



# TOÁN CHO KHOA HỌC MÁY TÍNH

## PHÂN TÍCH ĐỘ CHỆCH VÀ PHƯƠNG SAI

TS. Lương Ngọc Hoàng



# Nội dung

1. Giới thiệu
2. Ước lượng và phân tích lỗi của ước lượng
3. Phân tích độ chêch – phương sai
4. Sự đánh đổi giữa độ chêch và phương sai



# GIỚI THIỆU

## INTRODUCTION



# Ví dụ

Player	Height	Weight	Yrs Expr	2 Points	3 Points	Salary
1	...	...	...	...	...	...
2	...	...	...	...	...	...
3	...	...	...	...	...	...
...	...	...	...	...	...	...

- Dự đoán lương của vận động viên bóng rổ - biến ngẫu nhiên đầu ra, gọi là  $Y$ .
- Các biến còn lại ứng với các đặc trưng đầu vào – gọi là  $X_1, X_2, \dots, X_p$ .
- Ta giả sử tồn tại một **hàm đích (target function)**  $f()$ :

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

là một ánh xạ từ không gian các đặc trưng đầu vào  $\mathcal{X}$  tới không gian giá trị đích đầu ra  $\mathcal{Y}$ .

**Lưu ý: ta không có thông tin về hàm đích này.**

- Ta muốn “**đi tìm**” hàm đích này.



# Ví dụ

Player	Height	Weight	Yrs Expr	2 Points	3 Points	Salary
1	...	...	...	...	...	...
2	...	...	...	...	...	...
3	...	...	...	...	...	...
...	...	...	...	...	...	...

- Ta có một tập dữ liệu  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$  với  $\mathbf{x}_i$  là vector đặc trưng của vận động viên thứ  $i$ , và  $y_i$  là lương của vận động viên đó.
- Trên dữ liệu này, ta huấn luyện một mô hình, gọi là **mô hình giả thiết (hypothesis model)**  $\hat{f}()$ :

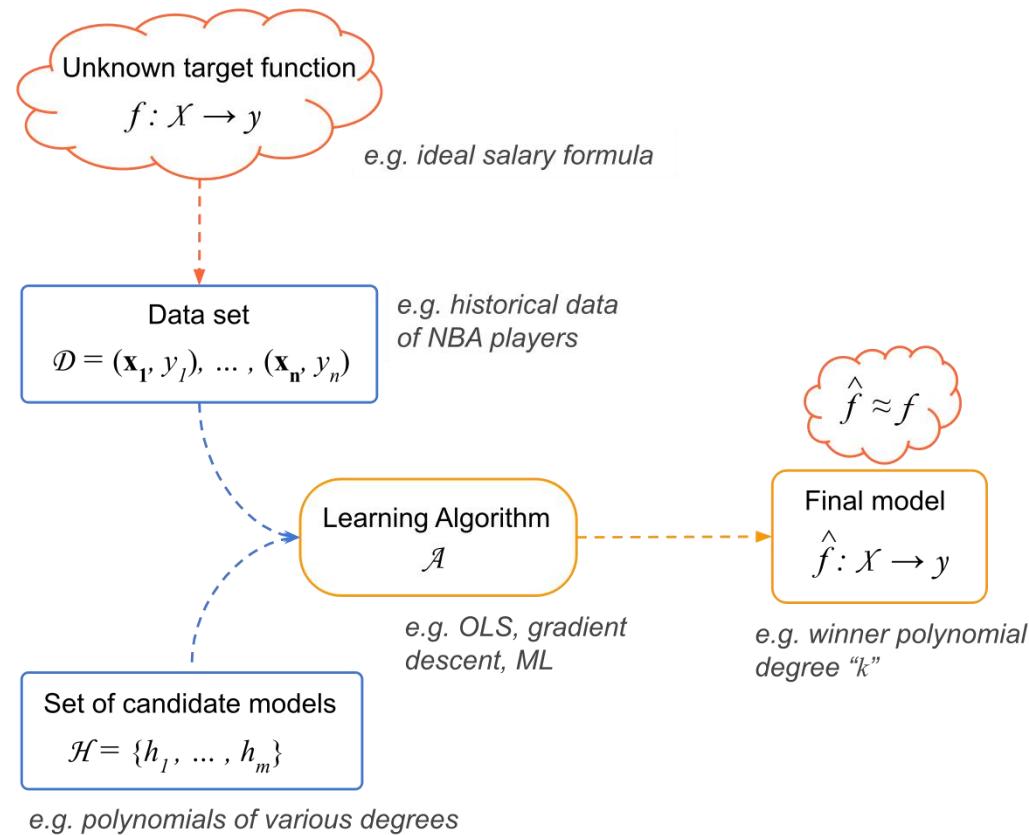
$$\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$$

có thể **xấp xỉ hàm đích (target function)**  $f()$ .

- Để tìm  $\hat{f}()$ , ta cần xem xét các mô hình ứng viên (candidate models), còn gọi là **tập giả thiết (hypothesis set)**  $\mathcal{H} = \{h_1, h_2, \dots, h_m\}$ .
- Mô hình giả thiết được chọn từ tập giả thiết  $h_m^*$  được dùng làm mô hình cuối cùng  $\hat{f}$ .



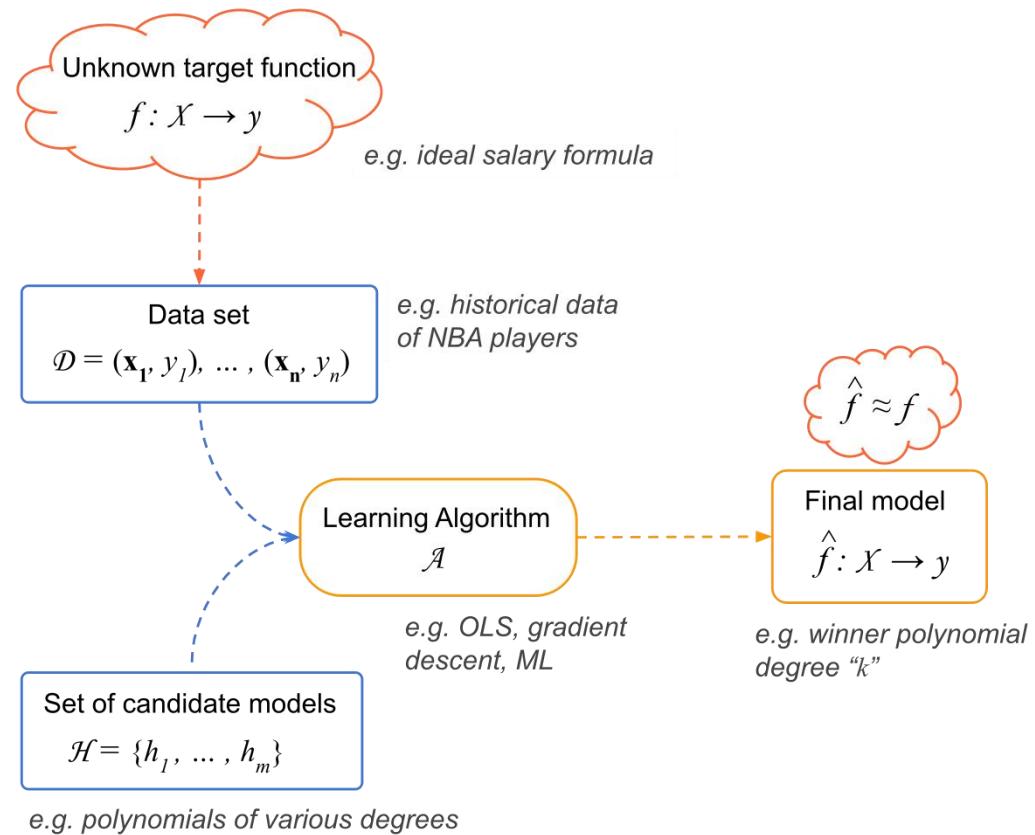
# Bộ khung lý thuyết



- **Hàm đích (target function)**  $f$  là thông tin ta không có, và ta không thể nào thực sự xác định được hàm  $f$ . Ta chỉ có thể tìm ra một xấp xỉ đủ tốt cho  $f$  bằng một ước lượng  $\hat{f}$ .
- Ý tưởng  $\hat{f} \approx f$  thực ra cũng mang tính lý thuyết vì ta không bao giờ biết chính xác  $f$ .



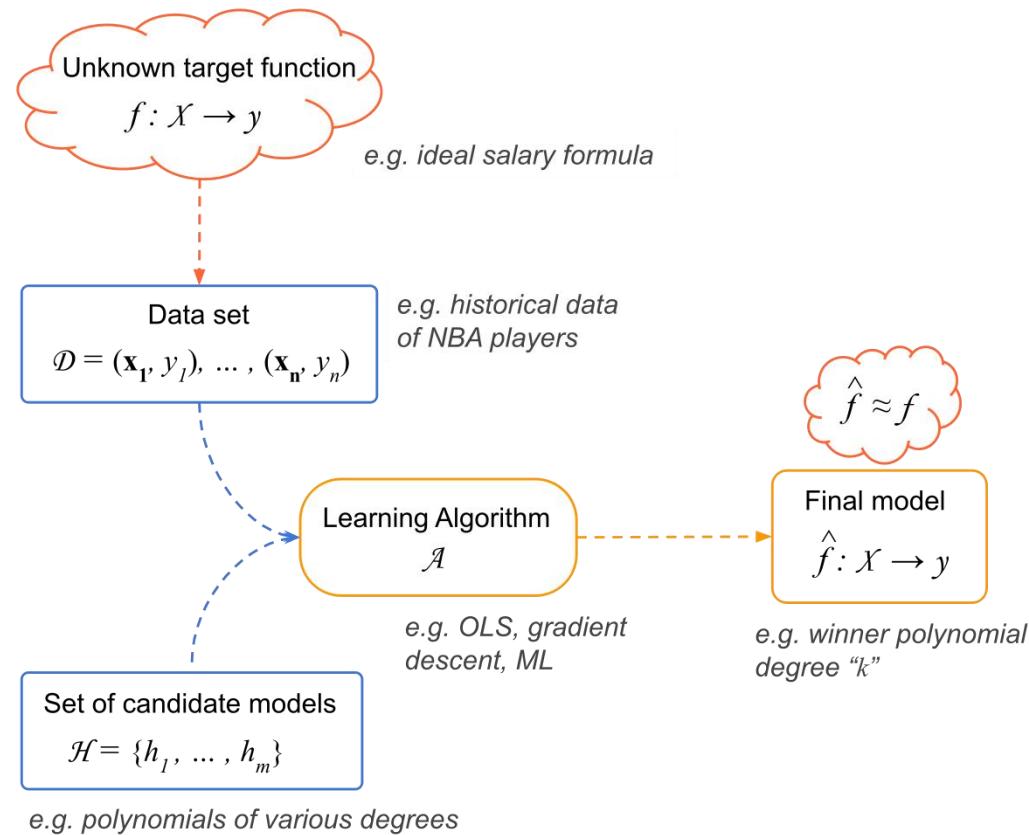
# Bộ khung lý thuyết



- Tập dữ liệu  $D$  chịu ảnh hưởng bởi hàm đích mà ta không có thông tin.
- Tập giả thiết  $\mathcal{H}$  là tập hợp các mô hình máy học mà ta muốn sử dụng (ví dụ: các mô hình tuyến tính, các mô hình đa thức, các mô hình phi tham số, v.v....).
- Thuật toán học  $\mathcal{A}$  bao gồm tập các câu lệnh được thực thi để thực hiện việc học từ dữ liệu.



# Bộ khung lý thuyết



- Mô hình kết quả  $\hat{f}$  được lựa chọn bởi thuật toán học từ tập hợp các mô hình giả thiết.
- Trường hợp lý tưởng:  $\hat{f}$  là một xấp xỉ tốt (good approximation) cho hàm đích (target function)  $f$ .



# Các loại dự đoán (types of predictions)

- *Thế nào là một mô hình tốt?*
- Ta muốn ước lượng một hàm đích  $f$  mà ta không có thông tin bằng một mô hình  $\hat{f}$  có thể thực hiện các dự đoán tốt.
- Một mô hình sau khi huấn luyện  $\hat{f}(x)$  có thể thực hiện 2 loại dự đoán:
  1. Cho các điểm dữ liệu ta đã thu thập được (**observed points**)  $x_i$ , ta có  $\hat{y}_i = \hat{f}(x_i)$ .  
**Lưu ý:**  $x_i$  nằm trong dữ liệu huấn luyện đã dùng để tìm ra  $\hat{f}$ .
  2. Cho các điểm dữ liệu trong tương lai (**unseen points**)  $x_0$ , ta có  $\hat{y}_0 = \hat{f}(x_0)$ .  
**Lưu ý:**  $x_0$  **không** nằm trong dữ liệu huấn luyện đã dùng để tìm ra  $\hat{f}$ .
- Ta có 2 loại tập dữ liệu (datasets):
  1. Dữ liệu trong mẫu (**in-sample data**): ký hiệu là  $D_{in}$ , sử dụng để huấn luyện mô hình.
  2. Dữ liệu ngoài mẫu (**out-of-sample data**): ký hiệu là  $D_{out}$ , sử dụng để đánh giá khả năng dự đoán của một mô hình.



# Các loại dự đoán (types of predictions)

- Với hai loại điểm dữ liệu, ta có 2 loại dự đoán tương ứng:
  1. Dự đoán  $\hat{y}_i$  cho các điểm dữ liệu đã thu thập  $x_i$
  2. Dự đoán  $\hat{y}_0$  cho các điểm dữ liệu mới  $x_0$
- Dự đoán trên các điểm dữ liệu đã thu thập (observed points)  $\hat{y}_i$  đánh giá khả năng ghi nhớ.
- Dự đoán trên các điểm dữ liệu tương lai (unobserved points)  $\hat{y}_0$  đánh giá khả năng **tổng quát hóa (generalization)**.
- Khả năng tổng quát hóa quan trọng hơn: Ta muốn tìm các mô hình có khả năng đưa ra dự đoán  $\hat{y}_0$  càng chính xác càng tốt so với giá trị đích thật sự  $y_0$ .
- Khả năng dự đoán tốt  $\hat{y}_i$  trên các điểm dữ liệu đã thu thập  $x_i$  là một điều kiện cần đối với một mô hình tốt, nhưng không phải là điều kiện đủ.
- Ta có thể huấn luyện một mô hình dự đoán hoàn hảo trên dữ liệu huấn luyện đã thu thập  $x_i$ , nhưng mô hình này có thể cho hiệu năng thấp với các điểm dữ liệu mới (unobserved data)  $x_0$ .



# Độ đo sai số (error measure)

- Ta cần có độ đo để đánh giá độ chính xác của các dự đoán.
- Ta cần có cơ chế định lượng sự khác biệt giữa dự đoán của mô hình sau khi huấn luyện  $\hat{f}()$  và hàm đích (target function)  $f()$ : **độ đo sai số tổng thể** (overall measure of error)  $E(\hat{f}, f)$ .
- Sai số tổng thể có thể được định nghĩa dựa trên sai số của từng điểm dữ liệu  $err_i(\hat{y}_i, y_i)$  định lượng sự khác biệt giữa giá trị đích quan sát được  $y_i$  và giá trị dự đoán  $\hat{y}_i = \hat{f}(x_i)$  của mô hình cho điểm dữ liệu thứ  $i$ .

$$E(\hat{f}, f) = \text{measure} \left( \sum_i err_i(\hat{y}_i, y_i) \right)$$

- Ta thường sử dụng sai số trung bình như một độ đo sai số tổng thể:

$$E(\hat{f}, f) = \frac{1}{n} \left( \sum_i err_i(\hat{y}_i, y_i) \right)$$



# Độ đo sai số (error measure)

- Sai số bình phương (squared error):  $err(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$
  - Sai số tuyệt đối (absolute error):  $err(\hat{y}_i, y_i) = |\hat{y}_i - y_i|$
  - Sai số phân lớp (misclassification error):  $err(\hat{y}_i, y_i) = \mathbf{1}[\hat{y}_i \neq y_i]$
  - ...
- Các độ đo sai số trên các điểm dữ liệu còn được gọi là hàm mất mát (loss function).



