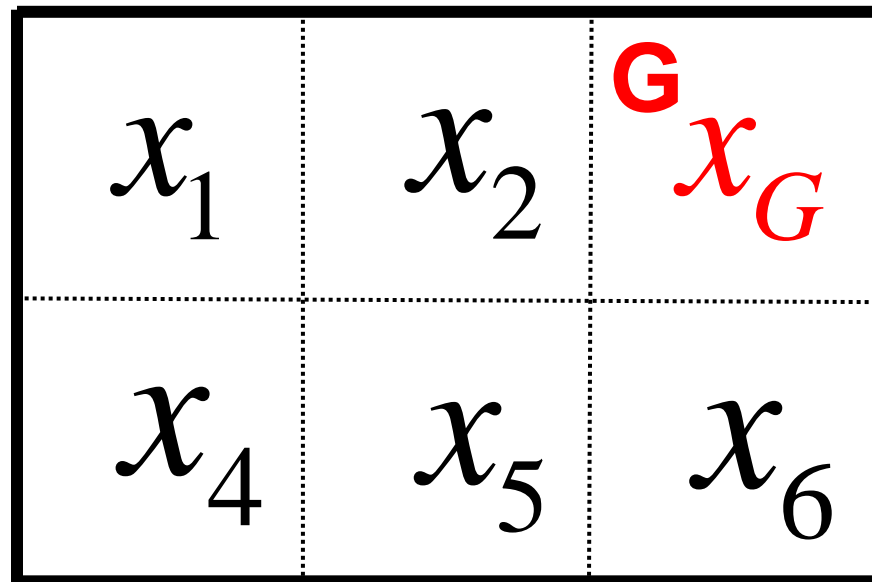
$$Q_{k+1}(s, a) = (1 - \alpha) Q_k(s, a) + \alpha \left(r + \gamma \max_a Q_k(s', a') \right)$$

$$Q_{k+1}(s, a) = Q_k(s, a) + \alpha \left(r + \gamma \max_a Q_k(s', a') - Q_k(s, a) \right)$$

Ví dụ 2

- Ví dụ 2: Tìm đường đi

Tìm chiến lược di chuyển tối ưu x_i đến đích x_G



Ví dụ 2

- Tập trạng thái: $\{x_1, x_2, \dots, x_6, x_G\}$ x_G : trạng thái đích
- Tín hiệu điều khiển u_k tại một trạng thái là tập con
$$U(x) = \{\text{Up, Down, Left, Right}\}$$
- Hàm chuyển trạng thái $f(x_k, u_k)$ xác định trạng thái tiếp theo nếu thực hiện tín hiệu điều khiển u_k
- Tín hiệu củng cố \sim reward tại bước lặp thứ i :

