## Khám phá không gian hành động liên tục

- **Vấn đề:** Trong các tác vụ điều khiển cổ điển như bài toán con lắc, hành động có thể là một giá trị liên tục (ví dụ: gia tốc góc trong khoảng từ -3 đến +3) thay vì chỉ ba mức rời rạc.
- **Hạn chế của hành động rời rạc:**
    - Không thể gán xác suất riêng lẻ cho từng hành động vì không gian hành động là vô hạn.
    - Việc giới hạn hành động trong một tập hợp con rời rạc (như đã làm trước đây) cho phép kiểm soát chi tiết nhưng lại hạn chế khả năng của tác nhân trong việc áp dụng gia tốc lớn khi cần thiết.
    - Chúng ta muốn tác nhân có thể áp dụng gia tốc lớn khi thích hợp và gia tốc nhỏ để điều chỉnh cân bằng.

---

## Tham số hóa chính sách bằng phân phối Gaussian

- **Giải pháp:** Tham số hóa chính sách dưới dạng một **phân phối liên tục**, chẳng hạn như **phân phối Gaussian**.

- Hàm mật độ xác suất Gaussian: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
    - $\mu$ (mean): Kiểm soát giá trị trung bình (tâm) của phân phối.
    - $\sigma$ (standard deviation): Kiểm soát độ phân tán của phân phối.
- Chính sách phụ thuộc trạng thái: Chúng ta đặt $\mu$ và $\sigma$ là các **hàm của trạng thái**, để tác nhân có thể gán các phân phối hành động khác nhau cho các trạng thái khác nhau.
    - Tham số hóa $\mu$: Có thể là bất kỳ hàm tham số nào, nhưng để đơn giản, nó được đặt là **hàm tuyến tính của các đặc trưng trạng thái** ($\mu = \theta_\mu^T \phi(s)$). Các tham số chính sách liên quan đến $\mu$ được ký hiệu là $\theta_\mu$.
    - Tham số hóa $\sigma$: Phải luôn **dương**. Để đảm bảo điều này, nó được tham số hóa dưới dạng **hàm mũ của một hàm tuyến tính** ($\sigma = \exp(\theta_\sigma^T \phi(s))$). Các tham số chính sách liên quan đến $\sigma$ được ký hiệu là $\theta_\sigma$.
    - Các tham số chính sách hiện tại bao gồm hai vector tham số đã được xếp chồng ($\theta_\mu$ và $\theta_\sigma$) có kích thước bằng nhau.

Video này khám phá cách áp dụng các phương pháp dựa trên chính sách để xử lý không gian hành động liên tục, cụ thể là bằng cách học một **phân phối Gaussian phụ thuộc trạng thái** trên các hành động liên tục trong thuật toán Actor-Critic.

## Khám phá không gian hành động liên tục

- **Vấn đề:** Trong các tác vụ điều khiển cổ điển như bài toán con lắc, hành động có thể là một giá trị liên tục (ví dụ: gia tốc góc trong khoảng từ -3 đến +3) thay vì chỉ ba mức rời rạc.
- **Hạn chế của hành động rời rạc:**
    - Không thể gán xác suất riêng lẻ cho từng hành động vì không gian hành động là vô hạn.
    - Việc giới hạn hành động trong một tập hợp con rời rạc (như đã làm trước đây) cho phép kiểm soát chi tiết nhưng lại hạn chế khả năng của tác nhân trong việc áp dụng gia tốc lớn khi cần thiết.
    - Chúng ta muốn tác nhân có thể áp dụng gia tốc lớn khi thích hợp và gia tốc nhỏ để điều chỉnh cân bằng.

---

## Tham số hóa chính sách bằng phân phối Gaussian

- **Giải pháp:** Tham số hóa chính sách dưới dạng một **phân phối liên tục**, chẳng hạn như **phân phối Gaussian**.
- **Hàm mật độ xác suất Gaussian:** $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
    - μ (mean): Kiểm soát giá trị trung bình (tâm) của phân phối.
    - σ (standard deviation): Kiểm soát độ phân tán của phân phối.
- **Chính sách phụ thuộc trạng thái:** Chúng ta đặt μ và σ là các **hàm của trạng thái**, để tác nhân có thể gán các phân phối hành động khác nhau cho các trạng thái khác nhau.
    - **Tham số hóa μ:** Có thể là bất kỳ hàm tham số nào, nhưng để đơn giản, nó được đặt là **hàm tuyến tính của các đặc trưng trạng thái** ($\mu = \theta_\mu^T \phi(s)$). Các tham số chính sách liên quan đến μ được ký hiệu là $\theta_\mu$.
    - **Tham số hóa σ:** Phải luôn **dương**. Để đảm bảo điều này, nó được tham số hóa dưới dạng **hàm mũ của một hàm tuyến tính** ($\sigma = \exp(\theta_\sigma^T \phi(s))$). Các tham số chính sách liên quan đến σ được ký hiệu là $\theta_\sigma$.
    - Các tham số chính sách hiện tại bao gồm hai vector tham số đã được xếp chồng ($\theta_\mu$ và $\theta_\sigma$) có kích thước bằng nhau.

---

## Chọn hành động và Khám phá

- **Cách chọn hành động:** Để chọn hành động với chính sách này, chúng ta **lấy mẫu từ phân phối Gaussian** đã xác định bởi μ và σ của trạng thái hiện tại.
- **Vai trò của σ trong khám phá:**
    - σ về cơ bản kiểm soát mức độ khám phá.
    - Ban đầu, phương sai thường được khởi tạo lớn để thử nhiều hành động khác nhau.
    - Khi quá trình học tiến triển, chúng ta kỳ vọng phương sai sẽ thu hẹp lại và chính sách tập trung vào hành động tốt nhất trong mỗi trạng thái.

    - Giống như nhiều chính sách được tham số hóa, tác nhân có thể giảm lượng khám phá theo thời gian thông qua học tập.

---

## Cập nhật Actor-Critic cho Hành động Liên tục

- **Kiến trúc Actor-Critic tương tự:** Mặc dù bây giờ chúng ta có hành động liên tục, chúng ta vẫn có thể sử dụng kiến trúc Actor-Critic tương tự.
- **Khác biệt chính:**
    - **Gradient của chính sách:** Khác biệt vì tham số hóa khác.
    - **Mục tiêu (lý thuyết):** Hơi khác vì chúng ta **tích phân** trên các hành động thay vì tính tổng.
    - **Lấy mẫu gradient:** Cuối cùng, việc lấy mẫu gradient vẫn tương tự.

- Cập nhật Actor: Tất cả những gì chúng ta phải làm là tìm ra gradient của logarit tự nhiên của chính sách ($\nabla_\theta \ln \pi(A_t|S_t,\theta)$).
    - Đối với chính sách Gaussian: $\nabla_\theta \ln \pi(a|s,\theta)$ có hai thành phần, một cho $\theta_\mu$ và một cho $\theta_\sigma$.
        - $\nabla_{\theta_\mu} \ln \pi(a|s,\theta) = \frac{(a-\mu)}{\sigma^2}\phi(s)$
        - $\nabla_{\theta_\sigma} \ln \pi(a|s,\theta) = \frac{(a-\mu)^2-\sigma^2}{\sigma^2}\phi(s)$ (đã sửa lỗi từ video)
    - Quy tắc cập nhật cho Actor sẽ sử dụng các thành phần gradient này nhân với TD error.