



DOI:10.22144/ctu.jvn.2023.063

## GIẢM THIỂU CỖ MẪU MÔ PHỎNG MONTE CARLO BẰNG PHƯƠNG PHÁP ENTROPY CHÉO

Trần Văn Lý<sup>1\*</sup>, Lý Ngọc Thanh<sup>2</sup>, Lê Phát Tài<sup>2</sup>, Lê Thị Thúy An<sup>3</sup>, Võ Thị Cẩm Tiên<sup>3</sup> và Bùi Thị Thùy Trang<sup>3</sup>

<sup>1</sup>Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ

<sup>2</sup>Lớp Toán ứng dụng – K45, Trường Đại học Cần Thơ

<sup>3</sup>Lớp cao học Lý thuyết xác suất và thống kê toán học – K28, Trường Đại học Cần Thơ

\*Người chịu trách nhiệm về bài viết: Trần Văn Lý (email: tvly@ctu.edu.vn)

### Thông tin chung:

Ngày nhận bài: 03/10/2022

Ngày nhận bài sửa: 29/10/2022

Ngày duyệt đăng: 03/11/2022

### Title:

Minimizing sample size of Monte Carlo simulations by means of Cross-Entropy method

### Từ khóa:

Đổi độ đo xác suất, entropy chéo, phương pháp Monte Carlo, sự kiện hiếm

### Keywords:

Change of probability measure, cross Entropy, Monte Carlo method, rare event

### ABSTRACT

The limitation of the method Monte Carlo Naïve (MCN) was presented in this paper. This method is used to estimate the probability of rare events (events with very small probability). In the MCN method, to be able to observe rare events, it is necessary to initialize very large simulation samples. This limitation can be solved by using the Cross Entropy (CE) algorithm. The numerical results presented in the last section will further clarify the superiority of this method.

### TÓM TẮT

Bài viết trình bày một số hạn chế của phương pháp mô phỏng Monte Carlo cơ bản (Monte Carlo Naïve - MCN). Phương pháp này được sử dụng để ước lượng xác suất sự kiện hiếm (các sự kiện có xác suất xảy ra rất bé). Trong phương pháp MCN, để có thể quan sát được những sự kiện hiếm cần phải khởi tạo các mẫu mô phỏng có kích thước rất lớn. Hạn chế này có thể được giải quyết bằng cách sử dụng một thuật toán Entropy chéo (Cross Entropy - CE). Kết quả áp dụng số được trình bày ở phần cuối cùng sẽ làm rõ hơn tính ưu việt của phương pháp này.

## 1. GIỚI THIỆU

Trong nghiên cứu phát triển khoa học công nghệ, có nhiều trường hợp cần phải thực hiện thực nghiệm các mẫu mô phỏng có kích thước rất lớn để quan sát, thử nghiệm các kịch bản có sự kiện hiếm (các sự kiện có xác suất xảy ra rất thấp). Chẳng hạn trong việc nghiên cứu phát triển hệ thống hỗ trợ lái xe nâng cao ADAS (Advanced Driver Assistance Systems), nếu muốn đánh giá tính hiệu quả của hệ thống khi đối mặt các kịch bản có thể dẫn đến rủi ro va chạm tai nạn (các sự kiện hiếm) thì ta cần phải thực nghiệm với một số lượng rất lớn các kịch bản thử nghiệm. Yêu cầu về cỡ mẫu mô phỏng lớn và cách tiếp cận có thể giảm thiểu cỡ mẫu mô phỏng là hai vấn đề được đặt ra thảo luận trong bài báo này.

Kỹ thuật Monte Carlo (Brémaud, 1999; Rubinstein & Kroese, 2017) thường được sử dụng để mô phỏng các kịch bản thử nghiệm trên các trường và các mô hình ngẫu nhiên. Chẳng hạn một số nghiên cứu đã xây dựng các mô hình ngẫu nhiên dựa trên dữ liệu thực lấy từ N-FOT (Naturalistic Field Operational Tests) để chạy mô phỏng Monte Carlo nhằm đánh giá thử nghiệm các thiết bị, hệ thống. Như được nêu trong Zhao et al. (2017), nhóm nghiên cứu của Woodrooffe đã mô phỏng 1,5 triệu kịch bản va chạm dựa trên dữ liệu xung đột thực tế của N-FOT và sử dụng chúng để đánh giá công nghệ cảnh báo va chạm và kỹ thuật phản ứng phanh thắng giúp giảm thiểu va chạm trên xe tải hạng nặng. Phương pháp mô phỏng thử nghiệm Monte Carlo

được sử dụng để giúp tiết kiệm chi phí thường rất lớn như các dự án N-FOT. Tuy nhiên, ngay cả đối với cách tiếp cận bằng mô phỏng máy tính, khi cần thực nghiệm thử nghiệm với một cỡ mẫu phải rất lớn (để nghiên cứu các sự kiện hiếm) thì các “chi phí”, “tốn thất” về thời gian, về dung lượng khởi tạo các mô phỏng, ... là một trở ngại không nhỏ đặt ra cho các dự án nghiên cứu.