$$r_i(x_G) = 0 \qquad r_i(x_2, \text{Right}) = r_i(x_6, \text{Up}) = 100$$

Còn lại

$$r_i(x_k, u_k) = 0 \qquad k = 1, 2, \dots, 6 \qquad u_k \in U(x_k)$$

Ví dụ 2

- Xét đơn giản

$\alpha = 1$: tốc độ học

$\gamma = 0,9$: hệ số chiết khấu

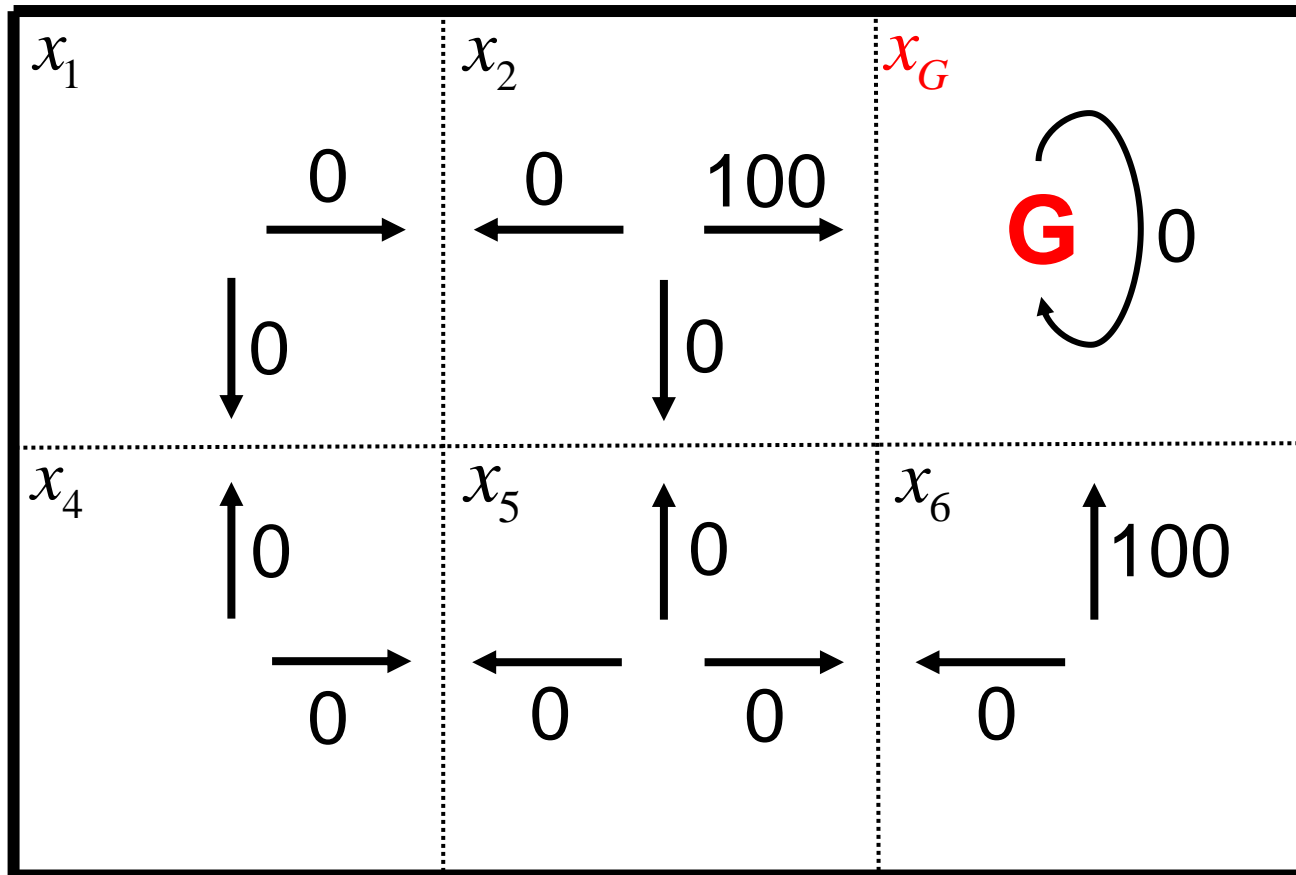
$$Q_{k+1}(s, a) = (1 - \alpha) Q_k(s, a) + \alpha \left(r + \gamma \max_a Q_k(s', a') \right)$$



