University of Science, VNU-HCM

Faculty of Information Technology

Introduction to Data Science Course

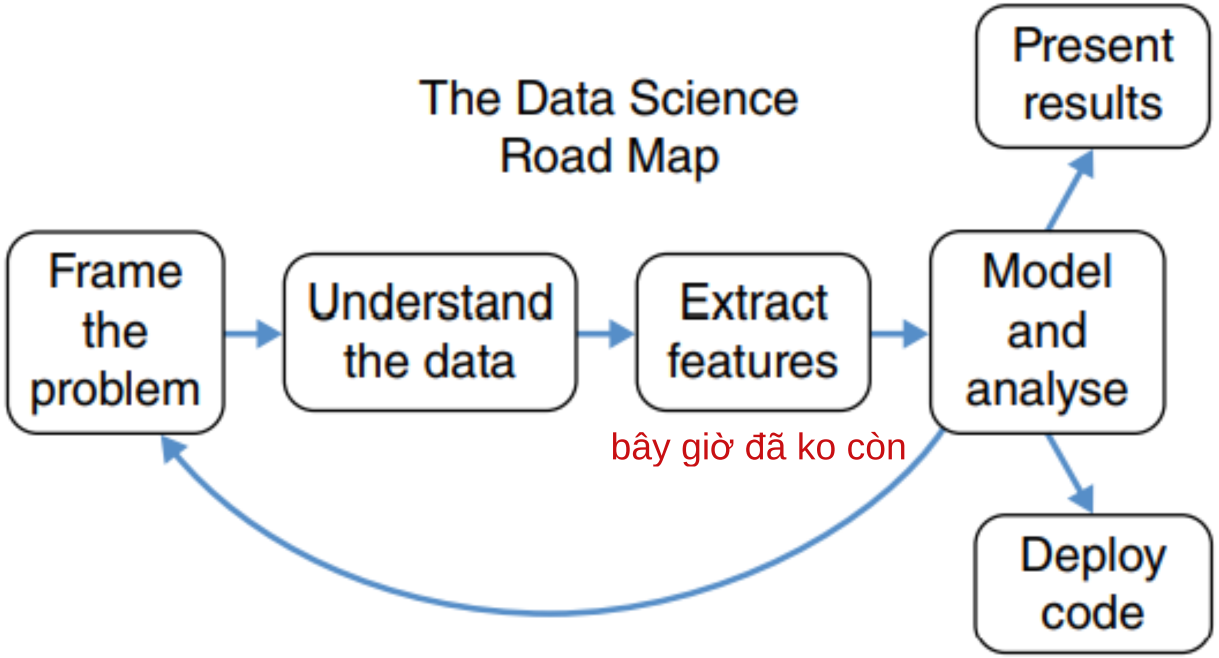
Data Modeling (Part 1)

Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn
Department of Computer Science

Ho Chi Minh City

Contents

Process



The Data Science Road Map
Frame the problem
Understand the data
Extract features
Model and analyse
Present results
Deploy code
bây giờ đã ko còn

After preprocessing

Data Science Process

- Give the question to answer
- Collecting data
- Data Discovery & preprocessing to obtain data that can be analyzed
- Data analysis (in visualizations, statistics,machine learning)
  → answers (hypotheses) for the question
- Evaluation
- Decision Making

Data Science vs. Machine Learning

Data Science

Field that determines the processes, systems, and tools needed to transform data into insights to be applied to various industries.

Skills needed:
- Statistics
- Data visualizatiom
- Coding skills (Python/R)
- Machine learning
- SQL/NoSQL
- Data wrangling

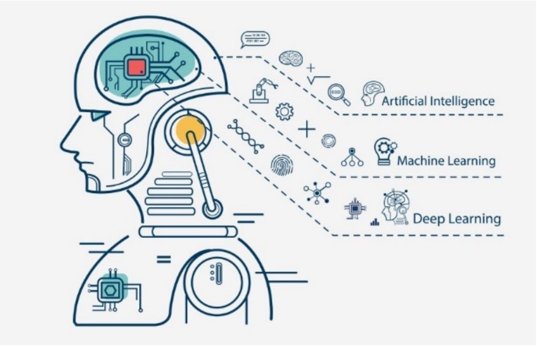
Machine learning is part of data science. Its algorithms train on data delivered by data science to "learn."

Skills needed:
- Math, statistics, and probability
- Comfortable working with data
- Programming skills

Machine Learning

Field of artificial intelligence (AI) that gives machines the human-like capability to learn and adapt through statistical models and algorithms.

Skills needed:
- Programming skills (Python, SQL, Java)
- Statistics and probability
- Prototyping
- Data modeling

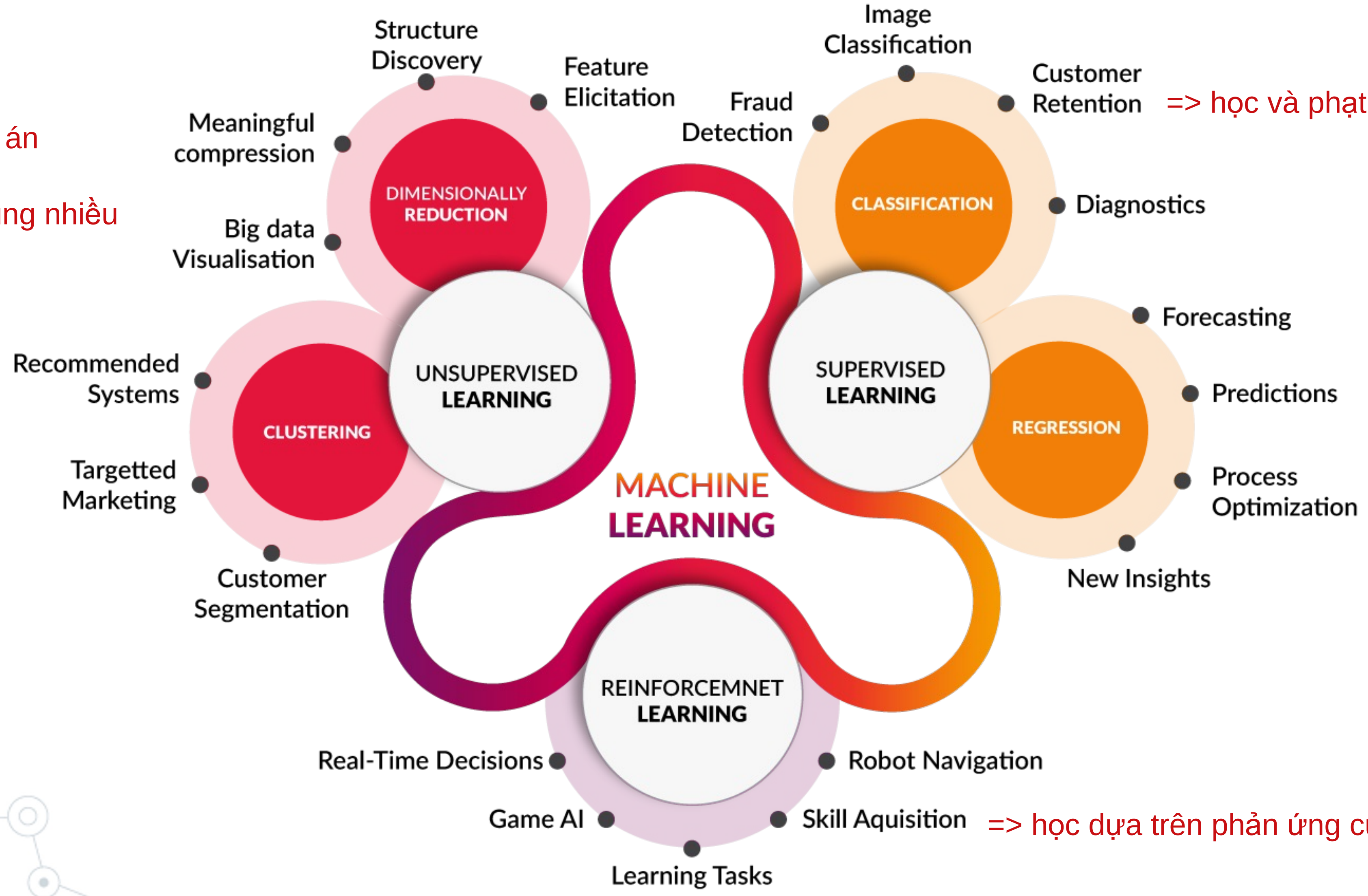


https://www.coursera.org/articles/data-science-vs-machine-learning

ML Tasks

=> UL: why choose
+ DL chưa có nhãn
+ Bài toán chưa có đáp án

=> giảm chiều DL => dùng nhiều

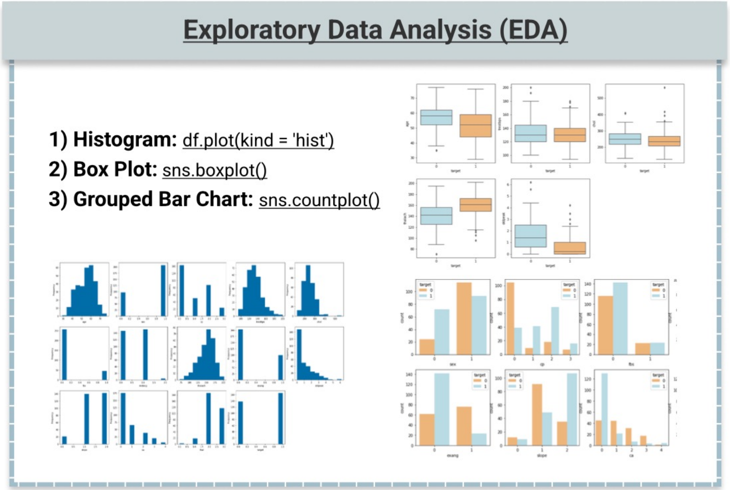


=> học và phạt (bài toán đã có đáp án)

