Foundations of Machine Learning (ECE 5984)

- Linear Regression and Gradient Descent -

Eunbyung Park

Assistant Professor

School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

Supervised Learning

Setup

$$D = \left\{ \left(x^{(i)}, y^{(i)} \right) \right\}$$

$x^{(i)}$ is the input (feature) vector of the i^{th} sample

$y^{(i)}$ is the label (target) of the i^{th} sample

$x \sim X, y \sim Y$

$h: X \rightarrow Y$

Regression – continuous target variable

Classification – discrete target variable

Setup

- Example

input	target
Living Area (sqft)	**Price ($)**
2000	400K
1500	330K
3700	600K
…	

Setup

- We would like to learn a function $h \in H$, that minimize the loss function L

$$h = \mathrm{argmin}_{h \in H} L(h)$$

- H is the hypothesis class
  - neural networks, linear models, …
- L is the loss function
  - Zero-one loss, squared loss, …

Loss

- Zero-one loss

$$L_{0/1} = \frac{1}{|D|} \sum_{(x,y)\in D} \delta_{h(x)\neq y}\,, \quad \text{where } \delta_{h(x)\neq y} = \begin{cases} 1, & h(x) \neq y \\ 0, & \text{otherwise} \end{cases}$$

- Squared loss (L2 loss)

$$L_{sq} = \frac{1}{|D|} \sum_{(x,y)\in D} (h(x) - y)^2$$

Generalization

- Machine learning is about 'prediction' to the unseen data
- We split the data into three subsets, $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$
  - Training (Learning) on $D_{\text{train}}, D_{\text{val}}$
  - Testing (Evaluation) on D_{test}

Linear Regression

Linear Models

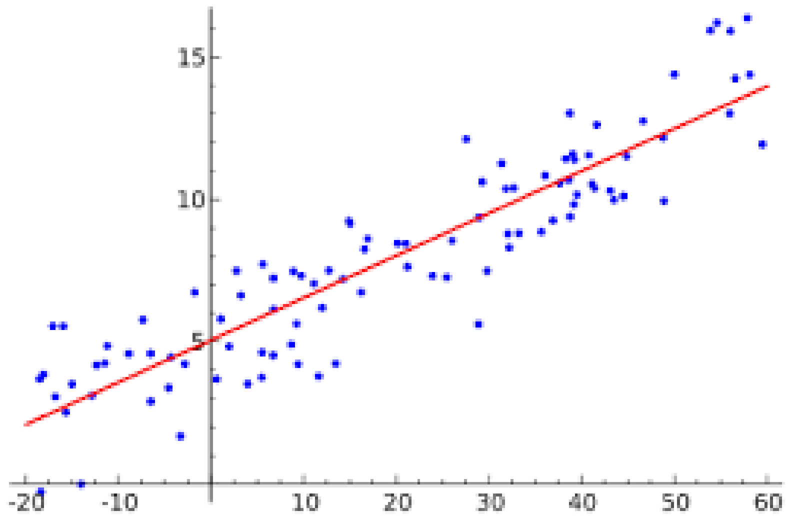
Linear Models

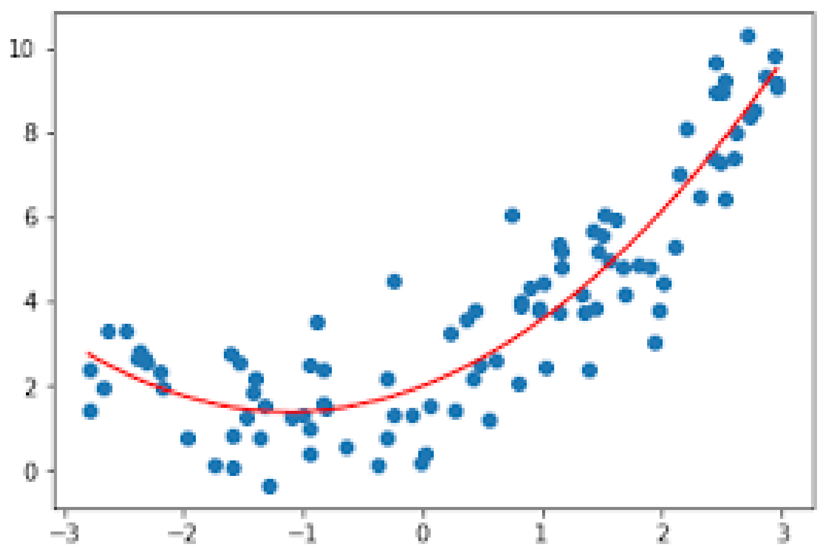
Linear (comes from 'Line')

Mathematical models

Regression

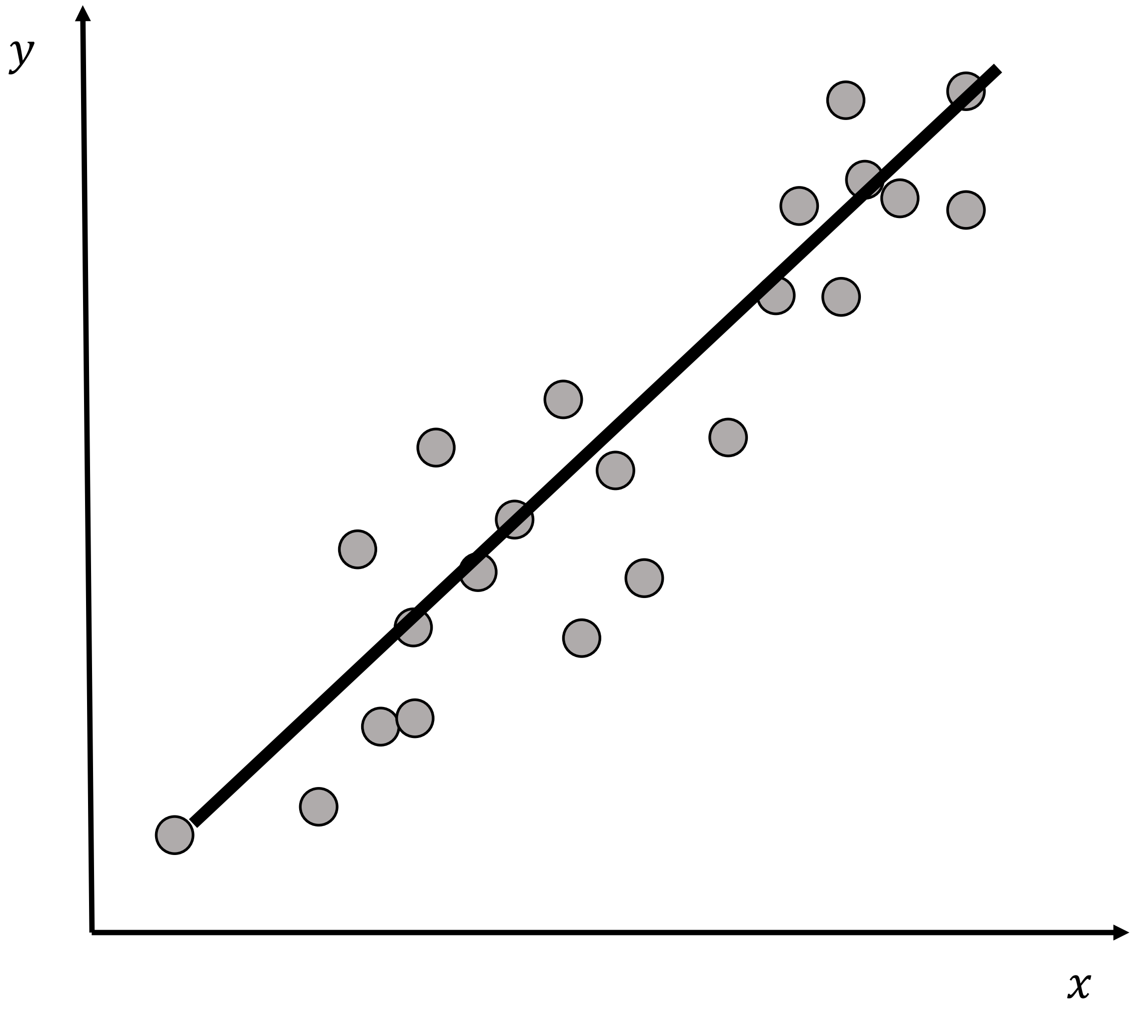
- Regression is a (statistical) method of fitting curves through data points
- The term “regression” was coined by Francis Galton in the 19th century to describe a biological phenomenon.
  - The taller the parents, the taller the children, but shorter than their parents
  - The shorter the parents, the shorter the children, but taller than their parents
  - “regression to the mean”





