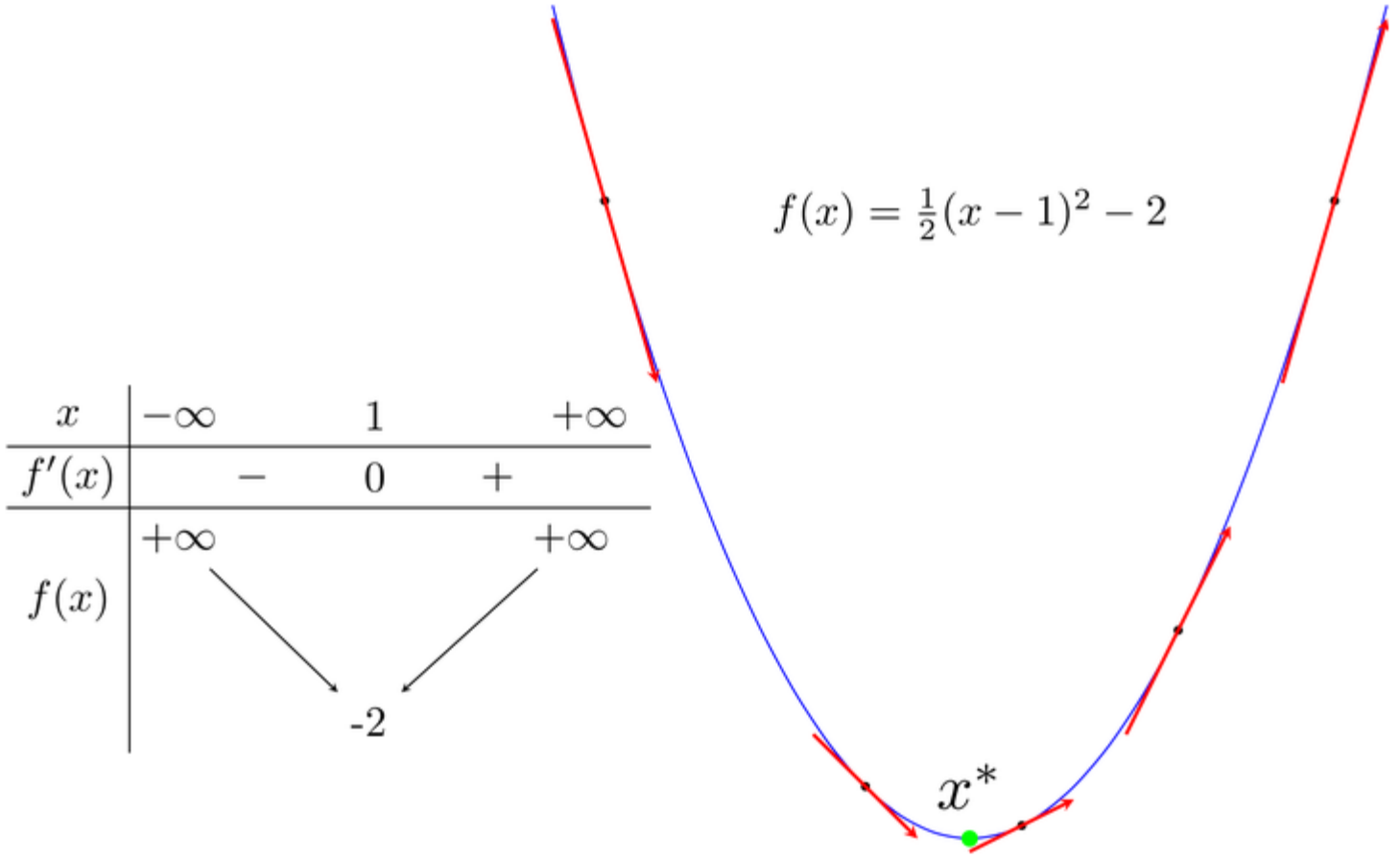
ITF: **Information Technology Faculty**
Thuyloi University

Giảm Gradient

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Bài giảng được dựa trên giáo trình machine learning cơ bản

Giới thiệu



x	$-\infty$		1		$+\infty$
$f'(x)$		$-$	0	$+$	
$f(x)$	$+\infty$	$\searrow$	-2	$\nearrow$	$+\infty$

Giới thiệu

Qui ước:

- *local minimum* cho *điểm cực tiểu*
- *global minimum* cho *điểm mà tại đó hàm số đạt giá trị nhỏ nhất*.
- global minimum là một trường hợp đặc biệt của local minimum.