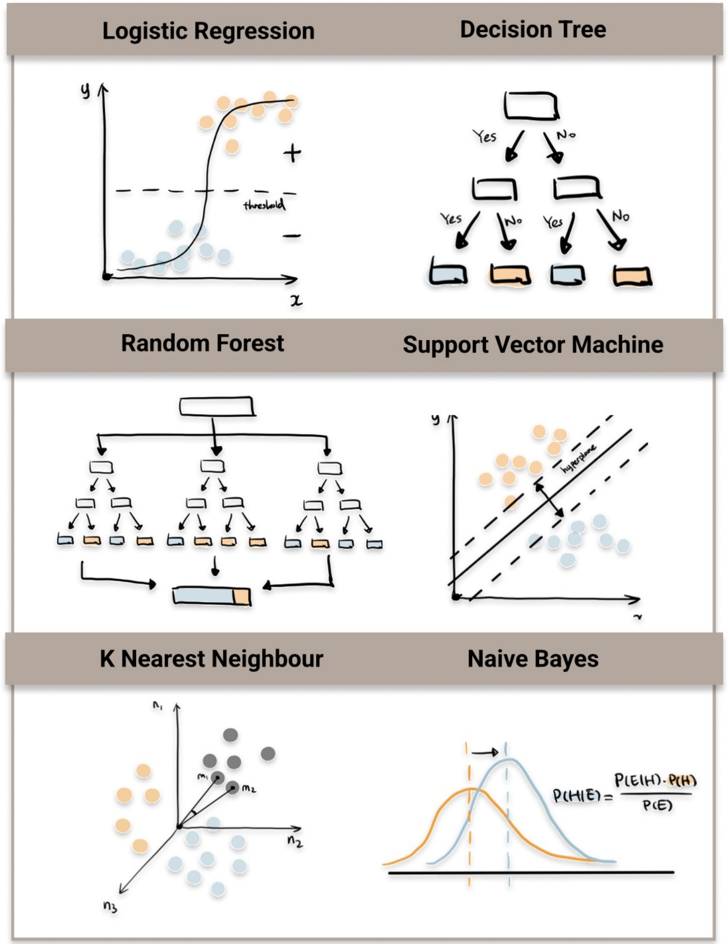
=> học dựa trên phản ứng của môi trường

Machine Learning Choice

- Before implementing the machine learning (ML) model, the data scientist needs to identify (several) branches in ML that can solve the given problem.



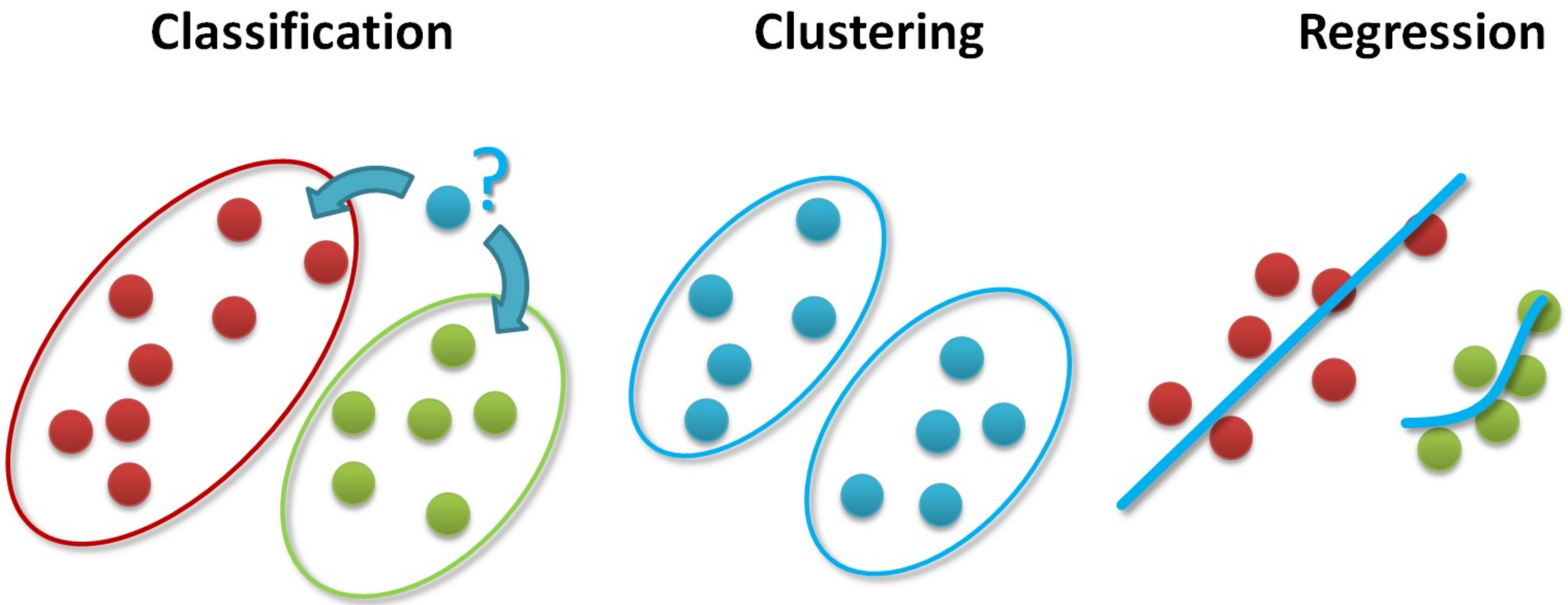
Visualization and Statistics



Machine Learning

The course's focus

- In this course, we focus on three main groups of ML:
  - Regression
  - Classification
  - Clustering



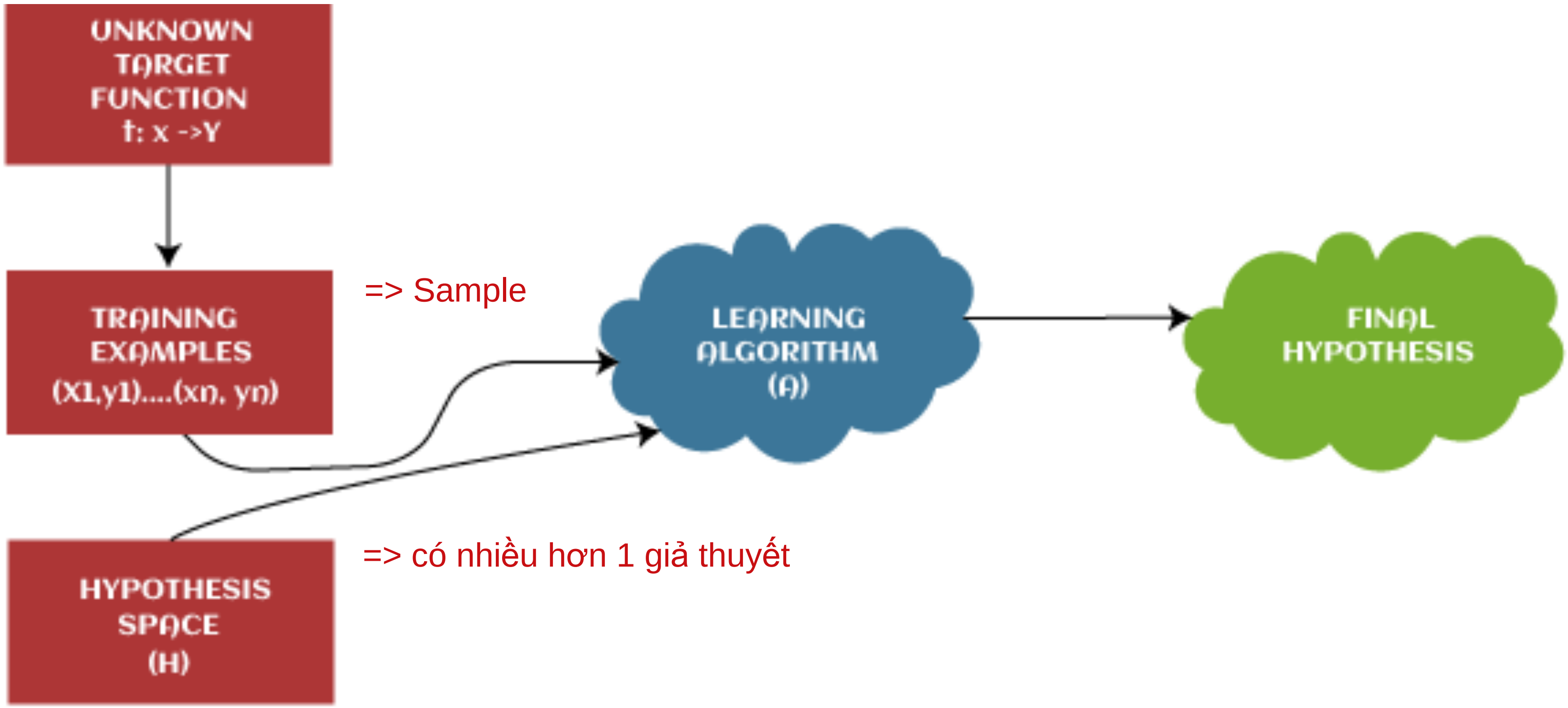
Contents

- Data science and machine learning
- **Machine learning architecture** Learning = identify the mapping (function)
- Regression model

How to identify the mapping => dùng giả thuyết (hypothesis) => hypothesis space

After hypothesis

- The job of a learning algorithm to find the best suitable hypothesis for a problem.



After hypothesis

Làm sao để biết hàm có error:
1. Đếm số điểm nằm ngoài đg thẳng (hàm) => ko cho bk nhiều thông tin
2. Tính khoảng cách từ điểm sai đến đg thẳng, tổng k/c nhiều thì ko chất lượng => cho biết thêm infor để ta tiệm cận đến hàm chính xác

- To choose the suitable hypothesis, we need to define the loss function.