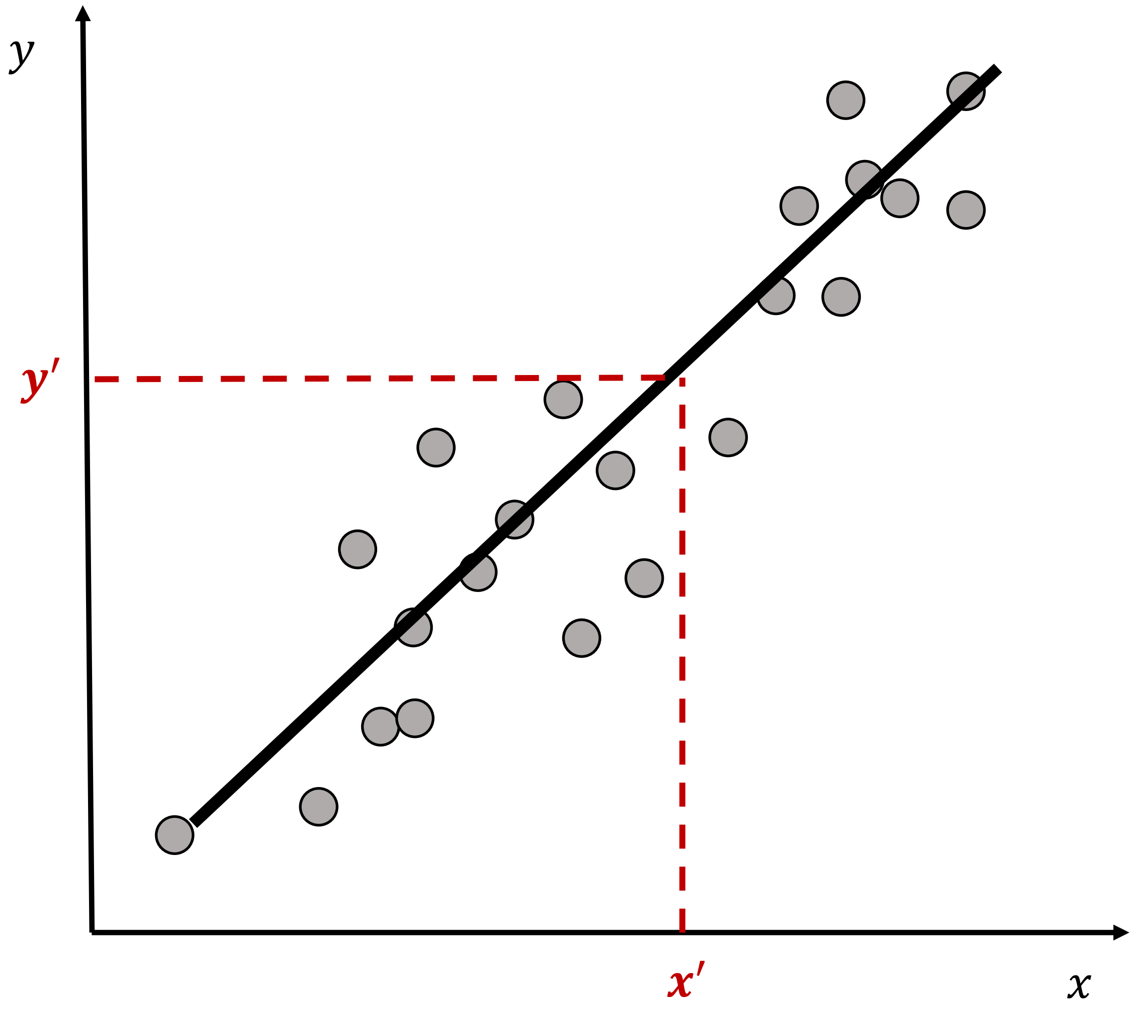
[JMPer Cable Summer 98: Why is it called Regression? (jhsph.edu)]