Chìa khóa có thể giúp giải quyết vấn đề phải mô phỏng mẫu kích thước lớn là làm lệch (đổi độ đo xác suất) các phân phối khởi tạo các mẫu Monte Carlo trong các bước lặp của thuật toán Entropy chéo (De Boer et al., 2005) nhưng vẫn có thể duy trì tính đúng đắn và sự chính xác của các tính toán tương ứng cần thiết. Phương pháp Entropy chéo được sử dụng để tìm kiếm cập nhật các tham số tối ưu sao cho tần số xảy ra các sự kiện hiếm trở nên thường xuyên hơn trong mẫu các mô phỏng, giúp gia tăng các cơ hội quan sát, thử nghiệm các thiết bị công nghệ trong các kịch bản có sự kiện hiếm mà không cần phải khởi tạo các mẫu có kích thước quá lớn.

Hai vấn đề thảo luận đã nêu ra, về yêu cầu cỡ mẫu mô phỏng lớn và phương pháp Entropy chéo giúp giảm thiểu cỡ mẫu, sẽ được xem xét trong bài báo theo cách tiếp cận thống kê toán học thông qua việc nghiên cứu các phương pháp ước lượng xác suất của sự kiện hiếm. Được xem xét trước hết là phương pháp ước lượng dựa trên các mô phỏng Monte Carlo cơ bản và sẽ có các thực nghiệm cụ thể để so sánh với cách ước lượng theo phương pháp Entropy chéo. Các áp dụng thực nghiệm số sẽ được trình bày sau cùng trong bài báo, sau khi các vấn đề về thuật toán Entropy chéo được xem xét chặt chẽ.

**2. ƯỚC LƯỢNG MONTE CARLO**

Xét  $X = (X1, X2, \dots, Xn)$  là một vector ngẫu nhiên với các thành phần  $Xi (i = 1, 2, \dots, n)$  có các lớp trạng thái được mã hóa sau cho trạng thái xảy ra với xác suất càng thấp thì sẽ được gán bởi các giá trị càng cao. Hàm tổng  $S(X) = X1 + X2 + \dots + Xn$  có giá trị càng lớn thì xác suất xảy ra càng nhỏ. Sự kiện  $\{S(X) > \gamma\}$  sẽ có xác suất xảy ra càng thấp (là một sự kiện hiếm) khi  $\gamma$  là một mức giá trị lớn.

Vấn đề đặt ra là cần ước lượng xác suất  $\ell$  của sự kiện hiếm  $\{S(X) > \gamma\}$ :

$$\ell = \mathbb{P}\{S(X) > \gamma\} = \mathbb{E}\{\mathbf{1}_{\{S(X) > \gamma\}}\}, \quad (2.1)$$

trong đó  $\mathbf{1}_{\{S(X) > \gamma\}}$  là hàm chỉ định của sự kiện  $\{S(X) > \gamma\}$ .

Ước lượng MCN dựa trên các mẫu độc lập  $X^{(1)}, X^{(2)}, \dots, X^{(N)}$  và có phân phối đồng nhất của  $X$  được tính bởi trung bình mẫu:

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{S(X^{(i)}) > \gamma\}}. \quad (2.2)$$

Với độ tin cậy  $(1 - \alpha)100\%$ , ước lượng khoảng tin cậy  $(\hat{\ell} - \varepsilon_\alpha, \hat{\ell} + \varepsilon_\alpha)$  của  $\ell$  có sai số tương đối  $\ell_r$  được xác định bởi

$$\ell_r = \frac{\varepsilon_\alpha}{\ell}, \quad (2.3)$$

trong đó một nửa khoảng tin cậy (độ chính xác của ước lượng) được cho bởi

$$\varepsilon_\alpha = z_\alpha \sigma(\hat{\ell}) = z_\alpha \sqrt{\text{Var}(\hat{\ell})}, \quad z_\alpha = \Psi^{-1}\left(1 - \frac{\alpha}{2}\right), \quad (2.4)$$

với  $\Psi^{-1}(\cdot)$  là hàm ngược của hàm phân phối xác suất chuẩn tắc  $\mathcal{N}(0,1)$ .