# Hai loại sai số

- Hai loại độ đo sai số tổng thể được định nghĩa trên loại dữ liệu được dùng để đánh giá sai số:

1. **Sai số trong mẫu (in-sample error)**: ký hiệu là  $E_{in}$ , là sai số trung bình của các điểm dữ liệu trong tập dữ liệu  $D_{in}$ :

$$E_{in}(\hat{f}, f) = \frac{1}{n} \sum_i err_i$$

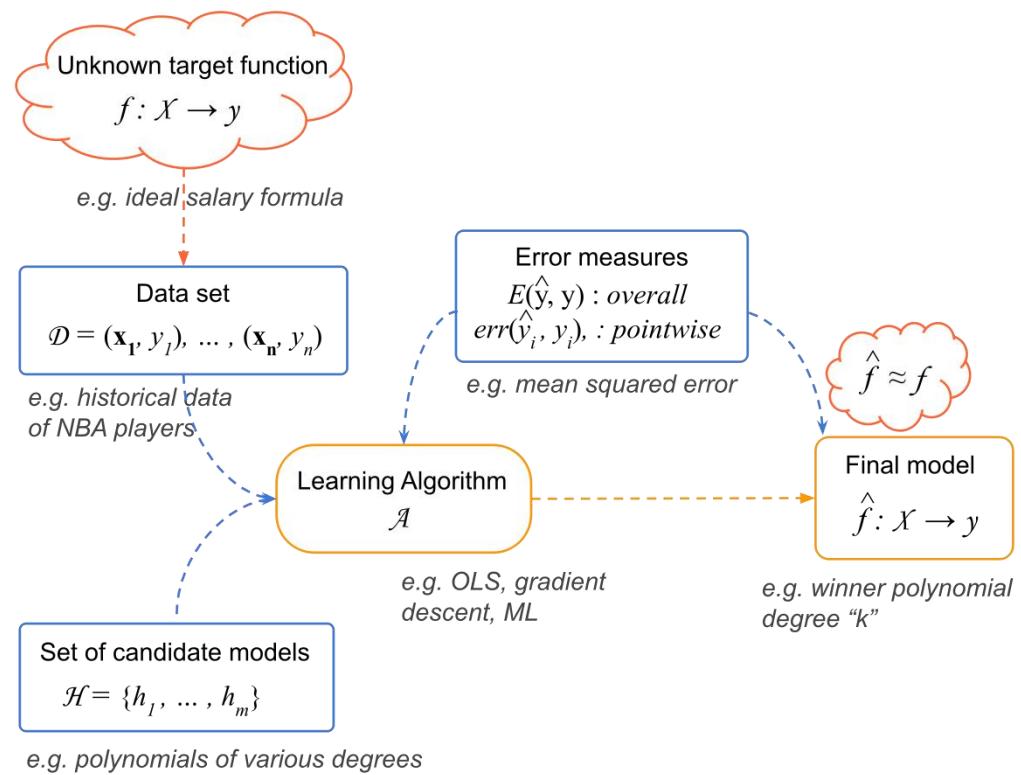
2. **Sai số ngoài mẫu (out-of-sample error)**: ký hiệu là  $E_{out}$ , là sai số trung bình lý thuyết, hoặc giá trị kỳ vọng của sai số, của các điểm dữ liệu xét trên toàn bộ không gian đầu vào (input space):

$$E_{out}(\hat{f}, f) = \mathbb{E}_{\mathcal{X}}[err(\hat{f}(x), f(x))]$$

- Điểm dữ liệu  $x$  đại diện cho một điểm dữ liệu tổng quát (a general data point) trong không gian đầu vào (input space)  $\mathcal{X}$ .
- Giá trị kỳ vọng được tính trên  $\mathcal{X}$ . Do đó, bản chất  $E_{out}$  mang tính lý thuyết (theoretical) vì ta không thể tính được cụ thể đại lượng này.



# Biểu đồ Học có giám sát (supervised learning)



- Thuật toán học  $A$  sử dụng độ đo sai số  $err()$ .
- Độ đo sai số tổng thể  $E()$  được dùng để xác định mô hình  $h()$  nào là xấp xỉ tốt nhất cho hàm đích (target function)  $f()$ .

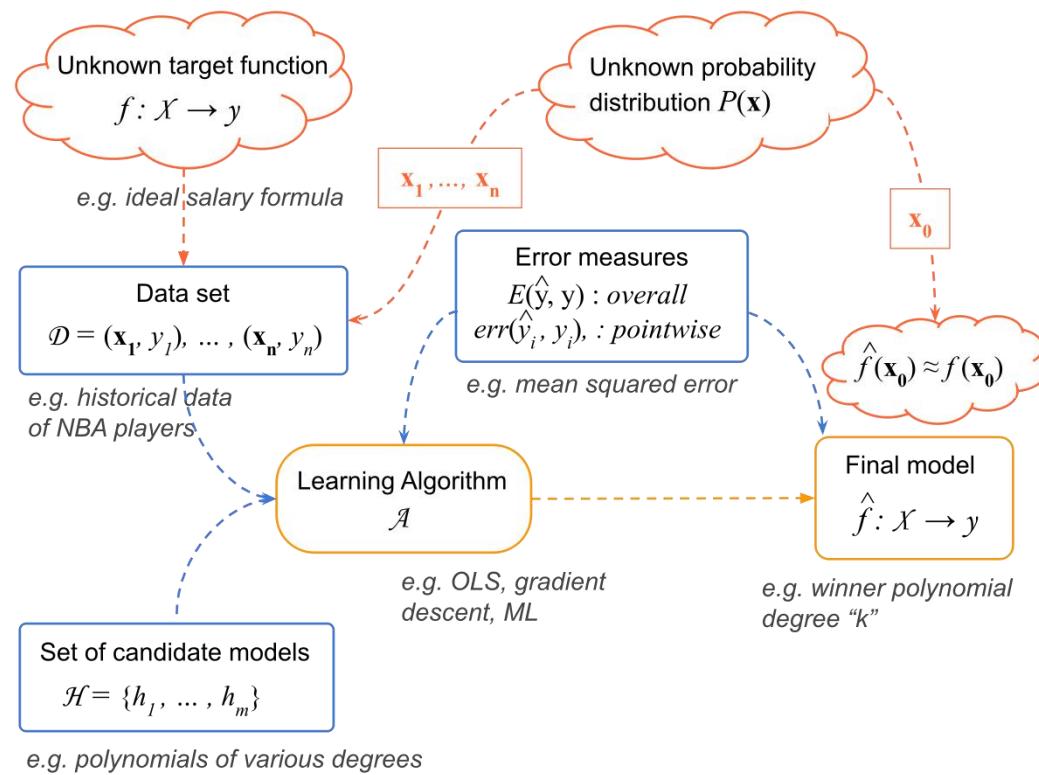


# Góc nhìn xác suất

- Ta muốn tìm ra một xấp xỉ tốt  $\hat{f} \approx f$ . Tức là ta muốn đạt được  $E_{out}(\hat{f}) \approx 0$ .
- Tuy nhiên, dữ liệu tổng quát (out-of-sample data) là khái niệm lý thuyết, và ta không thể thu thập được toàn bộ dữ liệu. Do đó, thực ra ta không thể tính được  $E_{out}$ .
- Ta chỉ có thể thu thập được một tập con của dữ liệu tổng quát (và ta gọi là **dữ liệu kiểm tra – test data**).
- Ta cần có một số giả định về phân phối xác suất  $P$  trên không gian đầu vào  $X$ . Các điểm dữ liệu (data points)  $x_1, x_2, \dots, x_n$  là độc lập với nhau, và có cùng phân phối (giả định iid), được lấy mẫu từ phân phối xác suất  $P$  này.
- Giả định này giúp liên kết sai số trong mẫu (in-sample error)  $E_{in}$  và sai số ngoài mẫu (out-of-sample error)  $E_{out}$ .



# Góc nhìn xác suất



$$E_{out}(\hat{f}) \approx 0 \Rightarrow \begin{cases} E_{in}(\hat{f}) \approx 0, & \text{kết quả thực tế} \\ E_{out}(\hat{f}) \approx E_{in}(\hat{f}), & \text{kết quả lý thuyết} \end{cases}$$



# Giá trị đích có nhiễu (noisy targets)

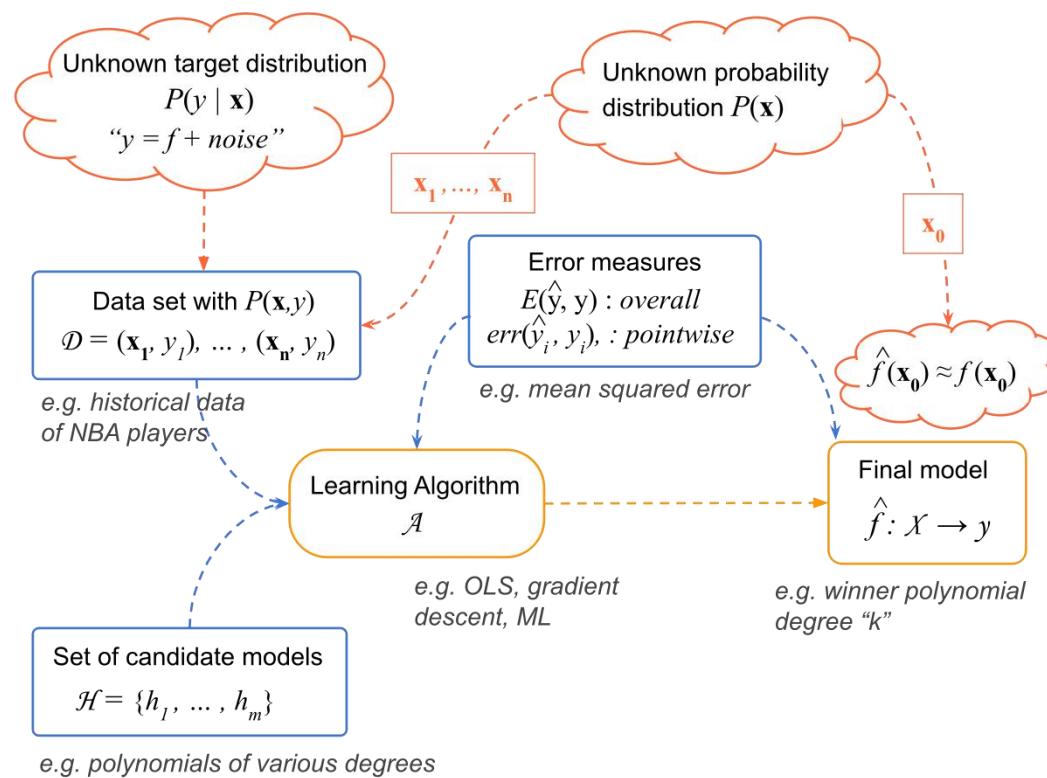
- Dữ liệu mà ta thu thập trong thực tế thường chứa một lượng **nhiễu (noise)**. Thay vì  $y = f(x)$  với  $f: \mathcal{X} \rightarrow \mathcal{Y}$  thì sẽ là:

$$y = f(x) + \varepsilon$$

- Ta có thể có nhiều vector đặc trưng khác nhau có cùng giá trị đích.
- Ta cũng có thể có 2 điểm dữ liệu với đặc trưng giống nhau  $x_A = x_B$  nhưng có giá trị đích khác nhau  $y_A \neq y_B$ .
- Ta sẽ cần mô hình hóa giá trị đích tuân theo một **phân phối xác suất có điều kiện (conditional distribution)**  $P(y | x)$ .
- Các điểm dữ liệu có thể được mô hình hóa với phân phối xác suất đồng thời  $P(x, y) = P(x)P(y|x)$ .



# Giá trị đích có nhiễu (noisy targets)



- Trong học có giám sát, ta muốn học được phân phối xác suất có điều kiện  $P(y|x)$  với  $y = f(x) + \varepsilon$ .
- Tập giả thiết  $\mathcal{H}$  và Thuật toán học  $\mathcal{A}$  được gọi là mô hình học (learning model).



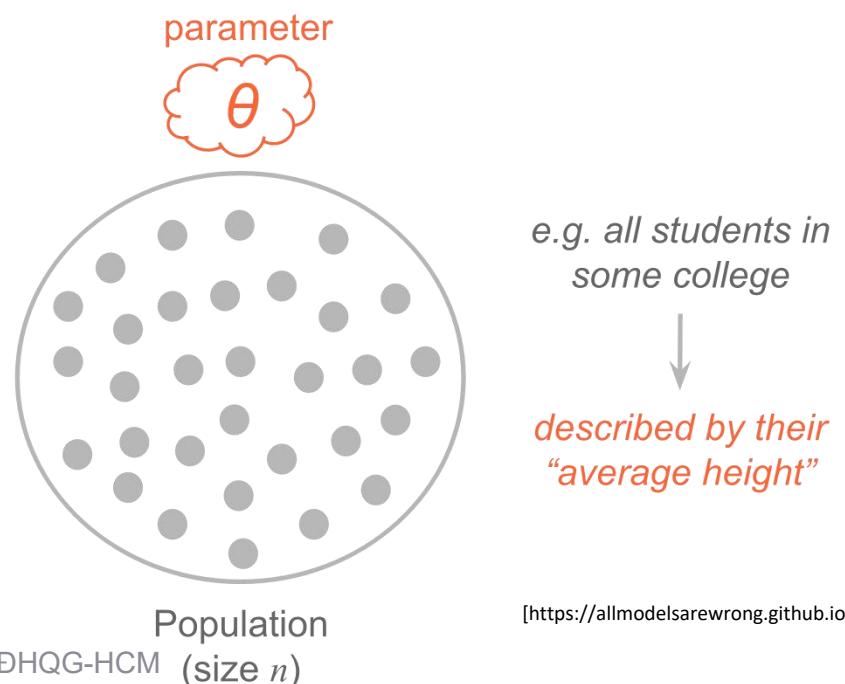
# **ƯỚC LƯỢNG VÀ PHÂN TÍCH LỖI**

**ESTIMATION AND ERROR DECOMPOSITION**



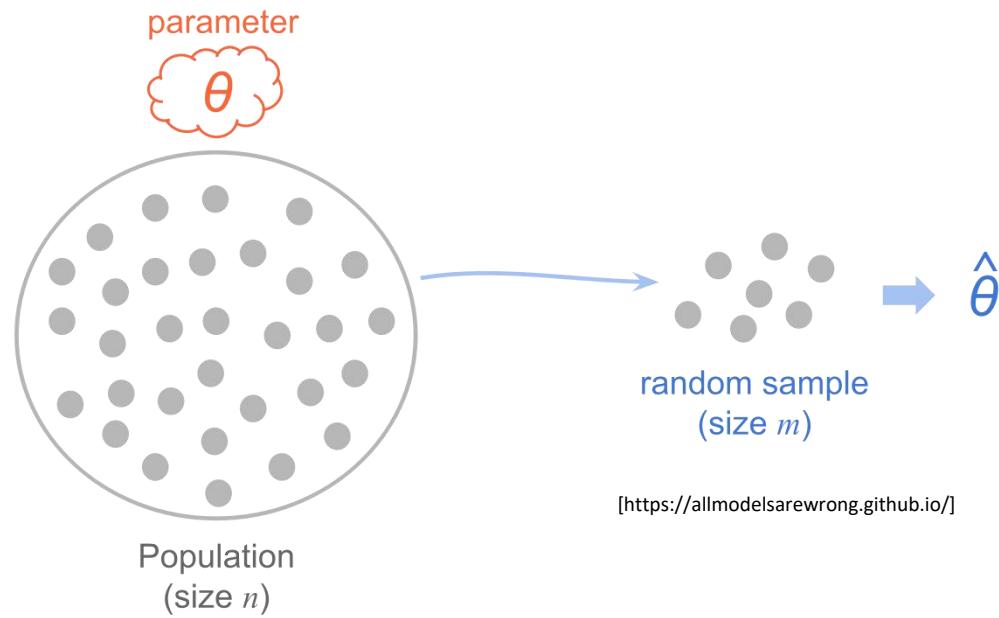
# Ước lượng (estimation)

- **Ước lượng** – tính toán giá trị xấp xỉ (approximate value) một tham số (parameter) nào đó của một tổng thể (population) sử dụng một mẫu ngẫu nhiên (random sample) các quan sát đến từ tổng thể đó.
- Ta có một tổng thể gồm  $n$  đối tượng, và ta muốn mô tả tổng thể này với một đại lượng  $\theta$ .
- Ví dụ, ta có một tổng thể gồm các sinh viên của một trường đại học, và ta muốn biết chiều cao trung bình. Giá trị trung bình (lý thuyết) này là một **tham số (parameter)**.





