Nhóm 13

ACTOR-CRITIC METHOD FOR DECISION MAKING

IS254.P11

Thành viên nhóm

La Hoài Nam	20521629
Vương Thanh Linh	21521082
Trần Lê Tứ	21522746
Lê Trần Anh Quí	21520094

Nội dung trình bày

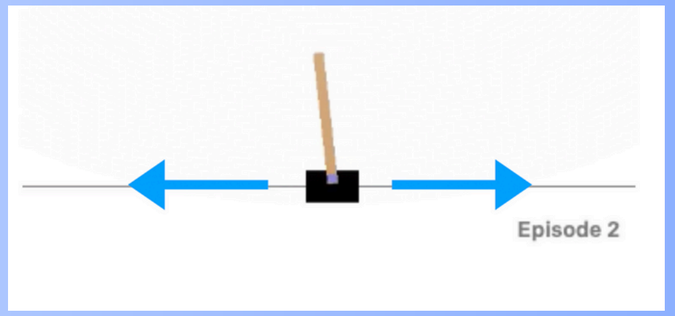
1. **Giới thiệu tổng quan**
2. **Cơ sở lý thuyết**
    a. **Khái niệm học tăng cường**
    b. **Tổng quan về Actor Critic**
    c. **Các phương pháp trong Actor Critic**
3. **Chương trình minh họa**
4. **Kết luận**

1. Giới thiệu tổng quan - Lí do chọn đề tài

- **Tầm quan trọng của Học Tăng Cường (Reinforcement Learning - RL):** Có vai trò quan trọng trong trí tuệ nhân tạo, giúp các tác nhân đưa ra quyết định tối ưu qua tương tác với môi trường, ứng dụng trong robotics và hệ thống ra quyết định.
- **Thuật toán Actor-Critic:** Kết hợp học dựa trên giá trị và chính sách. Kiến trúc với Actor và Critic giúp học chính sách liên tục, giảm độ biến thiên, cải thiện hiệu quả và tốc độ hội tụ.
- **Khả năng ứng dụng:** Actor-Critic đã thành công trong nhiều lĩnh vực như điều khiển robot và tối ưu hóa năng lượng.

1. Giới thiệu tổng quan - Mô tả bài toán

- **Bài toán CartPole-v1:** Tác nhân điều khiển một xe trượt để giữ một cây gậy đứng thăng bằng. Mục tiêu là giữ cây gậy không bị đổ càng lâu càng tốt bằng cách di chuyển xe sang trái hoặc phải.
- **Trạng thái:** Biểu diễn bằng vector gồm 4 giá trị: vị trí xe, vận tốc xe, góc nghiêng của gậy, và vận tốc góc.
- **Hành động:** Hai lựa chọn di chuyển xe sang trái hoặc phải.
- **Phần thưởng:** Cộng 1 điểm cho mỗi bước thời gian mà gậy không bị đổ.
- **Kết thúc:**
  - Cây gậy nghiêng quá 15 độ.
  - Xe trượt ra ngoài biên giới.
  - Hoặc sau 500 bước (được xem là thành công).



- Trong hiện thực mô hình, thuật toán được coi là đã giải quyết thành công bài toán khi đạt được phần thưởng trung bình động (running_reward) trên 475 điểm.

1. Giới thiệu tổng quan - Mô tả dữ liệu

- **Đặc Điểm Dữ Liệu**
  - Không phải tập dữ liệu cố định
  - Được sinh ra từ tương tác với môi trường
- **Chi Tiết Trạng Thái (State)**
  - x: Vị trí xe
  - x_dot: Vận tốc xe
  - theta: Góc nghiêng cột
  - theta_dot: Vận tốc góc cột
- **Hành Động (Action)**
  - 0: Sang trái
  - 1: Sang phải
- **Phần Thưởng (Reward)**
  - 1 điểm mỗi bước nếu cột vẫn đứng thăng bằng

2. Cơ sở lý thuyết - Học tăng cường

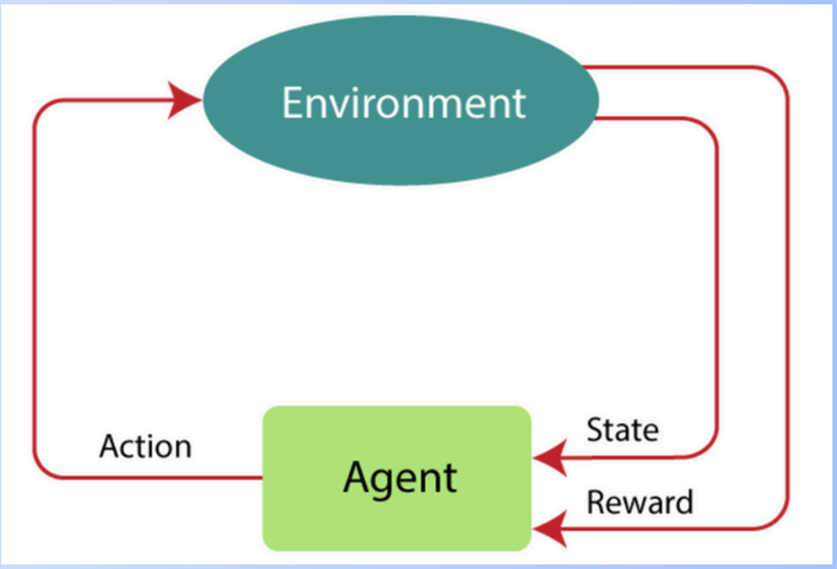
- **Khái niệm học tăng cường:** là một nhánh của học máy, nghiên cứu cách thức một tác nhân (agent) học cách tương tác với môi trường (enviroment) đang ở một trạng thái (state) thực hiện một hành động (action) và nhận phản hồi dưới dạng phần thưởng (reward) hoặc hình phạt (penalty).
- Mục tiêu của tác nhân là tối ưu hóa tổng phần thưởng nhận được theo thời gian bằng cách chọn những hành động tốt nhất trong từng tình huống.
- Gồm các thành phần sau:
  - **Tác nhân (Agent)**: Thực thể tương tác với môi trường và đưa ra quyết định.
  - **Môi trường (Environment):** Hệ thống bên ngoài hoặc thế giới mà tác nhân hoạt động trong đó. Môi trường cung cấp phản hồi dựa trên các hành động của tác nhân.
  - **Hành động (Action – A):** Tập các hành động của tác nhân.
  - **Trạng thái (State – S):** Tình trạng hiện tại của tác nhân trong môi trường.

