

The Central Limit Theorem

Định lý giới hạn trung tâm

Đặng Thanh Hải (Ph.D)

School of Engineering and Technology, VNUH

Email:hai.dang@vnu.edu.vn

The Road to the “Central Limit Theorem”

- Sums of random variables
 - ✓ Tổng các BNN
- The Weak and Strong laws of large numbers
 - ✓ Luật số lớn mạnh và yếu
- The “CTL”
 - ✓ Định lý giới hạn trung tâm

Sums of Random Variables

Tổng các BNN

- **Sums of random variables have many applications in a wide range of engineering and science problems**

Tổng các BNN có nhiều ứng dụng trong một phổ rộng lớn các vấn đề khoa học và kỹ thuật

- **One important case is the sum of: independent-identically-distributed (iid) random variables**

Một trường hợp quan trọng là tổng của: các BNN phân bố giống nhau - độc lập nhau (iid)

Sums of Random Variables

Tổng các BNN

■ For example, counting the number of occurrences of an event Ví dụ, đếm số lần xuất hiện 1 sự kiện

- this counting process is a "sum of random variables" each of which takes on "1" or "0"
- E.g. number of packets arriving at a router

Ví dụ, số gói tin đến 1 router

Quá trình
đếm này là
một tổng
các BNN,
mỗi biến
nhận giá trị
“1” hoặc “0”

■ Another example, accumulation of some parameter Ví dụ khác, sự tích luỹ của các tham số

- E.g. total (end-to-end) delay of a packet going through a large number of routers

Ví dụ, tổng độ trễ (từ đầu cuối đến đầu cuối) của 1 gói tin trên một lượng lớn các router

■ Third example, averaging a large number of measurements

Ví dụ thứ 3, tính trung bình trên một lượng lớn các giá trị đo đạc

Sums of Random Variables

Tổng các BNN

Với các BNN phân bố giống nhau - độc lập nhau (iid):

■ For independent-identically-distributed (iid) random variables:

- They have the same pdf \Rightarrow
Chúng có cùng hàm mật độ xác suất pdf
 - They have the same moments (e.g. means and variances)
Chúng có cùng mô men (v.d. kỳ vọng và phương sai)
- They are independent \Rightarrow
Chúng độc lập nhau
 - They are uncorrelated (zero correlation coefficient and zero covariance)
Chúng không tương quan đến nhau (hệ số tương quan 0 và đồng phương sai 0)

Sums of Random Variables

Tổng các BNN

- We start by looking at the sum:

Chúng ta bắt đầu bằng xem xét tổng:

$$S_n = X_1 + X_2 + \cdots + X_n$$

- And its: Và các giá trị đặc trưng sau của nó:

Expected value $E[S_n]$ Giá trị kỳ vọng E[Sn]

Variance(S_n) Phương sai (Sn)

Probability density function (pdf)

Hàm mật độ xác suất (pdf)

Sums of Random Variables

Tổng các BNN

$$S_n = X_1 + X_2 + \cdots + X_n$$

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$$

Kỳ vọng của tổng bằng tổng các kỳ vọng

- The expected value of the sum is the sum of the expected values
- This is true for both dependent and independent X_1, X_2, X_3, \dots

Điều này đúng cho cả trường hợp các BNN X_1, \dots phụ thuộc hoặc độc lập nhau

Variance of the Sum

Phương sai của tổng

- Let $Z=X+Y$, what is the variance of Z ?

Cho $Z=X+Y$, phương sai của Z là gì?

$$\text{VAR}(Z) = E[(Z - E[Z])^2]$$

$$= E[(X + Y - E[X] - E[Y])^2]$$

$$\text{VAR}(Z) = E\left[\{(X - E[X]) + (Y - E[Y])\}^2\right]$$

Variance of the Sum

Phương sai của tổng

$$\begin{aligned} \text{VAR}(Z) = & E [(X - E[X])^2 + (Y - E[Y])^2 \\ & + 2(X - E[X])(Y - E[Y])] \end{aligned}$$

$$\boxed{\text{VAR}(Z) = \text{VAR}(X) + \text{VAR}(Y) + 2\text{COV}(X, Y)}$$

Variance of the Sum

Phương sai của tổng

$$\text{VAR}(X_1 + X_2 + \dots + X_n)$$

$$= \sum_{k=1}^n \text{VAR}(X_k) + \sum_{j=1}^n \sum_{\substack{k=1 \\ j \neq k}}^n \text{cov}(X_j, X_k)$$