$$Q_{k+1}(s, a) = r + \gamma \max_a Q_k(s', a')$$

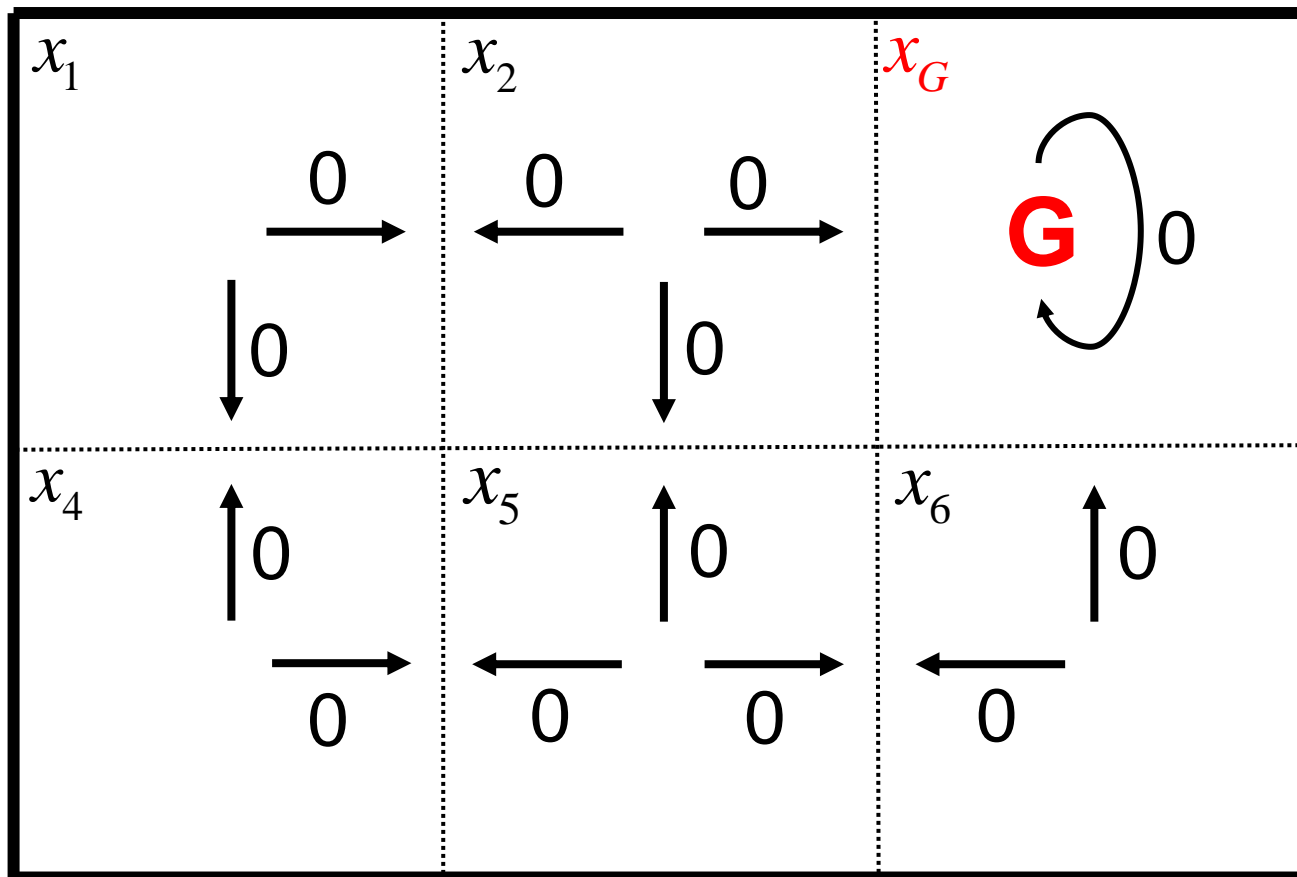
Ví dụ 2

- Ví dụ 2: Minh họa tín hiệu củng cố ~ reward



Ví dụ 2

- B1: Khởi tạo $Q_0(x_k, u_k)$



Ví dụ 2

- Cập nhật lần 1

Tại x_1

$$\begin{aligned} Q_1(x_1, \text{Left}) &= r + \gamma \max_a Q_0(x_2, a') \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

$$\begin{aligned} Q_1(x_1, \text{Down}) &= r + \gamma \max_a Q_0(x_4, a') \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

Ví dụ 2

- Cập nhật lần 1

Tại x_2

$$\begin{aligned} Q_1(x_2, \text{Left}) &= r_1(x_2, \text{Left}) + \gamma \max_a Q_0(x_1, a') \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

$$\begin{aligned} Q_1(x_2, \text{Right}) &= r_1(x_2, \text{Right}) + \gamma \max_a Q_0(x_G, a') \\ &= 100 + 0,9 \times 0 = 100 \end{aligned}$$

$$\begin{aligned} Q_1(x_1, \text{Down}) &= r_1(x_1, \text{Down}) + \gamma \max_a Q_0(x_4, a') \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

Ví dụ 2

- Cập nhật lần 1

Tại x_6

$$\begin{aligned} Q_1(x_6, \text{Left}) &= r_1(x_6, \text{Left}) + \gamma \max_a Q_0(x_5, a') \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

$$\begin{aligned} Q_1(x_6, \text{Up}) &= r_1(x_6, \text{Up}) + \gamma \max_a Q_0(x_G, a') \\ &= 100 + 0,9 \times 0 = 100 \end{aligned}$$