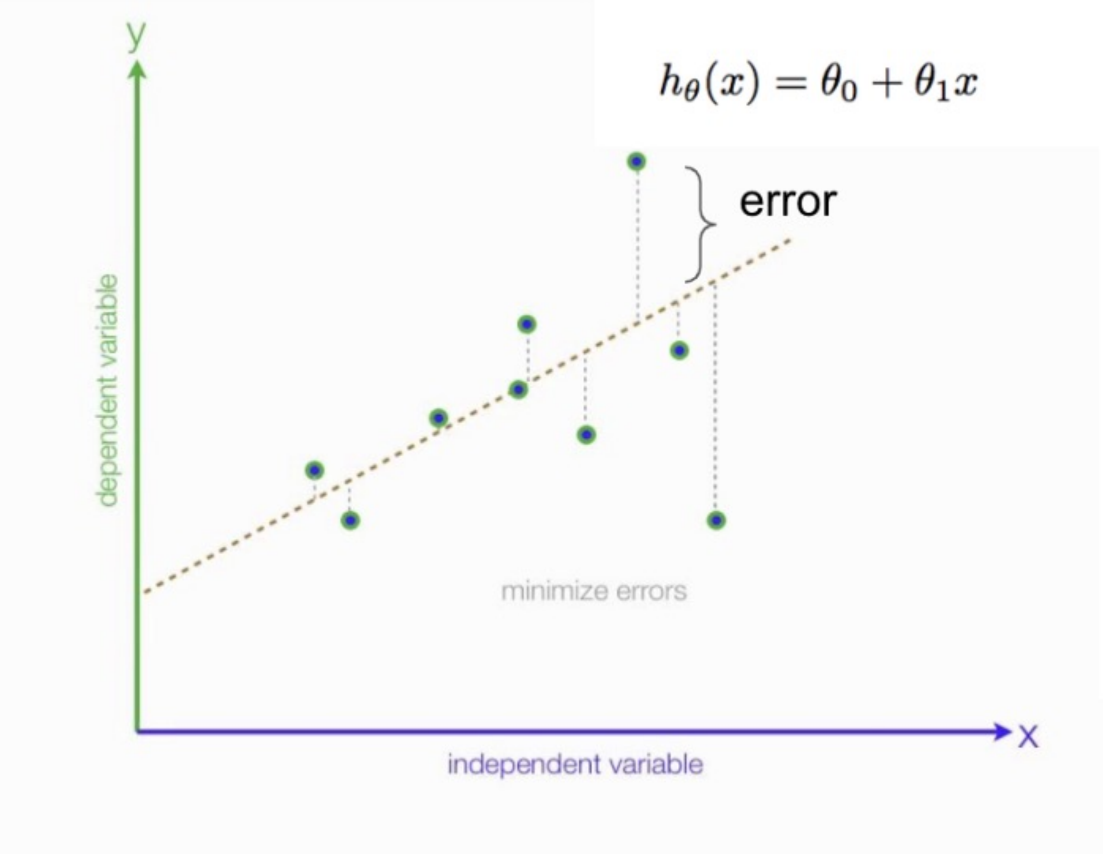
$$\mathcal{L}(y - \hat{y}) = \sum_{i=1}^{n} (y - \hat{y}_i)^2$$

Machine learning = iterative procedure to find a minimum of loss for the given data.

After loss function design

=> Ex about hypothesis: kiếm đường đi qua tất cả các điểm DL.
+ Nên bắt đầu với không gian tuyến tính
+ Nếu dùng KG tuyến tình mà ra KQ tệ thì có thể DL tuân theo KG phi tuyến (hàm uốn càng nhiều thì bậc càng cao => nhiều tham số => nhiều CPU)
+ Nếu ko giải đc thì tăng chiều lên (số chiều = số điểm DL thì vô nghĩa). Kernel là chuyển sang KG khác và tại đó ta dùng tuyến tính or phi tuyến đơn giản

- We are looking for what parameters to produce the lowest loss rate for given dataset, so we need the process to optimize the function (fitting).



Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

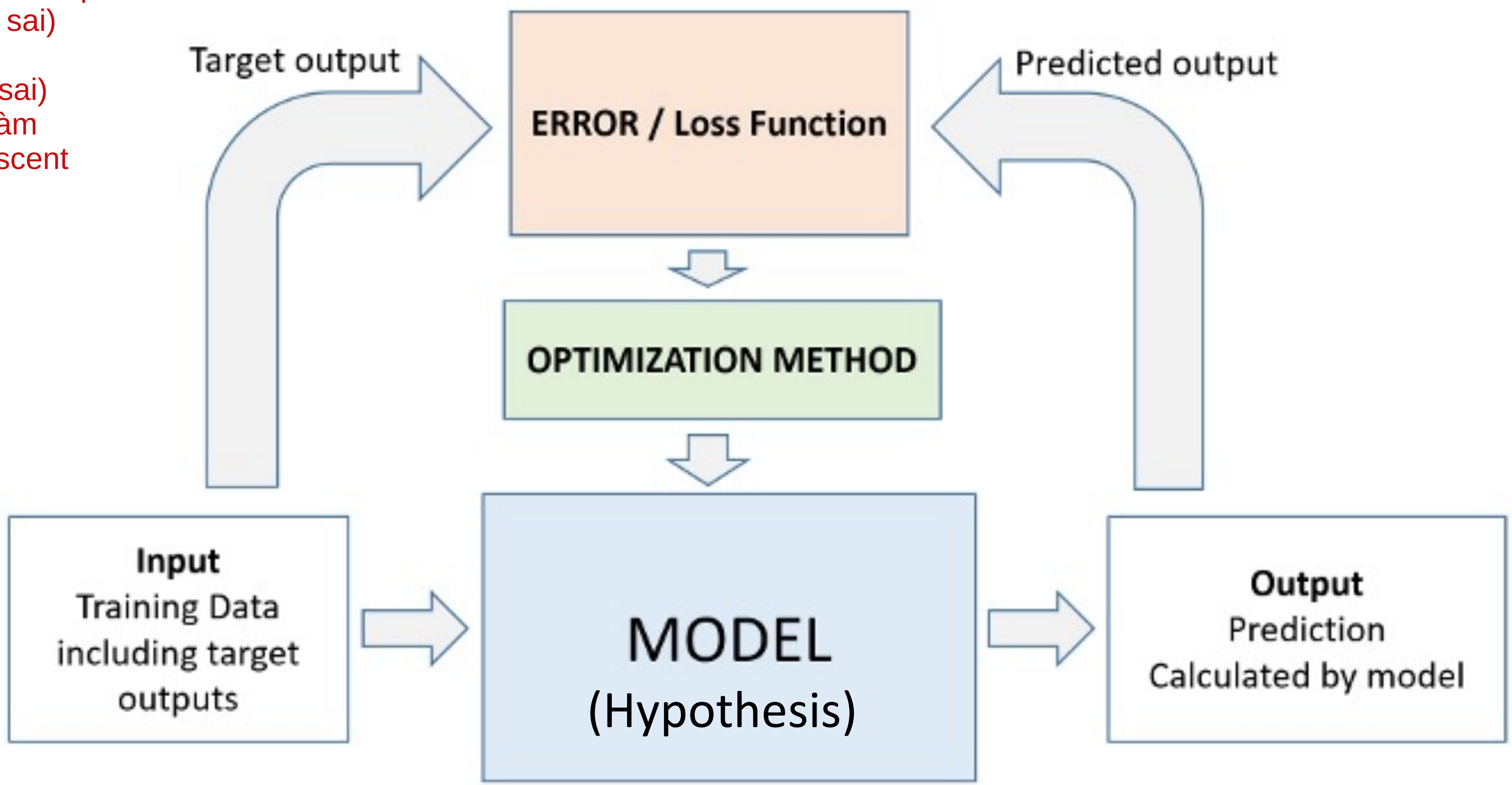
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Goal:

$$\underset{\theta_0, \theta_1}{minimize} J(\theta_0, \theta_1)$$

General model learning architecture

=>Sau khi có Loss rồi thì đi Optimization làm sao cho Loss tiệm cận 0 (thử và sai)
=> How to optimize
- Method 1: mò (thử và sai)
- Method 2: dùng đạo hàm
- Method 3: gradient descent



Contents

- Data science and machine learning
- Machine learning architecture
- **Regression model**
  - Linear regression
  - Non-linear regression
  - Over- and Under-Determined Systems
  - Model selection
  - Overfitting

For regression
+ MSE: những điểm ở xa sẽ bị nhiễu vì bình phương. Chỉ tốt cho những điểm gần. Chỉ dùng sau khi đã xử lý nhiễu or đẩy nhiều điểm vào (sample) với đk mẫu là tốt
+ MAE: chỉnh mô hình lâu do khoảng cách khá giống nhau
+ MBE: làm giảm lỗi

For classification:
+ BCE: dùng CrossEntropyLoss đo độ hỗn loạn
+ Hinge Loss: dùng SVMLoss chỉ lấy gtri >0 nếu ko thì nó bằng 0

Regression

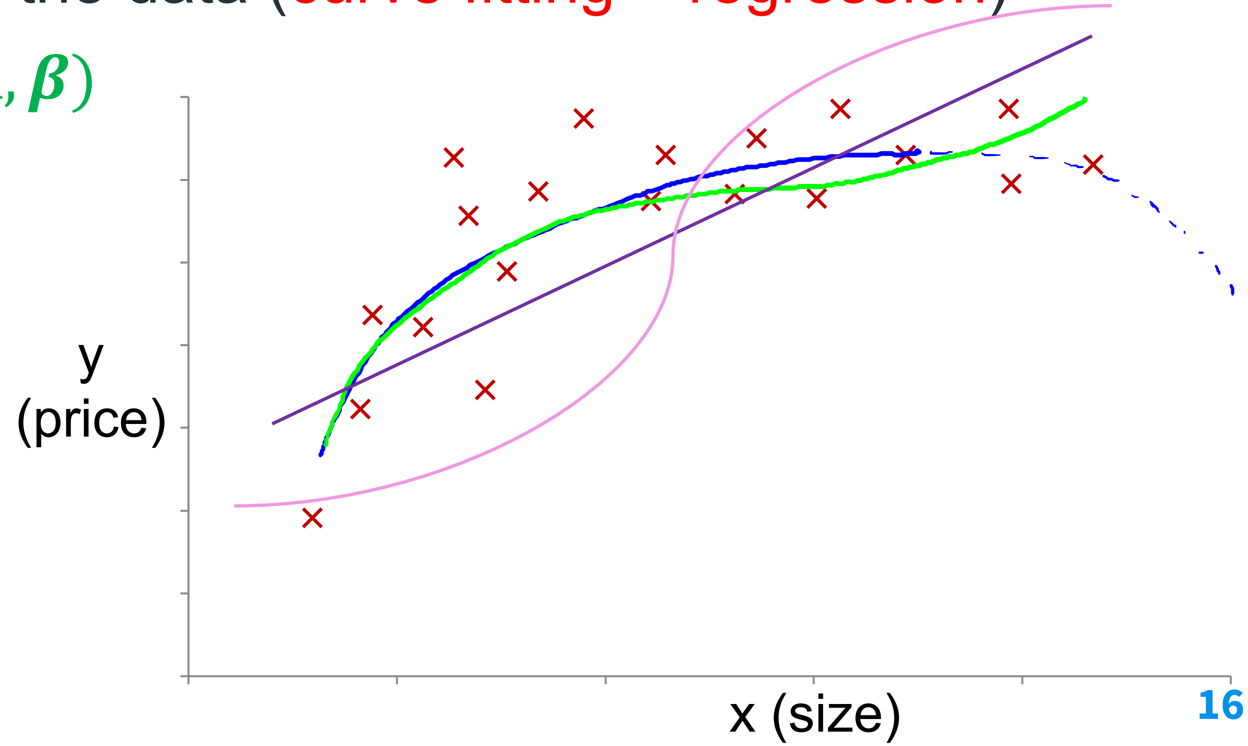
- Consider a set of n data points:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ \dots, (x_n, y_n)$$