1D Linear Regression



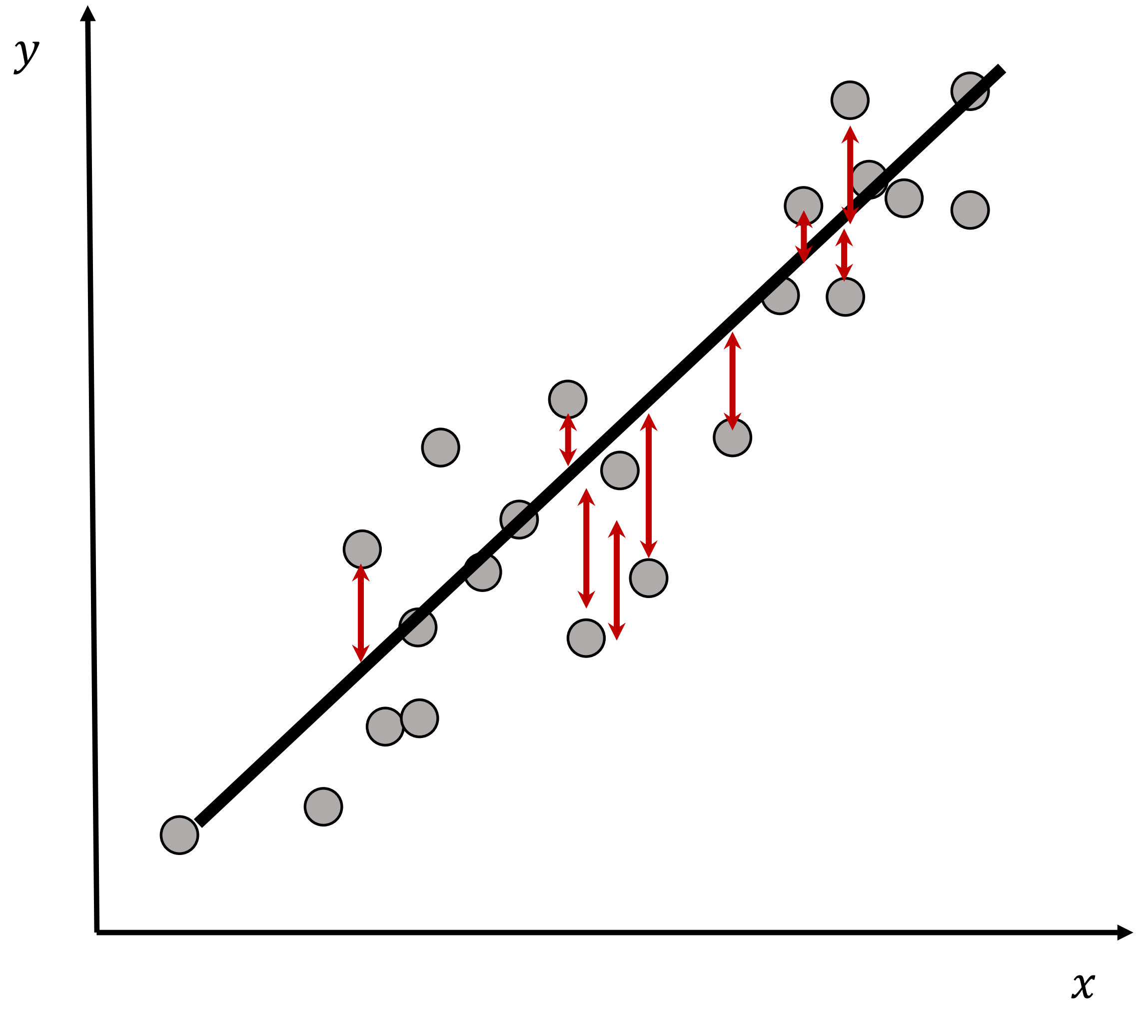
- Fitting a *line* that explains the data

1D Linear Regression



- Fitting a *line* that explains the data
- Given a new data $\boldsymbol{x'}$, predict $\boldsymbol{y'}$

1D Linear Regression



- Fitting a *line* that explains the data $\{(x^{(i)}, y^{(i)})\}$

$$f(x) = wx$$

- What is the best line?
  - A line that is close to all data points 'on average'
  - Mean squared error (MSE) loss

$$w^* = \arg\min_w \frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - wx^{(i)}\right)^2$$

1D Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)^2$$

$$w^* = \arg\min_{w} L(w)$$

- The least squares method
  - L2 Loss function
- N and $\{(x^{(i)}, y^{(i)})\}$ are constants (given), and only w is 'unknown'
- We are going to find w that minimizes the loss function $L(w)$
- Then, how?

1D Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)^2$$

1D Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)^2 = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)}\right)^2 + w^2\left(x^{(i)}\right)^2 - 2wx^{(i)}y^{(i)}$$

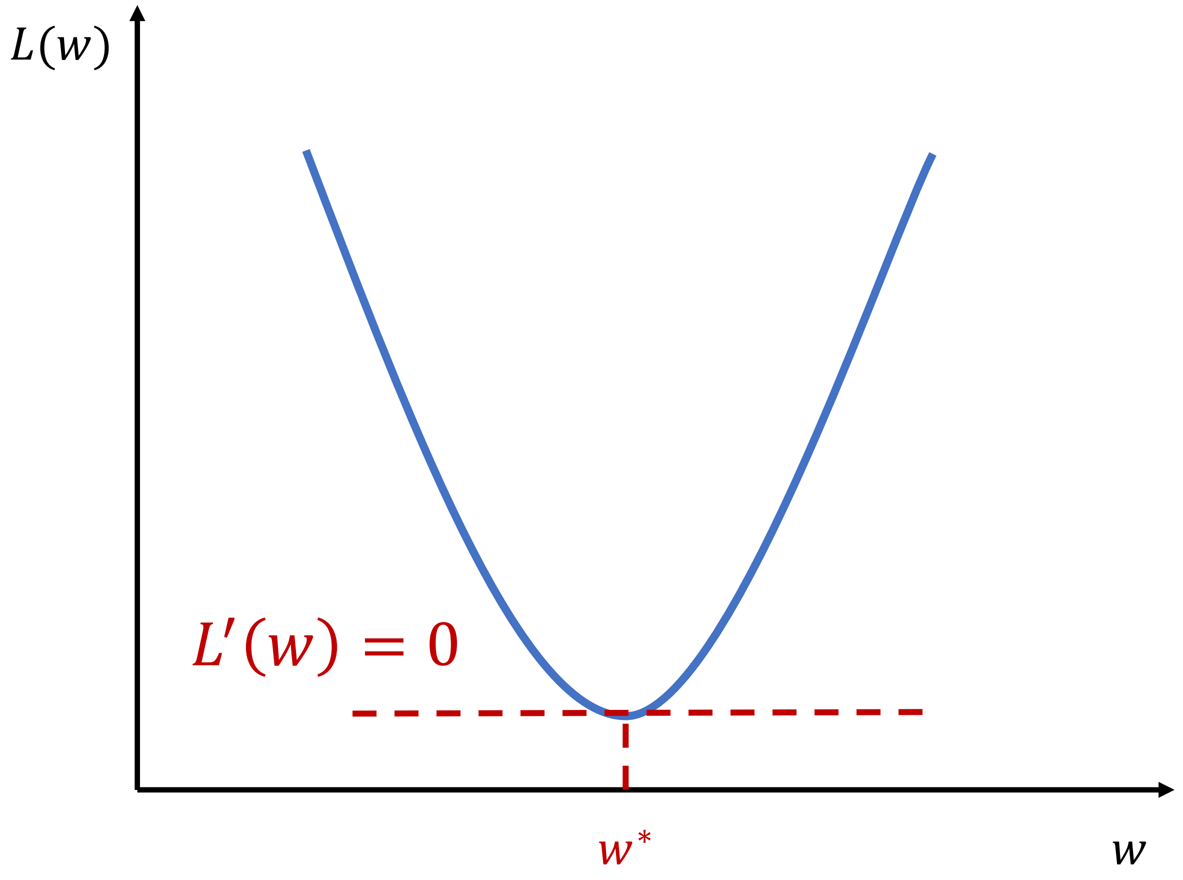
$$= \frac{1}{2}\left(\sum_{i=1}^{N}\left(x^{(i)}\right)^2\right){\color{red}w^2} + \left(\sum_{i=1}^{N}x^{(i)}y^{(i)}\right){\color{red}w} + \frac{1}{2}\left(\sum_{i=1}^{N}\left(y^{(i)}\right)^2\right)$$

$L(w)$ is a quadradic function

How to minimize a quadratic function?