Ví dụ 2

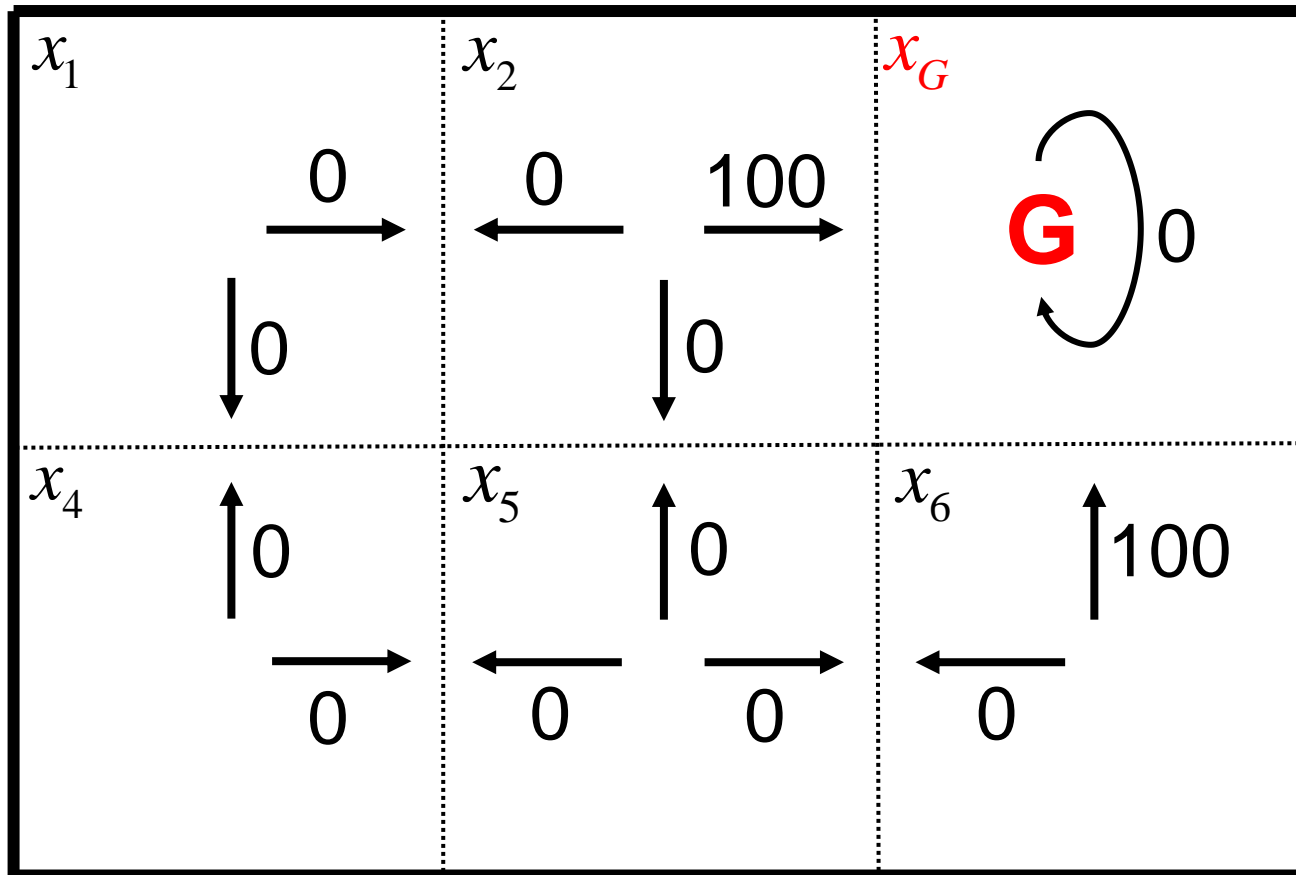
- Cập nhật lần 1

Tại x_G

$$\begin{aligned} Q_1(x_G) &= r_0(x_G) + 0,9 \times Q_0(x_G) \\ &= 0 + 0,9 \times 0 = 0 \end{aligned}$$