- Purpose:
  - Select a function f (·) and fit it to the data (curve fitting = regression)

$$\mathbf{Y} = f(\mathbf{A}, \boldsymbol{\beta})$$

Size in feet² (x)	Price ($) in 1000's (y)
100	10
800	150
1534	315
852	178



Linear regression

- Assume that a line is fitted through the points (hypothesis)

$$f(x) = \beta_1 x + \beta_2$$

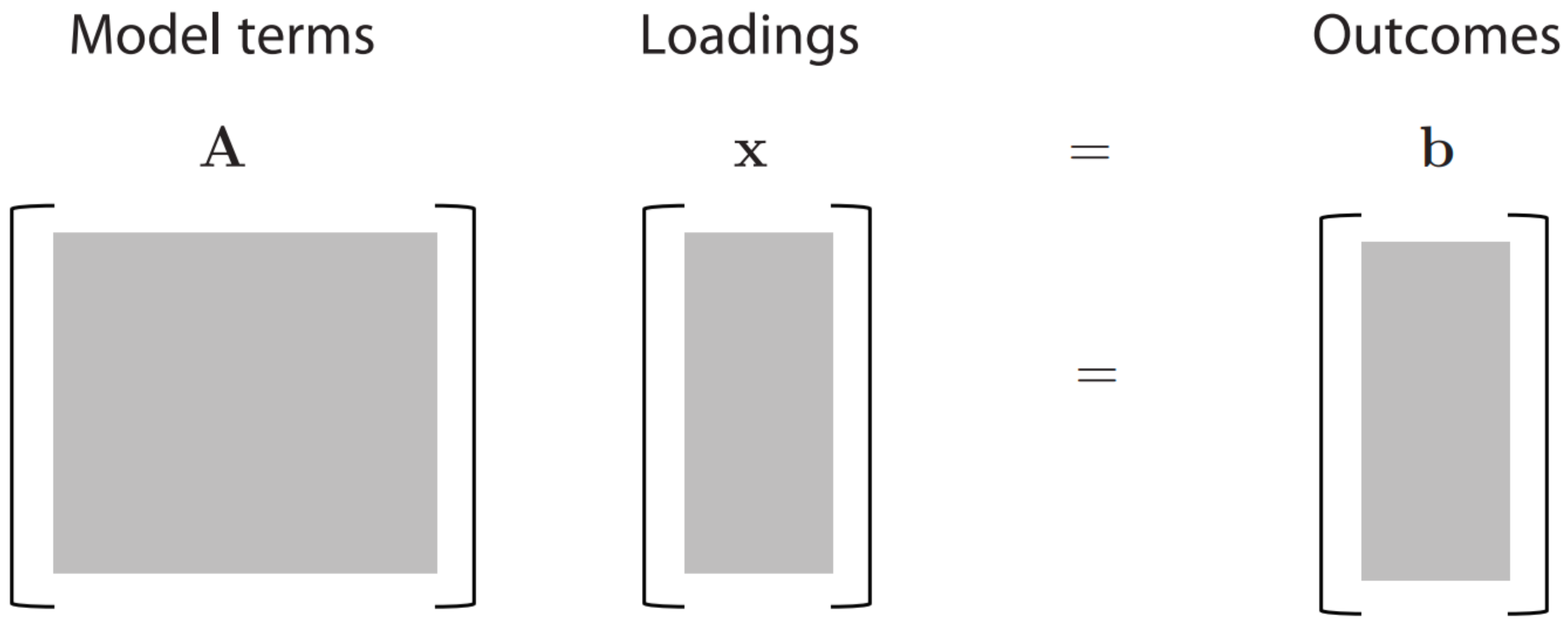
- The loss function is MSE (mean-squares error)

$$E(f) = \frac{1}{n}\sum_{k=1}^{n}(f(x_k) - y_k)^2 = \frac{1}{n}\sum_{k=1}^{n}(\beta_1 x_k + \beta_2 - y_k)^2$$

Linear regression

- The optimization method: derivatives
- Generalization, the 2×2 system:

$$\mathbf{Ax} = \mathbf{b}$$



Contents

Nonlinear regresstion

- How with nonlinear regresstion? For example:

$$f(x) = \beta_2 \exp(\beta_1 x)$$

- The MSE function:

$$E(\beta_1, \beta_2) = \sum_{k=1}^{n} (\beta_2 \exp(\beta_1 x_k) - y_k)^2$$

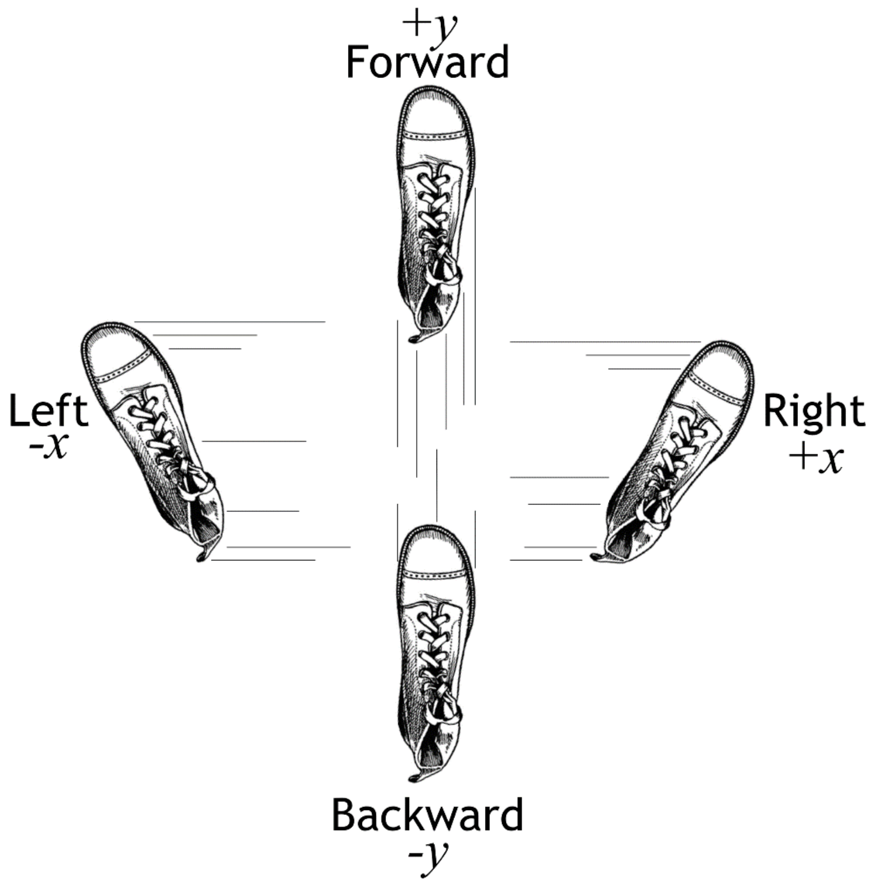
Contents

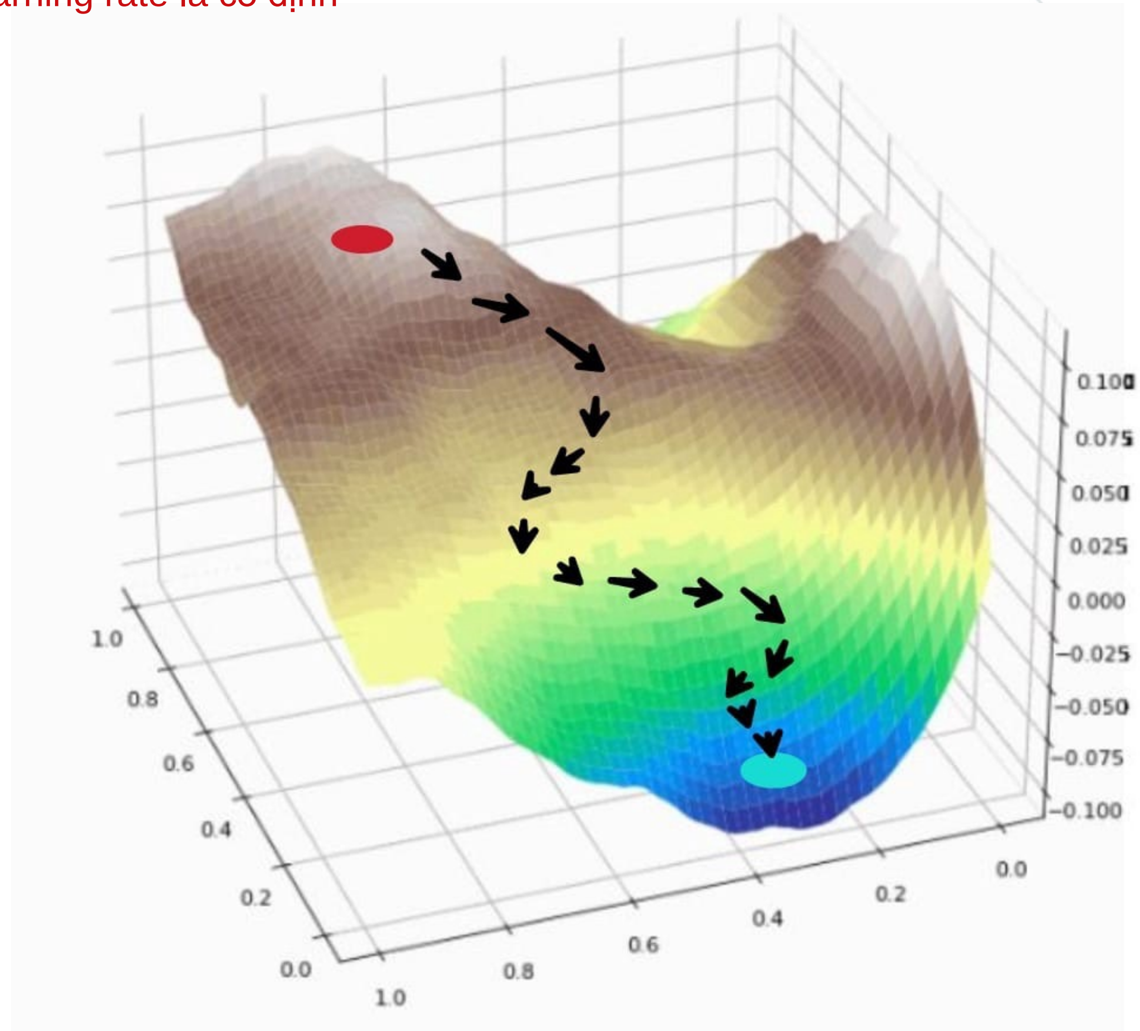
Go to downhill

Go to downhill

dùng đạo hàm riêng để cập nhật theo hướng và sẽ chọn hướng có độ dốc cao

=> khi đã biết hướng thì nhảy cóc (learning rate) nhưng có thể nhảy qua điểm đến thì phải quay lại. Trong các mô hình thì learning rate là cố định





Go to downhill

- What means if direction vector is:

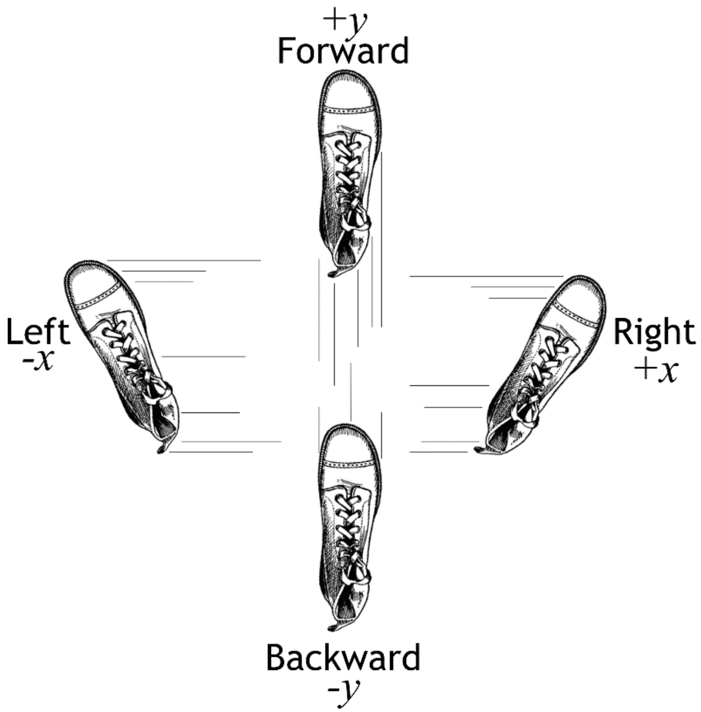
$[x, y]$

$= [which\ way\ is\ down\ in\ x\ direction, which\ way\ is\ down\ in\ y\ direction]$

$= [-1,1]$

- To actually move downhill, we move to:

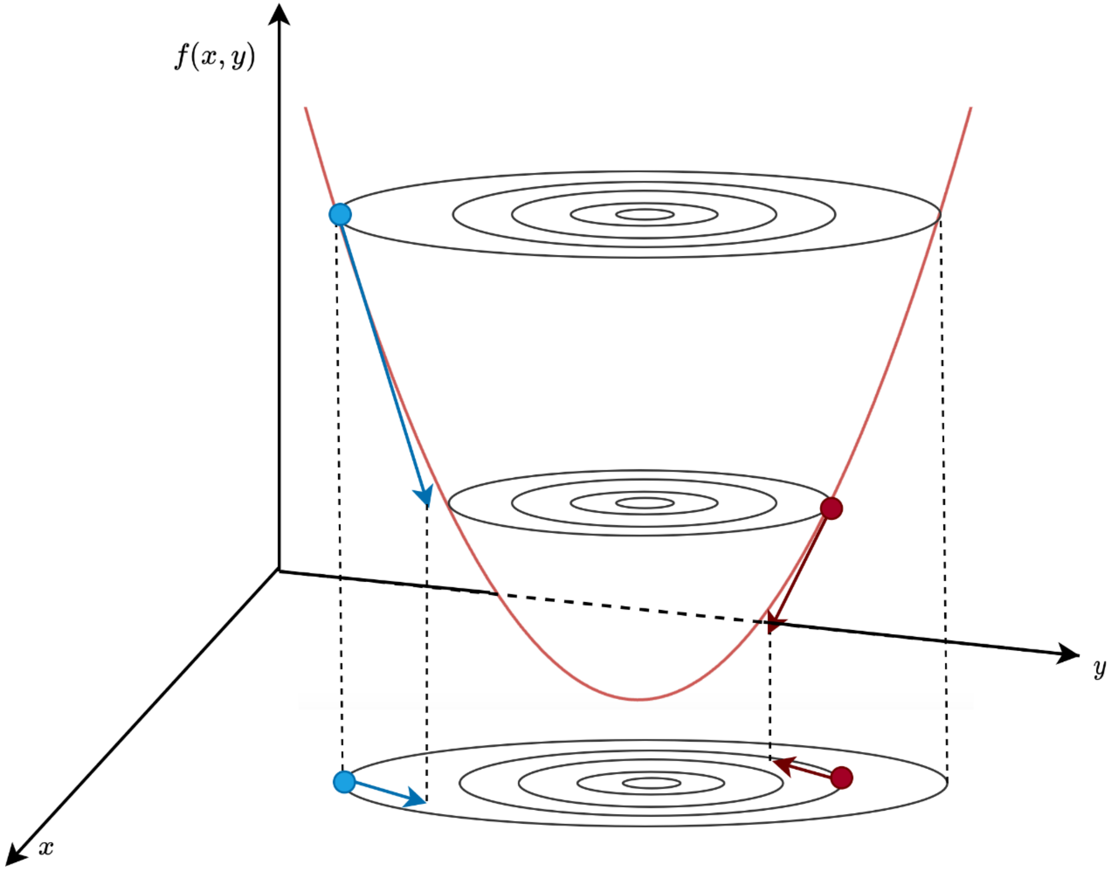
$\Rightarrow [x_{new}, y_{new}] = [x, y] + [-1, 1]$



Go to downhill

- Generally, to move in xy space toward the minimum point, we need identify:
  - Moving direction (increase/descrease x and y)
  - Rate of change (based on slope)

$\Rightarrow$ It is a direction vector

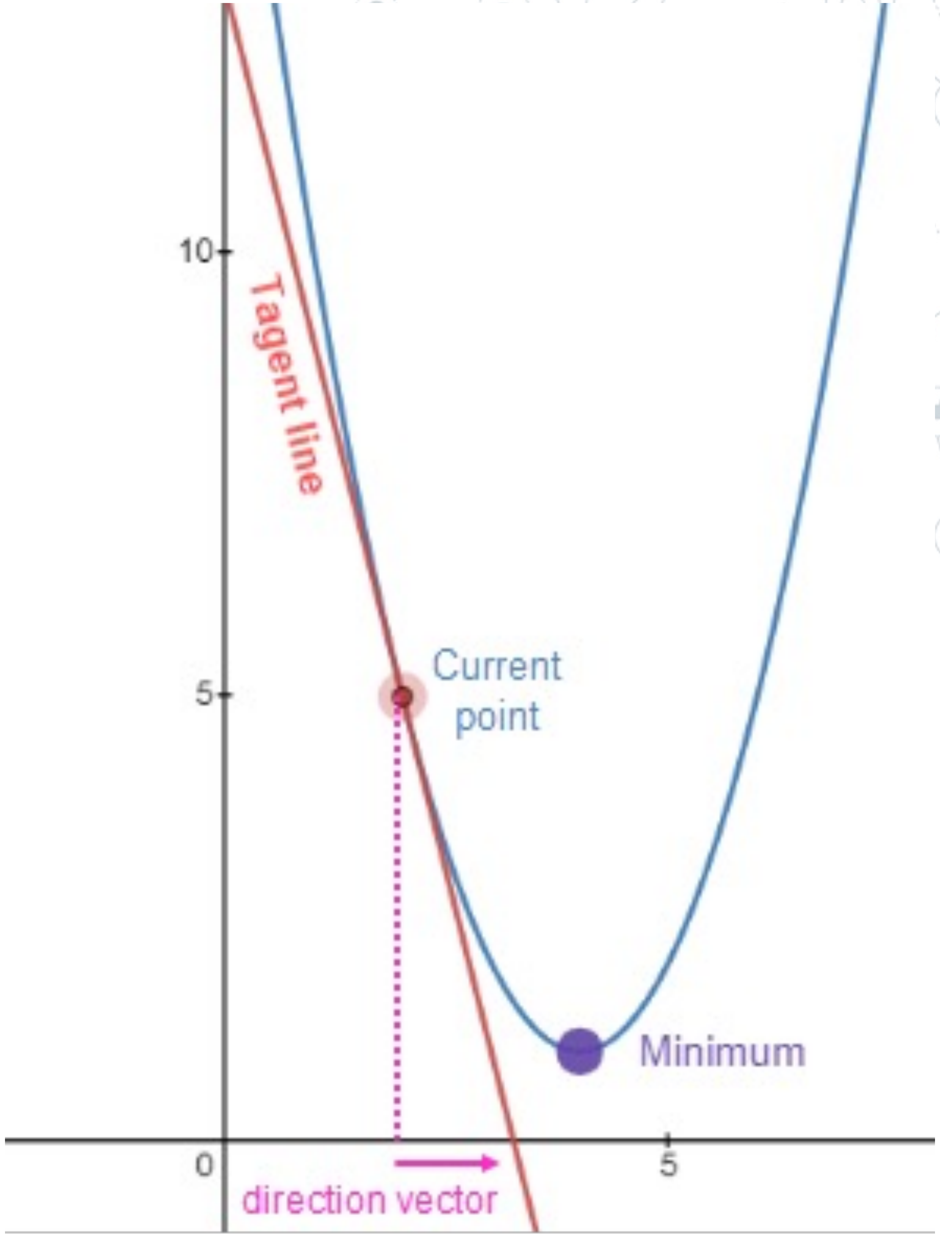


Direction vector

- The derivative of a function at a specific point gives the slope of the tangent line.

$$f'(x) = \lim_{(x_1 - x_0) \to 0} \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

- Why is the tangent line considered as a direction vector?