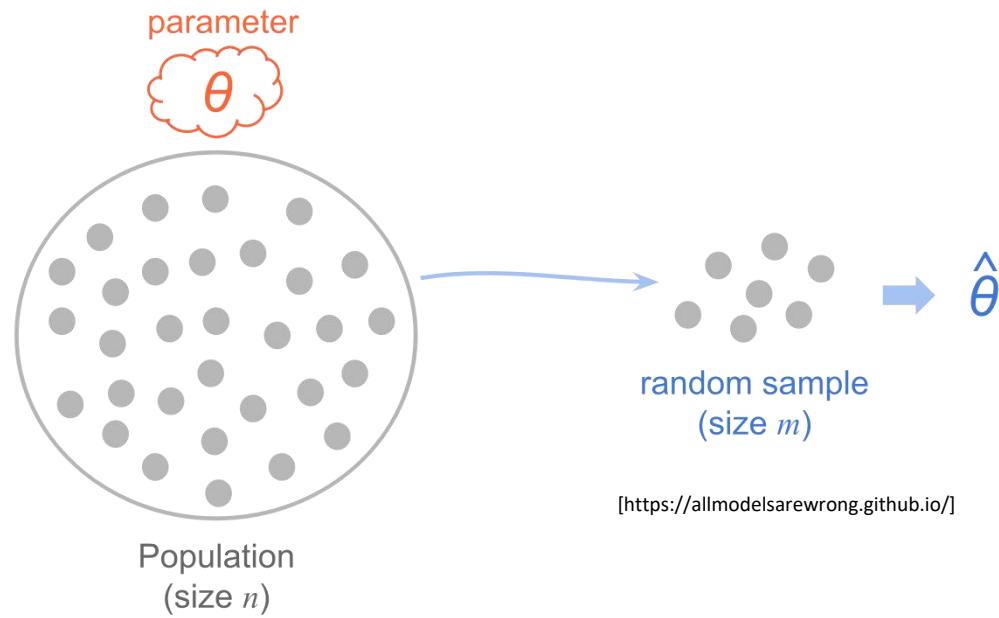
# Ước lượng (estimation)



- Để ước lượng giá trị của tham số, ta lấy mẫu ngẫu nhiên  $m < n$  sinh viên từ tổng thể, và thực hiện thống kê để tính ra giá trị xấp xỉ  $\hat{\theta}$ .
- Trong trường hợp lý tưởng, ta muốn có giá trị ước lượng  $\hat{\theta}$  xấp xỉ tốt giá trị thật sự của tham số  $\theta$ .



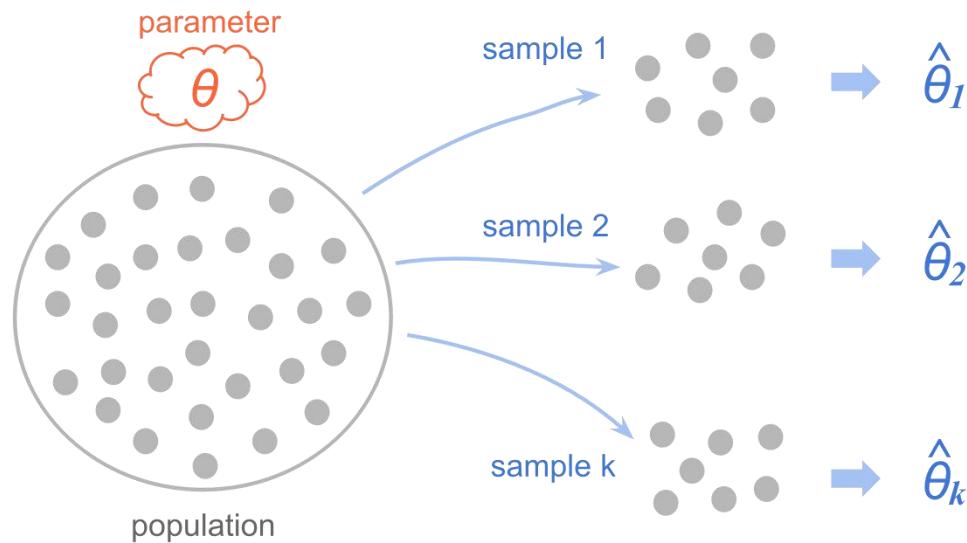
# Ước lượng (estimation)



1. Lấy mẫu ngẫu nhiên (random sample) từ tổng thể.
2. Sử dụng dữ liệu đã lấy mẫu và áp dụng các công thức cụ thể, tính ra giá trị xấp xỉ  $\hat{\theta}$  để ước lượng  $\theta$ .
3. Đánh giá độ tin cậy của ước lượng  $\hat{\theta}$ .



# Ước lượng (estimation)

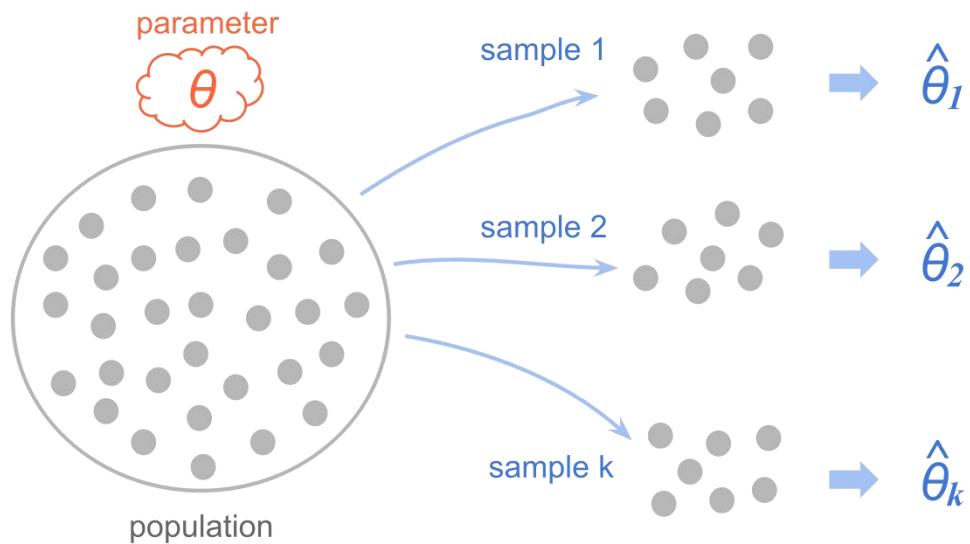


[<https://allmodelsarewrong.github.io/>]

- Giả sử ta có thể thực hiện lấy mẫu ngẫu nhiên nhiều lần từ tổng thể. Tất cả các mẫu có cùng kích thước là  $m$ .
- Với mỗi mẫu, ta tính ra một giá trị  $\hat{\theta}$ .



# Ước lượng (estimation)



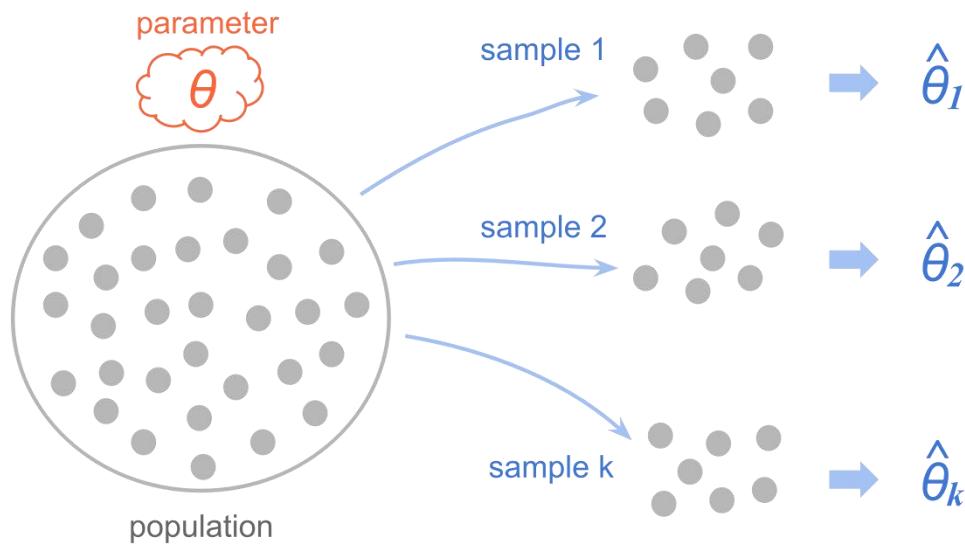
[<https://allmodelsarewrong.github.io/>]

Giá trị ước lượng là một biến ngẫu nhiên (random variable):

- Mẫu thứ 1 có kích thước  $m$  cho ta giá trị ước lượng  $\hat{\theta}_1$ .
- Mẫu thứ 2 có kích thước  $m$  cho ta giá trị ước lượng  $\hat{\theta}_2$ .
- Mẫu thứ 3 có kích thước  $m$  cho ta giá trị ước lượng  $\hat{\theta}_3$ .
-



# Ước lượng (estimation)



[<https://allmodelsarewrong.github.io/>]

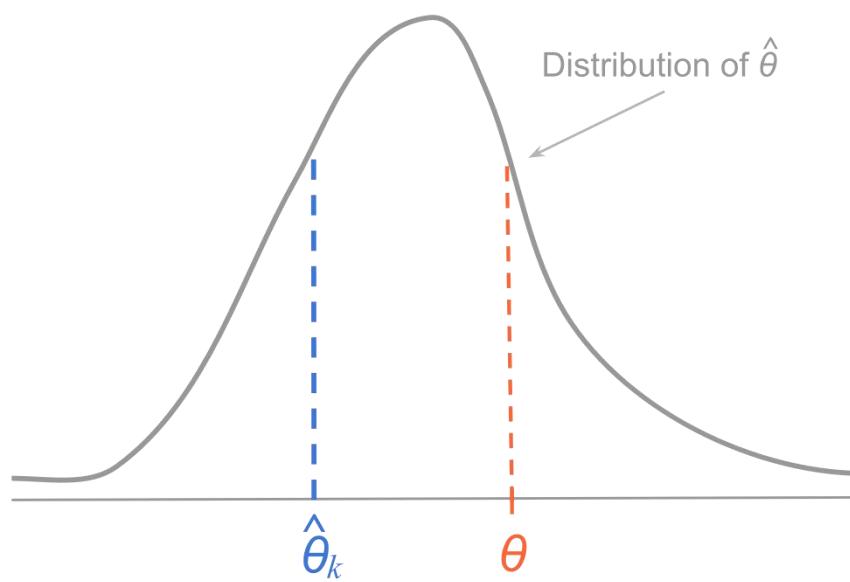
Giá trị ước lượng là một biến ngẫu nhiên (random variable):

- Một số mẫu cho ta giá trị ước lượng  $\hat{\theta}_k$  vượt quá giá trị thực sự  $\theta$ .
- Một số mẫu cho ta giá trị ước lượng  $\hat{\theta}_k$  thấp hơn giá trị thực sự  $\theta$ .
- Một số mẫu cho ta giá trị ước lượng  $\hat{\theta}_k$  khớp với giá trị thực sự  $\theta$ .



# Phân phối các giá trị ước lượng (estimators)

- Nếu ta lấy mẫu ngẫu nhiên rất nhiều lần, và thực hiện ước lượng trên các mẫu ngẫu nhiên này, ta có thể trực quan hóa phân phối các giá trị ước lượng  $\hat{\theta}$  như sau:



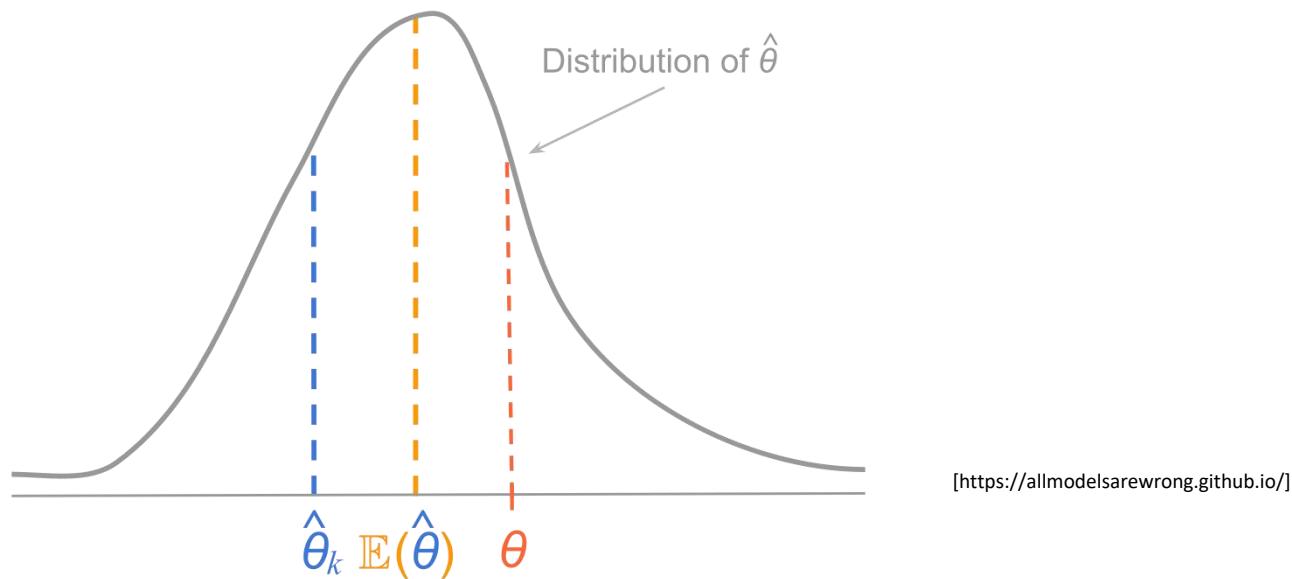
[<https://allmodelsarewrong.github.io/>]

- Một số giá trị ước lượng  $\hat{\theta}_k$  gần với giá trị thực sự  $\theta$ .
- Một số giá trị ước lượng  $\hat{\theta}_k$  cách xa giá trị thực sự  $\theta$ .



# Phân phối các giá trị ước lượng (estimators)

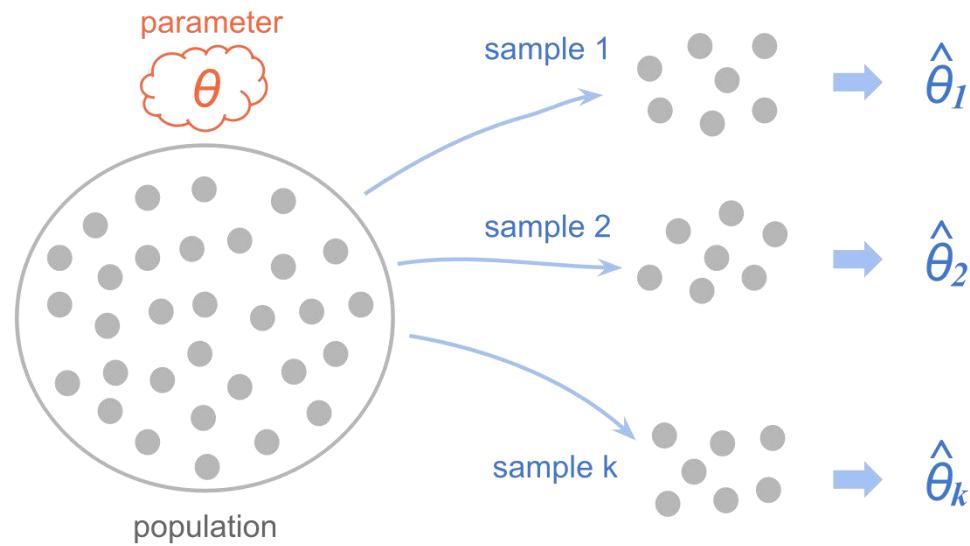
- Phân phối các giá trị ước lượng  $\hat{\theta}_k$  có **kỳ vọng (expected value)**  $\mathbb{E}(\hat{\theta})$  và **phương sai (variance)**  $var(\hat{\theta})$ .



- Các giá trị ước lượng  $\hat{\theta}$  có gần với giá trị thực sự  $\theta$  không? Tính trung bình, ta có thể kỳ vọng **giá trị ước lượng  $\hat{\theta}$**  chênh lệch với **tham số  $\theta$**  bao nhiêu?
- Ta cần một đô đo đánh giá khoảng cách giữa các giá trị ước lượng và tham số.