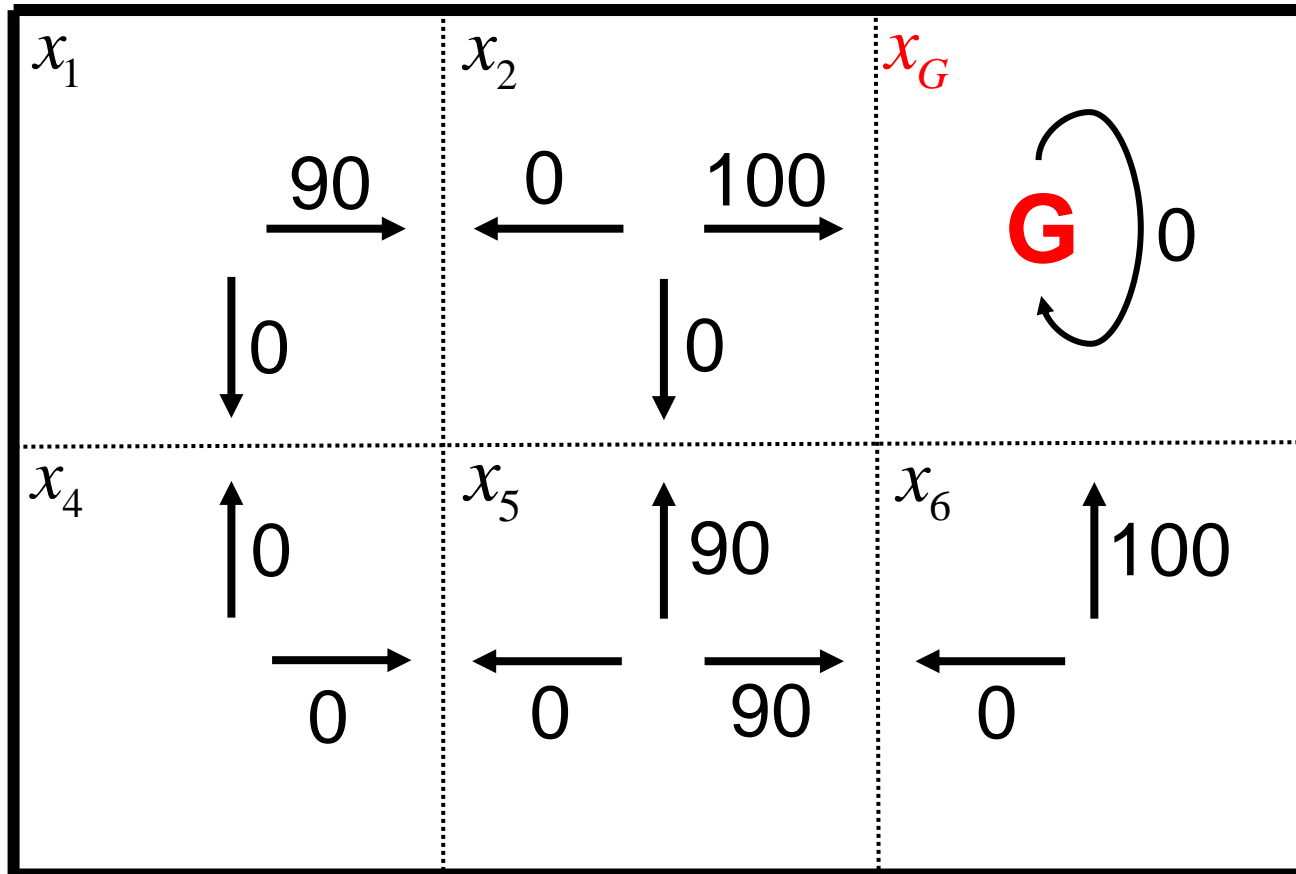
Ví dụ 2

- B2: Cập nhật lần 1 $Q_1(x_k, u_k)$



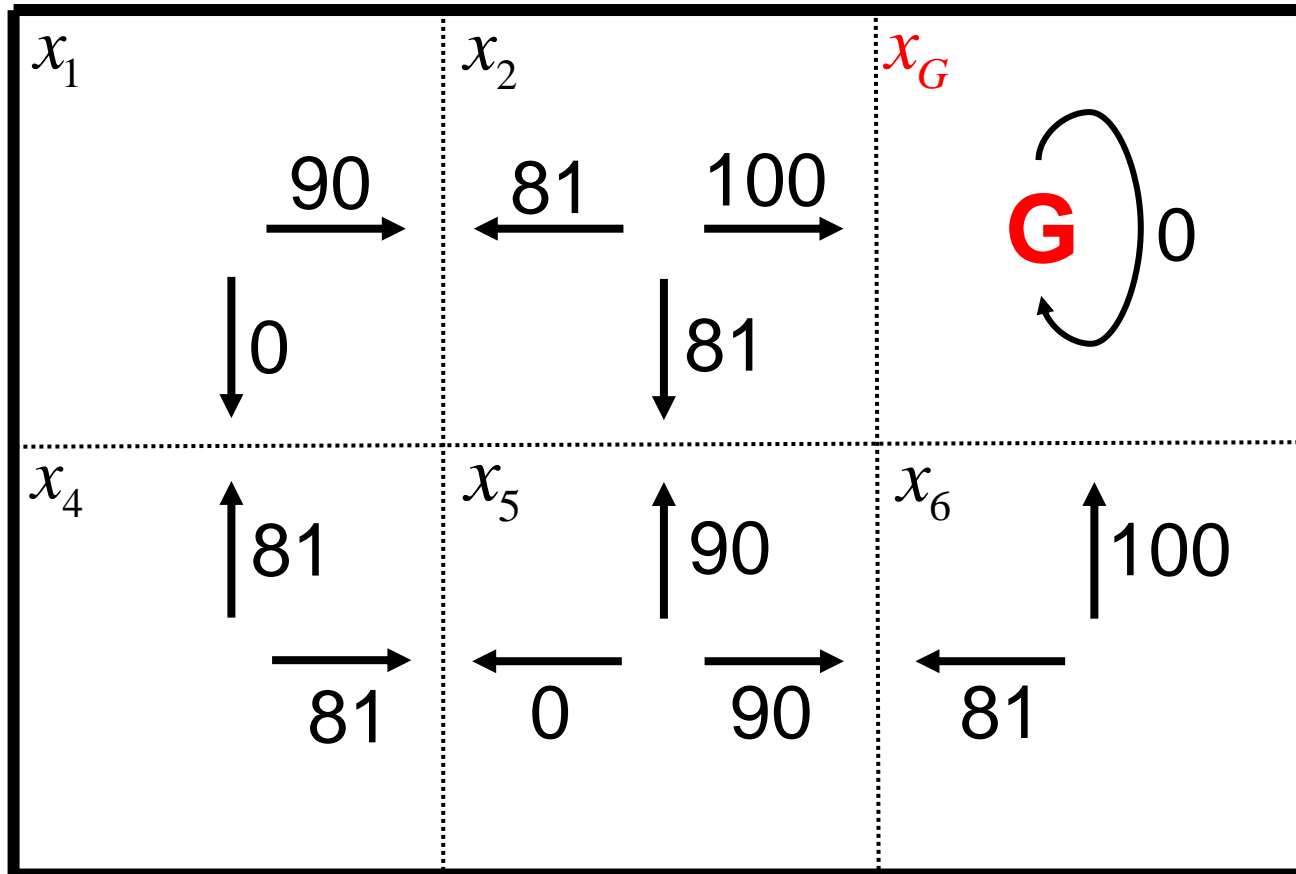
Ví dụ 2

- B2: Cập nhật lần 2 $Q_2(x_k, u_k)$



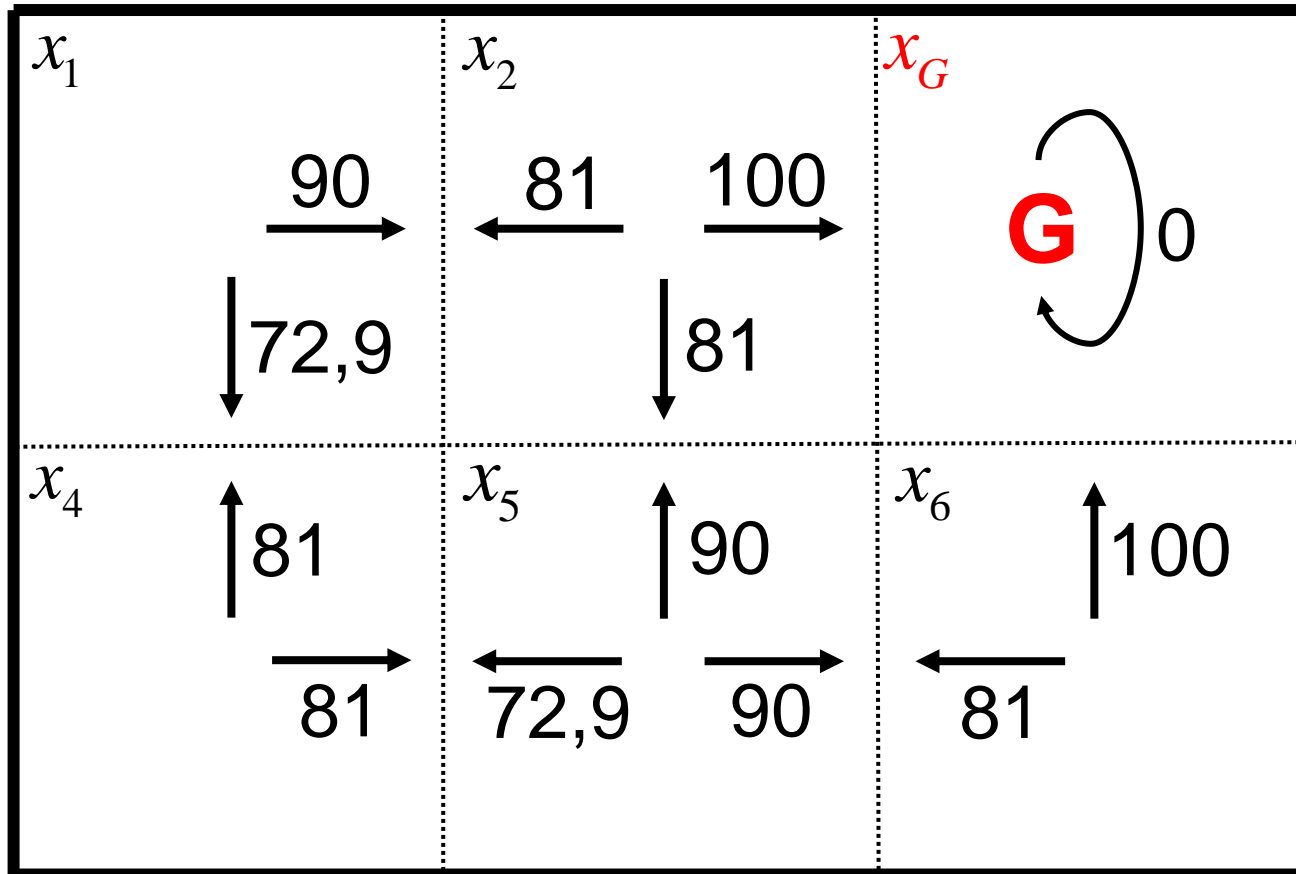
Ví dụ 2

- B2: Cập nhật lần 3 $Q_3(x_k, u_k)$



Ví dụ 2

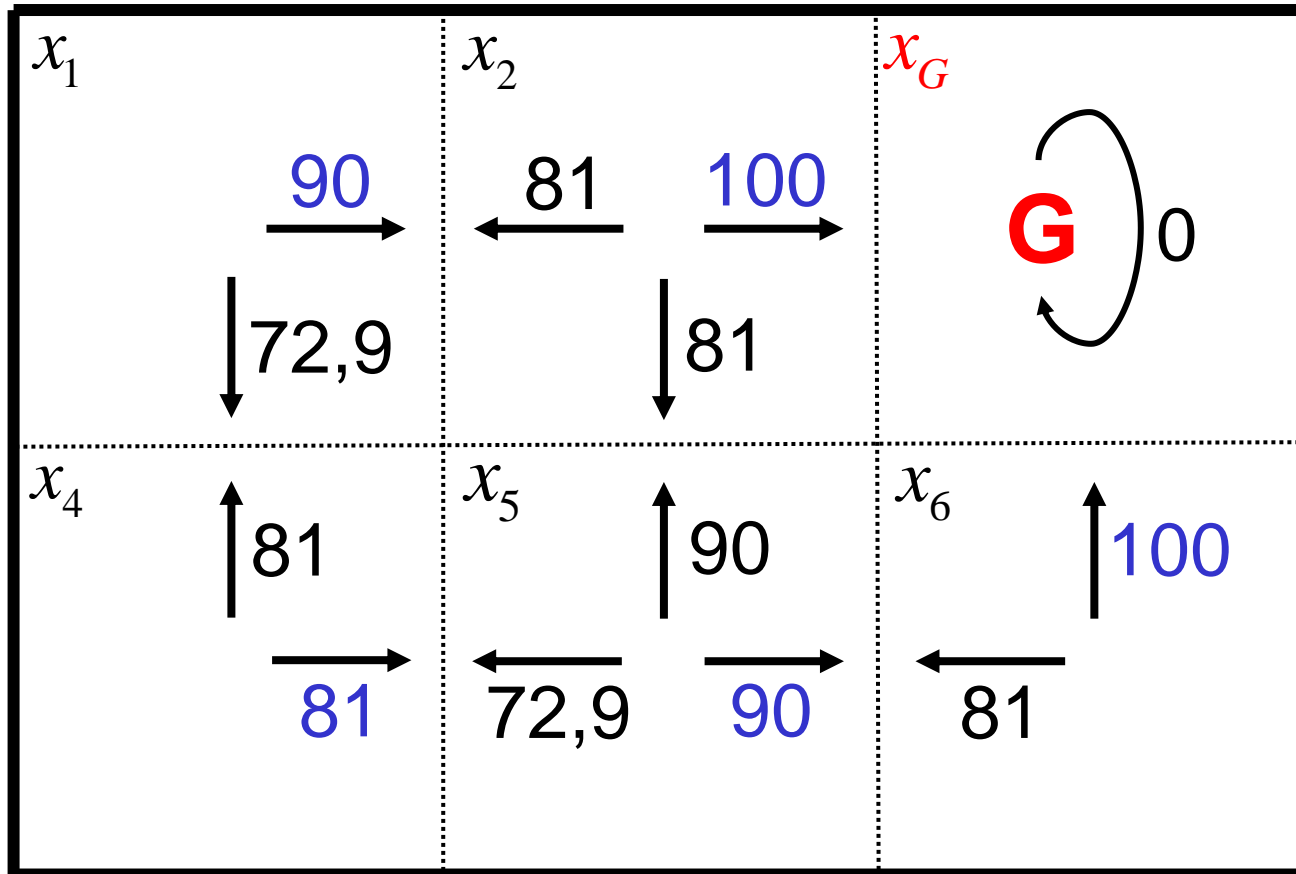