2. Cơ sở lý thuyết - Học tăng cường

- Gồm các thành phần sau:
  - **Phần thưởng (Reward – R)**: Đối với mỗi hành động được chọn bởi tác nhân, môi trường sẽ đưa ra một phần thưởng. Phần thưởng có giá trị dương, âm hoặc bằng không. Tác nhân hướng đến việc tối đa hóa phần thưởng này.
  - **Chính sách (Policy – π)**: Chiến lược (ra quyết định) mà tác nhân sử dụng để phản ứng trước môi trường giúp đạt được mục tiêu là tối đa hóa phần thưởng.
  - **Hàm giá trị (Value Function)**: Hàm ước tính phần thưởng tích lũy dự kiến từ một trạng thái nhất định, giúp tác nhân dự đoán giá trị dài hạn của các hành động.

2. Cơ sở lý thuyết - Học tăng cường

- **Quy trình hoạt động của quá trình Học tăng cường:**
  - Tác nhân thực hiện hành động (A) trong trạng thái (S) nhất định của môi trường.
  - Môi trường phản hồi bằng phần thưởng (R) và chuyển sang trạng thái mới (S').
  - Tác nhân sử dụng phản hồi này để cập nhật chiến lược (π) của mình, dần dần cải thiện khả năng ra quyết định bằng cách tối đa hóa phần thưởng đạt được tương lai.



2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Actor-Critic:** là một phương pháp trong lĩnh vực học tăng cường kết hợp giữa học chính sách (Policy Learning) và học hàm giá trị (Value Function Learning).
- **Học Chính Sách**
  - Tập trung tìm ra chính sách tối ưu $\pi(a|s)$, tức là 1 hàm xác định xác suất chọn hành động a khi ở trạng thái s.
  - Mục tiêu là tối ưu hóa để tối đa hóa phần thưởng kỳ vọng dài hạn.
- **Học Hàm Giá Trị**
  - Học hàm giá trị được hiện thực thông qua hai hàm cốt lõi sau.
  - **Hàm Giá Trị Trạng Thái V(s)**
    - Ước lượng giá trị kỳ vọng của trạng thái s.
    - Tính tổng phần thưởng dự kiến từ trạng thái s theo chính sách π.
  - **Hàm Giá Trị Hành Động Q(s,a)**
    - Ước lượng giá trị kỳ vọng của hành động a tại trạng thái s.
    - Tính tổng phần thưởng dự kiến khi thực hiện hành động a tại s và tiếp tục theo chính sách π.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Khái niệm Actor:** là thành phần trong thuật toán, thường được biểu diễn bằng một mạng nơ-ron, để học và tối ưu hóa chính sách πθ. Khi được huấn luyện, Actor sẽ cập nhật tham số θ sao cho chính sách πθ có thể chọn hành động tối ưu.
- Chính sách πθ là một hàm toán học hoặc mô hình xác suất định nghĩa cách lựa chọn hành động a dựa trên trạng thái s. Biểu thức toán học của chính sách:

$$\pi\theta(a|s) = P(a|s;\theta)$$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

$$\pi\theta(a|s) = P(a|s;\theta)$$

- $\pi\theta(a|s)$ là xác suất (hoặc mật độ xác suất) của hành động a khi ở trạng thái s, tham số hóa bởi θ.
- Trong thực tế, chính sách $\pi\theta(a|s)$ trong thuật toán Actor-Critic thường được thiết kế dưới dạng một mạng nơ-ron với tham số hóa θ, có thể sử dụng hàm softmax cho bài toán hành động rời rạc hoặc phân phối Gaussian cho hành động liên tục.
- Tham số θ là các giá trị số mà chúng ta điều chỉnh trong quá trình huấn luyện, nhằm tối ưu hóa hiệu suất của mô hình.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Mục tiêu của Actor** là tối đa hóa giá trị kỳ vọng U(θ), tức là cải thiện chính sách sao cho phần thưởng kỳ vọng là cao nhất.
- Với hàm giá trị kỳ vọng U(θ) là một hàm mục tiêu đại diện cho phần thưởng mà Actor có thể kỳ vọng nhận được từ việc thực hiện các hành động theo chính sách πθ.

$$U(\theta) = E[R] = E\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

$$U(\theta) = E[R] = E\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

- **E:** là giá trị kỳ vọng (trung bình) của tổng các phần thưởng
- **R**: Là tổng phần thưởng nhận được từ môi trường khi thực hiện các hành động theo chính sách πθ
- **y:** ·Là yếu tố giảm dần (discount factor), điều chỉnh trọng số của phần thưởng trong tương lai.
- **rt:** Phần thưởng nhận được tại thời điểm t.
- **T:** Thời gian kết thúc (hoặc số bước trong một episode)

--> Tuy nhiên, việc tính trực tiếp giá trị này qua toàn bộ các bước thời gian t là không khả thi trong thực tế khi làm việc với các hệ thống phức tạp hoặc không gian trạng thái lớn nên ta sử dụng các phương pháp xấp xỉ (Monte Carlo, Temporal-Difference,...) để ước lượng giá trị đó.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Khái niệm Critic:** là thành phần đảm nhận vai trò đánh giá chất lượng của các hành động do Actor thực hiện. Cụ thể, Critic ước lượng các đại lượng quan trọng sau, được tham số hóa bởi ϕ
  - **Hàm giá trị trạng thái V(s)**: Đại diện cho giá trị kỳ cọng khi thực hiện hành động a tại trạng thái s và tiếp tục theo chính sách hiện tại.
  - **Hàm giá trị hành động Q(s, a)** : Đại diện cho giá trị kỳ vọng khi thực hiện hành động a tại trạng thái s và tiếp tục theo chính sách hiện tại.
  - **Hàm lợi thế A(s, a)** : Đo lường lợi ích tương đối của hành động a tại trạng thái s so với giá trị trung bình của trạng thái s