1D Linear Regression

- Minimizing a quadratic function
  - Take a derivative, and set it to zero



Does it have a solution?

If so, is it the unique solution?

1D Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)^2$$

1D Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)^2$$

$$L'(w) = \frac{dL(w)}{dw} = \sum_{i=1}^{N}\left(y^{(i)} - wx^{(i)}\right)x^{(i)} = 0$$

$$w^* = \frac{\sum_1^N x^{(i)}y^{(i)}}{\sum_1^N (x^{(i)})^2}$$

Multivariable Calculus

Derivative

- The rate of change of a function with respect to a variable
- $f'(x) > 0$, what does it mean?
- $f'(x) < 0$, what does it mean?

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Partial Derivative

- A partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant
- It represents the instantaneous rates of change of the function f w.r.t one of its variables
  - $\frac{\partial f}{\partial x_i}$: how much f changes as x_i change while fixing other components at any given point

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

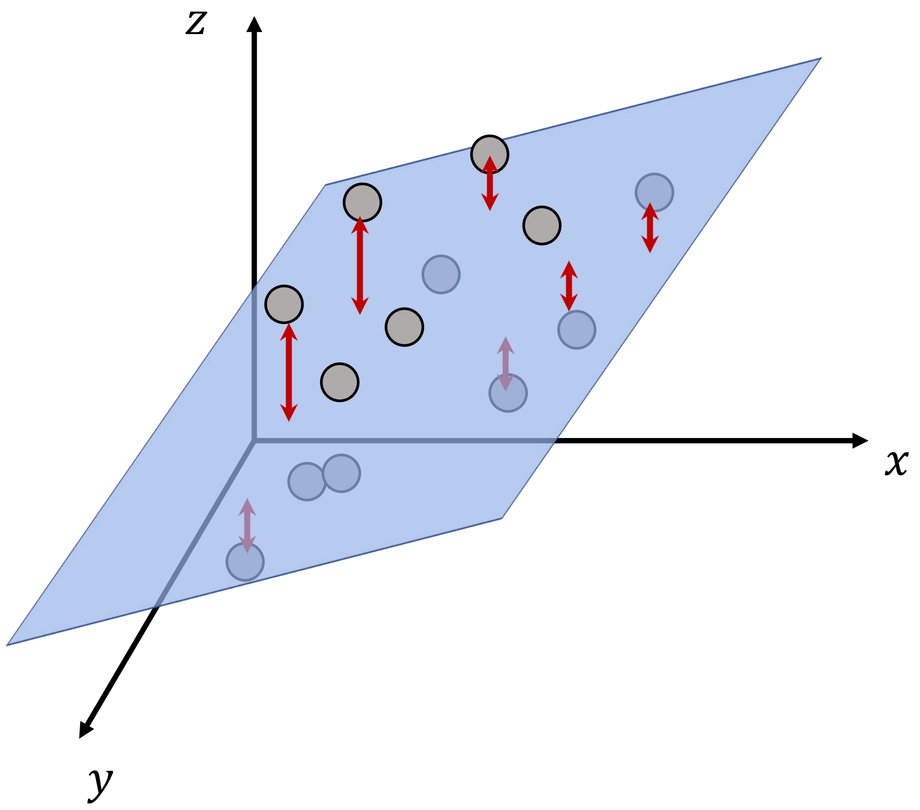
Gradient

- The gradient stores all the partial derivative information of a multivariable function

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

2D Linear Regression

2D Linear Regression



$$z = w_2 x + w_1 y$$

$$L(w_1, w_2) = \frac{1}{2} \sum_{i=1}^{N} \left(z^{(i)} - w_2 x^{(i)} - w_1 y^{(i)} \right)^2$$

2D Linear Regression

$$z = w_2 x + w_1 y$$

$$L(w_1, w_2) = \frac{1}{2}\sum_{i=1}^{N}\left(z^{(i)} - w_2 x^{(i)} - w_1 y^{(i)}\right)^2$$

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{N}\left(z^{(i)} - w_2 x^{(i)} - w_1 y^{(i)}\right)\left(-y^{(i)}\right) = 0$$

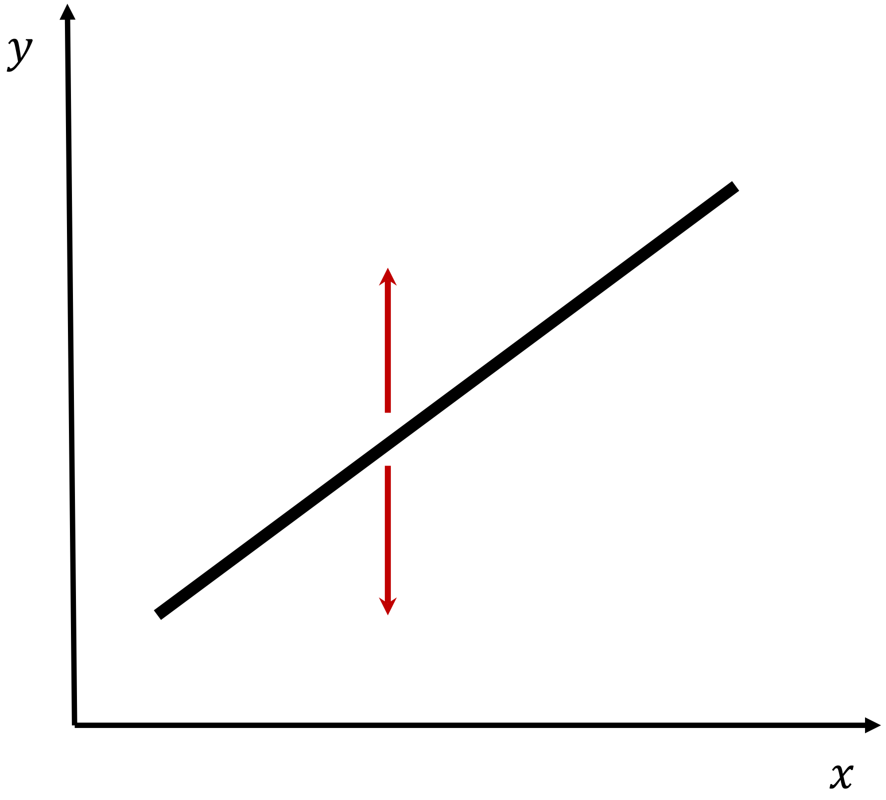
$$\frac{\partial L}{\partial w_2} = \sum_{i=1}^{N}\left(z^{(i)} - w_2 x^{(i)} - w_1 y^{(i)}\right)\left(-x^{(i)}\right) = 0$$

$$w_1 = \frac{\sum_{i=1}^{N} y^{(i)} z^{(i)} - w_2 \sum_{i=1}^{N} x^{(i)} y^{(i)}}{\sum_{i=1}^{N} (y^{(i)})^2}$$