Khi cỡ mẫu  $N$  đủ lớn, theo định lý giới hạn trung tâm, thống kê  $\hat{\ell}$  sẽ xấp xỉ phân phối chuẩn  $\mathcal{N}(\mathbb{E}(\hat{\ell}), \text{Var}(\hat{\ell}))$ , trong đó

$$\mathbb{E}(\hat{\ell}) = \mathbb{E}\left(\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{S(X^{(i)}) > \gamma\}}\right) = \ell, \quad (2.5)$$

$$\text{Var}(\hat{\ell}) = \text{Var}\left(\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{S(X^{(i)}) > \gamma\}}\right) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}\left(\mathbf{1}_{\{S(X^{(i)}) > \gamma\}}\right) = \frac{\ell(1-\ell)}{N}. \quad (2.6)$$

Từ (2.3), (2.4) và (2.6), để sai số tương đối  $\ell_r$  bị chặn bởi một mức  $\beta$  cho trước ( $\beta = 5\%, 10\%, \dots$ ) ta phải có

$$\frac{z_\alpha}{\ell} \sqrt{\frac{\ell(1-\ell)}{N}} \leq \beta. \quad (2.7)$$

Điều này yêu cầu cỡ mẫu phải thỏa mãn

$$N \geq \left(\frac{z_\alpha}{\beta}\right)^2 \frac{1-\ell}{\ell}. \quad (2.8)$$

Yêu cầu (2.8) chỉ ra rằng nếu xác suất  $\ell$  càng bé nghĩa là  $\ell \rightarrow 0$  thì cỡ mẫu sẽ dần về  $\infty$  rất nhanh. Tức là để sai số tương đối của ước lượng xác suất sự kiện hiếm không quá lớn thì yêu cầu cỡ mẫu mô phỏng phải cực lớn. Ví dụ để ước lượng được mức xác suất  $10^{-6}$  sao cho sai số tương đối không vượt quá 10% thì phải khởi tạo mẫu có kích thước tối thiểu phải là 384159616. Đây là lý do mà phương pháp MCN nếu được áp dụng trong trường hợp này sẽ rất chậm, sẽ tốn rất nhiều thời gian và dung lượng để khởi tạo mẫu.

**3. PHƯƠNG PHÁP ENTROPY CHÉO**

**3.1. Phương pháp đổi độ đo xác suất**

Đặt  $\pi(X)$  là hàm mật độ xác suất đồng thời của vector ngẫu nhiên  $X = (X1, X2, \dots, Xn)$ , trong đó sự kiện  $\{S(X) > \gamma\}$  có xác suất tính theo hàm mật độ

này là nhỏ. Ý tưởng ở đây (phương pháp đổi độ đo xác suất) là sẽ thay hàm mật độ xác suất  $\pi(X)$  bởi một hàm mật độ mới  $f(X)$ . Hàm mật độ mới  $f(X)$  được chọn sao cho, trong các mẫu mô phỏng từ hàm mật độ mới này, sự kiện  $\{S(X) > \gamma\}$  có tần suất xảy ra cao hơn.

Cụ thể cách tiếp cận có thể bắt đầu từ sự chuyển đổi cách tính xác suất như sau:

$$\begin{aligned} \mathbb{P}\{S(X) > \gamma\} &= \mathbb{E}_\pi\{\mathbf{1}_{\{S(X) > \gamma\}}\} \\ &= \int \mathbf{1}_{\{S(X) > \gamma\}} \pi(X) dX \\ &= \int \mathbf{1}_{\{S(X) > \gamma\}} L(X) f(X) dX \\ &= \mathbb{E}_f\{\mathbf{1}_{\{S(X) > \gamma\}} L(X)\}, \end{aligned} \quad (3.1)$$

trong đó  $L(X) = \frac{\pi(X)}{f(X)}$ .

Từ (3.1), xác suất  $\ell = \mathbb{P}\{S(X) > \gamma\}$  có thể ước lượng từ mẫu mô phỏng  $X^{(1)}, X^{(2)}, \dots, X^{(N)}$  được khởi tạo từ mật độ  $f(X)$  bởi

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{S(X^{(i)}) > \gamma\}} L(X^{(i)}). \quad (3.2)$$

Trong trường hợp này phương sai của ước lượng được tính bởi

$$\text{Var}_f(\hat{\ell}) = \frac{\mathbb{E}_f(\mathbf{1}_{\{S(X) > \gamma\}} L^2(X)) - \ell^2}{N}.$$

Để sai số tương đối của ước lượng bị chặn bởi một mức  $\beta$  cho trước thì

$$\frac{z_\alpha}{\ell} \sqrt{\frac{\mathbb{E}_f(\mathbf{1}_{\{S(X) > \gamma\}} L^2(X)) - \ell^2}{N}} \leq \beta.$$

Từ đó dẫn đến

$$N \geq \left(\frac{z_\alpha}{\beta}\right)^2 \left(\frac{\mathbb{E}_f(\mathbf{1}_{\{S(X) > \gamma\}} L^2(X))}{\ell^2} - 1\right). \quad (3.3)$$

Quan sát từ (3.3) ta thấy nếu hàm mật độ  $f(X)$  được chọn hợp lý, sao cho  $\mathbb{E}_f(\mathbf{1}_{\{S(X) > \gamma\}} L^2(X))$  có giá trị gần với  $\ell^2$ , thì cỡ mẫu mô phỏng khởi tạo cần thiết sẽ không cần phải quá lớn. Thuật toán Entropy chéo dưới đây sẽ giúp có được điều này.