# Phân phối các giá trị ước lượng (estimators)



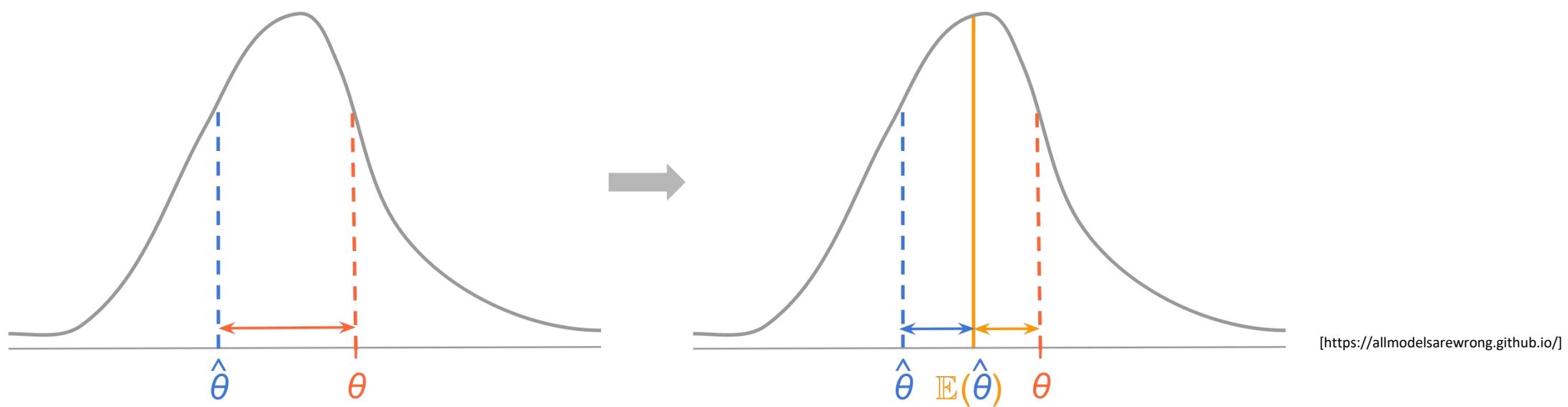
[<https://allmodelsarewrong.github.io/>]

Sự khác biệt giữa  $\hat{\theta} - \theta$  là **sai số ước lượng (estimation error)**. Sai số ước lượng cũng là một biến ngẫu nhiên.

- Mẫu thứ 1 có sai số  $\hat{\theta}_1 - \theta$ .
- Mẫu thứ 2 có sai số  $\hat{\theta}_2 - \theta$ .
- Thực hiện bởi Trường Đại học Công nghệ Thông tin, ĐHQG-HCM



# Phân phối các giá trị ước lượng (estimators)



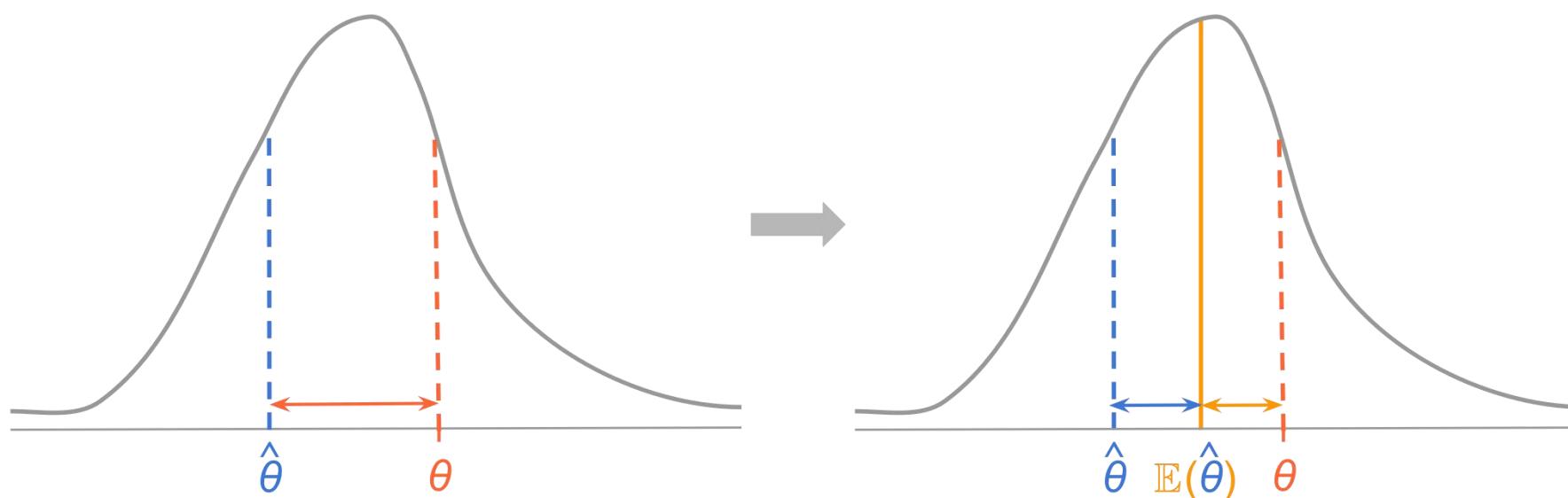
- Để đo kích thước của sai số ước lượng (estimation error), ta sử dụng hàm **sai số bình phương trung bình** (mean squared error – MSE).

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

- MSE là bình phương khoảng cách giữa giá trị ước lượng  $\hat{\theta}$  tới giá trị thực sự  $\theta$ , tính trung bình trên tất cả các mẫu ngẫu nhiên có thể (all possible samples).



# Phân phối các giá trị ước lượng (estimators)



$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2$$

$$= (\underbrace{\hat{\theta} - \mu_{\hat{\theta}}}_a + \underbrace{\mu_{\hat{\theta}} - \theta}_b)^2$$

$$= a^2 + b^2 + 2ab$$

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[a^2 + b^2 + 2ab]$$



# MSE của giá trị ước lượng

- $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$  có thể được phân rã như sau:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[a^2 + b^2 + 2ab] \\ &= \mathbb{E}(a^2) + \mathbb{E}(b^2) + 2\mathbb{E}(ab) \\ &= \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2] + \mathbb{E}[(\mu_{\hat{\theta}} - \theta)^2] + 2\mathbb{E}(ab) \end{aligned}$$

- Ta có:

$$\begin{aligned} \mathbb{E}(ab) &= \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})(\mu_{\hat{\theta}} - \theta)] = (\mu_{\hat{\theta}} - \theta)\mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})] \\ &= (\mu_{\hat{\theta}} - \theta)[\mathbb{E}(\hat{\theta}) - \mathbb{E}(\mu_{\hat{\theta}})] = 0 \end{aligned}$$

- Do đó:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2] + \mathbb{E}[(\mu_{\hat{\theta}} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2] + (\mu_{\hat{\theta}} - \theta)^2 \\ &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) \end{aligned}$$



# MSE của giá trị ước lượng

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2] + (\mu_{\hat{\theta}} - \theta)^2 \\ &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) \end{aligned}$$

MSE của ước lượng có thể được phân rã thành 2 thành phần: **độ chêch** (bias) và **phương sai** (variance).

- **Độ chêch (bias)**:  $\mu_{\hat{\theta}} - \theta$ , thể hiện giá trị ước lượng  $\hat{\theta}$  có khuynh hướng vượt quá (overestimate) hay thấp hơn (underestimate) giá trị thực sự của tham số  $\theta$ , xét trên tất cả các mẫu ngẫu nhiên có thể.
- **Phương sai (variance)**:  $\mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2]$ , thể hiện mức độ biến thiên trung bình (average variability) của các giá trị ước lượng xung quanh kỳ vọng của ước lượng  $\mu_{\hat{\theta}} = \mathbb{E}[\hat{\theta}]$ .



# Các trường hợp của độ chêch và phương sai

- Tùy thuộc vào phương pháp ước lượng  $\hat{\theta}$  và kích thước mẫu (sample size), ta có thể có những tình huống khác nhau liên quan đến sai số của ước lượng.





# PHÂN TÍCH ĐỘ CHỆCH – PHƯƠNG SAI

BIAS – VARIANCE DECOMPOSITION



# Học có giám sát (supervised learning)

- Tìm một mô hình  $\hat{f}$  xấp xỉ tốt hàm đích (target function)  $f$ .
- Ta muốn tìm một mô hình  $\hat{f}$  có khả năng dự đoán tốt trên cả 2 loại dữ liệu:
  1. Dữ liệu trong mẫu (**in-sample data**):  $\hat{y}_i = \hat{f}(x_i)$  với  $(x_i, y_i) \in D_{in}$
  2. Dữ liệu ngoài mẫu (**out-of-sample data**):  $\hat{y}_0 = \hat{f}(x_0)$  với  $(x_0, y_0) \in D_{out}$
- Hai loại dự đoán trên liên quan đến 2 loại sai số:
  1. Sai số trong mẫu (**in-sample error**):  $E_{in}(\hat{f})$
  2. Sai số ngoài mẫu (**out-of-sample error**):  $E_{out}(\hat{f})$
- Để đạt được  $\hat{f} \approx f$ , ta cần đạt được 2 mục tiêu:
  1. Sai số trong mẫu nhỏ:  $E_{in}(\hat{f}) \approx 0$
  2. Sai số ngoài mẫu tương tự như sai số trong mẫu:  $E_{out}(\hat{f}) \approx E_{in}(\hat{f})$
- Khảo sát sai số  $E_{out}(\hat{f})$  dưới góc độ bài toán hồi quy với **Sai số bình phương trung bình (mean squared error – MSE)**.

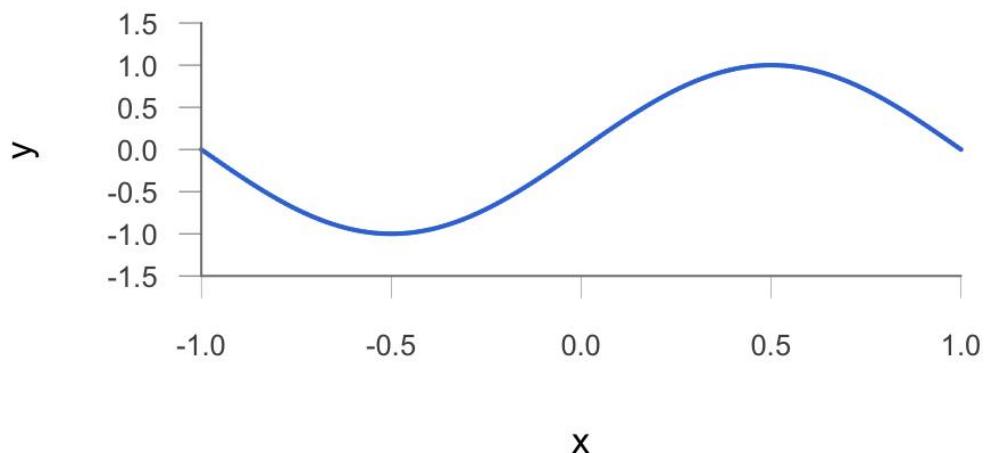


# Ví dụ

- Xét hàm đích không nhiễu (noiseless target function):

$$f(x) = \sin(\pi x)$$

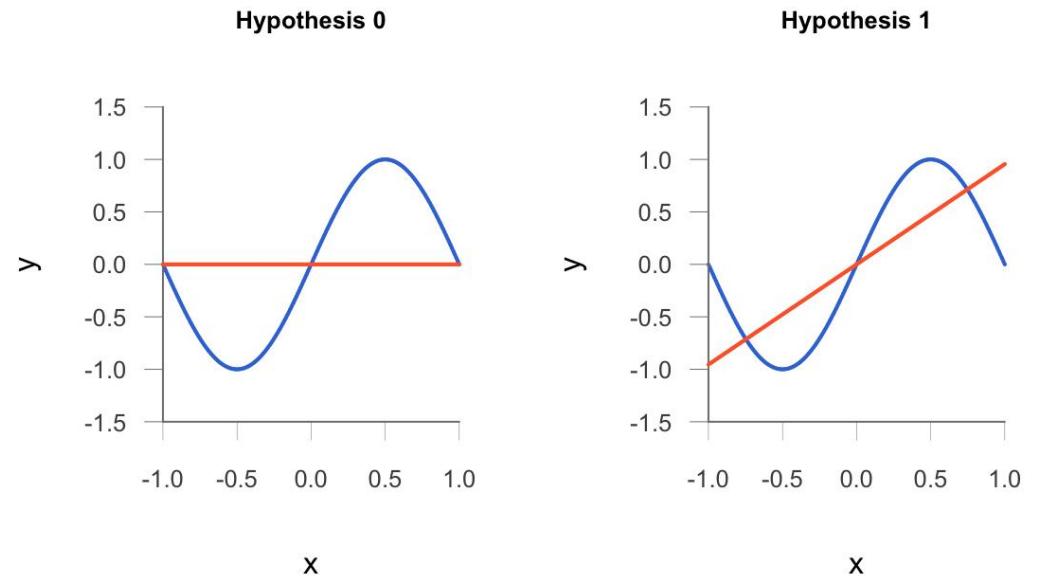
với đặc trưng đầu vào là biến  $x \in [-1,1]$ .





# Ví dụ - Hai giả thiết

- Cho một tập dữ liệu gồm  $n$  điểm dữ liệu, ta sẽ huấn luyện mô hình sử dụng 2 không gian giả thiết  $\mathcal{H}_0$  và  $\mathcal{H}_1$  như sau:
- $\mathcal{H}_0$ : tập hợp các đường thẳng có dạng  $h(x) = b$
- $\mathcal{H}_1$ : tập hợp các đường thẳng có dạng  $h(x) = b_0 + b_1x$





# Ví dụ - Học từ hai điểm dữ liệu

- Giả sử ta có một tập dữ liệu có kích thước  $n = 2$ ,  $D = \{(x_1, y_1), (x_2, y_2)\}$  với  $x_1, x_2 \in [-1, 1]$ .
- Với  $\mathcal{H}_0$ , ta cần chọn giả thiết khớp tốt nhất vào tập dữ liệu là đường thẳng nằm ngang đi qua điểm chính giữa:

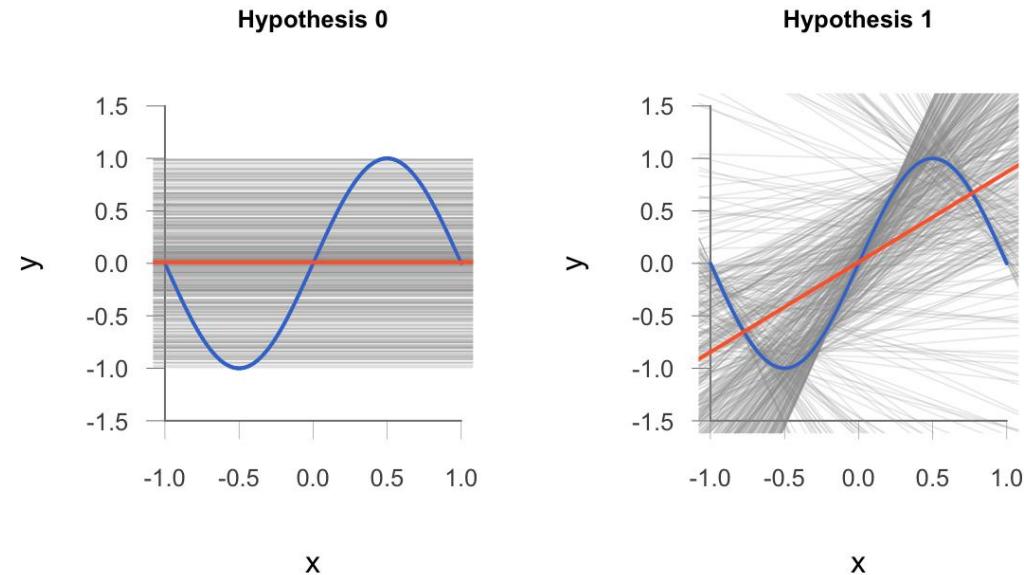
$$b = \frac{y_1 + y_2}{2}$$

- Với  $\mathcal{H}_1$ , ta cần chọn giả thiết khớp tốt nhất vào tập dữ liệu là đường thẳng đi qua hai điểm dữ liệu  $(x_1, y_1)$  và  $(x_2, y_2)$ .



# Ví dụ - Học từ hai điểm dữ liệu

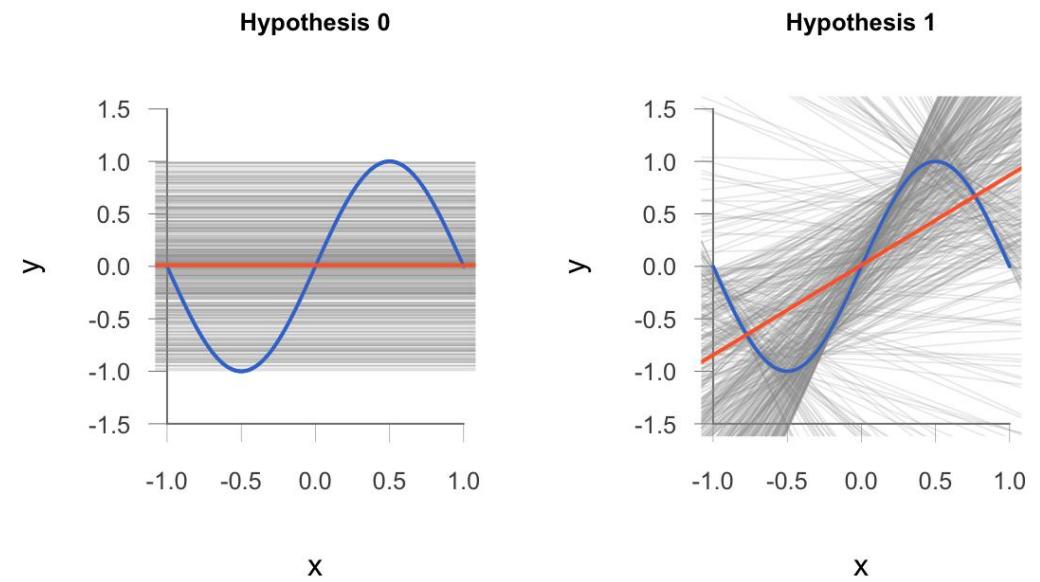
- Ta tiến hành thực hiện 500 thực nghiệm độc lập. Trong mỗi thực nghiệm ta lấy mẫu ngẫu nhiên 2 điểm dữ liệu trong  $[-1,1]$ , và huấn luyện 2 mô hình  $h_0$  và  $h_1$ .
- Ta có 2 giả thiết trung bình (average hypothesis)  $\bar{h}_0$  và  $\bar{h}_1$ .