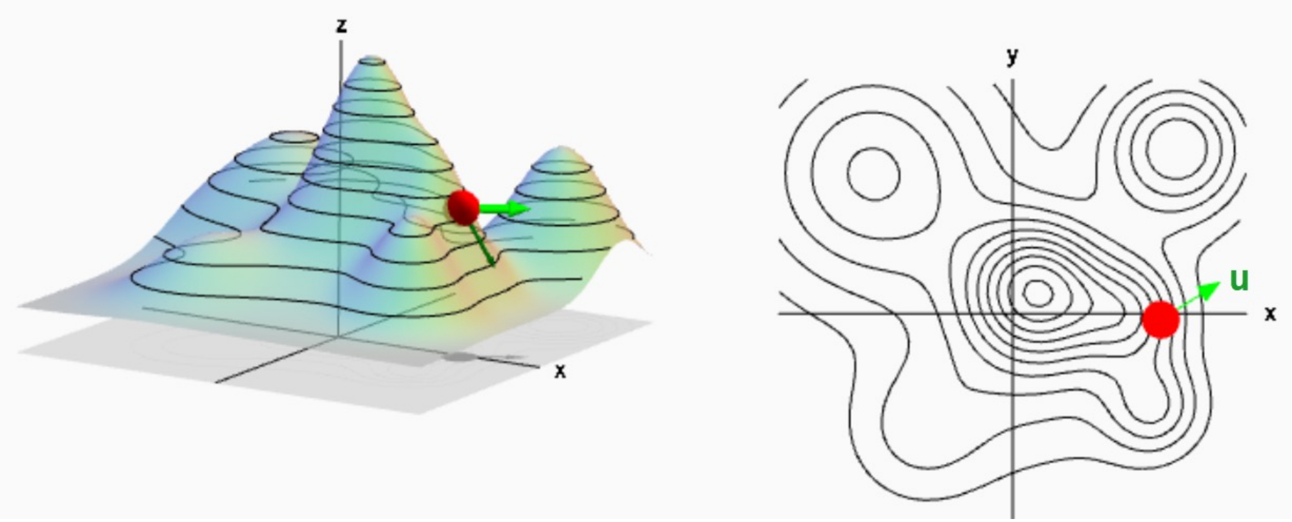


Directional derivative

- If you stand at some point $\mathbf{a} = (x_0, y_0)$, the slope of the ground in front of you will depend on the direction you are facing.
- To calculate the slope in any direction, we derivative in this direction.
  $\Rightarrow$ called the directional derivative.

$$D_{\mathbf{u}} f(x_0, y_0)$$

where $\mathbf{u} = (u_1, u_2)$ is an unit vector that points in the direction in which we want to compute the slope.



Gradient

- The gradient of f at any point tells you:
  - a direction is the steepest from that point with respect to the x,y plane
  - how steep it is (the slope of the hill in that direction)

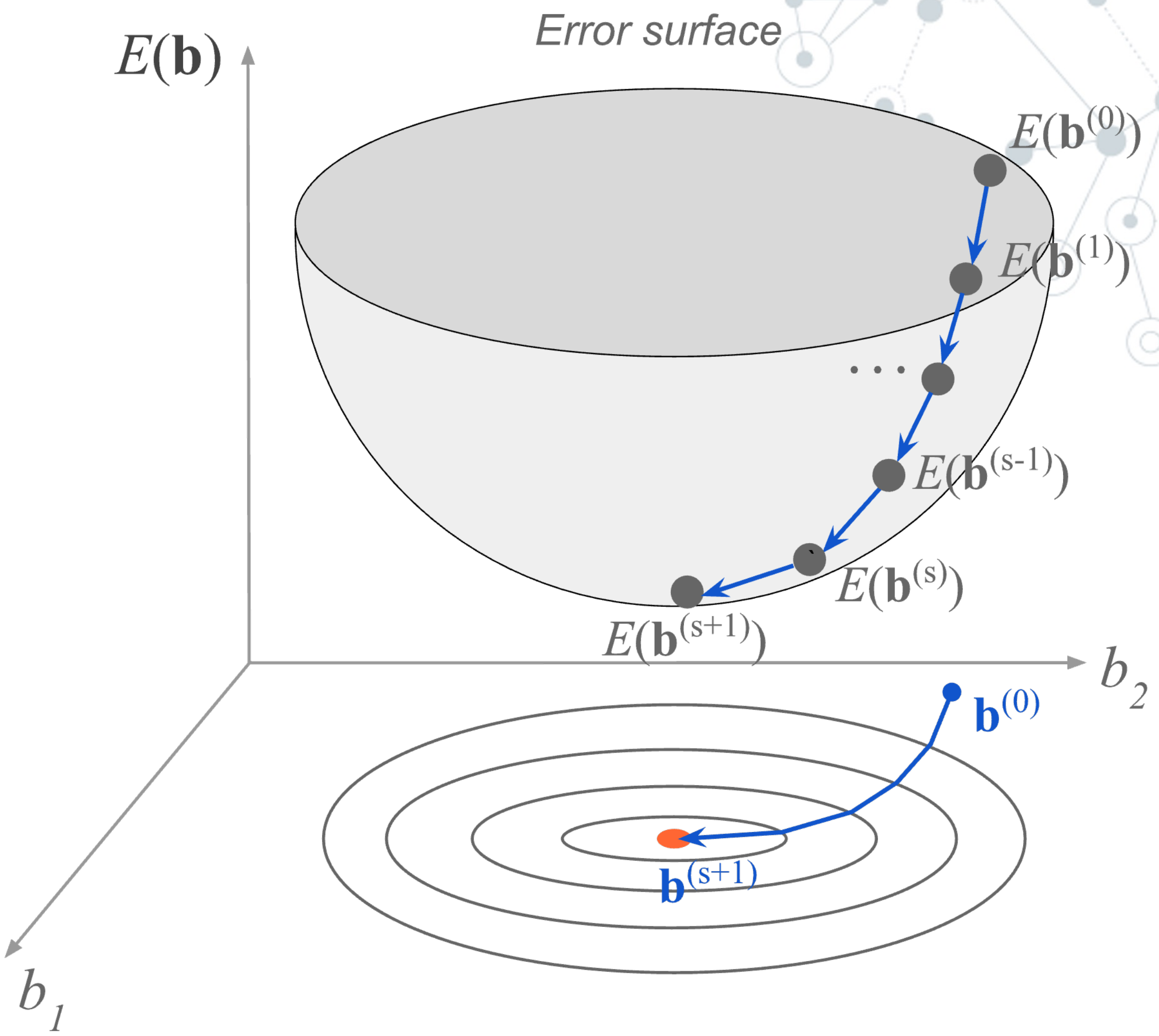
$$\nabla f(x,y) = \begin{bmatrix} \dfrac{\partial f(x,y)}{\partial x} \\ \dfrac{\partial f(x,y)}{\partial y} \end{bmatrix} = \frac{\partial f(x,y)}{\partial x}\hat{\mathbf{x}} + \frac{\partial f(x,y)}{\partial y}\hat{\mathbf{y}}$$

- The partial derivatives give the slope in the **positive** x direction and the slope in the **positive** y direction.

Gradient Descent

- As we update, we want the value of $f(x, y)$ to decrease.
  - When it stops decreasing, (x_0, y_0) will have arrived at the position giving the minimum value of $f(x, y)$.
- The next position at time step t:

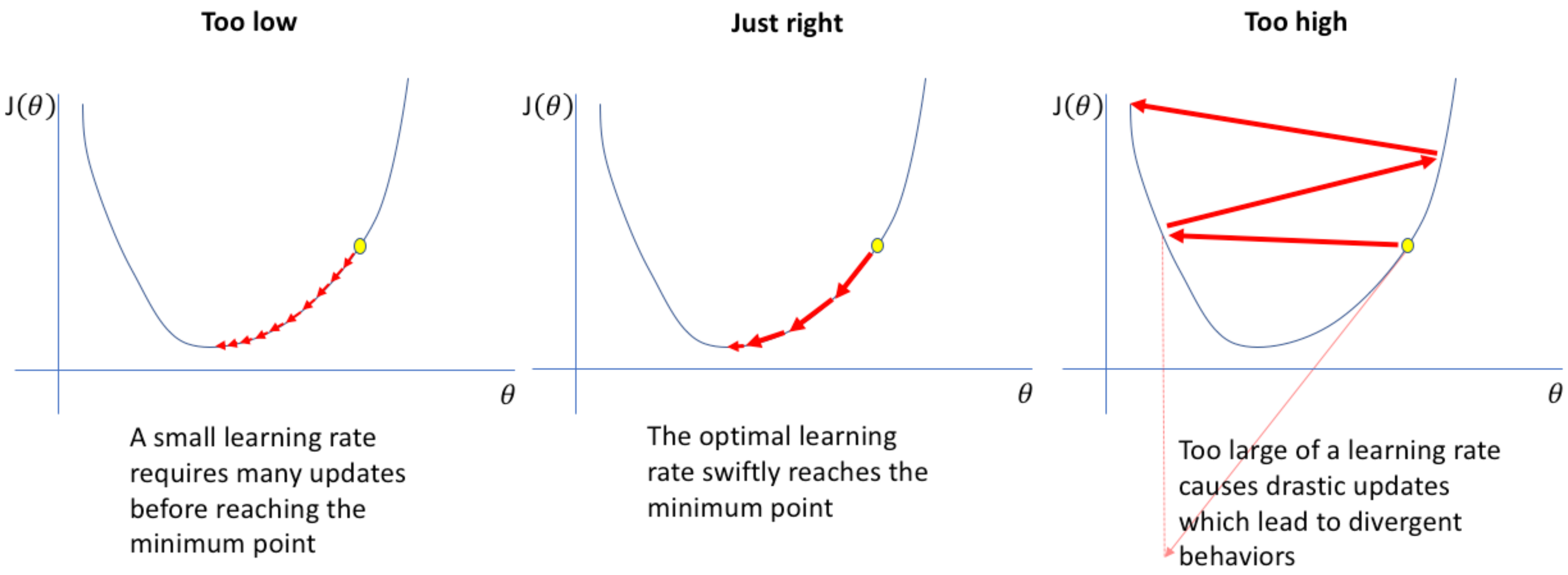
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nabla f(\mathbf{x}_t)$$