$$A_{\Phi}(s,a) = Q_{\Phi}(s,a) - V_{\Phi}(s)$$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Mối quan hệ giữa Actor và Critic:**
  - **Actor phụ thuộc vào Critic**: Critic đánh giá các hành động được Actor thực hiện. Dựa trên đánh giá của Critic, Actor sẽ điều chỉnh chính sách $\pi\theta$ để cải thiện xác suất chọn các hành động tốt hơn trong tương lai.
  - **Critic phụ thuộc vào Actor**: Critic cần các hành động từ Actor để đánh giá và học cách ước lượng giá trị trạng thái hoặc hành động.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Cơ chế hoạt động của Actor-Critic: Các bước hoạt động**
  - **Bước 1:** Khởi tạo với các tham số ban đầu cho **Actor (θ)** và **Critic (ϕ)**.
  - **Bước 2:** Thực hiện các rollout trajectories từ trạng thái ban đầu, sử dụng chính sách hiện tại πθ để thu thập dữ liệu (các trạng thái s, hành động a, và phần thưởng r)
  - **Bước 3:** Critic ước lượng giá trị của trạng thái hiện tại s và trạng thái mới s'. Bằng cách cập nhật tham số ϕ theo hướng gradient descent để giảm thiểu hàm mất mát: $\phi \leftarrow \phi - \beta \nabla \ell(\phi).$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Cơ chế hoạt động của Actor-Critic: Các bước hoạt động**
  - **Bước 4:** Actor sử dụng thông tin từ Critic để cập nhật chính sách. Bằng cách sử dụng gradient của hàm mục tiêu, thường là dựa trên hàm lợi ích hoặc hàm giá trị hành động để cập nhật tham số θ theo hướng gradient ascent: $\theta \leftarrow \theta + \alpha \nabla U(\theta)$
  - **Bước 5:** Lặp lại các bước trên cho đến khi hội tụ (Các tham số không thay đổi nhiều giữa các lần cập nhật, cho thấy mô hình đã đạt được hiệu suất tối ưu) hoặc đạt được số lần lặp tối đa (Đạt đến số lần lặp đã định trước, nhằm tránh việc chạy quá lâu mà không đạt được cải thiện đáng kể).

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Các công thức quan trọng trong Actor-Critic: Cập nhật Actor**
  - Quá trình cập nhật actor được thực hiện thông qua gradient ascent. Mục tiêu là tối đa hóa hàm giá trị kỳ vọng U(θ) với công thức

$$\nabla U(\theta) = E_\tau\left[\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta\left(a^{(k)}|s^{(k)}\right)\gamma^{k-1}A_\theta\left(s^{(k)}, a^{(k)}\right)\right]$$

Trong đó:

- **∇U(θ)**: Gradient của hàm mục tiêu đối với tham số chính sách θ.
- **Eτ**: Kỳ vọng trên các quỹ đạo (rollout trajectories) được thu thập từ môi trường. (tính giá trị trung bình)
- **∑k=1d**: Tổng từ bước 1 đến bước d trong một trajectory.
- $\nabla_\theta \log \pi_\theta\left(a^{(k)}|s^{(k)}\right)$: Gradient của log xác suất chọn hành động a(k) tại trạng thái s(k) theo chính sách πθ..
- **γ**: Hệ số chiết khấu tại bước k−1.
- $A_\theta(s, a)$**:** Hàm ước lượng lợi ích (advantage) khi thực hiện hành động a tại trạng thái s. Lợi thế dương có nghĩa là hành động tốt hơn kỳ vọng và cần được Actor ưu tiên, trong khi lợi thế âm cho thấy hành động không tốt và cần bị giảm trọng số.

$A\theta(s, a)$ có thể được tính theo công thức:

$$A_\theta(s, a) = E_{r,s'}\left[r + \gamma U_{\pi_\theta}(s') - U_{\pi_\theta}(s)\right]$$

- $E_{r,s'}$: Kỳ vọng trên phần thưởng r và trạng thái tiếp theo s′.
- **r**: Phần thưởng nhận được sau khi thực hiện hành động a tại trạng thái s.
- $U_{\pi_\theta}(s')$**:** Giá trị kỳ vọng của trạng thái s′ theo chính sách πθ
- $U_{\pi_\theta}(s)$**:** Giá trị kỳ vọng của trạng thái s theo chính sách πθ.
- $r + \gamma U_{\pi_\theta}(s') - U_{\pi_\theta}(s)$: được gọi là temporal difference residual, cho phép chúng ta đánh giá hiệu suất của hành động đã thực hiện.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Các công thức quan trọng trong Actor-Critic: Cập nhật Actor**
  - Tuy nhiên, trong thực tế, để giảm phương sai và cập nhật chính sách ổn định hơn, gradient của hàm mục tiêu cho Actor thường được ước lượng thông qua Critic, giúp đánh giá giá trị trạng thái và hành động một cách hiệu quả

$$\nabla U(\theta) \approx \mathbb{E}_{\tau}\left[\sum_{k=1}^{d} \nabla_{\theta} log \pi_{\theta}(a^{(k)}|s^{(k)}) \gamma^{k-1}\left(r^{(k)} + \gamma U_{\Phi}\left(s^{(k+1)}\right) - U_{\Phi}\left(s^{(k)}\right)\right)\right]$$

- $\boldsymbol{r^{(k)}}$: Phần thưởng nhận được tại bước k.
- $\boldsymbol{U_{\Phi}\left(s^{(k+1)}\right)}$: Giá trị ước lượng của trạng thái $\boldsymbol{s^{(k+1)}}$ bởi Critic.
- $\boldsymbol{U_{\Phi}\left(s^{(k)}\right)}$: Giá trị ước lượng của trạng thái $\boldsymbol{s^{(k)}}$ bởi Critic.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Các công thức quan trọng trong Actor-Critic: Cập nhật Critic**
  - Critic được cập nhật thông qua tối ưu hóa gradient. Mục tiêu là tìm ϕ sao cho giảm thiểu hàm mất mát

$$l(\phi) = \frac{1}{2} E_s \left[\left(U_\phi(s) - U_{\pi_\theta}(s) \right)^2 \right]$$