### 3.2. Thuật toán Entropy chéo

**Thuật toán 3.1.** Thuật toán Entropy chéo ước lượng xác suất sự kiện hiếm  $\ell = \mathbb{P}\{S(X) > \gamma\}$ .

**B1:** Bắt đầu đếm lần lặp bởi  $t = 1$ , chọn hàm mật độ xuất phát là hàm mật độ của mô hình  $f^{(0)}(X) = \pi(X)$ .

**B2:** Ở lần lặp thứ  $t \geq 1$ , khởi tạo mẫu mô phỏng  $X^{(1)}, X^{(2)}, \dots, X^{(N)}$  từ phân phối của hàm mật độ  $f^{(t-1)}(X)$  và tính các tổng  $S(X^{(1)}), S(X^{(2)}), \dots, S(X^{(N)})$  rồi sắp xếp thành dãy

giá trị từ nhỏ đến lớn  $S_{(1)} \leq S_{(2)} \leq \dots \leq S_{(N)}$ . Tính phân vị mức  $(1 - \rho)$  của mẫu  $\hat{\gamma}_t = S_{(\lceil(1-\rho)N\rceil)}$ .

**B3:** Sử dụng mẫu ở bước **B2** cập nhật tham số xác định hàm mật độ xác suất mới  $f^{(t)}(X)$  thỏa mãn

$$\max_{f(X)} \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{S(X^{(i)}) > \gamma\}} L(X^{(i)}) \log(f(X^{(i)})), \quad (3.4)$$

trong đó  $L(X^{(i)}) = \frac{\pi(X^{(i)})}{f^{(t-1)}(X^{(i)})}, i = 1, 2, \dots, N$ .

**B4:** Nếu  $\hat{\gamma}_t \geq \gamma$  thì chuyển sang bước **B5** rồi kết thúc; nếu ngược lại thì trở lại bước **B2**.

**B5:** Đặt  $T$  là chỉ số đếm ở bước lặp sau cùng, khởi tạo mẫu mô phỏng  $X^{(1)}, X^{(2)}, \dots, X^{(M)}$  từ phân phối của hàm mật độ  $f^{(T)}(X)$  và ước lượng xác suất  $\ell$  bởi

$$\hat{\ell} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}_{\{S(X^{(i)}) > \gamma\}} L(X^{(i)}). \quad (3.5)$$

**Chú ý 3.1.** Điều kiện (3.4) được xấp xỉ từ điều kiện

$$\begin{aligned} &\max_{f(X)} \mathbb{E}_{f^{(t-1)}} \left( \mathbf{1}_{\{S(X) > \gamma\}} \frac{\pi(X)}{f^{(t-1)}(X)} \log(f(X)) \right) \\ &\Leftrightarrow \max_{f(X)} \mathbb{E}_\pi \left( \mathbf{1}_{\{S(X) > \gamma\}} \log(f(X)) \right) = \\ &\max_{f(X)} \int \mathbf{1}_{\{S(X) > \gamma\}} \log(f(X)) \pi(X) dX. \end{aligned}$$

Điều này sẽ làm cho khoảng cách Kullback-Leibler  $\mathcal{D}(g(X), f(X))$  dưới đây nhỏ nhất

$$\begin{aligned} \mathcal{D}(g(X), f(X)) &= \mathbb{E}_g \log \left( \frac{g(X)}{f(X)} \right) \\ &= \int g(X) g(X) dX - \int \log(f(X)) g(X) dX, \end{aligned}$$

trong đó  $g(X) = \mathbf{1}_{\{S(X) > \gamma\}} \pi(X)$ . Tức là các bước lặp của thuật toán sẽ giúp khoảng cách Kullback-Leibler giữa hàm mật độ mới  $f^{(t)}(X)$  và hàm mục tiêu  $\pi(X)$  ngày càng gần hơn.

Khoảng cách Kullback-Leibler ở trên còn được gọi là Entropy chéo giữa 2 hàm  $g(X)$  và  $f(X)$  và do đó thuật toán trên được gọi là thuật toán Entropy chéo.

Sự tồn tại hàm mật độ tối ưu  $f^{(T)}(X)$  trong thuật toán đã được chỉ ra ở mệnh đề 5.1 trong Homem-de-Mello and Rubinstein (2002). Mục tiếp theo dưới đây sẽ xem xét sự đảm bảo về điều kiện dừng của thuật toán.

### 3.3. Điều kiện dừng của thuật toán Entropy chéo

Cơ sở đảm bảo về tính dừng của thuật toán Entropy chéo trong việc thực nghiệm số được thể hiện qua Định lý 3.1.

**Định lý 3.1.** Xét  $\hat{\gamma}_t$  là phân vị mức  $(1 - \rho)$  của mẫu mô phỏng  $X^{(1)}, X^{(2)}, \dots, X^{(N)}$  được khởi tạo từ

phân phối của hàm mật độ  $f^{(t)}(X)$  ở bước lặp  $t$  của thuật toán 3.1. Với mỗi giá trị  $\gamma$ , luôn tồn tại số  $N_t$  nguyên dương và một số dương  $\rho_t$  sao cho với mọi số nguyên dương  $N$  thỏa  $N \geq N_t$  và mọi số dương  $\rho$  thỏa  $\rho < \rho_t$  thì hầu chắc chắn (với xác suất 1) ta sẽ có  $\hat{\gamma}_t > \gamma$ , tức là thuật toán 3.1 hầu chắc chắn sẽ dừng.