Phương sai của tổng là tổng các phương sai cộng với tổng các đồng phương sai

- **The variance of the sum is the sum of the variances plus the sum of the covariances**

Variance of the Sum

Phương sai của tổng

- For independent X_1, X_2, X_3, \dots

Với các BNN X_1, X_2, X_3, \dots độc lập nhau

$$\text{COV}(X_j, X_k) = 0 \text{ for } j \neq k$$

$$\text{VAR}(X_1 + X_2 + \dots + X_n) =$$

$$\text{VAR}(X_1) + \dots + \text{VAR}(X_n)$$

IID Random Variables

Các BNN iid (phân bố giống nhau, độc lập nhau)

- If $X_1, X_2, X_3, \dots, X_n$ are iid RVs with:

Nếu $X_1, X_2, X_3, \dots, X_n$ là các BNN iid với:

$$E[X_j] = m_x \text{ for } j = 1, \dots, n$$

$$VAR(X_j) = \sigma_x^2 \text{ for } j = 1, \dots, n$$

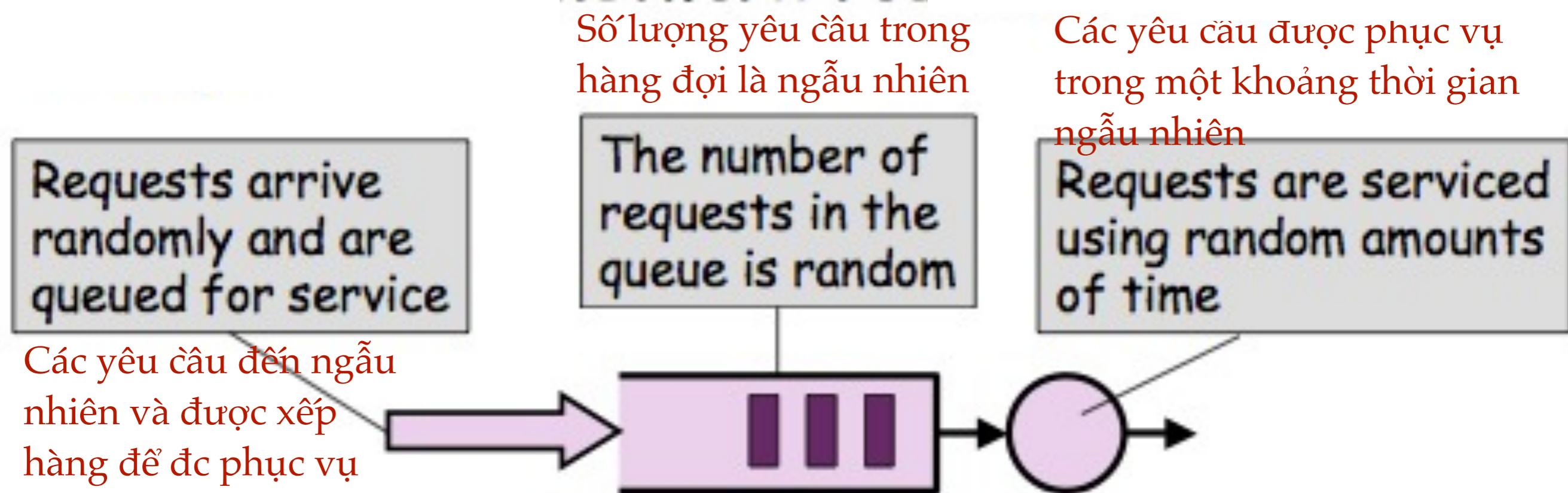
$$E[S_n] = E[X_1] + \dots + E[X_n] = nm_x$$

$$VAR(S_n) = nVAR(X_j) = n\sigma_x^2$$

Sum of a Random Number of RVs

Tổng của một số ngẫu nhiên các BNN

- Nay chúng ta thảo luận về 1 loại tổng quát hơn của tổng mà có nhiều ý nghĩa trong việc mô tả hành vi của nhiều hệ thống, như hệ hàng đợi (v.d. mang các router, máy tính, v.v.)



