**Bài tập 2: Nếu cài đặt Cross-Entropy method để huấn luyện agent leo dốc trong MountainCar-v0 environment, thì thời gian huấn luyện đến khi agent thực hiện được nhiệm vụ sẽ có thể mất nhiều giờ. Thời gian này dài hơn rất nhiều so với Bài tập 1 (chỉ dưới vài chục phút) dù sử dụng cùng reward bounds và network architectures. Hãy giải thích lý do.**

**\* Giải:**

**\* Giải thích vì sao huấn luyện agent với CEM trên MountainCar-v0 lâu hơn rất nhiều so với LunarLander-v2:**

**- MountainCar-v0** yêu cầu agent thực hiện chiến lược phức tạp: phải đi lùi để lấy đà rồi mới leo lên đỉnh dốc được. Tuy nhiên, môi trường này **chỉ cho điểm thưởng khi agent lên tới đích** cuối cùng, không có thưởng cho các bước chuẩn bị (như lấy đà).

-> Vì vậy, agent gần như không có tín hiệu phản hồi gì khi thử các hành động giữa chừng, làm việc học rất khó khăn, đặc biệt lúc đầu khi agent chọn hành động hoàn toàn ngẫu nhiên.

- Ngược lại, **LunarLander-v2** cung cấp nhiều tín hiệu thưởng/phạt rõ ràng, ví dụ:

* + Thưởng khi gần hạ cánh đúng, giữ thăng bằng, giảm tốc độ.
  + Phạt khi va chạm mạnh hoặc mất kiểm soát.

-> Nhờ đó, agent dễ học và điều chỉnh hành vi từng bước trong quá trình huấn luyện.

**- Cross-Entropy Method (CEM)** dựa vào việc chọn ra các tập hành động (episode) tốt nhất để cập nhật chính sách. Nếu phần thưởng hiếm hoặc không phản ánh rõ sự tiến bộ (như MountainCar-v0), CEM rất khó tìm ra các episode tốt để học, dẫn đến tốn nhiều thời gian khám phá và huấn luyện.

**\* Ví dụ minh họa**

Giả sử agent phải leo dốc trong MountainCar-v0:

* Agent ban đầu đi lùi vài bước để lấy đà (nhưng không được thưởng).
* Sau đó mới tiến lên dốc và đạt phần thưởng lớn nhất.
* Vì agent không biết lùi là tốt, nó phải thử rất nhiều lần ngẫu nhiên mới phát hiện ra cách làm này.

Trong khi đó, ở LunarLander-v2:

* Agent được thưởng từng bước nhỏ khi bay đúng hướng, giảm tốc hợp lý.
* Agent có tín hiệu rõ ràng giúp nhanh chóng học được cách hạ cánh an toàn.

**\* Kết luận:** MountainCar-v0 có phần thưởng rất **thưa thớt và khó nắm bắt** chiến lược đúng, nên huấn luyện với CEM mất nhiều thời gian hơn. LunarLander-v2 có phần thưởng phong phú, giúp agent học nhanh và hiệu quả hơn nhiều.