**Chứng minh.** Trước hết chúng ta lưu ý rằng ở mỗi bước lặp  $t$  của thuật toán 3.1 phân vị mức  $(1 - \rho)$  của tổng  $S(X)$  có thể nhận được từ lời giải tối ưu của min  $\mathbb{E}_t\{\phi[S(X), \xi]\}$  với

$$\phi[S(X), \xi] = \begin{cases} (1 - \rho)[S(X) - \xi] & \text{nếu } S(X) \geq \xi \\ \rho[\xi - S(X)] & \text{nếu } S(X) \leq \xi \end{cases} \quad (3.6)$$

Thật vậy, theo (3.6), kỳ vọng của  $\phi[S(X), \xi]$  được tính bởi

$$\mathbb{E}_t\{\phi[S(X), \xi]\} = (1 - \rho)[S(X) - \xi] \mathbb{P}_t\{S(X) \geq \xi\} + \rho[\xi - S(X)] \mathbb{P}_t\{S(X) \leq \xi\}.$$

Do đó

$$\frac{\partial \mathbb{E}_t\{\phi[S(X), \xi]\}}{\partial \xi} = \rho - \mathbb{P}_t\{S(X) \geq \xi\}.$$

Đặt là  $\gamma_t$  phân vị mức  $(1 - \rho)$  của  $S(X)$  tức là  $\mathbb{P}_t\{S(X) \leq \gamma_t\} = 1 - \rho$  hay  $\rho - \mathbb{P}_t\{S(X) \geq \gamma_t\} = 0$  thì đồng thời ta cũng có

$$\begin{cases} \rho - \mathbb{P}_t\{S(X) \geq \xi\} \leq 0 & \text{với } \xi \leq \gamma_t \\ \rho - \mathbb{P}_t\{S(X) \geq \xi\} \geq 0 & \text{với } \xi \geq \gamma_t \end{cases} \quad (3.7)$$

Điều này đồng nghĩa  $\gamma_t$  cũng là lời giải tối ưu của min  $\mathbb{E}_t\{\phi[S(X), \xi]\}$ .

Tương tự, nếu ký hiệu  $\hat{\gamma}_t$  là phân vị mức  $(1 - \rho)$  của mẫu mô phỏng  $X^{(1)}, X^{(2)}, \dots, X^{(N)}$  từ phân phối của hàm mật độ  $f^{(t)}(X)$  thì  $\hat{\gamma}_t$  là lời giải tối ưu xấp xỉ của min  $\frac{1}{N} \sum_{i=1}^N \phi[S(X^{(i)}), \xi]$ .

Do  $\mathbb{E}_t\{\phi[S(X), \xi]\}$  là hàm lồi nên khoảng cách giữa  $\gamma_t$  và  $\hat{\gamma}_t$  sẽ dần về 0 khi  $N$  dần về  $\infty$ . Nghĩa là tồn tại số nguyên dương  $N_t$  sao cho  $\gamma_t \approx \hat{\gamma}_t$ , với  $\hat{\gamma}_t$  là phân vị mức  $(1 - \rho)$  của mẫu mô phỏng có kích thước  $N > N_t$ .

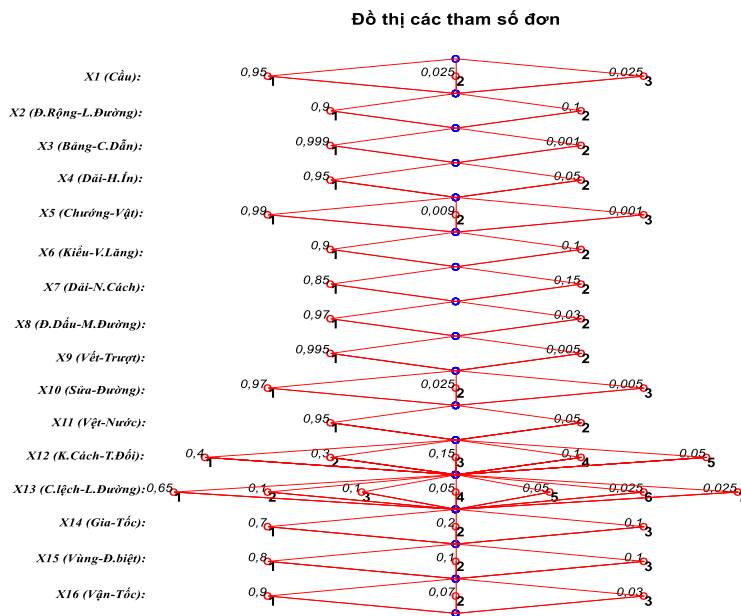
Giả sử  $\gamma < \max X$  khi đó  $\rho_t = \mathbb{P}_t\{S(X) \geq \gamma\} > 0$  và có thể dễ dàng chỉ ra rằng  $\gamma_t > \gamma$  là phân vị mức  $(1 - \rho)$  của  $S(X)$ ,  $\rho \in (0; \rho_t)$ .