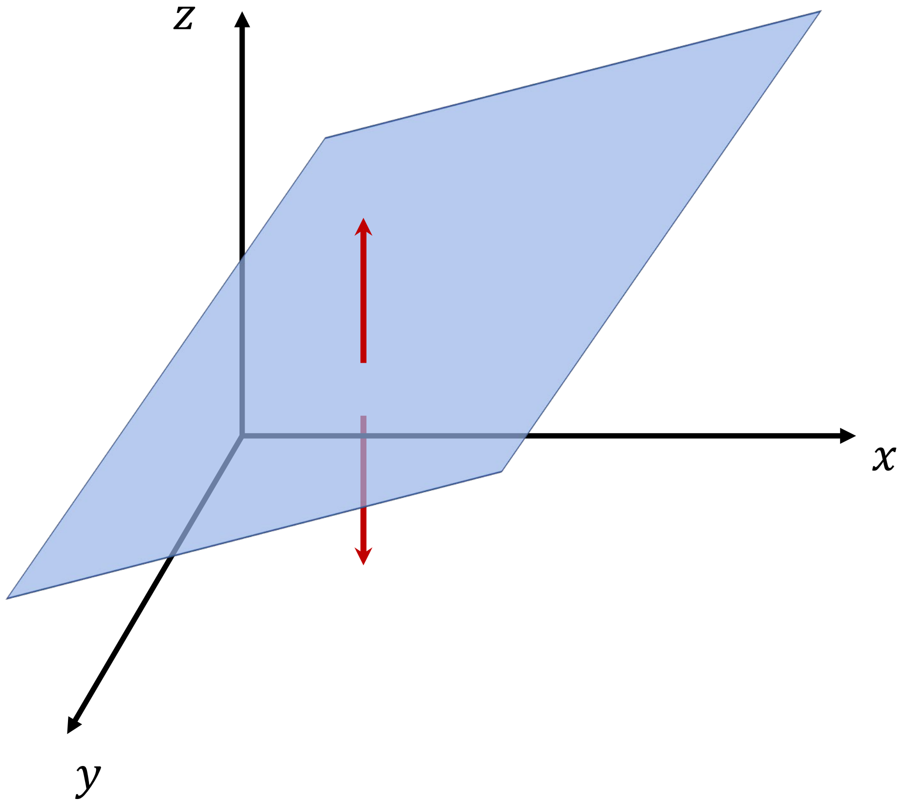
$$w_2 = \frac{\sum_{i=1}^{N} x^{(i)} z^{(i)} - w_1 \sum_{i=1}^{N} x^{(i)} y^{(i)}}{\sum_{i=1}^{N} (x^{(i)})^2}$$

Bias term and Higher Dimension

$$y = w_1 x + w_0$$



$$z = w_2 x + w_1 y + w_0$$



Linear Algebra Review

Basic Concepts

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \\ -2x_1 + 3x_2 &= 9 \end{aligned}$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \qquad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

$$Ax = b$$

Basic Notation

$$A \in \mathbb{R}^{m \times n}$$

$$x \in \mathbb{R}^{n}$$

Inner Products

$$x, y \in \mathbb{R}^n$$

$$x^\top y \in \mathbb{R}$$

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

Outer Products

$$x, y \in \mathbb{R}^n$$

$$xy^\top \in \mathbb{R}^{n \times n}$$

Matrix Vector Products

$A \in \mathbb{R}^{m \times n}$

$x \in \mathbb{R}^n$

$Ax \in$

Matrix Matrix Products

$$A \in \mathbb{R}^{m \times n}$$

$$AB \in$$

$$B \in \mathbb{R}^{n \times p}$$

Matrix Matrix Products

- Associative
  - $(AB)C = A(BC)$
- Distributive
  - $A(B + C) = AB + AC$
- Not commutative
  - $AB \neq BA$

Identity Matrix and Diagonal Matrices

$$AI = A = IA$$

$$D = \mathrm{diag}(d_1, d_2, \ldots, d_n) =$$

The Transpose

$$(A^\top)_{ij} = A_{ji}$$

$$(A^\top)^\top = A$$
$$(AB)^\top = B^\top A^\top$$
$$(A+B)^\top = A^\top + B^\top$$

Norms

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^\top x}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_\infty = \max_i |x_i|$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad \|A\|_F = \sqrt{\sum_{i=1}^{n} A_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)}$$

Linear Independence and Rank

- A set of vectors $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^m$ is said to be *linearly independent* if no vector can be represented as a linear combination of the remaining vectors

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i \qquad \text{(linearly dependent)}$$

- Geometrical interpretation

Linear Independence and Rank

- The *column rank* of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of *columns* that constitute a linearly independent set
- The *row rank* of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of *rows* that constitute a linearly independent set
- For any matrix $A \in \mathbb{R}^{m \times n}$ the *column rank* is equal to the *row rank*, so both quantities are referred to collectively as the *rank of A*.
- For $\mathrm{A} \in \mathbb{R}^{m \times n}$, $\mathrm{rank}(A) \leq \min(m, n)$. If $\mathrm{rank}(A) = \min(m, n)$, then A is said to be *full rank*

The Inverse of a Square Matrix

- The inverse of a square matrix
  - Non-square matrices do not have inverses by definition
  - A^{-1} may not exist: non-invertiable or singular (not full rank)

$$A^{-1}A = I = AA^{-1}$$

$$(A^{-1})^{-1} = A$$
$$(AB)^{-1} = B^{-1}A^{-1}$$
$$(A^{-1})^{\top} = (A^{\top})^{-1} = A^{-\top}$$

- For standard linear system, $Ax = b, x = A^{-1}b$
- What if A is not square?

