Chứng minh Phương trình Bellman và tính chất hội tụ của giải thuật Q-Learning

Nhat Minh Nguyen [1]

Abstract

Trong lĩnh vực học tăng cường (Reinforcement Learning), một trong những phương pháp cơ bản và cốt lõi nhất là học dựa trên giá trị (value-based). Các phương pháp thường dựa trên phương trình Bellman để xây dựng Q-table hoặc Deep Q-Network. Trong bài báo này, chúng tôi sẽ chứng minh tính đúng đắn của phương trình Bellman và cơ sở hội tụ của thuật toán Q-learning.

1. Cơ sở Toán học

Định nghĩa giá trị kỳ vọng (Expectation) Giá trị kỳ vọng của một biến ngẫu nhiên X được xác định bởi:

$$E[X] = E_{x\sim P}[x] = \sum_x xP(X = x)$$

Tính chất tuyến tính của giá trị kỳ vọng Ta xét tính chất sau:

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

Chứng minh:

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_x \sum_y (ax + by)P(X = x, Y = y) \\
&= \sum_x \sum_y axP(X = x, Y = y) \\
&+ \sum_x \sum_y byP(X = x, Y = y) \\
&= a\sum_x \sum_y xP(X = x, Y = y) \\
&+ b\sum_x \sum_y yP(X = x, Y = y) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y)
\end{aligned}$$

Vậy: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

[1] 25CTT3, Faculty of Information Technology, VNUHCM-University of Science, Ho Chi Minh city, Vietnam. Correspondence to: Nguyen, N. M. <2512021580@student.hcmus.edu.vn>.

2. Phương trình Bellman

Định nghĩa phần thưởng tích lũy (Cumulative Reward)
Gọi phần thưởng tại thời điểm t: R_t
Hệ số chiết khấu (Discount factor): $\gamma \in [0, 1]$
Khai triển tính chất đệ quy của G_t:

$$\begin{aligned}
G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \\
&= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots) \\
&= R_{t+1} + \gamma G_{t+1}
\end{aligned}$$

Công thức tổng quát:

$$G_t = R_{t+1} + \gamma G_{t+1} \qquad (3)$$

Chứng minh tính hội tụ của phần thưởng tích lũy Ta cần chứng minh chuỗi G_t hội tụ. Vì chỉ khi chuỗi này hội tụ, việc tính toán kỳ vọng và xấp xỉ giá trị mới có ý nghĩa về mặt toán học.
Giả thiết:
Ta có: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$
Đặt $R_{max} = \max\{R_{t+1}, R_{t+2}, R_{t+3}, \dots\}$ là giá trị phần thưởng lớn nhất có thể nhận được.
Khai triển chứng minh:

$$\begin{aligned}
|G_t| &\le (1 + \gamma + \gamma^2 + \dots) \cdot R_{max} \\
&= \left(\sum_{k=0}^{\infty} \gamma^k\right) \cdot R_{max} \quad \text{với } \gamma \in [0, 1)
\end{aligned}$$

Xét chuỗi cấp số nhân lùi vô hạn $\sum_{k=0}^{\infty} \gamma^k$:

$$\begin{aligned}
\sum_{k=0}^{\infty} \gamma^k &= \lim_{n\to\infty} \frac{\gamma^n - 1}{\gamma - 1} \\
&= \frac{1}{1-\gamma} \quad \text{(đây là một giá trị hữu hạn)}
\end{aligned}$$

Kết luận:
Chuỗi $\sum_{k=0}^{\infty} \gamma^k \cdot R_{max}$ hội tụ về giá trị $\frac{R_{max}}{1-\gamma}$.
Do đó, $|G_t|$ hội tụ, kéo theo G_t hội tụ tuyệt đối.

Định nghĩa hàm giá trị $V(s)$ Hàm giá trị trạng thái $V(s)$ được định nghĩa là giá trị kỳ vọng của phần thưởng tích lũy G_t khi bắt đầu tại trạng thái s.

$$V(s) = \mathbb{E}[G_t \mid S_t = s]$$

Dựa vào tính chất đệ quy của G_t (mục 3) và tính chất tuyến tính của kỳ vọng (mục 2), ta có:

$$\begin{aligned} V(s) &= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}[G_{t+1} \mid S_t = s] \end{aligned}$$

Để giải quyết phương trình này, ta cần xác định hai thành phần:
$\mathbb{E}[R_{t+1} \mid S_t = s]$: Kỳ vọng phần thưởng tức thì.
$\mathbb{E}[G_{t+1} \mid S_t = s]$: Kỳ vọng phần thưởng tích lũy trong tương lai.