Các điều trên chỉ ra rằng tồn tại số dương  $\rho_t = \mathbb{P}_t\{S(X) \geq \gamma\}$  và số nguyên dương  $N_t$  sao cho với mọi  $\rho \in (0; \rho_t)$  và mọi số nguyên dương  $N > N_t$  luôn có hầu chắc chắn (với xác suất 1)  $\hat{\gamma}_t > \gamma$ .  $\square$

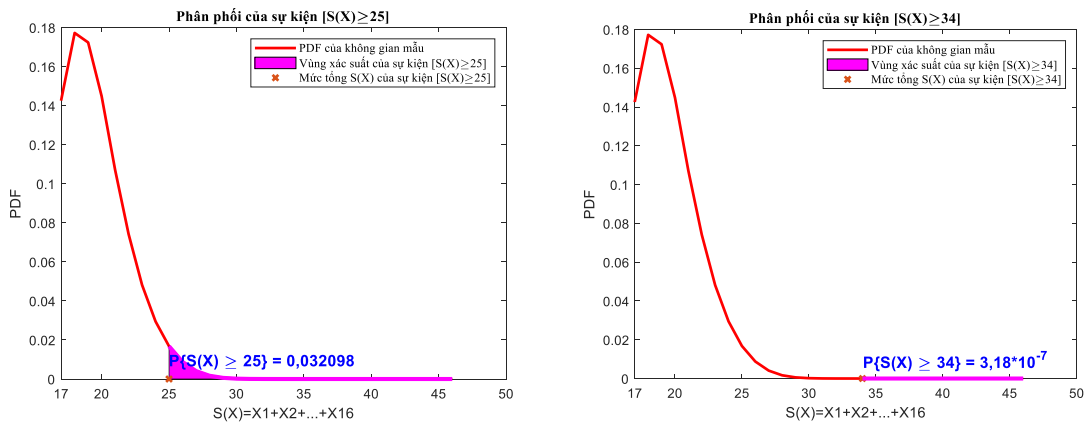
**Chú ý 3.2.** Trong thực nghiệm số của thuật toán Entropy chéo, các phân vị mức  $(1 - \rho)$  thường được lấy với  $\rho \in [0,01; 0,1]$ ; nếu  $\rho$  lớn thì cỡ mẫu  $N$  ở các bước lặp không cần phải quá lớn, khi đó thực nghiệm thuật toán cần nhiều bước lặp hơn; ngược lại, nếu  $\rho$  nhỏ thì cỡ mẫu khởi tạo cần phải lấy đủ lớn.

**Bảng 1. Mã hóa các trạng thái các tham số thành phần của vector  $X = (X1, X2, \dots, X16)$**

Ký hiệu	Tham số	Mã hóa các trạng thái
X1	Cầu	1: Không-Có-Cầu, 2: Dưới-Cầu, 3: Trên-Cầu
X2	Độ-Rộng-Đường	1: Đường-Thường, 2: Đường-Có-Vĩa-Hè
X3	Bảng-Chỉ-Dẫn	1: Không-Bảng-Chỉ-Dẫn, 2: Có-Bảng-Chỉ-Dẫn
X4	Dài-Hắc-Ấn	1: Không-Dài-Hắc-Ấn, 2: Có-Dài-Hắc-Ấn
X5	Chướng-Vật	1: Không-Chướng-Vật, 2: Chướng-Vật-Bất-Ngờ, 3: Chướng-Vật-Thường
X6	Kiểu-Vô-Lăng	1: Vô-Lăng-Phải, 2: Vô-Lăng-Trái
X7	Dây-Ngăn-Cách	1: Có-Dây-Ngăn-Cách, 2: Không-Dây-Ngăn-Cách
X8	Đánh-Dấu-Mặt-Đường	1: Có-Đ.Dấu-M.Đường, 2: Không-Đ.Dấu-M.Đường
X9	Vết-Trượt	1: Có-Vết-Trượt, 2: Không-Vết-Trượt
X10	Sửa-Đường	1: Có-Sửa-Đường, 2: Không-Sửa-Đường
X11	Vết-Nước	1: Có-Vết-Nước, 2: Không-Vết-Nước
X12	Khoảng-Cách-Tương-Đối	1: $K.C \geq 100m$ , 2: $50 \leq K.C < 100$ , 5: $K.C < 10$ , 3: $25 \leq K.C < 50$ , 4: $10m \leq K.C < 25m$ 1: $-0,5m \leq C.L < -1m$ , 2: $-1m \leq C.L < -1,5m$ , 3: $-1,5m \leq C.L < -2m$ , 4: Không-Chênh-Lệch, 5: $0,5m \leq C.L < 1m$ , 6: $1m \leq C.L < 1,5m$ , 7: $1,5m \leq C.L < 2m$
X13	Chênh-Lệch-Làn-Đường	1: Bình-Thường, 2: Yếu, 3: Mạnh
X14	Vùng-Đặc-Biệt	1: Tăng-S.Làn, 2: Không-Vùng-Đ.Biệt, 3: Giảm-S.Làn
X16	Vận-Tốc-Xe-Tự-Hành	1: Nhanh, 2: Trung-Bình, 3: Chậm