Orthogonal Matrices

- If all its columns are orthogonal to each other and are normalized

$$x^\top y = 0$$ (orthogonal)

$$\|x\|_2 = 1$$ (normalized)

$$U^\top U = I = UU^\top$$ (orthogonal)

$$U^\top = U^{-1}$$

Quadratic Forms

- Given a square matrix $A \in \mathbb{R}^{n\times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^\top Ax$ is a quadratic form

$$x^\top Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

Positive Semidefinite Matrices

- A symmetric matrix $A \in \mathbb{S}^n$ is *positive definite (PD)*, a.k.a $\mathbb{S}^n_{++}$
  - $x^\top A x > 0,\qquad$ for all non-zero vectors $x \in \mathbb{R}^n$
- A symmetric matrix $A \in \mathbb{S}^n$ is *positive semidefinite (PSD)*, a.k.a $\mathbb{S}^n_{+}$
  - $x^\top A x \geq 0,\qquad$ for all non-zero vectors $x \in \mathbb{R}^n$
- A symmetric matrix $A \in \mathbb{S}^n$ is *negative definite (ND)*
  - $x^\top A x < 0,\qquad$ for all non-zero vectors $x \in \mathbb{R}^n$
- A symmetric matrix $A \in \mathbb{S}^n$ is *seminegative definite (ND)*
  - $x^\top A x \leq 0,\qquad$ for all non-zero vectors $x \in \mathbb{R}^n$
- A symmetric matrix $A \in \mathbb{S}^n$ is *indefinite*
  - If there exists $x, y \in \mathbb{R}^n$ such that $x^\top A x > 0$ and $y^\top A y \leq 0$

Positive Semidefinite Matrices

- Positive definite matrices (or negative definite) are always full rank, invertible
- Prove by contradiction

$$a_j = \sum_{i \neq \mathrm{j}} x_i a_i \qquad \text{(linearly dependent)}$$

If $x_{\mathrm{j}} = -1$, then $Ax = 0$, so $x^\top A x = 0$

Gram Matrix

- For any matrix $A \in \mathbb{R}^{m\times n}$, gram matrix is symmetric
- And, *always positive semidefinite*

$$G = A^\top A$$

$$x^\top G x = \sum_{i=1}^{n}\sum_{j=1}^{n} G_{ij}x_i x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i^\top a_j x_i x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} (x_i a_i)^\top (x_j a_j)$$
$$= \left(\sum_{i=1}^{n} x_i a_i\right)^\top \left(\sum_{j=1}^{n} x_j a_j\right) = \left\|\sum_{i=1}^{n} x_i a_i\right\|^2 \geq 0$$

Linear Regression (High-dim)

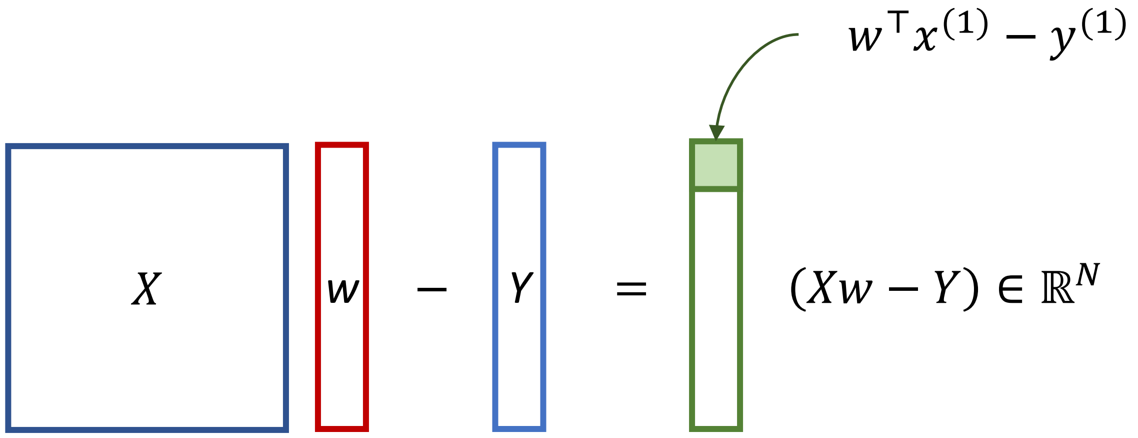
Linear Algebra

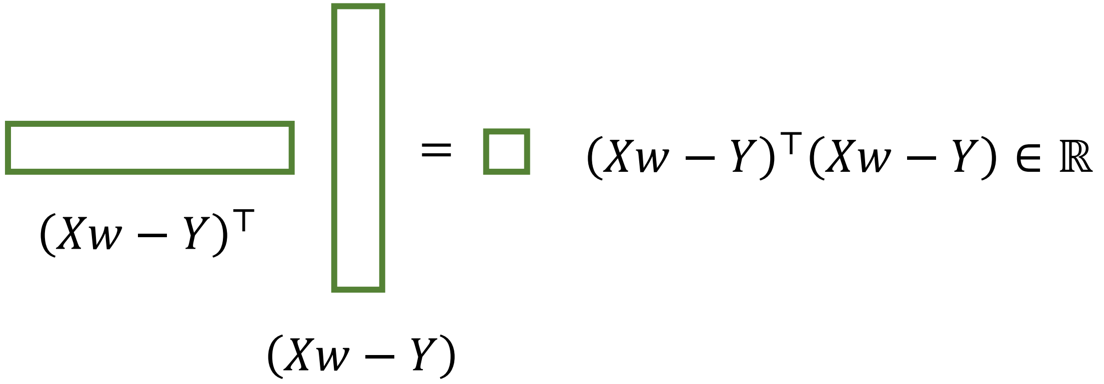
- Linear algebra comes to the rescue!
- Problem setup

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$$
$$x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d$$
$$X \in \mathbb{R}^{N \times d}, Y \in \mathbb{R}^N$$

$$L(w) = \frac{1}{2}(Xw - Y)^\top (Xw - Y)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$





Linear Algebra

- Linear algebra comes to the rescue!
- Problem setup

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$$
$$x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d$$
$$X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^N$$

$$L(w) = \frac{1}{2}(Xw - Y)^\top (Xw - Y)$$

Linear Algebra

- Linear algebra comes to the rescue!
- Problem setup

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(N)}, y^{(N)}\right)\right\}$$
$$x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d$$
$$X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^N$$

$$L(w) = \frac{1}{2}(Xw - Y)^\top (Xw - Y)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

$$\arg\min_w L(w)$$

$$L(w) = \frac{1}{2}(w^\top X^\top X w - w^\top X^\top Y - Y^\top X w + Y^\top Y)$$