- $l(\phi)$: Hàm mất mát cho Critic, đo lường sự khác biệt giữa giá trị ước lượng của Critic và giá trị thực tế của chính sách.
- E_s: Kỳ vọng trên các trạng thái s.
- $U_\phi(s)$: Giá trị ước lượng của trạng thái s bởi Critic với tham số ϕ.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

- **Các công thức quan trọng trong Actor-Critic: Cập nhật Critic**
  - Vì chúng ta không biết chính xác , nên có thể ước lượng nó bằng cách sử dụng phần thưởng còn lại dọc theo các quỹ đạo, dẫn đến:

$$\nabla \mathrm{l}(\phi) = E_\tau \left[\sum_{k=1}^{d} \left(U_\phi\left(s^{(k)}\right) - r_{to-go}^{(k)} \right) \nabla_\phi \, U_\phi\left(s^{(k)}\right) \right]$$

- $\boldsymbol{r_{to-go}^{(k)}}$ là phần thưởng từ bước k trở đi trong một quỹ đạo.
- $\boldsymbol{\nabla_\phi U_\phi\left(s^{(k)}\right)}$: Gradient của giá trị ước lượng $U_\phi\left(s^{(k)}\right)$ đối với tham số ϕ. Được tính theo công thức sau:

$$\nabla_\phi U_\phi\left(s(k)\right) = \frac{\partial U_\phi\left(s(k)\right)}{\partial \phi}$$

$\frac{\boldsymbol{\partial U_\phi(s(k))}}{\boldsymbol{\partial \phi}}$:là đạo hàm riêng của $U_\phi\left(s(k)\right)$ theo ϕ

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

Ví dụ minh họa thuật toán

Giả sử, một robot đang ở trong một mê cung. Điều kiện của mê cung như sau:

- Robot bắt đầu ở trạng thái s = 1 và cần đến trạng thái cuối cùng (ví dụ: s = 5) để nhận phần thưởng tối đa.
- Các trạng thái trong mê cung được đánh số từ 1 đến 5.
- Robot có 2 hành động có thể thực hiện tại mỗi trạng thái:
  - a = 1: Đi sang trái (trạng thái thấp hơn).
  - a = 2: Đi sang phải (trạng thái cao hơn).

Thông tin về phần thưởng:

- Nếu robot đến trạng thái cuối cùng (s = 5), nhận phần thưởng r = 10.
- Mỗi bước di chuyển khác đều nhận phần thưởng r = -1 (trừ năng lượng vì robot đang di chuyển không hiệu quả).

Mục tiêu của robot là:

- Học được cách di chuyển hiệu quả nhất để đến s = 5 nhanh nhất và nhận phần thưởng cao nhất.

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

Ví dụ minh họa thuật toán

- Khởi Tạo Tham Số:
  - Actor được biểu diễn bằng một vector tham số θ với giá trị ban đầu : $\theta = [0.5, -0.3]$
  - Critic được biểu diễn bằng một vector tham số ϕ với giá trị ban đầu: $\phi = [0.2, 0.4]$
- Giả sử robot đang ở trạng thái $s = 1$, và Actor chọn hành động $a = 2$ (di chuyển sang phải). Robot thu thập được dữ liệu sau:
  - Trạng thái hiện tại: $s = 1$
  - Hành động thực hiện: $a = 2$
  - Phần thưởng nhận được: $r = 1$ (vì chưa đến đích nhưng đang di chuyển).
  - Trạng thái mới: $s' = 2$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

Ví dụ minh họa thuật toán

Cập nhật Critic:

- Giả định hàm Uϕ(s) là một hàm tuyến tính đơn giản: $U_\phi(s) = \phi_1 s + \phi_2 s$
- Ước lượng giá trị của trạng thái hiện tại và trạng thái mới:
  - $U_\phi(s) = 0.2 \cdot 1 + 0.4 \cdot 1 = 0.6$
  - $U_\phi(s') = 0.2 \cdot 2 + 0.4 \cdot 2 = 1.2$
- Tính toán sai số (TD error): $\delta = r + \gamma U_\phi(s') - U_\phi(s)$

  Giả sử γ=0.9: $\delta = 1 + 0.9 \cdot 1.2 - 0.6 = 1.48$
- Cập nhật tham số của Critic: $\phi \leftarrow \phi - \beta\delta\nabla_\phi U_\phi(s)$

  Giả sử β=0.01 và ∇ϕUϕ(s)=[1,1]:

$$\phi = [0.2,0.4] - 0.01 \cdot 1.48 \cdot [1,1] = [0.2,0.4] - [0.0148,0.0148]$$

$$= [0.1852,0.3852]$$

2. Cơ sở lý thuyết - Tổng quan về Actor Critic

Ví dụ minh họa thuật toán

Cập nhật Actor:

- Giả định rằng gradient của log chính sách đã được tính toán trước:

  $\nabla_\theta \log \pi_\theta(a \mid s) = [0.1, -0.1]$

- Cập nhật tham số của Actor: $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a \mid s)\, \delta$

  Giả sử α=0.01:

$$\theta = [0.5, -0.3] + 0.01 \cdot 1.48 \cdot [0.1, -0.1]$$

$$= [0.5, -0.3] + [0.00148, -0.00148] = [0.50148, -0.30148]$$

- Lặp lại đến khi tối ưu.
- Cuối cùng, robot sẽ có thể di chuyển qua mê cung một cách hiệu quả nhất, tối ưu hóa phần thưởng nhận được trong quá trình di chuyển.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Generalized Advantage Estimation (GAE)

- Là một phương pháp trong học tăng cường
- Được phát triển để cải thiện ước lượng lợi ích (advantage) trong các thuật toán **Actor-Critic**.
- Không chỉ đơn thuần ước lượng lợi ích tại một thời điểm mà còn kết hợp thông tin từ nhiều bước thời gian

→ Giúp cân bằng giữa độ thiên lệch (bias) và phương sai (variance) trong ước lượng.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh GAE với Actor Critic cơ bản

Ước lượng Advantage

Actor Critic cơ bản

- Sử dụng một ước lượng đơn giản cho advantage, thường là dựa vào giá trị **Temporal Difference Residual**.
- Điều này có thể dẫn đến bias cao nhưng variance thấp.