# Ví dụ - Học từ hai điểm dữ liệu

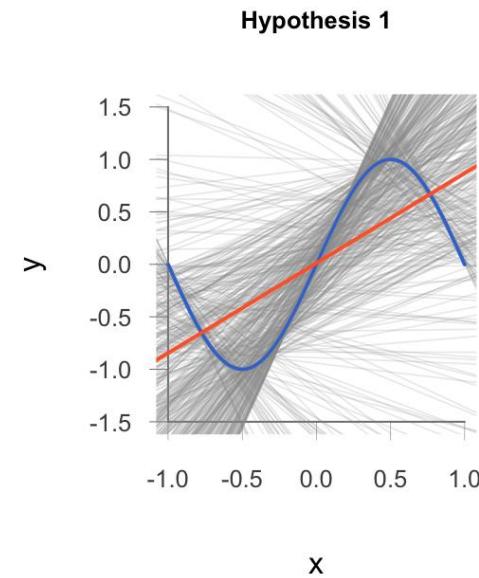
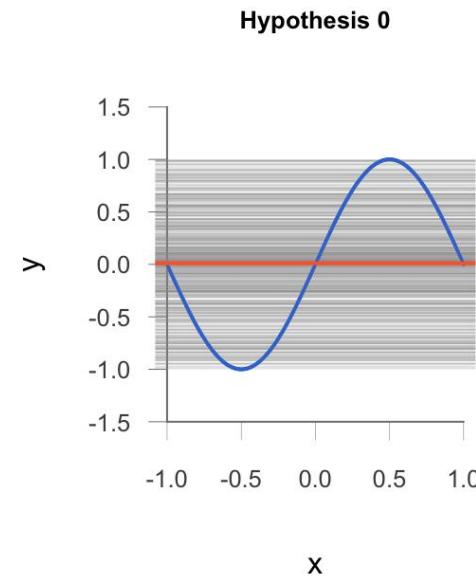
- Với các mô hình thuộc  $\mathcal{H}_0$ , nếu ta tính trung bình trên 500 mô hình sau khi huấn luyện (500 bộ giá trị tham số), ta thu được **giả thiết trung bình**  $\bar{h}_0$  là đường thẳng nằm ngang  $y = 0$ .
- Tất cả các mô hình trong  $\mathcal{H}_0$  sau khi huấn luyện đều là các đường thẳng có cùng độ dốc với giả thiết trung bình  $\bar{h}_0$  và chỉ khác giá trị cắt với trục tung (intercept).
- Lớp giả thiết  $\mathcal{H}_0$  có **phương sai thấp (low variance)** và **độ chêch cao (high bias)**.





# Ví dụ - Học từ hai điểm dữ liệu

- Với các mô hình thuộc  $\mathcal{H}_1$ , **giả thiết trung bình**  $\bar{h}_1$  có độ dốc dương chứng tỏ đa số các mô hình sau khi huấn luyện cũng có độ dốc dương.
- Giả thiết trung bình  $\bar{h}_1$  thể hiện được khuynh hướng chính của hàm đích (target function)  $f()$  trong khoảng giá trị của đặc trưng  $x \in [-0.5, 0.5]$ .
- Lớp giả thiết  $\mathcal{H}_1$  có **phương sai cao (high variance)** và **độ chêch thấp (low bias)**.





# Phân tích độ chêch – phương sai

- MSE của giá trị ước lượng  $\hat{\theta}$  có thể được phân rã thành độ chêch (bias) và phương sai (variance) như sau:

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2] + (\mu_{\hat{\theta}} - \theta)^2$$

với  $\mu_{\hat{\theta}} = \mathbb{E}[\hat{\theta}]$ .

- Độ chêch (bias):**  $\mu_{\hat{\theta}} - \theta$ , thể hiện giá trị ước lượng  $\hat{\theta}$  có khuynh hướng vượt quá (overestimate) hay thấp hơn (underestimate) giá trị thực sự của tham số  $\theta$ , xét trên tất cả các mẫu ngẫu nhiên có thể.
- Phương sai (variance):**  $\mathbb{E}[(\hat{\theta} - \mu_{\hat{\theta}})^2]$ , thể hiện mức độ biến thiên trung bình (average variability) của các giá trị ước lượng xung quanh kỳ vọng của ước lượng  $\mu_{\hat{\theta}} = \mathbb{E}[\hat{\theta}]$ .
- Ta có  $\hat{\theta}$  là ước lượng trong trường hợp tổng quát.
- Ta sẽ xét cụ thể trường hợp ta cần xây dựng mô hình  $\hat{f}()$  để ước lượng hàm đích  $f()$ .



# Phân tích độ chêch – phương sai

- Để phân tích MSE dưới góc độ một giá trị kỳ vọng lý thuyết (chứ không phải một giá trị trung bình thực nghiệm), ta xét một điểm dữ liệu tổng quát  $x_0$ .
- Với một tập dữ liệu huấn luyện  $D$  gồm  $n$  điểm dữ liệu, một không gian giả thiết  $\mathcal{H}$  gồm các mô hình ứng viên  $h(x)$ , kỳ vọng (expectation) của sai số bình phương cho điểm dữ liệu  $x_0$  xét trên **tất cả các trường hợp có thể xảy ra của tập huấn luyện  $D$**  là:

$$\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - f(x_0) \right)^2 \right]$$

- Giả sử hàm đích (target function)  $f()$  **không có nhiễu (noiseless)**.
- Ta có  $h^{(D)}$  khi huấn luyện mô hình trên một tập dữ liệu  $D$  cụ thể.
- $h^{(D)}(x_0)$  là giá trị dự đoán của mô hình sau khi huấn luyện trên một điểm dữ liệu mới  $x_0$  nằm ngoài tập dữ liệu huấn luyện  $D$ .
- $h^{(D)}(x_0)$  đóng vai trò như  $\hat{\theta}$  và  $f(x_0)$  đóng vai trò như  $\theta$ .



# Phân tích độ chêch – phương sai

- Ta xét giả thiết trung bình (average hypothesis)  $\bar{h}(x_0)$  đóng vai trò như  $\mu_{\hat{\theta}} = \mathbb{E}[\hat{\theta}]$ :

$$\bar{h}(x_0) = \mathbb{E}_D[h^{(D)}(x_0)]$$

- Sai số cho một điểm dữ liệu tổng quát (out-of-sample)  $x_0$  là:

$$h^{(D)}(x_0) - f(x_0) = h^{(D)}(x_0) - \bar{h}(x_0) + \bar{h}(x_0) - f(x_0)$$

$$h^{(D)} - f = h^{(D)} - \bar{h} + \bar{h} - f$$

- Tương tự, ta có thể phân tích giá trị kỳ vọng của bình phương sai số như sau:

$$\begin{aligned}\mathbb{E}_D[(h^{(D)} - f)^2] &= \mathbb{E}_D[(\underbrace{h^{(D)} - \bar{h}}_a + \underbrace{\bar{h} - f}_b)^2] \\ &= \mathbb{E}_D[(a + b)^2] = \mathbb{E}_D[a^2 + b^2 + 2ab] \\ &= \mathbb{E}_D[a^2] + \mathbb{E}_D[b^2] + \mathbb{E}_D[2ab]\end{aligned}$$



# Phân tích độ chêch – phương sai

- $\mathbb{E}_D[a^2] = \mathbb{E}_D[(h^{(D)} - \bar{h})^2] = Variance(h)$
- $\mathbb{E}_D[b^2] = \mathbb{E}_D[(\bar{h} - f)^2] = (\bar{h} - f)^2 = Bias^2(h)$
- Ta có:

$$\begin{aligned}\mathbb{E}_D[2ab] &= \mathbb{E}_D[2(h^{(D)} - \bar{h})(\bar{h} - f)] \\ &= 2\mathbb{E}_D[h^{(D)}\bar{h} - h^{(D)}f - \bar{h}^2 + \bar{h}f] \\ &\propto \bar{h}\mathbb{E}_D[h^{(D)}] - f\mathbb{E}_D[h^{(D)}] - \mathbb{E}_D[\bar{h}^2] + f\mathbb{E}_D[\bar{h}] = \bar{h}^2 - f\bar{h} - \bar{h}^2 + f\bar{h} = 0\end{aligned}$$

- Giả sử hàm đích (target function)  $f()$  là **không có nhiễu (noiseless)**, ta có kỳ vọng của bình phương sai số trên một điểm dữ liệu tổng quát (out-of-sample)  $x_0$ , tính trên **tất cả các trường hợp có thể xảy ra** của tập dữ liệu  $D$  là:

$$\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - f(x_0) \right)^2 \right] = \underbrace{\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - \bar{h}(x_0) \right)^2 \right]}_{\text{variance}} + \underbrace{\left( \bar{h}(x_0) - f(x_0) \right)^2}_{\text{bias}^2}$$



# Giá trị đích có nhiễu (noisy targets)

- Khi có **nhiễu (noise)** trong dữ liệu, thì giá trị đích mà ta thu thập được thực tế là:

$$y = f(x) + \varepsilon$$

- Giả sử  $\varepsilon$  là các nhiễu có kỳ vọng 0 và phương sai  $\sigma^2$ , phân tích sai số trở thành:

$$\mathbb{E}_D \left[ (h^{(D)}(x_0) - y_0)^2 \right] = \text{Variance} + \text{Bias}^2 + \sigma^2$$

- **Lưu ý:** Công thức trên ta chỉ xét Sai số bình phương (squared error) liên quan đến một điểm dữ liệu tổng quát (out-of-sample)  $(x_0, y_0)$  (chính là một **điểm dữ liệu kiểm tra bất kỳ - test point**).



# Các loại sai số bình phương trung bình (MSE)

- MSE liên quan đến **một điểm dữ liệu tổng quát bất kỳ** (out-of-sample)  $x_0$ : đánh giá hiệu năng của một lớp các giả thiết  $h \in \mathcal{H}$  trên tất cả các trường hợp có thể xảy ra của tập dữ liệu huấn luyện  $D$  - **kỳ vọng** của MSE trên dữ liệu kiểm tra (expected test MSE):

$$\mathbb{E}_D \left[ (h^{(D)}(x_0) - f(x_0))^2 \right]$$

- MSE liên quan đến **một giả thiết**  $h()$ : đánh giá hiệu năng của  $h$  tính trung bình trên tất cả các trường hợp có thể xảy ra của các điểm dữ liệu tổng quát (out-of-sample)  $x_0$ . Lưu ý: ta xét  $h()$  được huấn luyện trên một tập dữ liệu cụ thể  $D$ :

$$\mathbb{E}_x \left[ (h(x_0) - f(x_0))^2 \right]$$

- MSE đánh giá hiệu năng của một lớp các giả thiết  $h \in \mathcal{H}$  trên tất cả các trường hợp có thể xảy ra của tập dữ liệu huấn luyện  $D$ , trên tất cả các trường hợp có thể xảy ra của các điểm dữ liệu tổng quát (out-of-sample)  $x_0$  - **overall expected test MSE**:

$$\mathbb{E}_x \left[ \mathbb{E}_D \left[ (h^{(D)}(x_0) - f(x_0))^2 \right] \right]$$



# Các loại sai số bình phương trung bình (MSE)

- Các độ đo MSE này mang tính chất lý thuyết.
- Ta không biết chính xác hàm đích (target function)  $f$ .
- Ta không thể thu thập tất cả dữ liệu trong trường hợp tổng quát  $D_{out}$ .
- Ta không thể xét tất cả các trường hợp có thể xảy ra của tập dữ liệu huấn luyện  $D_{in}$  để tính được giả thiết trung bình  $\bar{h}$ .
- Ta chỉ có thể tính giá trị xấp xỉ (chính là một ước lượng) của MSE sử dụng một **tập dữ liệu kiểm tra (test dataset)**  $D_{test}$ .
- $D_{test}$  được xem như một tập con đại diện (chính là một mẫu không thiên lệch – unbiased sample) của dữ liệu ngoài mẫu (out-of-sample data)  $D_{out}$ .

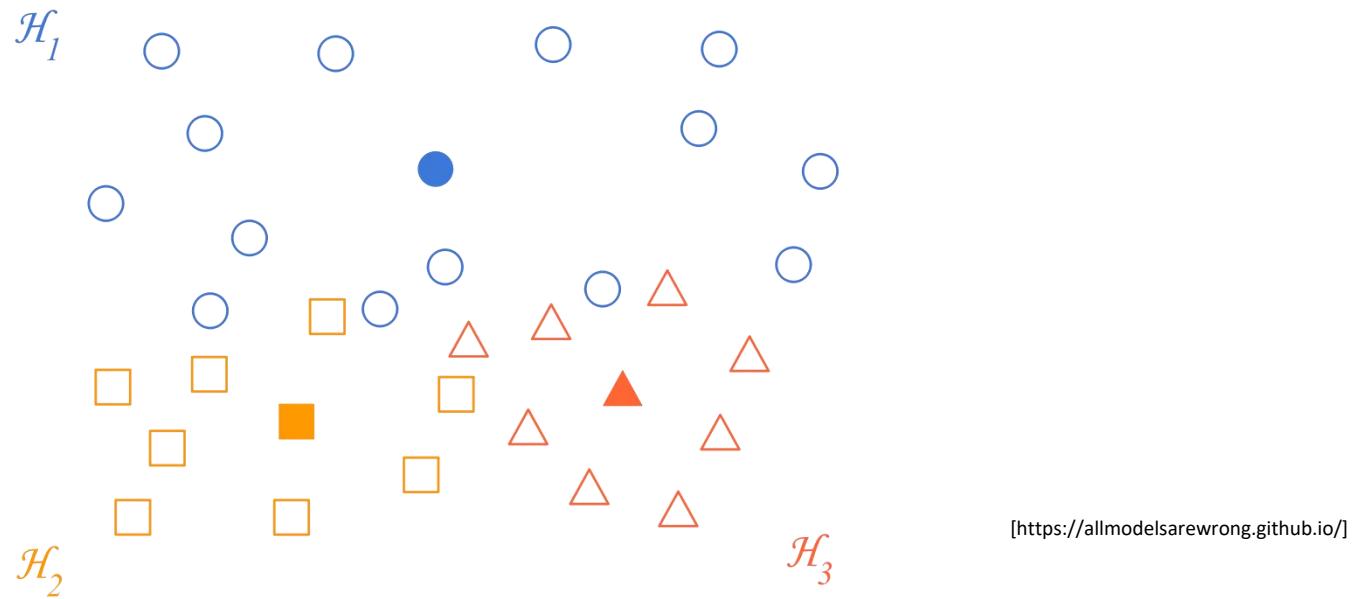


# ĐÁNH ĐỒI ĐỘ CHỆCH – PHƯƠNG SAI

## BIAS-VARIANCE TRADE-OFF



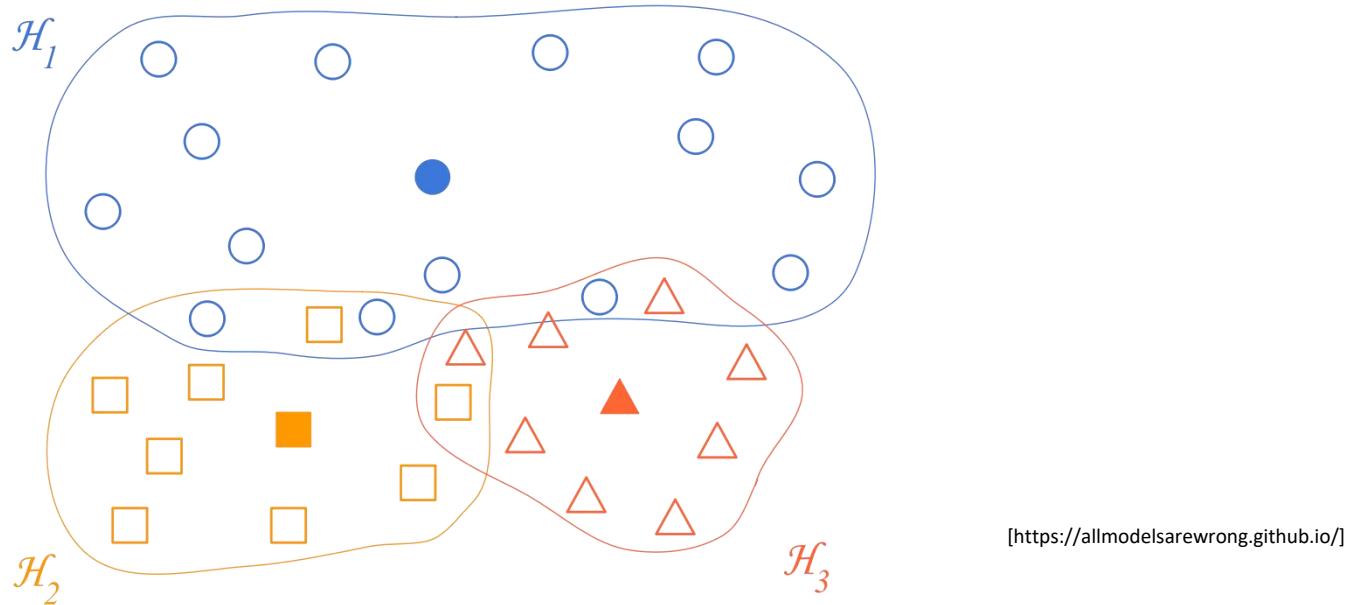
# Ví dụ: Đánh đổi độ chệch – phuong sai



- Giả sử ta có 3 lớp giả thiết sau:
  - $\mathcal{H}_1$ : các mô hình hồi quy đa thức bậc 3
  - $\mathcal{H}_2$ : các mô hình hồi quy đa thức bậc 2
  - $\mathcal{H}_3$ : các mô hình hồi quy tuyến tính (bậc 1)