Tính giá trị kỳ vọng phần thưởng tức thì. Ta xét:

$$\mathbb{E}[R_{t+1} \mid S_t = s]$$

Trong đó:
Giả sử ta chọn hành động a với xác suất $\pi(a \mid s)$ tại trạng thái s.
Môi trường trả về phần thưởng r và chuyển đến trạng thái s' với xác suất $P(s', r \mid s, a)$.
Theo định nghĩa giá trị kỳ vọng, ta có:

$$\mathbb{E}[R_{t+1} \mid S_t = s] = \sum_a \pi(a \mid s) \sum_{s',r} P(s', r \mid s, a) \cdot r \tag{4}$$

Tính giá trị kỳ vọng phần thưởng tích lũy trong tương lai. Ta xét:

$$\mathbb{E}[G_{t+1} \mid S_t = s]$$

Tương tự như trên, ta xét sự chuyển dịch từ trạng thái hiện tại sang trạng thái kế tiếp:
Xác suất chuyển sang trạng thái s' khi thực hiện hành động a là $\pi(a \mid s) \cdot P(s', r \mid s, a)$.
Giá trị kỳ vọng của G_{t+1} khi bắt đầu tại trạng thái s' chính là $V(s')$.
Suy ra:

$$\mathbb{E}[G_{t+1} \mid S_t = s] = \sum_a \pi(a \mid s) \sum_{s',r} P(s', r \mid s, a) \cdot V(s') \tag{5}$$

Thay công thức (4) và (5) vào định nghĩa của $V(s)$ ở mục 5, ta được:

$$\begin{aligned} V(s) &= \mathbb{E}[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}[G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a \mid s) \sum_{s',r} P(s', r \mid s, a) \cdot r \\ &+ \gamma \sum_a \pi(a \mid s) \sum_{s',r} P(s', r \mid s, a) \cdot V(s') \\ &= \sum_a \pi(a \mid s) \sum_{s',r} P(s', r \mid s, a) \left[r + \gamma V(s')\right] \end{aligned} \tag{6}$$

Công thức (6) chính là Phương trình Bellman kỳ vọng (Bellman Expectation Equation). Ta cũng có thể viết dưới dạng thu gọn hơn:

$$V(s) = \mathbb{E}\left[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s\right] \tag{7}$$

Xây dựng công thức cập nhật $Q(s,a)$ trong Q-learning. Trong Q-learning, thay vì tính toán giá trị của một trạng thái $V(s)$, ta tập trung vào giá trị của một cặp trạng thái - hành động $Q(s,a)$.

Cấu trúc dữ liệu: Sử dụng một bảng tra cứu gọi là Q-table.
Định nghĩa: $Q(s,a)$ là hàm tính tổng kỳ vọng nhận được khi thực hiện hành động a tại trạng thái s.
Mục đích: Tìm giá trị tối ưu $Q^*(s,a)$. Khi đó, giá trị trạng thái tối ưu được xác định bởi:

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

Khai triển toán học: Giả sử từ trạng thái tiếp theo (s') trở về sau, thuật toán luôn chọn được các hành động tối ưu để đạt được Q^* tối ưu. Với a cố định cho bước hiện tại, áp dụng phương trình Bellman, ta có:

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}[R + \gamma V^*(s')] \\ &= \sum_{s',r} P(s', r \mid s, a) \left[r + \gamma V^*(s')\right] \\ &= \sum_{s',r} P(s', r \mid s, a) \left[r + \gamma \max_{a'} Q^*(s', a')\right] \end{aligned}$$

Thuật toán xấp xỉ (Temporal Difference Learning). Trong thực tế, ta thường không biết xác suất chuyển trạng thái P. Do đó, ta sử dụng cơ chế cập nhật dựa trên trải nghiệm thực tế. Ta đặt:

$$\text{Target} = r + \gamma \max_{a'} Q(s', a')$$

Công thức cập nhật Q-learning (quy tắc Delta) được viết như sau:

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{Giá trị cũ}} + \alpha \underbrace{\left[\left(r + \gamma \max_{a'} Q(s', a')\right) - Q(s,a)\right]}_{\text{Sai số giữa Target và Giá trị cũ}} \tag{5}$$

Trong đó:
$\alpha \in (0, 1]$ là Tốc độ học (Learning rate).
Phần trong ngoặc vuông được gọi là TD Error (Temporal Difference Error).

Tài liệu

Carvalho, D. S., Santos, P. A., and Melo, F. S. Multi-bellman operator for convergence of q-learning with linear function approximation, 2023. URL https://arxiv.org/abs/2309.16819.

Chadi, M.-A. and Mousannif, H. Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization, 2023. URL https://arxiv.org/abs/2304.00026.