GAE

- Kết hợp nhiều bước ước lượng advantage thông qua một tham số λ, cho phép điều chỉnh giữa bias và variance.
- Điều này giúp tạo ra các ước lượng advantage chính xác hơn và ổn định hơn.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh GAE với Actor Critic cơ bản

Tính đến nhiều bước

Actor Critic cơ bản	GAE
• Thường chỉ xem xét 1 bước trong quá trình cập nhật, dẫn đến việc sử dụng thông tin hạn chế từ các bước trước đó.	• Tính toán advantage dựa trên nhiều bước trong một chuỗi rollout, giúp cải thiện khả năng học tập và hiệu suất tổng thể.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh GAE với Actor Critic cơ bản

Tham số điều chỉnh

Actor Critic cơ bản	GAE
• Không có tham số điều chỉnh cho bias và variance.	• Tham số λ cho phép điều chỉnh mức độ ảnh hưởng của các ước lượng khác nhau, từ đó tối ưu hóa hiệu suất giữa bias và variance.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh GAE với Actor Critic cơ bản

Khả năng ổn định

Actor Critic cơ bản	GAE
• Thường có thể gặp phải vấn đề về độ ổn định trong quá trình học do sự tương quan giữa các cập nhật của actor và critic.	• Cung cấp một cách tiếp cận ổn định hơn nhờ vào việc kết hợp các ước lượng từ nhiều bước, giúp giảm thiểu sự biến động.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của GAE

- **Bước 1:** Thực hiện các hành động theo chính sách hiện tại $\pi\theta$ và thu thập các trạng thái, hành động, và phần thưởng.
- **Bước 2:** Sử dụng công thức phần dư sai số thời gian để tính toán phần dư sai số thời gian cho mỗi bước.
- **Bước 3:** Sử dụng công thức để ước lượng hàm lợi ích từ k bước rollout.
- **Bước 4:** Sử dụng công thức để kết hợp các ước lượng hàm lợi ích từ nhiều bước rollout với trọng số mũ λ.
- **Bước 5:** Sử dụng ước lượng hàm lợi ích tổng quát để cập nhật tham số của Actor và Critic.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích tổng quát

$$A_\theta(s,a) = \mathbb{E}_{r_1,\dots,r_d}[r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{d-1} r_d - U^{\pi\theta}(s)]$$
$$= \mathbb{E}_{r_1,\dots,r_d}\left[-U^{\pi\theta}(s) + \sum_{\ell=1}^{d} \gamma^{\ell-1} r_\ell\right]$$

Trong đó:

+ $A_\theta(s,a)$: Đo lường giá trị lợi ích của hành động a trong trạng thái s so với việc thực hiện hành động trung bình

+ $\mathbb{E}_{r_1,\dots,r_d}$: Kỳ vọng tính toán giá trị trung bình của một chuỗi các phần thưởng $r_1, \dots, r_d$ mà một tác nhân nhận được sau khi thực hiện hành động a trong trạng thái s

+ γ: Hệ số chiết khấu nằm trong khoảng [0, 1] và dùng để giảm giá trị của các phần thưởng trong tương lai, nhằm phản ánh rằng phần thưởng gần hơn có giá trị hơn phần thưởng xa hơn

+ $U^{\pi\theta}(s)$: Hàm giá trị cho trạng thái s. Thể hiện giá trị mà tác nhân mong đợi nhận được từ trạng thái đó nếu không thực hiện hành động nào

- Một phiên bản tổng quát của hàm lợi ích cơ bản
- Đo lường sự khác biệt giữa tổng phần thưởng tích lũy theo các rollout r1,r2,...,rd (có tính đến hệ số chiết khấu γ) và giá trị kỳ vọng của trạng thái ban đầu Uπθ(s).

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích tổng quát

- Cân bằng giữa sai số (bias) và phương sai (variance) trong quá trình huấn luyện.
- Đánh giá hiệu quả của hành động tại trạng thái , với các phần thưởng được lấy từ các quỹ đạo rollout thực tế.
- Có thể đạt được một ước lượng không sai số (unbiased) thông qua các quỹ đạo rollout.
- Tuy nhiên, ước lượng có phương sai cao, do đó cần nhiều mẫu để đạt được độ chính xác cao.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Phần dư sai số thời gian (Temporal Difference Residual):

$$\delta_t = r_t + \gamma U(s_{t+1}) - U(s_t)$$

- Với st, rt, và st+1 lần lượt là trạng thái, phần thưởng và trạng thái tiếp theo trong một quỹ đạo đã được mẫu
- Phần dư sai số thời gian đo lường sự khác biệt giữa phần thưởng nhận được cộng với giá trị kỳ vọng của trạng thái tiếp theo và giá trị kỳ vọng của trạng thái hiện tại
- Sử dụng để cập nhật hàm giá trị U.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích từ k bước rollout

$$\hat{A}^{(k)}(s,a) = \mathbb{E}_{r_1,\ldots,r_k,s'}[r_1 + \gamma r_2 + \cdots + \gamma^{k-1} r_k + \gamma^k U^{\pi_\theta}(s') - U^{\pi_\theta}(s)]$$

$$= \mathbb{E}_{r_1,\ldots,r_k,s'}\left[-U^{\pi_\theta}(s) + \gamma^k U^{\pi_\theta}(s') + \sum_{\ell=1}^{k} \gamma^{\ell-1} r_\ell\right]$$

- Ước lượng hàm lợi ích từ k bước rollout và giá trị kỳ vọng của trạng thái kết quả.
- Cân bằng giữa độ chệch và phương sai bằng cách sử dụng một số bước rollout cố định.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích tổng quát với phần dư sai số thời gian

$$\hat{A}^{(k)}(s,a) = \mathbb{E}\left[\sum_{\ell=1}^{k} \gamma^{\ell-1}\delta_{\ell}\right]$$