- B2: Cập nhật lần 4 $Q_4(x_k, u_k)$



Ví dụ 2

- Ví dụ 2: Tìm đường đi

Tìm chiến lược di chuyển tối ưu x_i đến đích x_G



4.2.4 Các thuật toán khác

- Approximate Q-learning
- Deep Q-learning
- Actor-critic algorithms
- Asynchronous advantage actor-critic (A3C)
- Advantage actor-critic (A2C)
- ...

Tổng kết

- Sinh viên nắm được các thuật toán cơ bản trong học tăng cường

Hoạt động sau buổi học

- Sinh viên tìm hiểu thêm về các thuật toán khác trong học tăng cường

Chuẩn bị cho buổi học tiếp theo

- Tìm hiểu thêm về các bài toán học tăng cường, ứng dụng của học tăng cường
- Sinh viên ôn tập chuẩn bị cho việc thi cuối kỳ

Tài liệu tham khảo

- [B12TLTK1] Richard S. Sutton and Andrew G. Barto, Reinforcement Learning, second edition: An Introduction, 2nd Edition, MIT Press, 2020
 - <http://incompleteideas.net/book/the-book-2nd.html>
- [B12TLTK1] Giáo Trình Học Máy Và Ứng Dụng Điều Khiển Thông Minh (NXB Đại Học Công Nghiệp 2019) - Nguyễn Tấn Lũy
 - <https://opac.iuh.edu.vn:8000/search/detail.asp?aID=68&ID=110136>