Issues: Learning rate

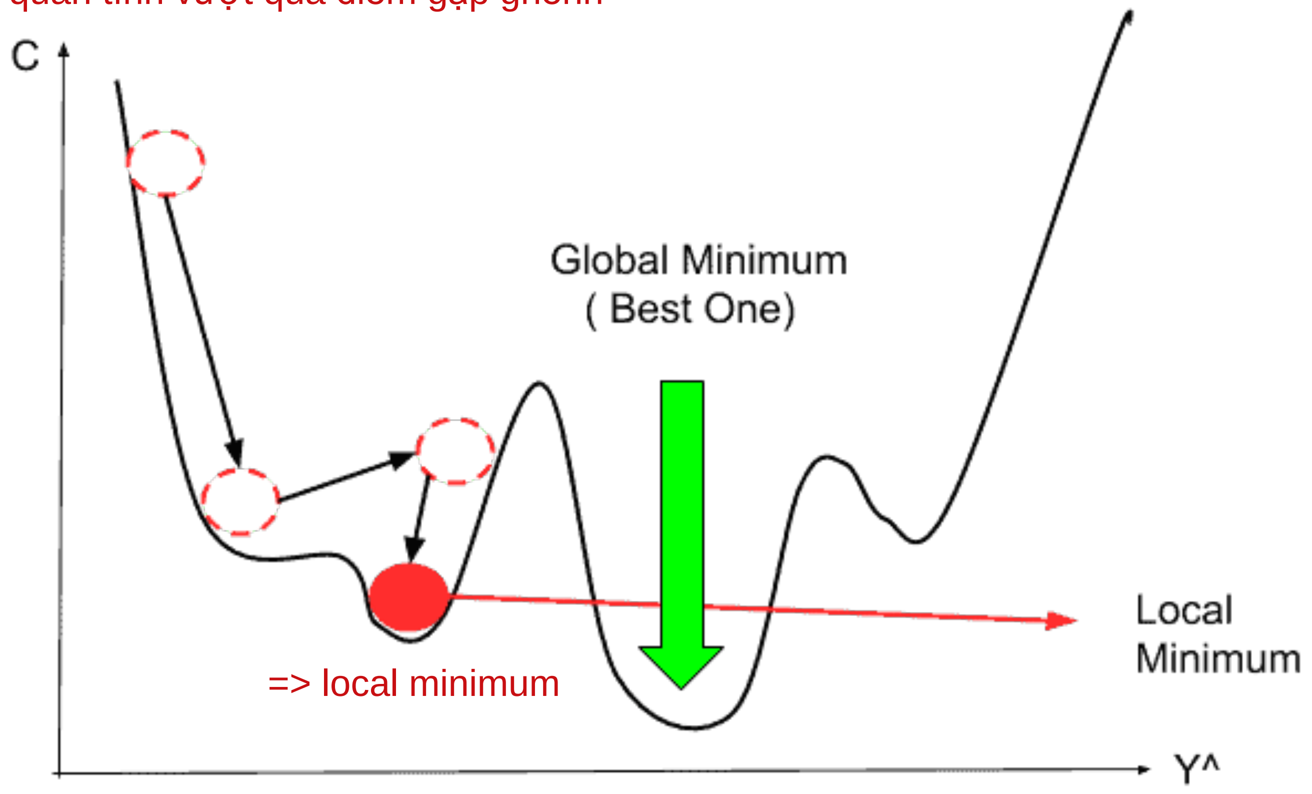
- Need to restrict the size of the steps by shrinking the direction vector using a learning rate η, whose value is less than 1:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$$

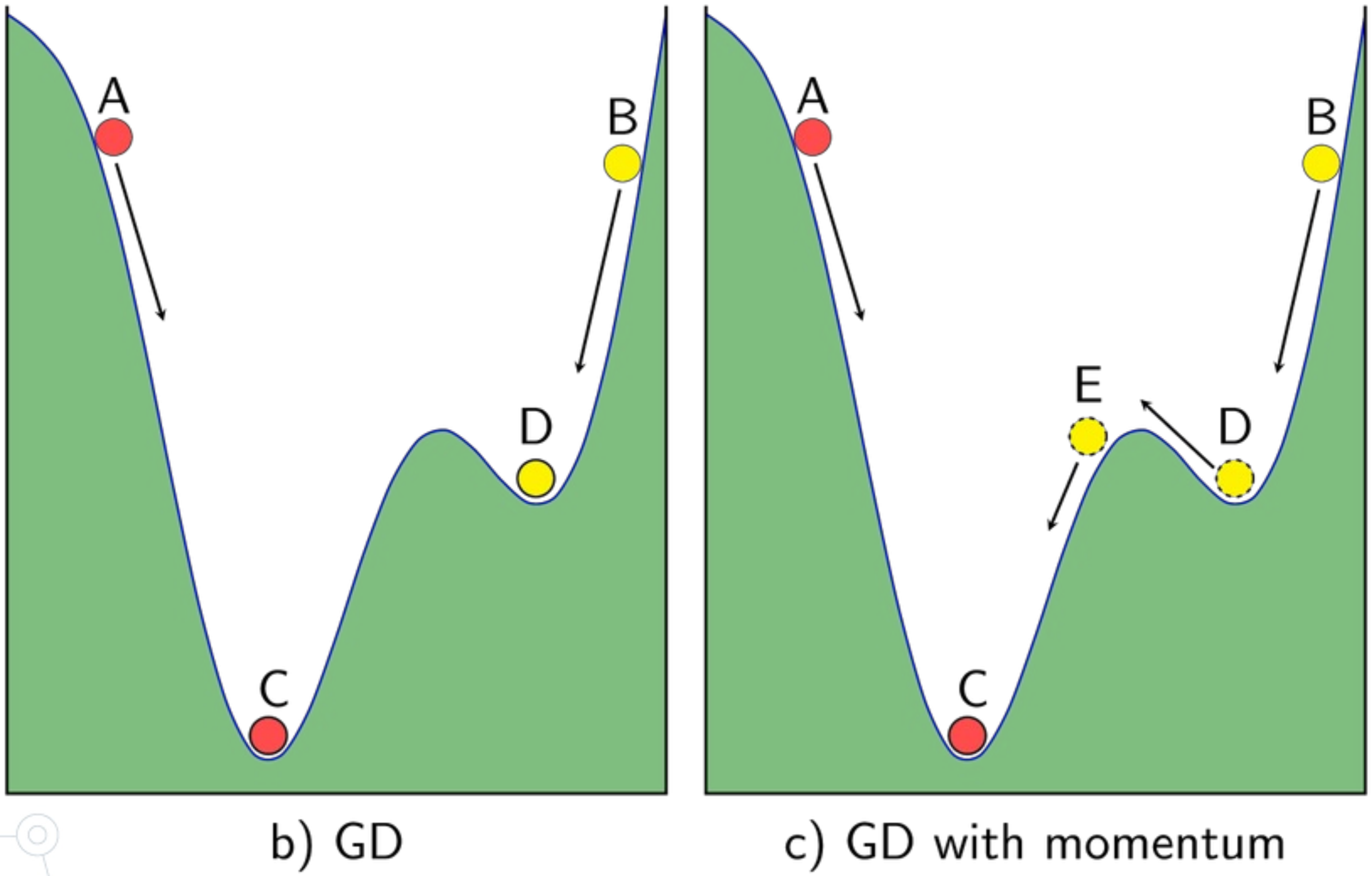


Issues: Starting point (non-linear function)

=> Cách khắc phục:
+ thiết kế lại hàm Loss thành hàm lồi để nó ko lõm
+ Thêm quán tính, cho nó đi 1 cơ hội để quán tính vượt qua điểm gặp ghềnh



Momentum



Summary for nonlinear regression

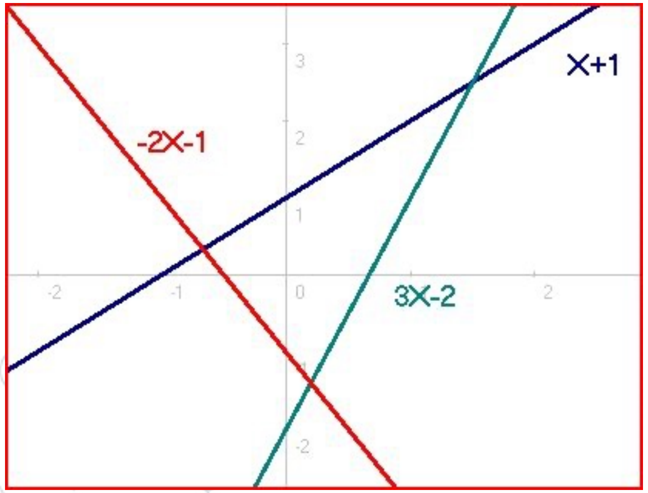
- The nonlinear optimization procedure:
  - The initial guess
  - Step size η
  - Computing the gradient efficiently

Contents

- Data science and machine learning
- Machine learning architecture
- **Regression model**
  - Linear regression
  - Non-linear regression
  - Over- and Under-Determined Systems
  - Model selection
  - Overfitting