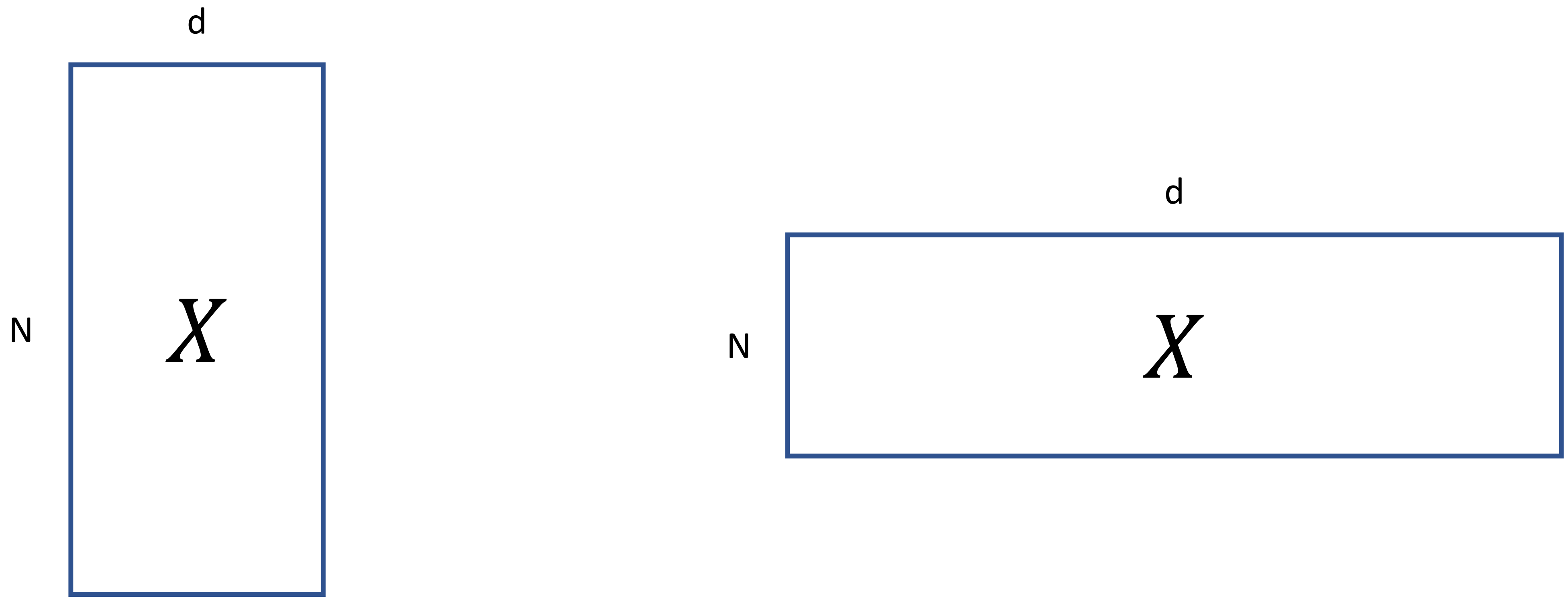
$$= \frac{1}{2}(w^\top X^\top X w - 2w^\top X^\top Y + Y^\top Y)$$

$$\nabla_w L(w) = \frac{1}{2}(2X^\top X w - 2X^\top Y) = -X^\top Y + X^\top X w = 0$$

$$w^* = (X^\top X)^{-1} Y^\top X = \boxed{(X^\top X)^{-1} X^\top} Y$$

(pseudo-inverse) (normal equation)

Overdetermined vs Underdetermined



(over-determined)

(under-determined)

What’s Wrong with It?

$$w^* = (X^\top X)^{-1} X^\top Y$$

1. Invertible?
2. When d is large?
3. Accuracy?

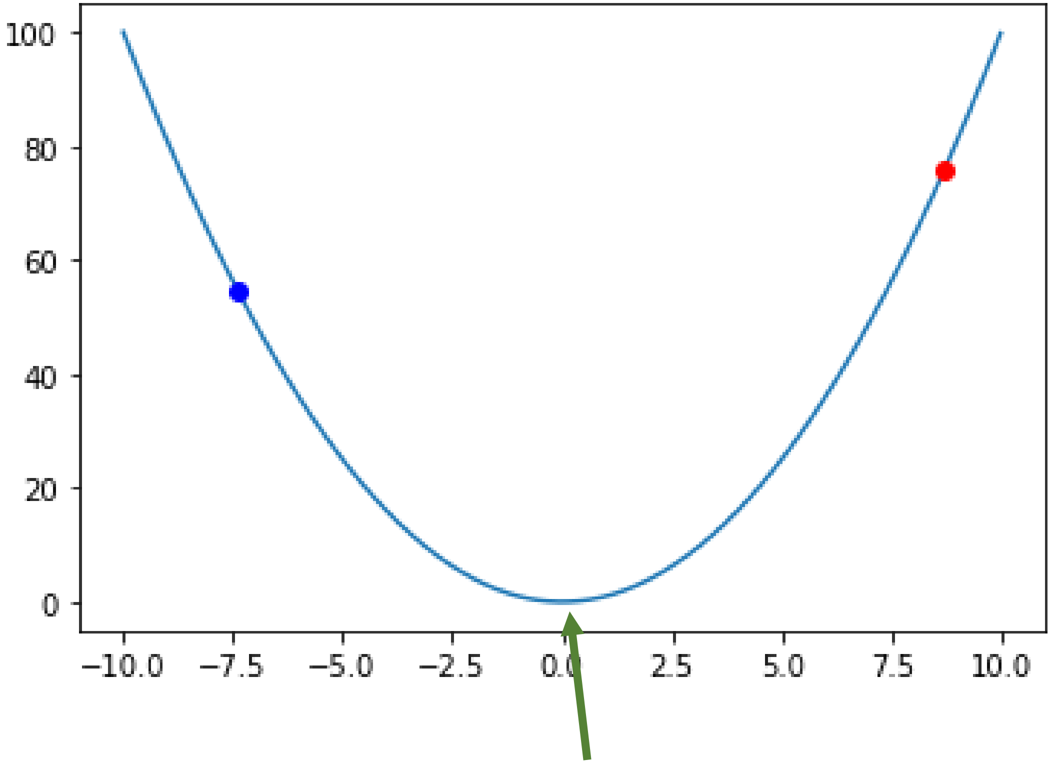
Gradient Descent

Derivative?