Giới thiệu

- G/S chúng ta đang quan tâm đến một hàm số một biến có đạo hàm mọi nơi:
  - Điểm local minimum x^* của hàm số là điểm có đạo hàm $f'(x^*)$ bằng 0.
  - Hơn thế nữa, trong lân cận của nó, đạo hàm của các điểm phía bên trái x^* là không dương, đạo hàm của các điểm phía bên phải x^* là không âm.
  - Đường tiếp tuyến với đồ thị hàm số tại 1 điểm bất kỳ có hệ số góc chính bằng đạo hàm của hàm số tại điểm đó.

Gradient Descent

- Trong Machine Learning, chúng ta thường xuyên phải tìm giá trị nhỏ nhất (hoặc đôi khi là lớn nhất) của một hàm số nào đó. Ví dụ như:
  - các hàm mất mát trong hai bài Linear Regression và K-means Clustering.
- Việc tìm global minimum của các hàm mất mát trong Machine Learning là rất phức tạp, thậm chí là bất khả thi.
- Người ta thường cố gắng tìm các điểm local minimum, và ở một mức độ nào đó, coi đó là nghiệm cần tìm của bài toán.

Gradient Descent

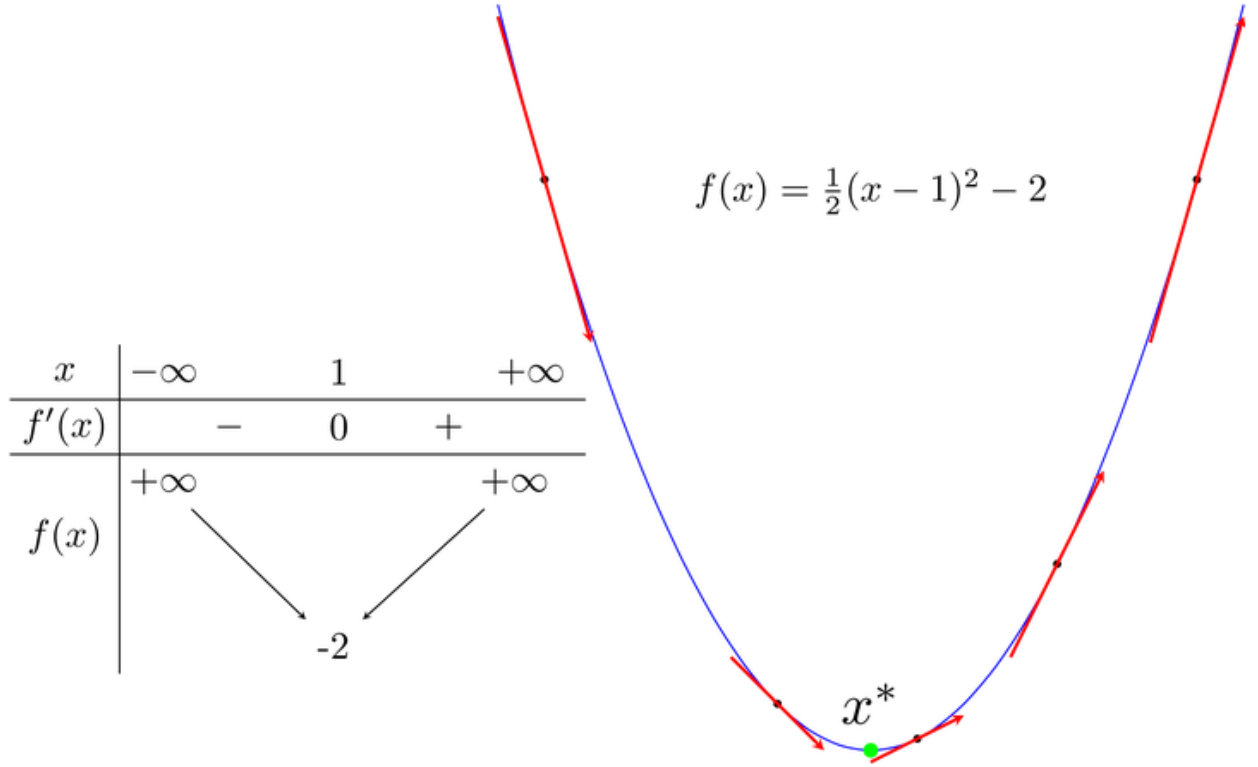
- Các điểm local minimum là nghiệm của phương trình đạo hàm bằng 0.
- Nếu bằng một cách nào đó có thể tìm được toàn bộ (hữu hạn) các điểm cực tiểu, ta chỉ cần thay từng điểm local minimum đó vào hàm số rồi tìm điểm làm cho hàm có giá trị nhỏ nhất.
- Tuy nhiên, trong hầu hết các trường hợp, việc giải phương trình đạo hàm bằng 0 là bất khả thi, nguyên nhân:
  - có thể do sự phức tạp của dạng của đạo hàm,
  - do việc các điểm dữ liệu có số chiều lớn,
  - hoặc do việc có quá nhiều điểm dữ liệu.