- Biểu diễn ước lượng hàm lợi ích tổng quát dưới dạng tổng của các phần dư sai số thời gian.
- Giúp giảm phương sai của ước lượng bằng cách sử dụng thông tin từ nhiều bước thời gian.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích tổng quát với phần dư sai số thời gian

$$\hat{A}_{GAE}(s,a) = (1-\lambda)\left(\hat{A}^{(1)} + \lambda\hat{A}^{(2)} + \lambda^2\hat{A}^{(3)} + \cdots\right)$$

- Khi λ=0: GAE trở thành ước lượng dựa trên residual tạm thời với độ thiên lệch cao và phương sai thấp.
- Khi λ=1: GAE chuyển thành ước lượng dựa trên toàn bộ quỹ đạo với độ thiên lệch thấp nhưng phương sai cao hơn.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của GAE

Ước lượng hàm lợi ích tổng quát với phần dư sai số thời gian

$$\hat{A}_{GAE}(s,a) = (1-\lambda)\left(\hat{A}^{(1)} + \lambda\hat{A}^{(2)} + \lambda^2\hat{A}^{(3)} + \cdots\right)$$

- Sử dụng trọng số mũ λ để kết hợp các ước lượng hàm lợi ích từ nhiều bước rollout khác nhau.
- Cân bằng giữa độ chệch và phương sai bằng cách điều chỉnh giá trị λ.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Ví dụ - GAE

Giả sử chúng ta có một robot di chuyển trong một mê cung. Chúng ta sẽ sử dụng GAE để cải thiện hiệu quả của thuật toán Actor-Critic.

- **Bước 1:** Thu thập dữ liệu. Robot di chuyển và ghi lại các trạng thái, hành động, phần thưởng, và trạng thái mới.
- **Bước 2:** Tính toán phần dư sai số thời gian.

Trạng thái hiện tại: $s=1$

Hành động: $a=2$

Phần thưởng nhận được: $r=1$

Trạng thái mới: $s'=2$

Giá trị kỳ vọng của trạng thái hiện tại: $U(s)=0.6$

Giá trị kỳ vọng của trạng thái mới: $U(s')=1.2$

Phần dư sai số thời gian: $\delta=1+0.9\cdot1.2-0.6=1.48$

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Ví dụ - GAE

Bước 3: Ước lượng hàm lợi ích từ các bước rollout

- k=1: $\hat{A}^{(1)}(s,a) = \delta = 1.48$
- k=2: $\hat{A}^{(2)}(s,a) = \delta_1 + \gamma\delta_2 = 1.48 + 0.9 \cdot 1.48 = 2.812$

Bước 4: Kết hợp các ước lượng hàm lợi ích

- λ=0.5: $\hat{A}_{GAE}(s,a)$ =(1−0.5)(1.48+0.5·2.812)=1.852

Bước 5: Cập nhật Actor và Critic

- Sử dụng $\hat{A}_{GAE}(s,a)$ =1.852 để cập nhật tham số của Actor và Critic.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Deterministic Policy Gradients (DPG)

- Là một phương pháp trong học tăng cường
- Được thiết kế để tối ưu hóa các chính sách xác định (deterministic policies) trong các bài toán có không gian hành động liên tục
- Tập trung vào việc tối ưu hóa một chính sách xác định $\pi\theta(s)$ mà không cần sử dụng phân phối xác suất cho hành động, giúp cải thiện tốc độ hội tụ và tính ổn định

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của DPG

Deterministic Policy Gradients hoạt động theo cách tương tự với Actor Critic cơ bản

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh DPG với Actor Critic cơ bản

Ứng Dụng

Actor Critic cơ bản	DPG
• Thích hợp cho các bài toán có không gian hành động rời rạc	• Đặc biệt hiệu quả cho các bài toán điều khiển với không gian hành động liên tục, như trong robot hoặc các hệ thống điều khiển phức tạp.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh DPG với Actor Critic cơ bản

Loại Chính Sách

Actor Critic cơ bản	DPG
• Thường sử dụng chính sách ngẫu nhiên (stochastic policy), nơi hành động được chọn dựa trên phân phối xác suất	• Sử dụng chính sách xác định (deterministic policy), trong đó hành động được quyết định trực tiếp từ đầu vào mà không có ngẫu nhiên. Điều này giúp cải thiện hiệu suất trong các bài toán có không gian hành động liên tục

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh DPG với Actor Critic cơ bản

Gradient Tính Toán

Actor Critic cơ bản	DPG
• Cập nhật chính sách dựa trên gradient của giá trị hàm U, thường sử dụng phương pháp gradient ascent dựa trên các giá trị ước lượng từ critic	• Tính toán gradient của chính sách thông qua giá trị Q, sử dụng công thức gradient của Q để tối ưu hóa chính sách. Điều này cho phép DPG tận dụng thông tin từ critic một cách hiệu quả hơn

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh DPG với Actor Critic cơ bản

Cách Cập Nhật

Actor Critic cơ bản

- Cập nhật cả actor và critic đồng thời, thường thực hiện cập nhật theo cách song song

DPG

- Tập trung vào việc tối ưu hóa chính sách trước, sau đó cập nhật giá trị hàm Q. Quy trình này có thể giúp cải thiện độ ổn định và hiệu suất

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

So sánh DPG với Actor Critic cơ bản

Khám phá

Actor Critic cơ bản	DPG
• Khám phá hành động thường dựa vào tính ngẫu nhiên của chính sách, có thể không hiệu quả trong một số trường hợp	• Thường sử dụng noise (chẳng hạn như Gaussian noise) để khám phá, giúp cải thiện khả năng tìm kiếm trong không gian hành động liên tục

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Định nghĩa hàm mất mát: Để tối ưu hóa chính sách và hàm giá trị hành động:

$$\ell(\phi) = \frac{1}{2} E_{s,a,r,s'} \left[\left(r + \gamma Q_\phi(s', \pi_\theta(s')) - Q_\phi(s,a) \right)^2 \right]$$

ϕ: Tham số của hàm giá trị hành động $Q_\phi(s,a)$, cần tối ưu.

s: Trạng thái hiện tại,

a: Hành động hiện tại, được thực hiện bởi Actor $\pi_\theta(s)$

r: Phần thưởng từ môi trường khi thực hiện hành động a tại trạng thái s.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Định nghĩa hàm mất mát: Để tối ưu hóa chính sách và hàm giá trị hành động:

$$\ell(\phi) = \frac{1}{2} E_{s,a,r,s'} \left[\left(r + \gamma Q_\phi(s', \pi_\theta(s')) - Q_\phi(s,a) \right)^2 \right]$$

s': Trạng thái tiếp theo sau hành động a.