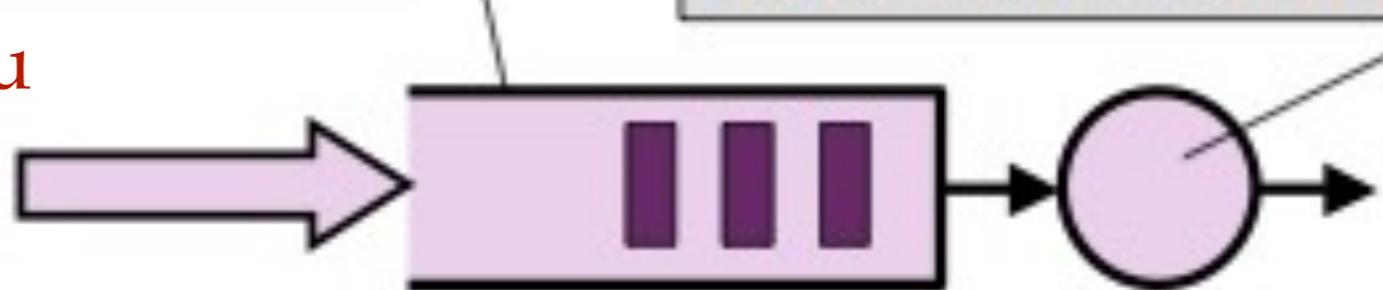
Sum of a Random Number of RVs

Tổng của một số ngẫu nhiên các BNN

The number of requests in the queue is a random variable N

The time it takes to service a packet (k) is a random variable X_k

Số lượng các yêu cầu trong hàng đợi là một BNN
 N



Thời gian để xử lý một yêu cầu (k) là một BNN X_k

Sum of a Random Number of RVs

Tổng của một số ngẫu nhiên các BNN

The total amount of time it takes to service all of the packets in the queue is:

Tổng thời gian để xử lý tất cả các yêu cầu trong hàng đợi là:

$$S_N = \sum_{k=1}^N X_k$$

In this case, N is a random variable as well as X_1, X_2, \dots

Trong trường hợp này, N là một BNN, X1, X2, cũng vậy

Sum of a Random Number of RVs

Tổng của một số ngẫu nhiên các BNN

- Therefore, when N and X_1, X_2, \dots, X_N are random variables, then for:

Do đó, khi N và X_1, X_2, \dots, X_N là các BNN, ta có:

$$S_N = \sum_{k=1}^N X_k$$

$$E[S_N] = E[N]E[X]$$

Trung bình mẫu

- X là một BNN với kỳ vọng $E[X] = \mu$, chưa biết
- X_1, X_2, \dots, X_n : n phép đo độc lập của X

$$M_n = \frac{1}{n} \sum_{j=1}^n X_j$$

- Ta có:

$$E[M_n] = E\left[\frac{1}{n} \sum_{j=1}^n X_j\right] = \frac{1}{n} \sum_{j=1}^n E[X_j] = \mu$$

$$\text{VAR}[M_n] = \frac{1}{n^2} \text{VAR}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Trung bình mẫu

- Áp dụng bất đẳng thức Chebyshev với M_n

$$P[|M_n - E[M_n]| \geq \varepsilon] \leq \frac{\text{VAR}[M_n]}{\varepsilon^2}$$