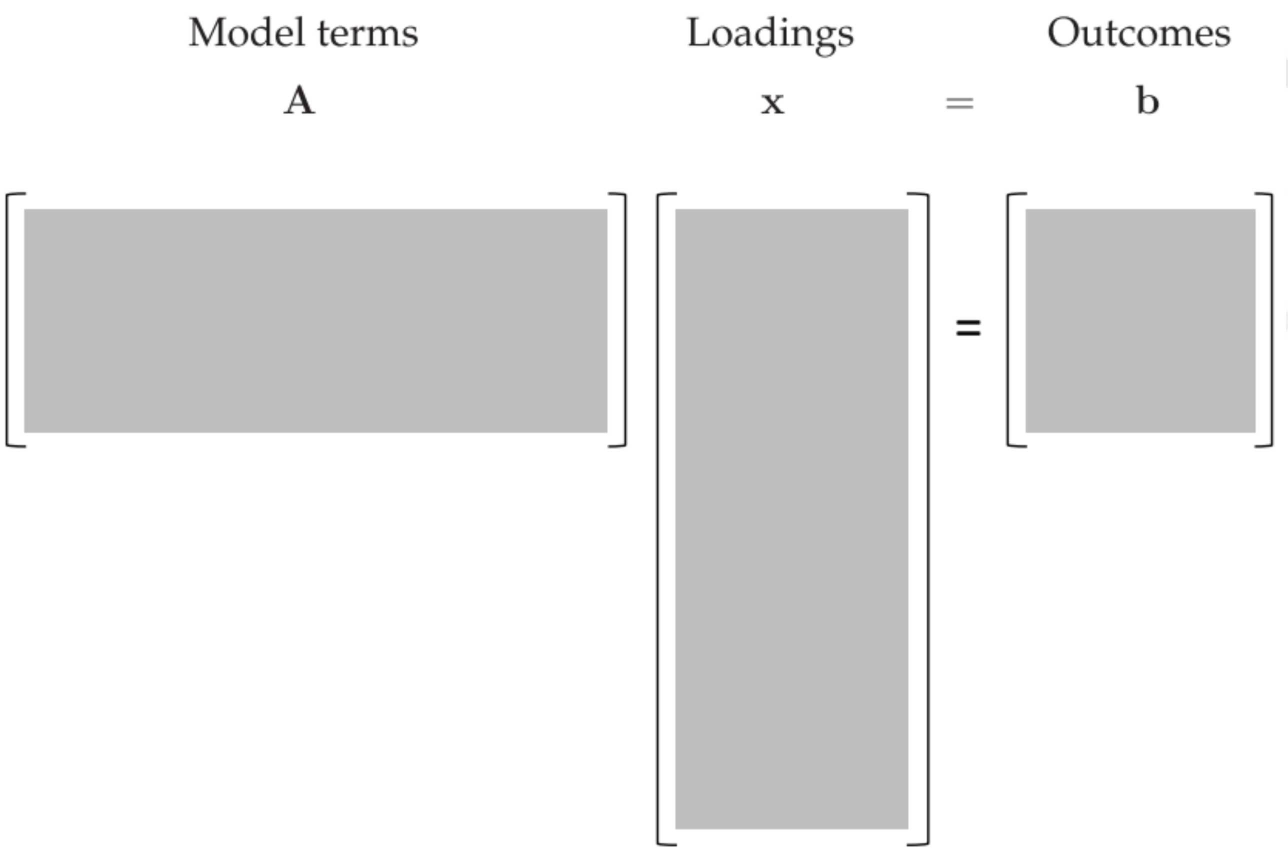
Over-determined systems

- Over-determined systems have more constraints (equations) than unknown variables.
  - No solutions satisfying the linear system.
  - Approximate solutions to minimize a given error.



Under-Determined Systems

- Under-determined systems have more unknowns than constraints.
  - an infinite number of solutions.
  - some choice of constraint must be made.

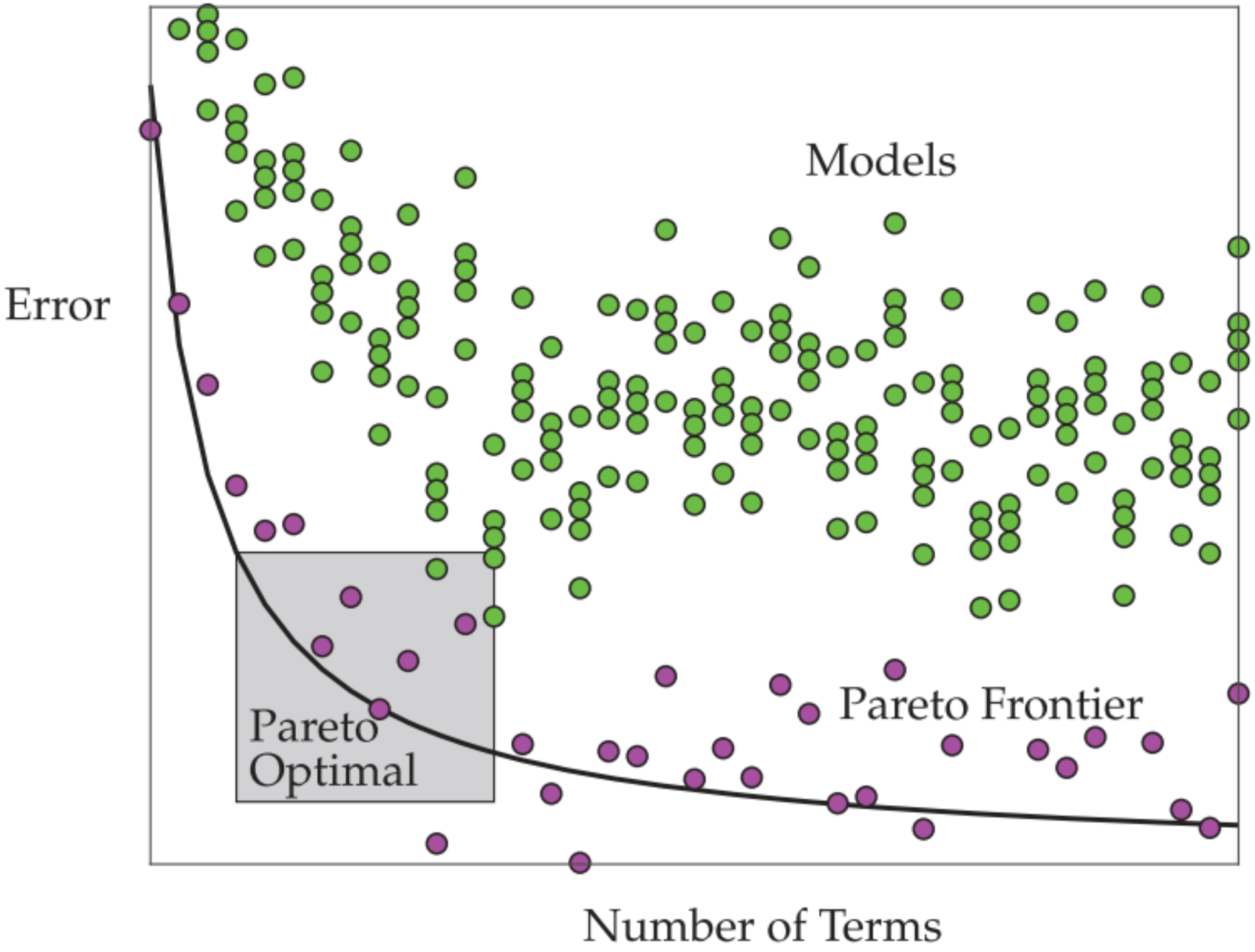


Contents

- Data science and machine learning
- Machine learning architecture
- **Regression model**
  - Linear regression
  - Non-linear regression
  - Over- and Under-Determined Systems
  - Model selection
  - Overfitting

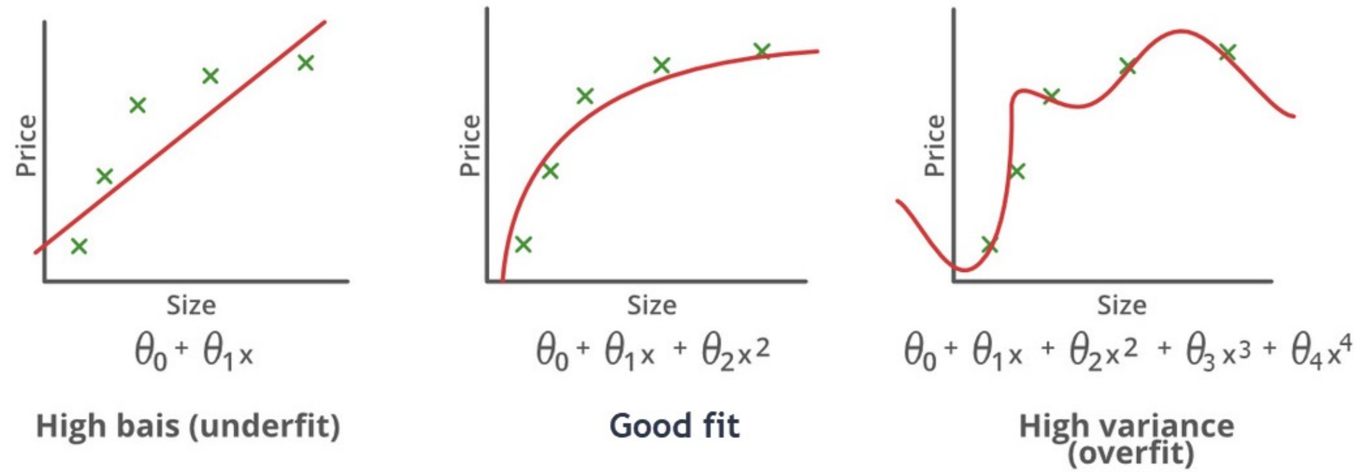
Model Selection

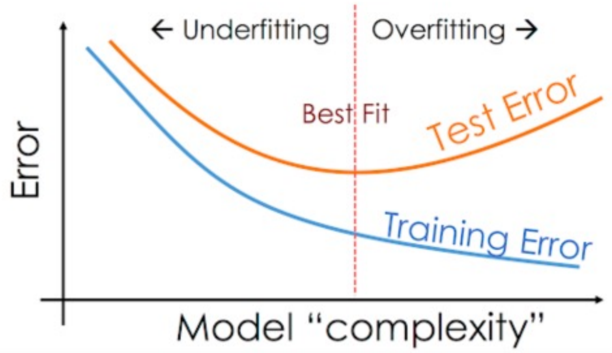
- Model selection is not simply about reducing error, it is about producing a model that has a high degree of interpretability, generalization and predictive capabilities.



Overfitting

- The production is too closely to a particular set of data, and may therefore fail to fit to predict future observations reliably.
  - Overfitting does not allow for generalization.





The End