Gradient Descent

- Hướng tiếp cận phổ biến nhất là:
  - xuất phát từ một điểm mà chúng ta coi là *gần* với nghiệm của bài toán,
  - sau đó dùng một phép toán lặp để *tiến dần* đến điểm cần tìm, tức đến khi đạo hàm gần với 0.
- Gradient Descent (viết gọn là GD) là một trong những phương pháp được dùng nhiều nhất.

Gradient Descent cho hàm 1 biến

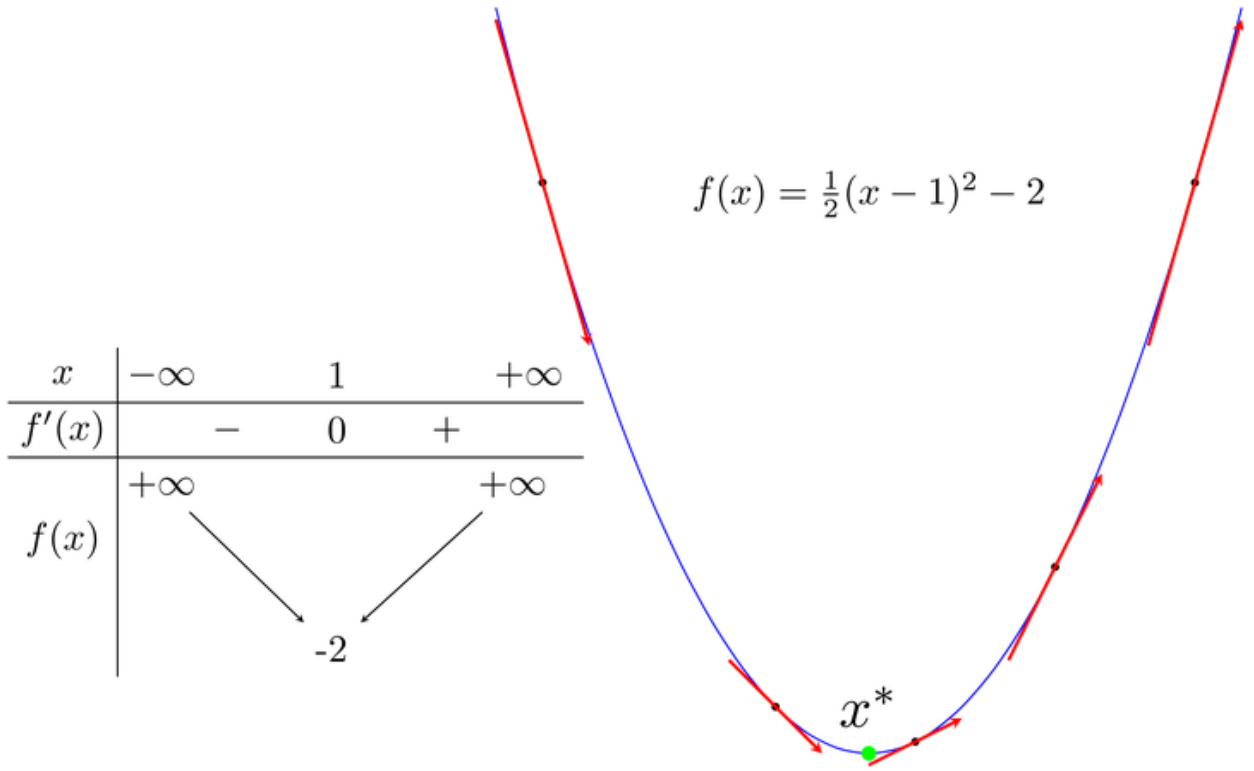
Giả sử x_t là điểm ta tìm được sau vòng lặp thứ t. Ta cần tìm một thuật toán để đưa x_t về càng gần x^* càng tốt.