γ: Yếu tố giảm dần (discount factor), điều chỉnh tầm quan trọng của phần thưởng tương lai.

$Q_\phi(s', \pi_\theta(s'))$: Giá trị dự đoán của hành động ở trạng thái tiếp theo s', hành động theo Actor π_θ.

$Q_\phi(s,a)$: Giá trị dự đoán của hành động a ở trạng thái hiện tại s.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Tính Gradient của hàm mất mát: Để cập nhật các tham số ϕ của hàm giá trị hành động Qϕ, chúng ta cần tính gradient của hàm mất mát

$$\nabla_{\phi}\ell(\phi) = E_{s,a,r,s'}\left[r + \gamma Q_{\phi}\left(s', \pi_{\theta}(s')\right) - Q_{\phi}(s,a)\right]\left[\gamma\nabla_{\phi}Q_{\phi}\left(s', \pi_{\theta}(s')\right) - \nabla_{\phi}Q_{\phi}(s,a)\right]$$

Gradient này cho phép chúng ta cập nhật tham số ϕ theo hướng tối ưu hóa hàm giá trị hành động.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Tối Ưu Hàm Mục tiêu

- Để tối ưu hóa chính sách, chúng ta cần tìm giá trị của θ sao cho hàm giá trị kỳ vọng U(θ) được tối đa hóa:

$$U(\theta) = E_{s\sim b}[Q_\phi(s, \pi_\theta(s))]$$

- Ở đây, b là phân phối trạng thái đầu vào. Để tối ưu hóa θ, chúng ta sử dụng gradient ascent:

$$\nabla U(\theta) = E_s[\nabla_\theta Q_\phi(s, \pi_\theta(s))]$$

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Tính Gradient của Hàm mục tiêu: tính toán để cải thiện chính sách

$$\nabla U(\theta) = E_s\left[\nabla_\theta \pi_\theta(s) \nabla_a Q_\phi(s,a)|_{a=\pi_\theta(s)}\right]$$

Trong đó, $\nabla\theta\pi\theta(s)$ là ma trận Jacobian, thể hiện sự thay đổi của hành động a khi thay đổi tham số θ tại trạng thái s

Trong đó, $\nabla_a Q_\phi(s,a)$ là gradient của hàm giá trị Q_ϕ với hành động a, mô tả cách Q_ϕ, thay đổi khi thay đổi hành động.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Giải thích ma trận Jacobian

$$\nabla_\theta \pi_\theta(s) = \begin{bmatrix} \frac{\partial \pi_i}{\partial \theta_j} & \cdots & \frac{\partial \pi_1}{\partial \theta_j} \\ \vdots & \ddots & \vdots \\ \frac{\partial \pi_m}{\partial \theta_1} & \cdots & \frac{\partial \pi_m}{\partial \theta_n} \end{bmatrix}$$

$\pi_\theta(s)$: Đầu ra của Actor, là một hành động a.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Các công thức của DPG

Giải thích ma trận Jacobian

Hàng i: Biểu diễn gradient của hành động π_i (thành phần thứ i của hành động) với từng tham số θ_j.

Cột j: Biểu diễn anhr hưởng của θ_j lên các thành phần π_i của hành động.

Đây là ma trận mxn, với m là số chiều của hành động a và n là tham số θ.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Monte Carlo Tree Search (MCTS)

- Là một kỹ thuật kết hợp giữa học tăng cường và quy hoạch trực tuyến
- Được thiết kế để tối ưu hóa các chính sách trong các bài toán quyết định Markov (MDP)
- chính sách $\pi\theta(a|s)$ là một hàm xác suất quyết định hành động a dựa trên trạng thái s
- Hàm giá trị $U\phi(s)$ thể hiện giá trị kỳ vọng của trạng thái s theo chính sách $\pi\theta$
- Đưa ra các hành động tối ưu bằng cách xây dựng một cây tìm kiếm, trong đó mỗi nút đại diện cho một trạng thái và mỗi nhánh đại diện cho một hành động.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Giải thích phương pháp MCTS

- MCTS tìm kiếm hành động tốt nhất bằng cách xây dựng một cây tìm kiếm trong quá trình được mở rộng dựa trên các mô phỏng ngẫu nhiên (Monte Carlo rollouts).

- Ý tưởng: lặp lại các bước mô phỏng để đánh giá các hành động, dần dần cải thiện độ chính xác của cây tìm kiếm.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Quy trình hoạt động của MCTS

Gồm 4 bước chính được lặp lại nhiều lần để xây dựng cây tìm kiếm

- **Bước 1:** Selection (Chọn hành động trong cây)
- **Bước 2:** Expansion (Mở rộng cây)
- **Bước 3:** Simulation (Mô phỏng ngẫu nhiên)
- **Bước 4:** Backpropagation (Lan truyền ngược)

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Một số công thức được sử dụng trong MCTS

Công thức cho hành động tối ưu trong tìm kiếm cây Monte Carlo:

$$a = \underset{a}{\mathrm{argmax}} Q(s,a) + c\pi_\theta(a|s)\frac{\sqrt{N(s)}}{1+N(s,a)}$$

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Một số công thức được sử dụng trong MCTS