# Ví dụ: Đánh đổi độ chệch – phuong sai

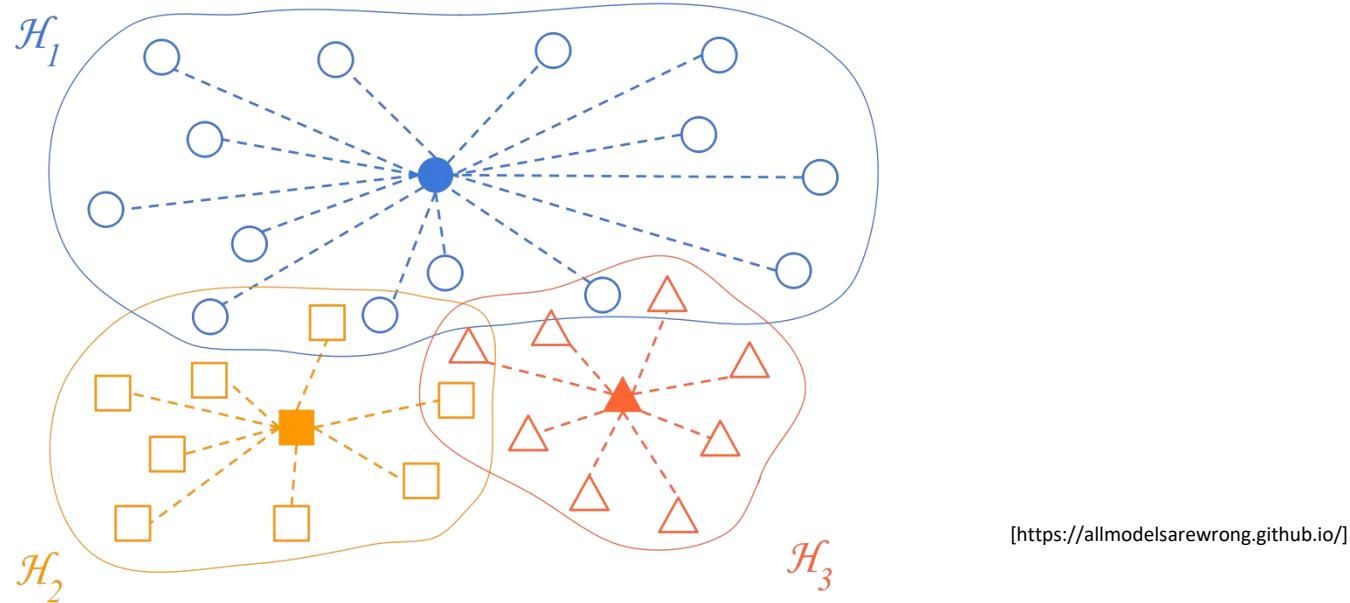


[<https://allmodelsarewrong.github.io/>]

- Mỗi điểm không tô màu là một mô hình sau khi huấn luyện  $h^{(D)}$  trên một tập dữ liệu  $D$ .
- Mỗi điểm có tô màu là một mô hình trung bình của mỗi lớp giả thiết  $\mathcal{H}$ .
- Ví dụ,  $\mathcal{H}_3$  đại diện cho tập hợp các mô hình hồi quy tuyến tính. Mỗi điểm hình tam giác là một đường thẳng  $ax + b$  với các tham số  $a, b$  được xác định trên một tập dữ liệu  $D$ .



# Ví dụ: Đánh đổi độ chệch – phương sai

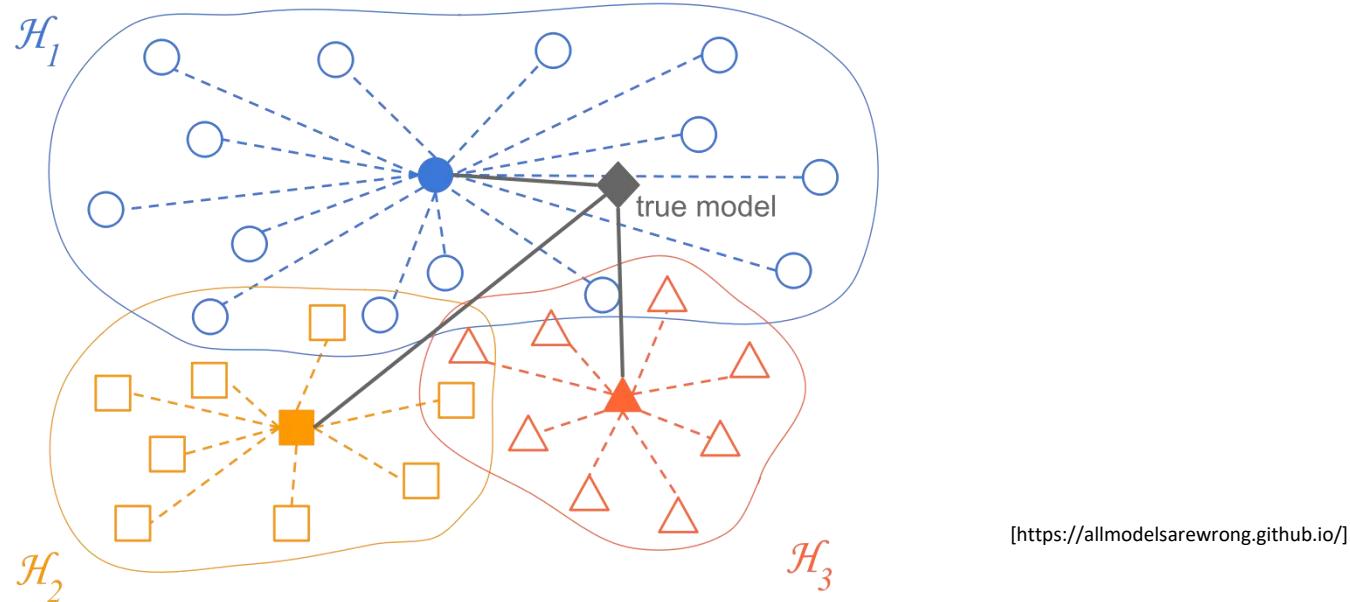


[<https://allmodelsarewrong.github.io/>]

- Ta đo **mức độ biến thiên (variability)** trong mỗi lớp giả thiết.
- Các đường nét đứt biểu diễn khoảng cách giữa từng mô hình sau khi huấn luyện trên một tập dữ liệu cụ thể  $D$  đến mô hình trung bình của lớp giả thiết.
- Tập hợp tất cả các đường nét đứt thể hiện **phương sai (variance)** trong mỗi lớp giả thiết, cho thấy mức độ phân tán của các mô hình trong từng lớp giả thiết.



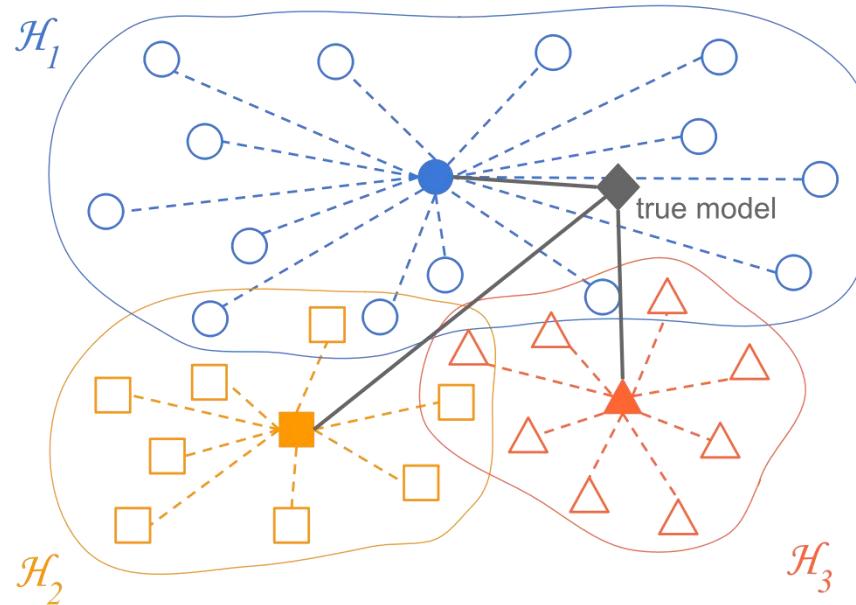
# Ví dụ: Đánh đổi độ chêch – phuong sai



- Giả sử ta có thể xác định được **hàm đích thực sự (target function)**  $f$  trong không gian này. Giả sử  $f()$  là một mô hình nằm trong lớp giả thiết  $\mathcal{H}_1$ .
- Các đường thẳng nét liền giữa các giả thiết trung bình của mỗi lớp giả thiết và hàm đích (target function) thể hiện **độ chêch (bias)** của mỗi lớp giả thiết.
- **Lưu ý:** trong thực tế, ta không thể xác định được giả thiết trung bình  $\bar{h}$  và hàm đích  $f$ .



# Ví dụ: Độ chêch (bias)



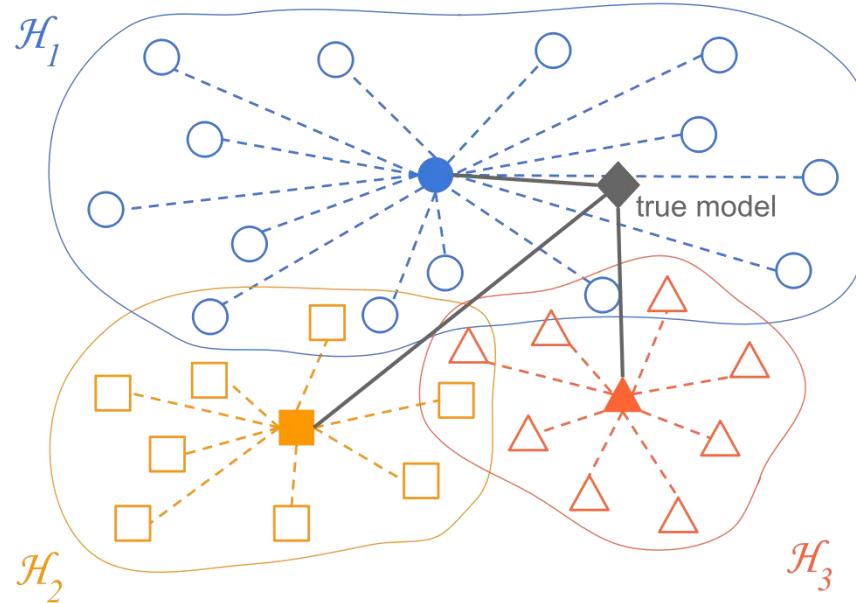
[<https://allmodelsarewrong.github.io/>]

- **Độ chêch** liên quan tới  $\bar{h}(x) - f(x)$ . Mỗi lớp giả thiết  $\mathcal{H}$  (các mô hình tuyến tính, các mô hình bậc 2, các mô hình bậc 3) có một mô hình trung bình (giả thiết trung bình)  $\bar{h}$ .
- $\bar{h}$  có thể xem như là **mô hình tiêu biểu** cho một lớp giả thiết  $\mathcal{H}$ .
- **Độ chêch (bias)** đánh giá khả năng một lớp giả thiết  $\mathcal{H}$  xấp xỉ hàm đích  $f$  như thế nào:

$$MSE = \text{Variance} + \text{Bias}^2 + \text{Noise}$$



# Ví dụ: Phương sai (variance)



[<https://allmodelsarewrong.github.io/>]

- **Phương sai (variance)**  $\mathbb{E}_D \left[ (h^{(D)}(x) - \bar{h}(x))^2 \right]$  đánh giá khoảng cách giữa một mô hình (giả thiết)  $h^{(D)}(x)$  và mô hình (giả thiết) trung bình  $\bar{h}$ .
- Phương sai (variance) thể hiện sự biến thiên của các mô hình  $h^{(D)}(x)$  trong một lớp giả thiết so với mô hình trung bình  $\bar{h}$  của lớp giả thiết.



# Đánh đổi độ chêch và phương sai

- Một mô hình tốt cần có **độ chêch thấp và phương sai thấp**. Giữa 2 thành phần lỗi này có một sự đánh đổi (trade-off).
- Để phân tích độ chêch – phương sai, ta cần có  $\bar{h}$ . Nhưng để tính được  $\bar{h}$ , ta cần phải xét tất cả các mô hình ứng viên sau khi huấn luyện trong một lớp giả thiết  $\mathcal{H}$ .
- Các mô hình phức tạp, có tính linh hoạt cao, thường có **độ chêch thấp (low bias)**, và do đó có tiềm năng xấp xỉ tốt hàm đích  $f(x)$ . Các mô hình phức tạp thường có sai số trong mẫu (in-sample error) thấp  $E_{in} \approx 0$ .
- Các mô hình phức tạp thường có **phương sai cao (high variance)**. Do đó, các mô hình này có rủi ro sai số ngoài mẫu (out-of-sample error) lớn  $E_{out} \gg 0$ . Ta cần nhiều tài nguyên để xử lý các mô hình phức tạp (nhiều dữ liệu huấn luyện, tài nguyên tính toán).
- Các mô hình đơn giản thường có **phương sai thấp (low variance)**, nhưng **độ chêch cao (high bias)**.  $E_{in} \approx E_{out}$  nhưng  $E_{in} \approx E_{out} \gg 0$ .



# Đánh đổi độ chêch và phương sai

- Để thực sự giảm độ chêch (bias), ta cần có thông tin về hàm đích (target function)  $f$ .
- Nhưng hàm đích (target function)  $f$  là gì thì ta không biết chính xác được. Do đó, gần như ta **không thể có độ chêch bằng 0** (zero bias).
- Để giảm **độ chêch (bias)**, ta thường cần phải sử dụng các mô hình phức tạp, có tính linh hoạt cao. Do đó, ta cần tìm cách giảm **phương sai (variance)**:
  - Tăng cường thêm dữ liệu huấn luyện (more training data).
  - Giảm chiều dữ liệu (ví dụ: kỹ thuật phân tích thành phần chính – PCA).
  - Áp dụng các kỹ thuật điều chỉnh (regularizations) để giảm độ lớn các tham số của mô hình.
  - ...



# Đánh đổi độ chêch và phương sai – Tóm tắt

- Dữ liệu mà ta thu thập được thường có nhiễu  $y = f(x) + \varepsilon$ .
- Ta cần tìm một mô hình  $h(x)$  xấp xỉ tốt hàm mục tiêu (target function)  $f$ .
- Với mỗi tập dữ liệu huấn luyện  $D$  gồm  $n$  điểm dữ liệu, và một mô hình (giả thiết)  $h(x)$ ,  
**kỳ vọng của bình phương sai số** của mô hình trên một điểm dữ liệu tổng quát (out-of-sample)  $x_0$ , xét trên tất cả các tập dữ liệu có thể có, là:

$$\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - f(x_0) \right)^2 \right] = \underbrace{\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - \bar{h}(x_0) \right)^2 \right]}_{\text{variance}} + \underbrace{\left( \bar{h}(x_0) - f(x_0) \right)^2}_{\text{bias}^2} + \underbrace{\sigma^2}_{\text{noise}}$$

với  $\bar{h}(x_0) = \mathbb{E}_D[h^{(D)}(x_0)]$  là mô hình trung bình (average hypothesis).