Hình 1. Phân bố xác suất các trạng thái của các thành phần vector  $X = (X_1, X_2, \dots, X_{16})$



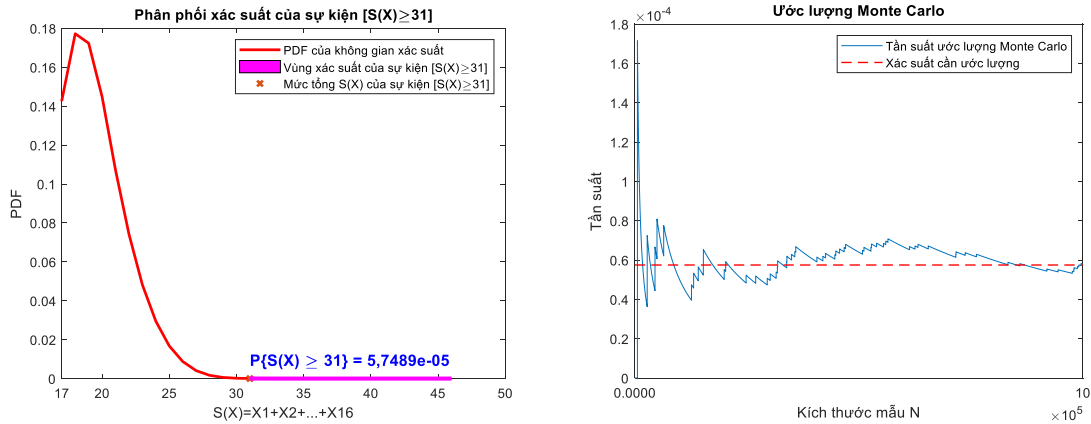
Hình 2. Phân bố xác suất của sự kiện  $[S(X) \geq 25]$  (trái) và sự kiện  $[S(X) \geq 34]$  (phải)

## 4. ÁP DỤNG SỐ

### 4.1. Mô hình ngẫu nhiên

Xét một nhóm gồm 16 tham số độc lập, là các tham số đơn giản nhất trong mô hình mô phỏng các kịch bản thảm định hệ thống hỗ trợ lái xe nâng cao ADAS sử dụng cho xe tự hành trình (xe tự lái). Các tham số  $X_1, X_2, \dots, X_{16}$  được xét trong mô hình vector ngẫu nhiên  $X$  có 16 biến độc lập  $X = (X_1, X_2, \dots, X_{16})$ . Các tham số thành phần này có các trạng thái được mã hóa như trong Bảng 1 với các xác suất tương ứng được giới thiệu ở Hình 1.

Do các trạng thái có xác suất xảy ra nhỏ được mã hóa bởi các giá trị lớn nên xác suất để tổng  $S(X) = X_1 + X_2 + \dots + X_{16}$  nhận giá trị lớn sẽ rất nhỏ, tức là các sự kiện có dạng  $[S(X) \geq \gamma]$  sẽ có xác suất  $P\{S(X) \geq \gamma\}$  rất nhỏ khi  $\gamma$  lớn. Chẳng hạn Hình 2 (trái) thể hiện vùng xác suất  $P\{S(X) \geq 25\} = 0,032$ , Hình 3 (trái) thể hiện vùng xác suất  $P\{S(X) \geq 31\} = 5,75 \cdot 10^{-5}$ , còn Hình 2 (phải) biểu diễn cho vùng xác suất  $P\{S(X) \geq 34\} = 3,18 \cdot 10^{-7}$ . Sự kiện  $[S(X) \geq 34]$  được xem là sự kiện hiếm do có xác suất xảy ra rất nhỏ.



Hình 3. Phân bố xác suất (trái) và ước lượng MCN (phải) của sự kiện  $[S(X) \geq 31]$

4.2. Thử nghiệm ước lượng

Sử dụng mẫu mô phỏng MCN với kích thước  $N = 10^6$ , xác suất  $P\{S(X) \geq 31\}$  ước lượng được là  $5,8 \cdot 10^{-5}$  (Hình 3) với sai số tương đối là 0,26 và thời gian khởi tạo là 1892 giây. Theo (3.3), nếu muốn ước lượng xác suất này theo MCN với sai số tương đối là không quá 0,25 thì cỡ mẫu Monte Carlo tối thiểu cần khởi tạo là  $N = 1069113$ . Nếu sử dụng phương pháp CE thì chỉ cần mẫu với kích thước  $N = 1000$ , qua 3 vòng lặp của thuật toán, xác suất ước lượng nhận được là  $5,72 \cdot 10^{-5}$  với sai số tương đối 0,24 và thời gian khởi tạo chỉ 1,8 giây.