Gradient Descent cho hàm 1 biến

- Nếu đạo hàm của hàm số tại x_t: $f'(x_t) > 0$ thì x_t nằm phía phải so với x^* và ngược lại
- Để điểm tiếp theo x_{t+1} gần với x^* hơn, ta cần di chuyển x_t về phía trái (phía âm): ngược dấu đạo hàm

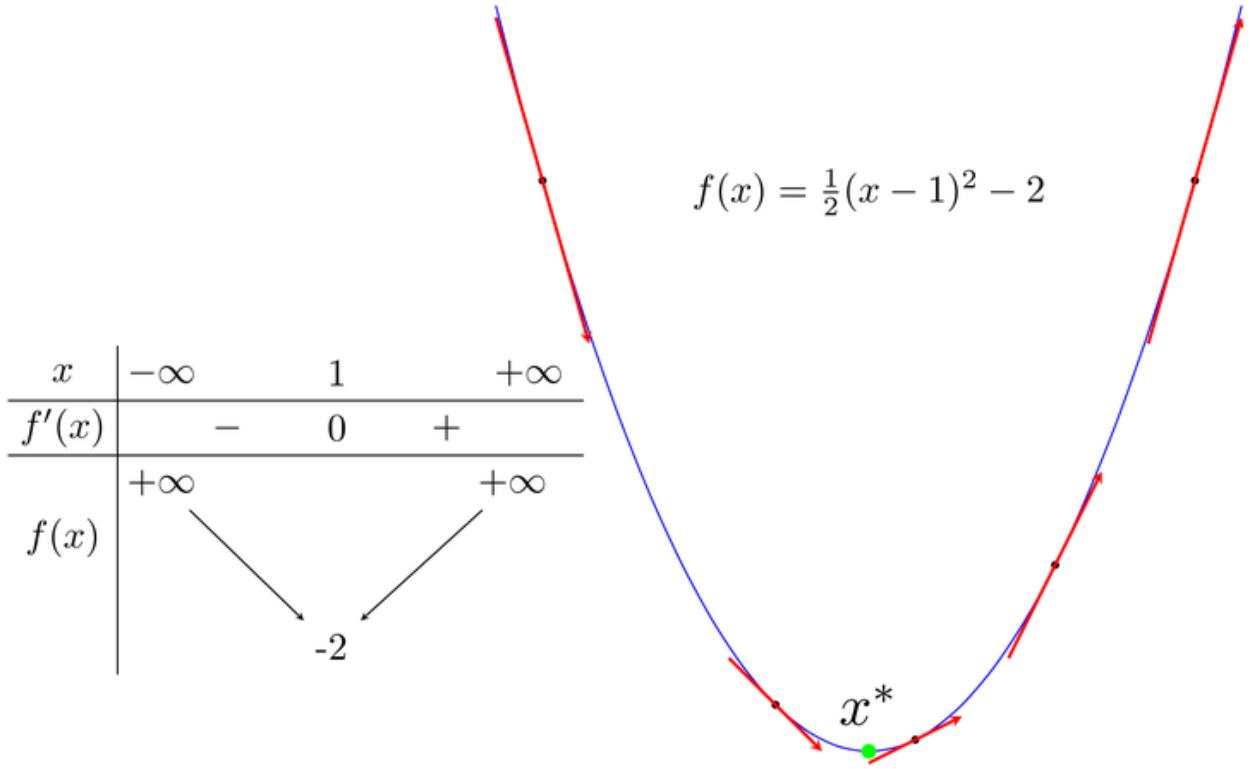
$$x_{t+1} = x_t - f'(x_t)$$



Gradient Descent cho hàm 1 biến

- x_t càng xa x^* về phía phải thì $f'(x_t)$ càng lớn hơn 0 (và ngược lại).
- Lượng di chuyển là $-f'(x_t)$
- Dẫn đến chúng ta có công thức cập nhật:

$$x_{t+1} = x_t - \eta f'(x_t)$$



Gradient Descent cho hàm nhiều biến

- Giả sử ta cần tìm global minimum cho hàm $f(\theta)$ trong đó θ là một vector
- Đạo hàm của hàm số đó tại một điểm θ bất kỳ được ký hiệu là $\nabla_\theta f(\theta)$
- Giống như hàm một biến:
  - Thuật toán GD cho hàm nhiều biến cũng bắt đầu bằng một điểm khởi tạo θ_0
  - Qui tắc cập nhật ở vòng lặp thứ t là: $\theta_{t+1} = \theta_t - \eta\nabla_\theta f(\theta)$ hoặc $\theta = \theta - \eta\nabla_\theta f(\theta)$
  -

Tối ưu hàm mất mát của Linear Regression bằng GD

- Hàm mất mát của Linear Regression là:

$$\mathcal{L}(w) = \frac{1}{2N} \|\mathrm{y} - \mathrm{Xw}\|_2^2$$

- Đạo hàm của hàm mất mát là:

$$\nabla_w \mathcal{L}(w) = \frac{1}{N} \mathrm{X}^T (\mathrm{Xw} - \mathrm{y})$$