# Đánh đổi độ chêch và phương sai – Tóm tắt

- **Độ chêch (bias) lớn:** Lớp giả thiết  $\mathcal{H}$  có các mô hình đơn giản, khả năng hạn chế trong việc xấp xỉ hàm đích (target function).
- **Phương sai (variance) lớn:** Mô hình có độ phức tạp quá cao, có khả năng tập trung quá nhiều vào các chi tiết trong tập dữ liệu, và có rủi ro học thuộc lòng các giá trị nhiều hơn là các quy luật quan trọng, dẫn đến tình trạng quá khớp.
- **Nhiễu (noise) lớn:** Dữ liệu dùng để huấn luyện tạo ra  $h^{(D)}$  chất lượng không tốt, chứa nhiều thông tin sai, bị thiếu dữ liệu.

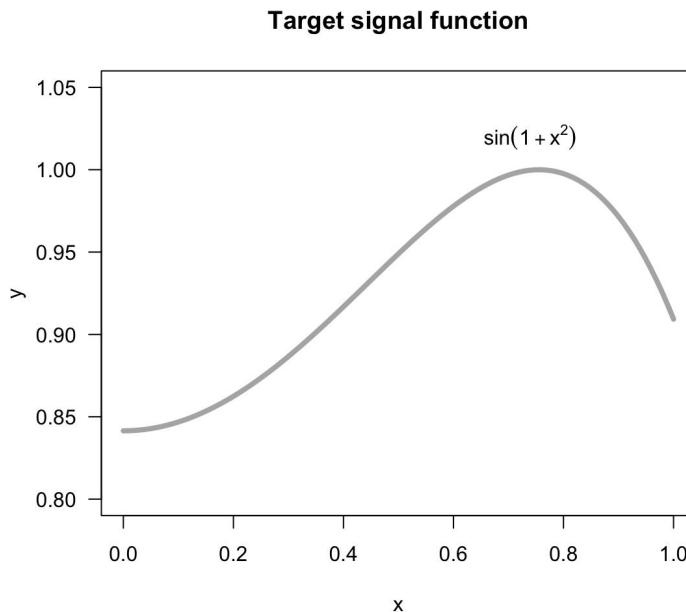


# Ví dụ

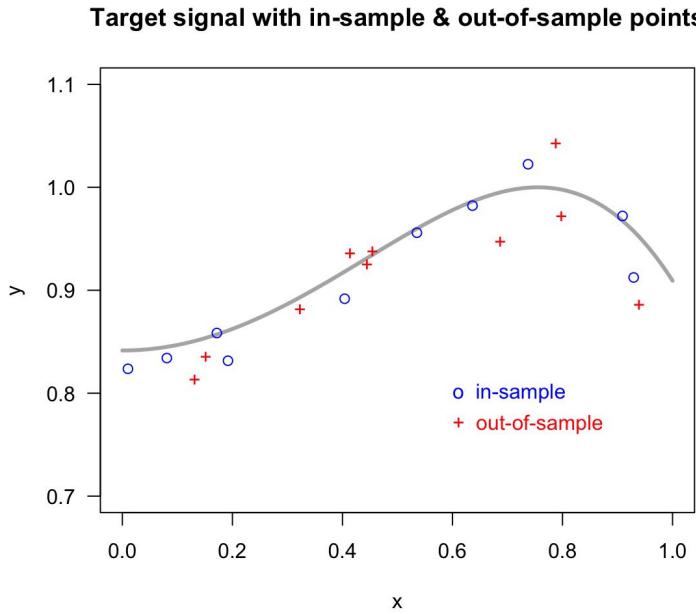
- Xét một hàm đích (target function) với nhiễu như sau:

$$y = f(x) + \varepsilon = \sin(1 + x^2) + \varepsilon$$

với đặc trưng đầu vào  $x \in [0,1]$  và thành phần nhiễu  $\varepsilon \sim N(\mu = 0, \sigma = 0.03)$ .



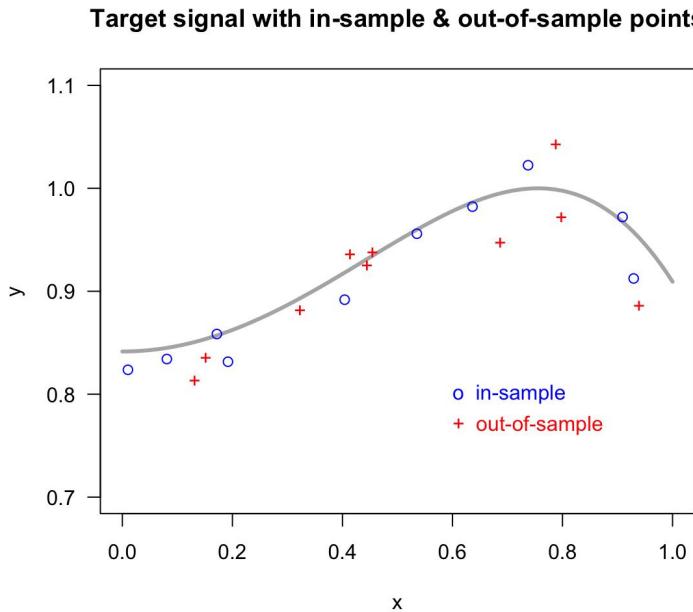
# Ví dụ



- Ta tạo một tập dữ liệu huấn luyện (in-sample dataset) gồm 10 điểm dữ liệu  $(x_i, y_i)$  và một tập dữ liệu kiểm tra (out-of-sample dataset) gồm 10 điểm dữ liệu  $(x_0, y_0)$ .
- **Lưu ý:** một tập dữ liệu ngoài mẫu (out-of-sample) thực sự sẽ chứa tất cả các giá trị  $x \in [0,1]$ .



# Ví dụ

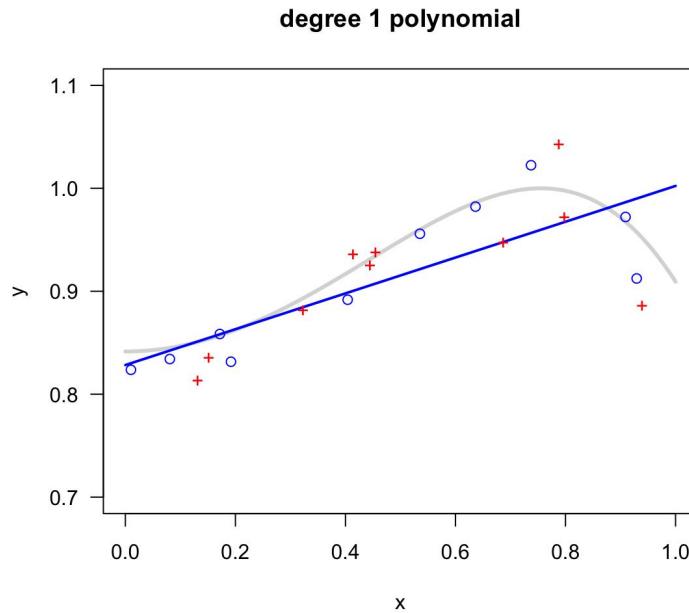


Sử dụng tập dữ liệu huấn luyện (in-sample dataset) gồm 10 điểm dữ liệu để:

- Huấn luyện một mô hình tuyến tính (đa thức bậc 1).
- Huấn luyện một mô hình hồi quy đa thức bậc 2.
- ...
- Với 10 điểm dữ liệu, ta có thể khớp hoàn hảo với một mô hình hồi quy đa thức bậc 9.



# Ví dụ

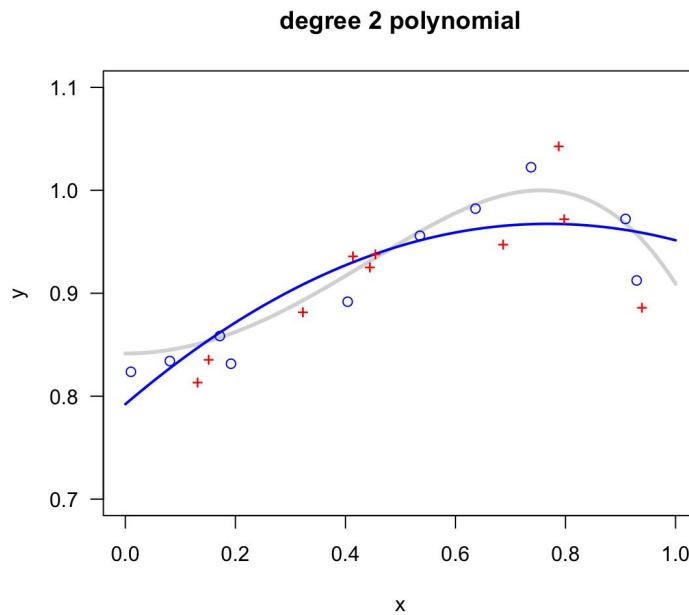


- Xét mô hình hồi quy tuyến tính đơn giản có dạng:

$$h_1(x) = b_0 + b_1x$$

- Mô hình hồi quy tuyến tính sau khi huấn luyện là **đường thẳng màu xanh** như hình vẽ.
- $E_{in} = 0.00147$  và  $E_{out} = 0.00215$ .

# Ví dụ

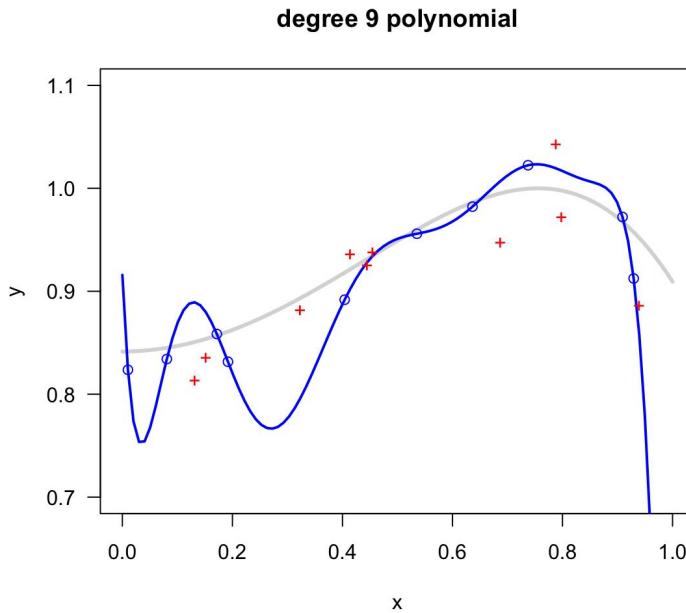


- Xét mô hình hồi quy đa thức bậc 2 có dạng:

$$h_2(x) = b_0 + b_1x + b_2x^2$$

- Mô hình hồi quy đa thức bậc 2 sau khi huấn luyện là **đường cong màu xanh**.
- $E_{in} = 0.00093$  và  $E_{out} = 0.00137$ .

# Ví dụ



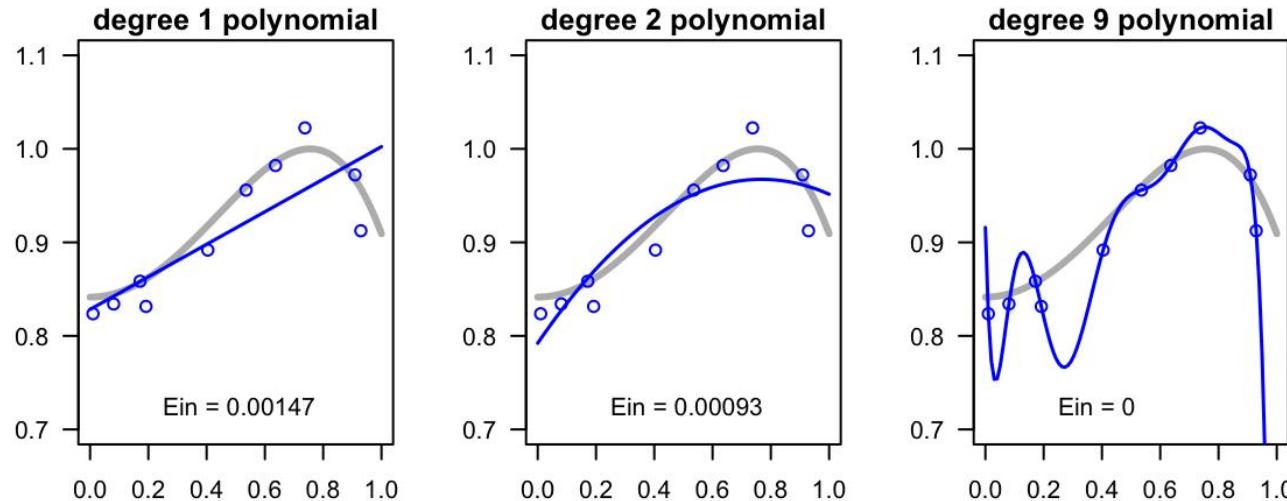
- Xét mô hình hồi quy đa thức bậc 9 có dạng:

$$h_9(x) = b_0 + b_1x + b_2x^2 + \dots + b_8x^8 + b_9x^9$$

- Mô hình hồi quy đa thức bậc 9 sau khi huấn luyện là **đường cong màu xanh**.
- $E_{in} = 0.00000$  và  $E_{out} = 0.00231$ .



# Ví dụ - Chọn mô hình nào?

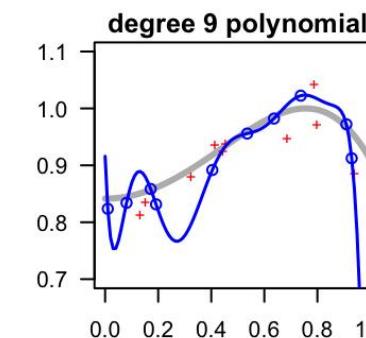
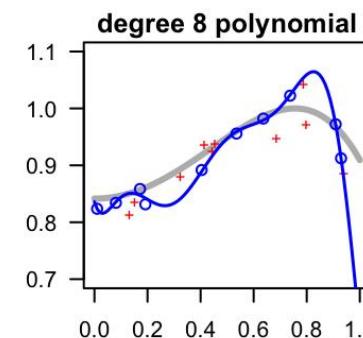
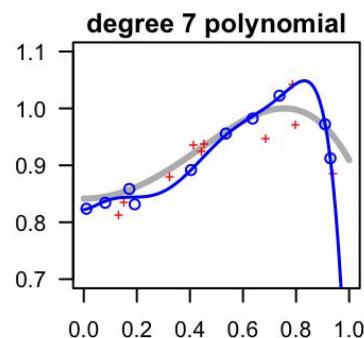
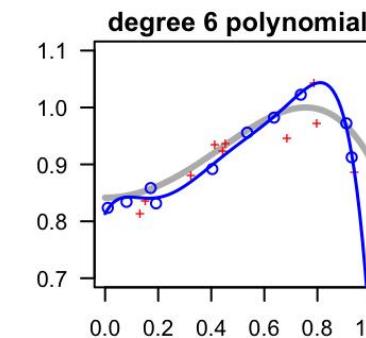
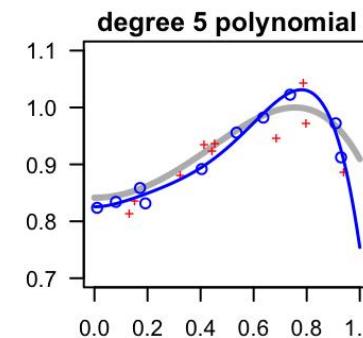
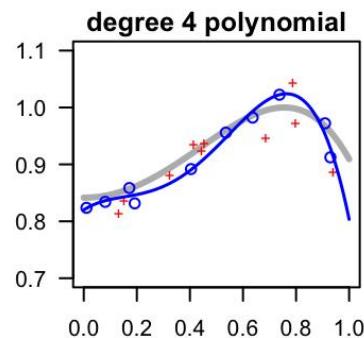
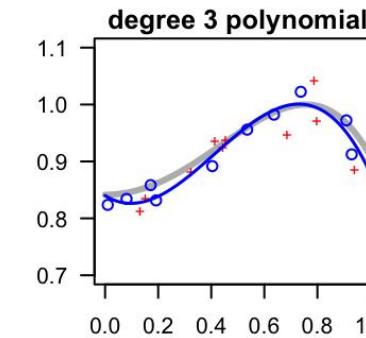
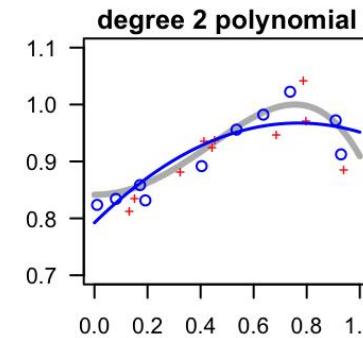
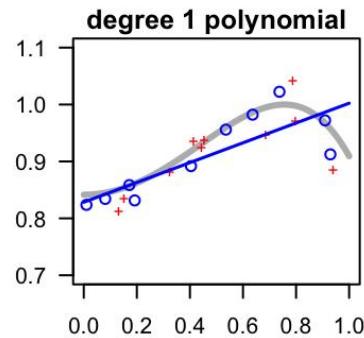


[<https://allmodelsarewrong.github.io/>]

- Mô hình hồi quy đa thức bậc 9 đạt được  $E_{in} = 0$ .
- Ta có nên chọn mô hình  $h_9(x)$  là mô hình cuối cùng  $\hat{f}$  xấp xỉ hàm đích  $f$  hay không?  
Đây là mô hình khớp tốt nhất vào tập dữ liệu huấn luyện.
- Không!** Ta cần chọn lựa mô hình dựa trên sai số ngoài mẫu (out-of-sample error)  $E_{out}$ .