Đối với xác suất  $P\{S(X) \geq 34\}$ , qua 4 vòng lặp của thuật toán CE với cỡ mẫu khởi tạo là  $N = 20000$ , giá trị ước lượng nhận được là  $2,71 \cdot 10^{-7}$  với sai số tương đối là 0,254 và thời gian khởi tạo là 13,7 giây. Để ước lượng xác suất này với sai số tối đa là 0,254 thì khi tính toán theo (3.3), nếu sử dụng phương pháp MCN cần khởi tạo mẫu có kích thước tối thiểu là  $N = 1,93 \cdot 10^8$ .

Các thử nghiệm đã thực hiện ở trên cho thấy phương pháp CE giúp giảm thiểu ít đi hơn 1000 lần về cỡ mẫu khởi tạo và thời gian khởi tạo so với phương pháp MCN. Những thử nghiệm này được chạy mô phỏng từ các chương trình được mã hóa trong phần mềm Matlab 2020a và thực hiện chạy trên máy tính Processor Core i5-4460 CPU 3.20 GHz RAM 8G.

TÀI LIỆU THAM KHẢO

Brémaud, P. (1999). Markov chains – Gibbs Fields, Monte Carlo Simulation and Queues. Springer – New York. <https://doi.org/10.1007/978-1-4757-3124-8>

5. KẾT LUẬN

So với phương pháp ước lượng dựa trên các mẫu mô phỏng Monte Carlo cơ bản, khi cần ước lượng xác suất rất nhỏ của các sự kiện hiếm, phương pháp sử dụng thuật toán Entropy chéo giúp giảm rất nhiều lần kích thước mẫu cần khởi tạo. Đây là điều đã được phân tích và xem xét chặt chẽ về nền tảng cơ sở lý thuyết, đồng thời cũng đã được thể hiện trên các kết quả thực nghiệm. Về mặt đóng góp ứng dụng, kết quả được khẳng định trong bài báo đặc biệt rất có ý nghĩa trong các trường hợp cần thực hiện thử nghiệm, thẩm định các thiết bị, công nghệ mới, nhất là các thử nghiệm cho những kịch bản hiếm xảy ra theo các điều kiện thông thường. Chẳng hạn như thử nghiệm các thiết bị hỗ trợ tránh va chạm tai nạn trong nghiên cứu phát triển xe tự lái, hay các thử nghiệm những tính huống nghẽn mạng, các mô phỏng thử nghiệm hạt nhân, ... Về những nghiên cứu có thể phát triển tiếp theo, thứ nhất có thể nghiên cứu cải tiến thuật toán 3.1 để có thể tối ưu cỡ mẫu khởi tạo và tối ưu mức phân vị trong từng bước lặp của thuật toán; khả năng thứ hai là có thể nghiên cứu kết hợp với các thuật toán tạo mẫu Gibbs (Levine & Casella, 2006) nhằm làm tăng tính hiệu quả của phương pháp khi áp dụng cho các cấu trúc mô hình phức tạp.

LỜI CẢM ƠN

Các tác giả xin trân trọng cảm ơn Đại học Mines-ParisTech đã tài trợ kinh phí và cung cấp dữ liệu nghiên cứu thông qua Dự án EUFRA03321NCTN.

De Boer, P.T., Kroese, D.P., Mannor, S. & Rubinstein, R. Y. (2005). A Tutorial on the Cross- Entropy Method. Annals of Operations Research, 134,19-67.

- <https://doi.org/10.1007/s10479-005-5724-z>  
Homem-de-Mello T. & Rubinstein, R. Y. (2002). Estimation of rare event probabilities using cross-entropy. *Proceedings of the Winter Simulation Conference*, vol.1, 310-319. <https://doi.org/10.1109/WSC.2002.1172900>
- Levine, R. A., & Casella, G. (2006). Optimizing random scan gibbs samplers. *Journal of Multivariate Analysis*, 97, 2071-2100. <https://doi.org/10.1016/j.jmva.2006.05.008>
- Rubinstein, R. Y., & Kroese, D. P. (2017). *Simulation and the Monte Carlo method* (Wiley Series in Probability and Statistics). John Wiley & Sons, Inc., Hoboken, NJ. <https://doi.org/10.1002/9781118631980>
- Zhao, D., Lam, H., Peng, H., Bao, S., LeBlanc, D. J., Nobukawa, K., & Pan, C. S. (2017). Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques. *IEEE Transactions on Intelligent Transportation Systems*, 18(3), 595-607. <https://doi.org/10.1109/TITS.2016.2582208>