Công thức hàm mất mát chính sách:

$$l(\theta) = -E_s\left[\sum_a \pi_{MCTS}(a|s)\log\pi_\theta(a|s)\right]$$

$$\nabla l(\theta) = -E_s\left[\sum_a \pi_{MCTS}(a|s)\nabla_\theta\pi_\theta(a|s)\right]$$

Công thức hàm mất mát giá trị:

$$l(\phi) = \frac{1}{2}E_s\left[\left(U_\phi(s) - U_{MCTS}(s)\right)^2\right]$$

$$\nabla l(\phi) = E_s\left[\left(U_\phi(s) - U_{MCTS}(s)\right)\nabla_\phi U_\phi(s)\right]$$

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của MCTS

Sử dụng Actor làm "Prior" cho MCTS

- Chính sách π(s, a) từ Actor được sử dụng làm prior probability (xác suất ban đầu) để hướng dẫn MCTS.
- Thay vì khởi tạo các xác suất hành động một cách ngẫu nhiên, MCTS sử dụng các xác suất từ Actor để ưu tiên các hành động khả thi hơn trong bước tìm kiếm.
- Điều này giúp MCTS tập trung tìm kiếm vào các nhánh tiềm năng, giảm số lượng mô phỏng cần thiết.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của MCTS

MCTS thực hiện tìm kiếm

- Selection (Chọn nhánh): Tại mỗi nút trong cây tìm kiếm, các hành động được chọn dựa trên công thức cho hành động tối ưu trong tìm kiếm cây Monte Carlo.
- Expansion (Mở rộng nhánh): Khi một trạng thái mới được gặp trong cây, MCTS mở rộng các nhánh con. Các hành động khả thi được gán xác suất ban đầu $P(s,a) \propto \pi\theta(a|s)$. Công thức này cho biết xác suất $P(s,a)$ của hành động a tại trạng thái s trong (MCTS) được xác định tỷ lệ với $\pi\theta(a|s)$, chính sách được cung cấp bởi Actor. Xác suất $P(s,a)$ không cần phải bằng chính xác $\pi\theta(a|s)$ nhưng sẽ tăng hoặc giảm theo tỉ lệ thuận với $\pi\theta(a|s)$.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của MCTS

MCTS thực hiện tìm kiếm

- Simulation (Mô phỏng): Từ trạng thái hiện tại, MCTS thực hiệnmột hoặc nhiều mô phỏng ngẫu nhiên (Monte Carlo rollouts) đến khi đạt trạng thái kết thúc. Chính sách Actor có thể được sử dụng làm chính sách mô phỏng để tăng độ chính xác của mô phỏng.
- Backpropagation (Lan truyền ngược): Kết quả từ các mô phỏng (reward) được lan truyền ngược trong cây để cập nhật giá trị Q(s,a) và số lần thăm N(s,a).
- Critic có thể hỗ trợ trong việc định lượng giá trị Q(s,a) tốt hơn, đặc biệt trong các trạng thái chưa được khám phá đầy đủ.

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của MCTS

Hành động từ MCTS

Sau một số lượng vòng lặp (hoặc thời gian giới hạn), MCTS chọn hành động tốt nhất từ nút gốc dựa trên: Số lần thăm N(s,a) (hành động được khám phá nhiều nhất) $a* = argamaxN(s_0, a)$. Hoặc giá trị trung bình Q(s,a) (hành động có phần thưởng dự kiến cao nhất)

$$a* = argamaxN(s_0, a)$$

2. Cơ sở lý thuyết - Các phương pháp Actor Critic

Cách hoạt động của MCTS

Cập nhật Actor-Critic

- Cập nhật Critic: Giá trị từ MCTS: $UMCTS(s) = maxQ(s,a)$ Critic được cập nhật để hàm giá trị $U\phi(s)$ gần với UMCTS(s) thông qua hàm mất mát giá trị.
- Cập nhật Actor: Chính sách từ MCTS: $\pi_{MCTS}(a|s) \propto N(s,a)^{\eta}$
  - η Siêu tham số điều chỉnh mức độ ngẫu nhiên
- Công thức này biểu thị chính sách $\pi_{MCTS}(a|s)$ từ MCTS, được xác định tỷ lệ với số lần hành động a được chọn ở trạng thái s trong quá trình tìm kiếm, nâng lên lũy thừa η

3. Chương trình minh họa

4. Kết luận

Ưu điểm

- Hiệu quả trong việc học chính sách: Actor-Critic có khả năng học chính sách trực tiếp thông qua thành phần Actor, giúp tối ưu hóa hành động trong môi trường phức tạp.
- Giảm phương sai: Thành phần Critic giúp ước lượng giá trị hành động, giảm phương sai trong quá trình cập nhật chính sách, làm cho việc học ổn định hơn.
- Khả năng áp dụng rộng rãi: Thuật toán này có thể áp dụng cho nhiều loại môi trường khác nhau, từ các bài toán đơn giản đến các bài toán phức tạp hơn như trò chơi hoặc robot học

4. Kết luận

Hạn chế

- Phức tạp trong việc triển khai: Actor-Critic yêu cầu việc triển khai hai mạng nơ-ron riêng biệt (Actor và Critic), điều này có thể làm tăng độ phức tạp và yêu cầu tài nguyên tính toán cao hơn.
- Khả năng hội tụ chậm: Trong một số trường hợp, thuật toán có thể hội tụ chậm hoặc không ổn định nếu không được điều chỉnh đúng cách.
- Cần điều chỉnh nhiều siêu tham số: Việc điều chỉnh các siêu tham số như tốc độ học, hệ số chiết khấu, và các tham số khác có thể phức tạp và tốn thời gian.

4. Kết luận

Hướng phát triển

- Tối ưu hóa hiệu suất và ổn định:
  - Giảm phương sai và tăng tốc độ hội tụ
  - Học không giám sát và bán giám sát:
- Kết hợp với các công nghệ khác:
  - Học sâu (Deep Learning)
  - Học tăng cường đa tác tử (Multi-Agent Reinforcement Learning)
- Cải thiện khả năng giải thích và minh bạch:
  - Giải thích quyết định
  - Đảm bảo đạo đức và trách nhiệm

THANK YOU!