# Ví dụ - Chọn mô hình nào?



[<https://allmodelsarewrong.github.io/>]



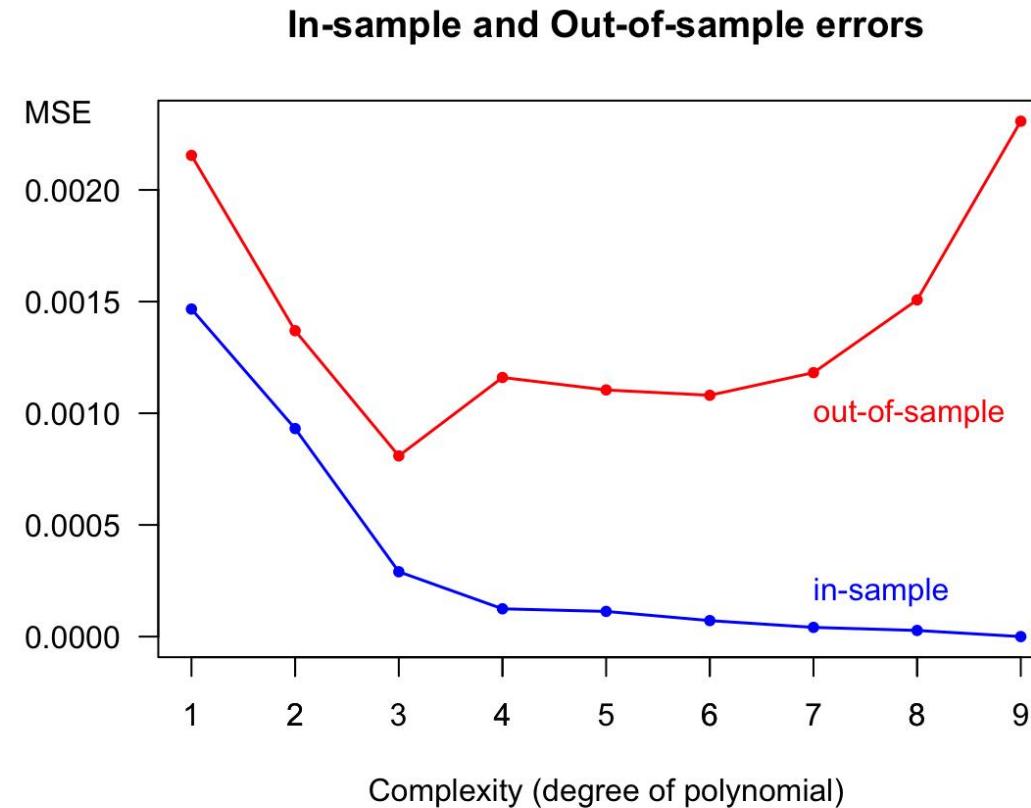
# Ví dụ - Chọn mô hình nào?

Degree	$E_{in}$	$E_{out}$
1	0.00147	0.00215
2	0.00093	0.00137
3	0.00029	0.00081
4	0.00012	0.00116
5	0.00011	0.00110
6	0.00007	0.00108
7	0.00004	0.00118
8	0.00003	0.00151
9	0.00000	0.00231

- Mô hình hồi quy tuyến tính (bậc 1) quá đơn giản, dẫn đến tình trạng **underfit** trên tập dữ liệu. Trong số các mô hình, độ lỗi  $E_{in}$  của mô hình tuyến tính là lớn nhất. Các mô hình phức tạp hơn sẽ giảm được  $E_{in}$  và  $E_{out}$ .
- Các mô hình hồi quy đa thức bậc 3, 4, 5 khớp tốt vào dữ liệu huấn luyện, và cũng cho kết quả tốt trên dữ liệu kiểm tra → “**okayfit**.”
- Mô hình hồi quy đa thức bậc 9 quá phức tạp và có độ linh hoạt cao. Độ lỗi trên dữ liệu huấn luyện rất thấp  $E_{in} = 0.0$  nhưng độ lỗi trên dữ liệu kiểm tra cũng rất lớn. Mô hình này bị **overfit**.



# Ví dụ - Chọn mô hình nào?

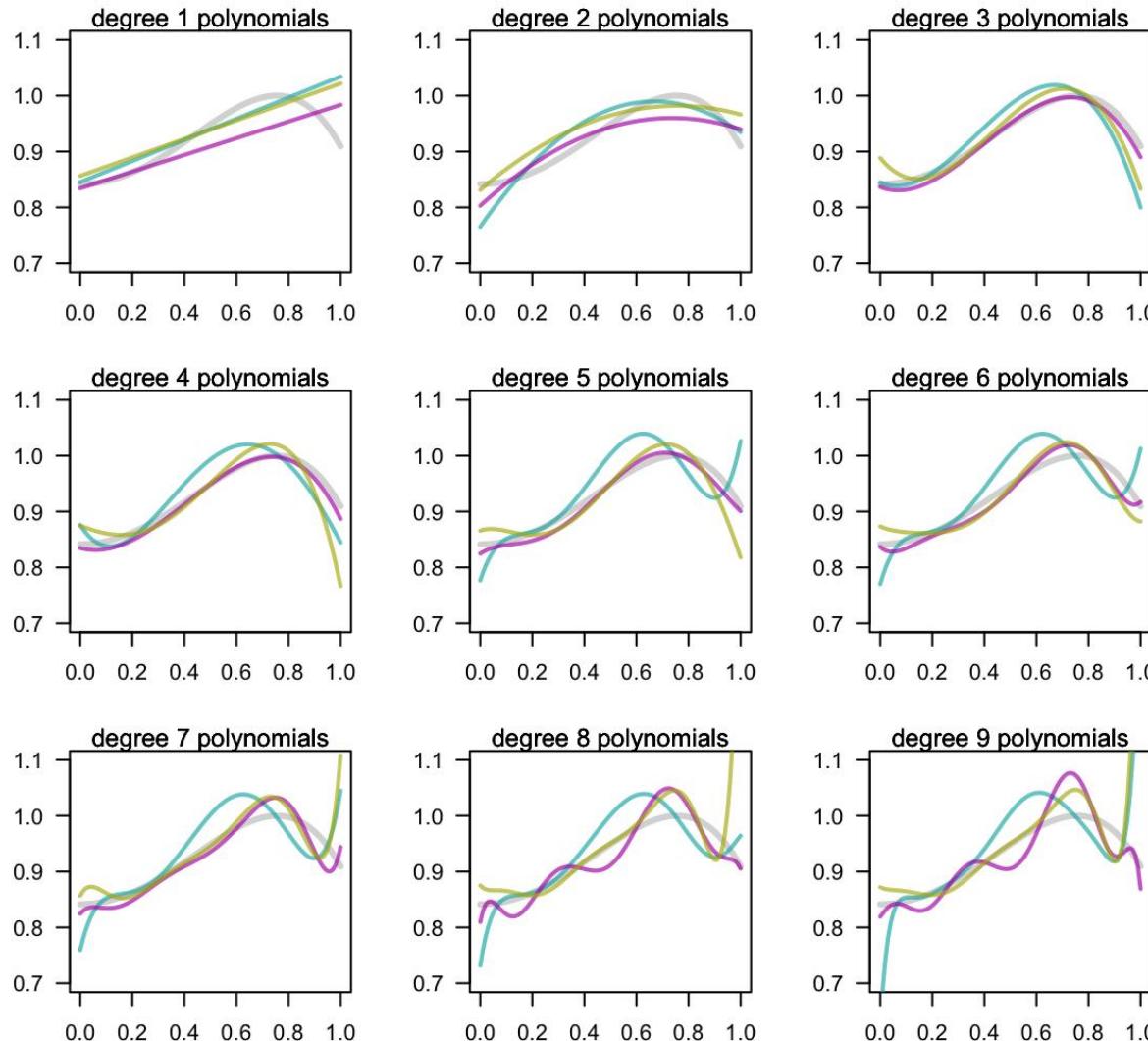


[<https://allmodelsarewrong.github.io/>]

- **Quá khớp (overfitting)** là hiện tượng khi độ lỗi trên dữ liệu huấn luyện  $E_{in}$  rất nhỏ không còn là dấu hiệu tốt phản ánh hiệu năng thực sự của mô hình trên dữ liệu tổng quát.



# Ví dụ - Chọn mô hình nào?



[<https://allmodelsarewrong.github.io/>]

- Phát sinh 3 tập dữ liệu huấn luyện, mỗi tập dữ liệu có  $n = 10$  điểm dữ liệu.
- Với mỗi tập dữ liệu, ta huấn luyện 9 mô hình hồi quy đa thức (từ hồi quy tuyến tính bậc 1 → đa thức bậc 9).
- Với mỗi tập dữ liệu, mô hình hồi quy đa thức bậc 9 khớp hoàn hảo vào các điểm dữ liệu. Nhưng các mô hình bậc 9 này rất biến động theo mỗi tập dữ liệu. Phương sai cao → overfit.
- Các mô hình bậc 1, 2, 3 thì ổn định hơn. Các mô hình này có phương sai thấp.



# Overfitting và Underfitting

- **Overfitting** xảy ra khi mô hình được huấn luyện khớp quá mức vào dữ liệu huấn luyện. Ta chọn một mô hình có độ lỗi trên dữ liệu huấn luyện  $E_{in}$  thấp, nhưng mô hình này lại cho ra độ lỗi trên dữ liệu kiểm tra  $E_{out}$  cao. Độ lỗi trên dữ liệu huấn luyện  $E_{in}$  không còn là dấu hiệu tốt để phản ánh khả năng tổng quát hóa của mô hình. Các mô hình này có **phương sai cao (high variance)**.
- **Underfitting** xảy ra với những mô hình có hiệu năng thấp trên dữ liệu kiểm tra / dữ liệu tổng quát bởi mô hình không đủ khả năng / tính linh hoạt để học các tri thức quan trọng từ dữ liệu. Các mô hình underfit thường do tập giả thiết lựa chọn không phù hợp. Do đó, các mô hình này có **độ chêch cao (high bias)**.
- Các mô hình có độ phức tạp thấp thường có độ chêch cao. Sử dụng các mô hình phức tạp hơn ta có thể giảm được độ chêch. Tuy nhiên, ta thường không thể biết được chính xác hàm đích (target function) nằm trong những tập giả thiết nào.



# Overfitting và Underfitting

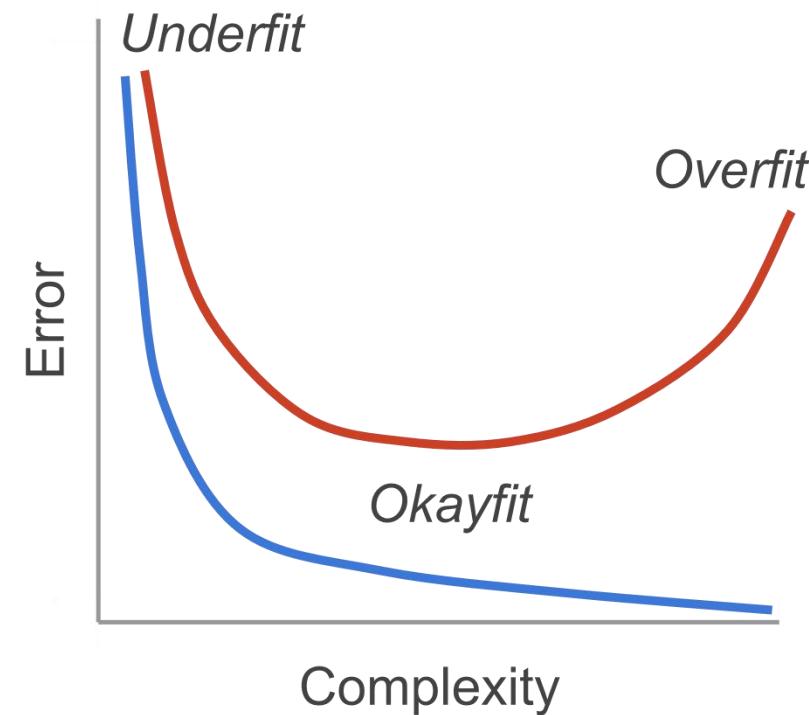
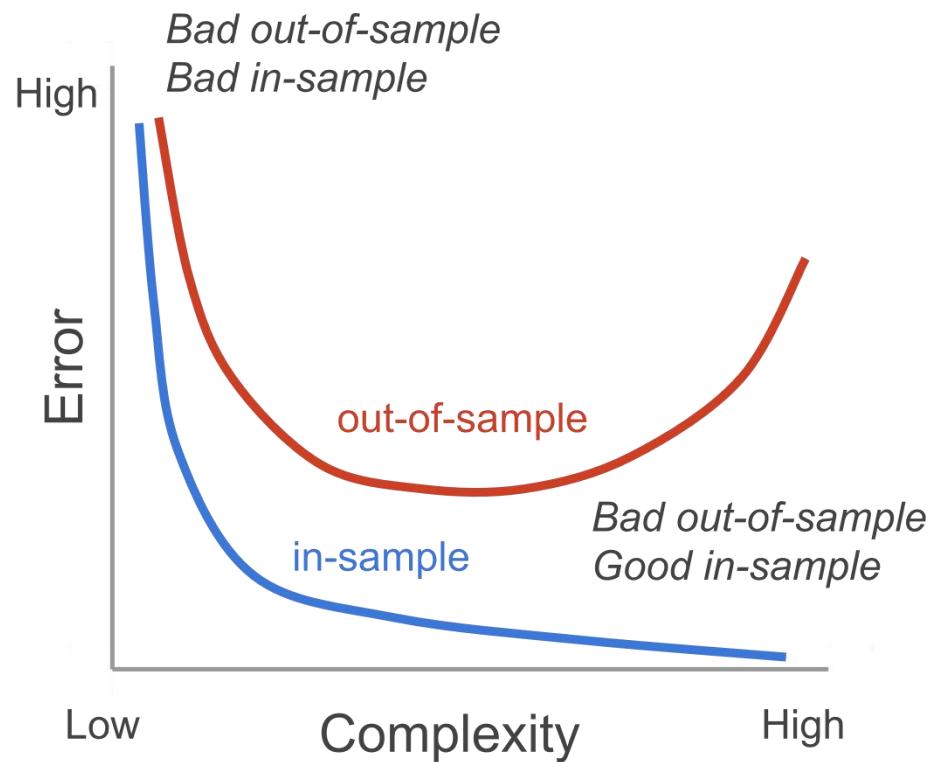
- Độ lớn của độ chêch (bias) không phụ thuộc vào kích thước tập dữ liệu huấn luyện  $D_{in}$ .
$$\bar{h}(x_0) - f(x_0)$$
- Điều này có nghĩa là nếu ta tăng số điểm dữ liệu huấn luyện lên thì cũng không giúp ta giảm được độ chêch.
- Độ lớn của phương sai (variance) phụ thuộc vào số lượng điểm dữ liệu huấn luyện  $n$ .

$$\mathbb{E}_D \left[ \left( h^{(D)}(x_0) - \bar{h}(x_0) \right)^2 \right]$$

- Khi kích thước tập dữ liệu huấn luyện tăng lên, ta giảm được sự biến động (variability) giữa các mô hình sau khi huấn luyện. Khi đó, tính linh hoạt của mô hình sẽ trở thành một lợi thế.



# Overfitting và Underfitting



[<https://allmodelsarewrong.github.io/>]