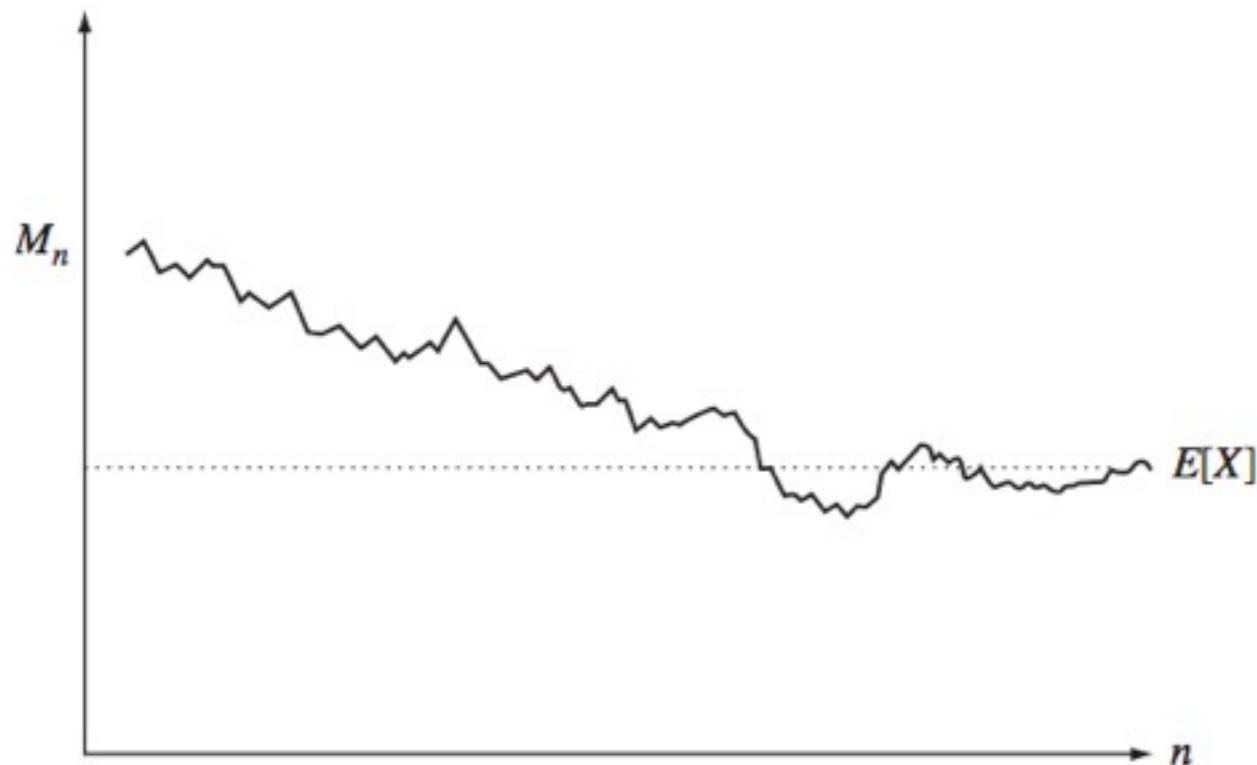
$$P[|M_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$P[|M_n - \mu| < \varepsilon] \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Luật số lớn

- X_1, X_2, \dots, X_n là chuỗi các biến ngẫu nhiên giống nhau và độc lập nhau, có kỳ vọng hữu hạn $E[X] = \mu$, và phương sai hữu hạn, khi đó:

$$P\left[\lim_{n \rightarrow \infty} M_n = \mu \right] = 1$$



Luật số lớn: ví dụ

In order to estimate the probability of an event A , a sequence of Bernoulli trials is carried out and the relative frequency of A is observed. How large should n be in order to have a .95 probability that the relative frequency is within 0.01 of $p = P[A]$?

Let $X = I_A$ be the indicator function of A . From Table 3.1 we have that the mean of I_A is $\mu = p$ and the variance is $\sigma^2 = p(1 - p)$. Since p is unknown, σ^2 is also unknown. However, it is easy to show that $p(1 - p)$ is at most 1/4 for $0 \leq p \leq 1$. Therefore, by Eq. (7.19),

$$P[|f_A(n) - p| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

The desired accuracy is $\varepsilon = 0.01$ and the desired probability is

$$1 - .95 = \frac{1}{4n\varepsilon^2}.$$

We then solve for n and obtain $n = 50,000$

Định lý giới hạn trung tâm (Central Limit Theorem)

- Gọi S_n là tổng của n biến ngẫu nhiên phân bố giống nhau, độc lập nhau có kỳ vọng hữu hạn $E[X] = \mu$, và phương sai hữu hạn σ^2 , gọi Z_n là biến ngẫu nhiên được định nghĩa như sau:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \text{ hay } Z_n = \sqrt{n} \frac{M_n - \mu}{\sigma}.$$

Khi đó ta có

$$\lim_{n \rightarrow \infty} P[Z_n \leq z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

Định lý giới hạn trung tâm

- Suppose that orders at a restaurant are iid random variables with mean $\mu = \$8$ and standard deviation $\sigma = \$2$. Estimate the probability that the first 100 customers spend a total of more than \$840. Estimate the probability that the first 100 customers spend a total of between \$780 and \$820.
- After how many orders can we be 90% sure that the total spent by all customers is more than \$1000?
- The time between events in a certain random experiment is iid exponential random variables with mean m seconds. Find the probability that the 1000th event occurs in the time interval $(11000 \pm 50)m$.