- The rate of change of a function with respect to a variable
- $f'(x) > 0$, what does it mean?
- $f'(x) < 0$, what does it mean?
- Our purpose is to minimize a function (loss function) w.r.t model parameter
  - $L(\theta)$

1D Gradient Descent

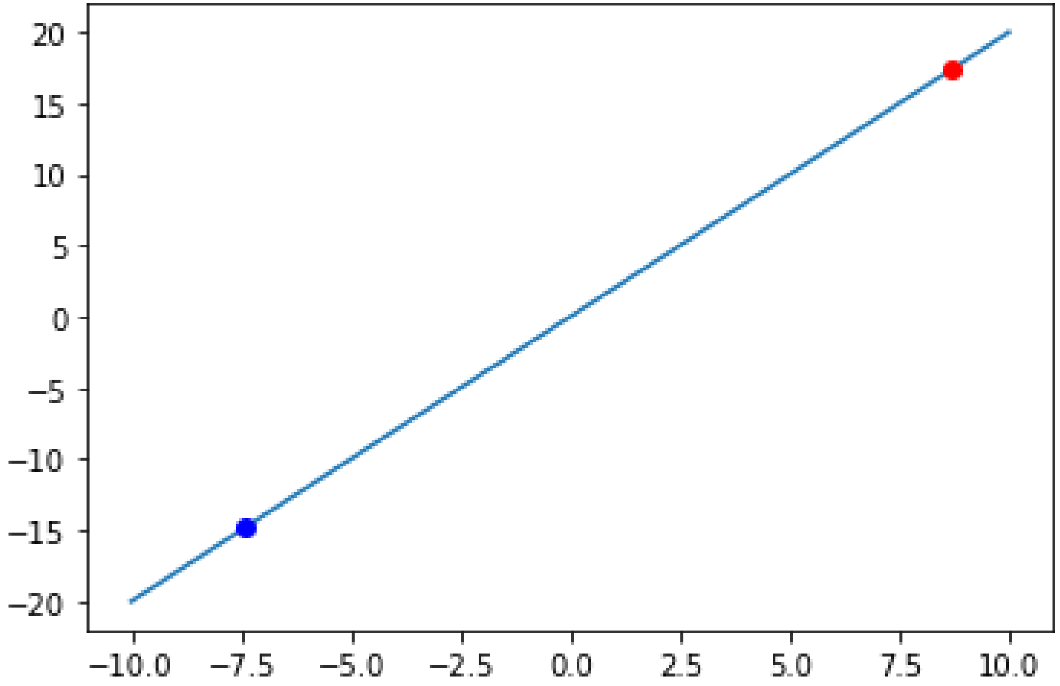
$$f(x) = x^2$$



$$x^* = 0 = \arg\min_x f(x)$$

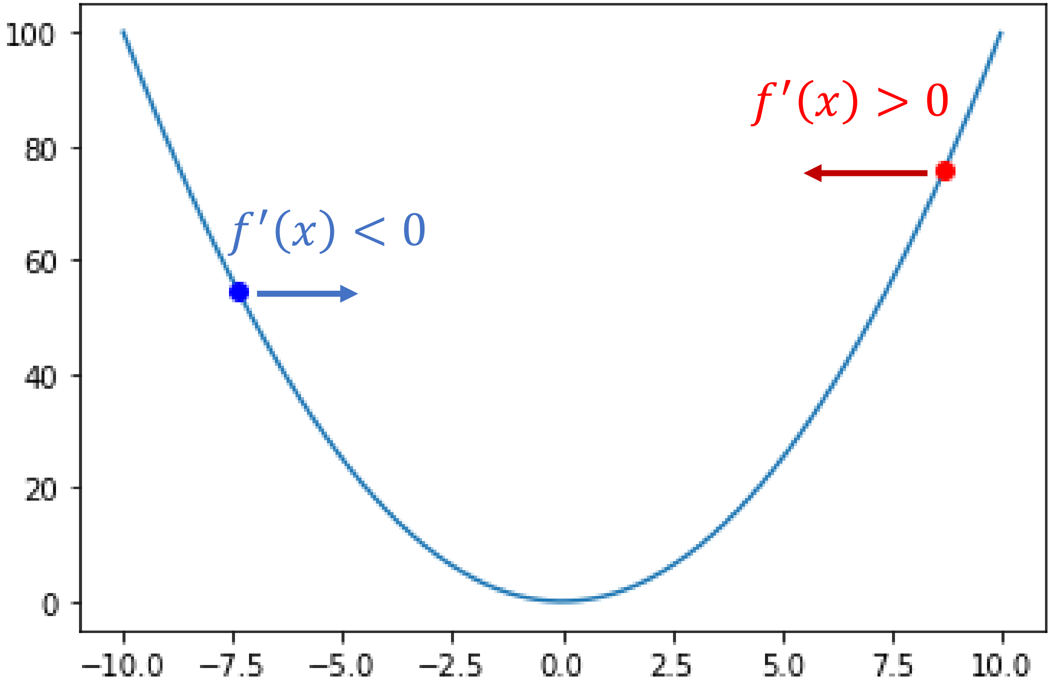
$$f(x^*) = 0$$

$$f'(x) = 2x$$

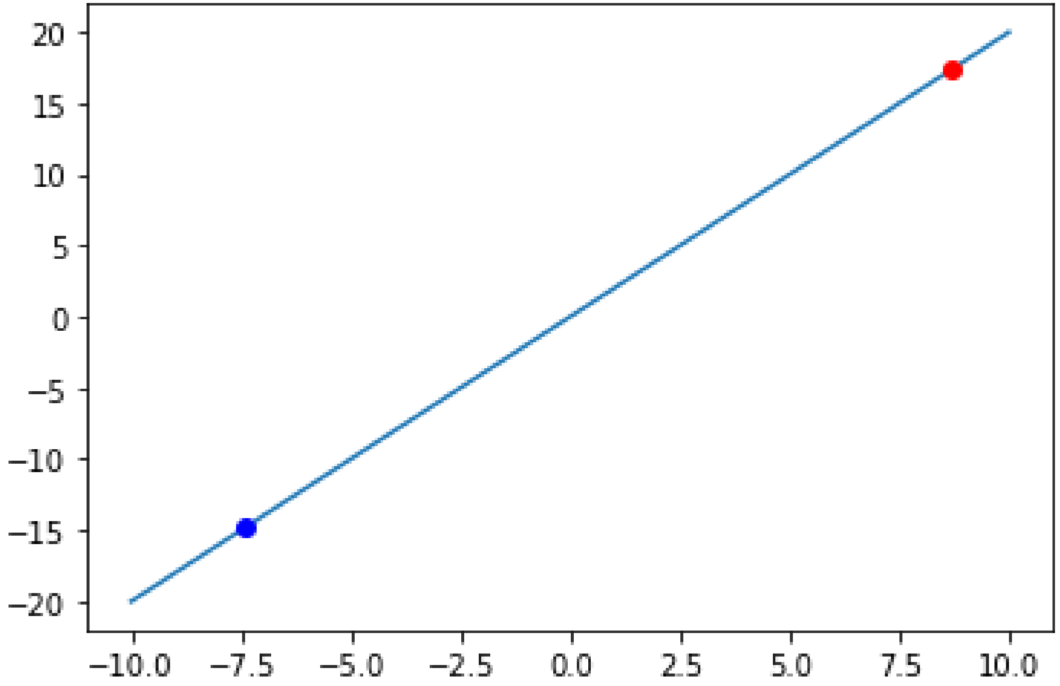


1D Gradient Descent

$$f(x) = x^2$$



$$f'(x) = 2x$$



$$x \leftarrow x - \alpha f'(x)$$

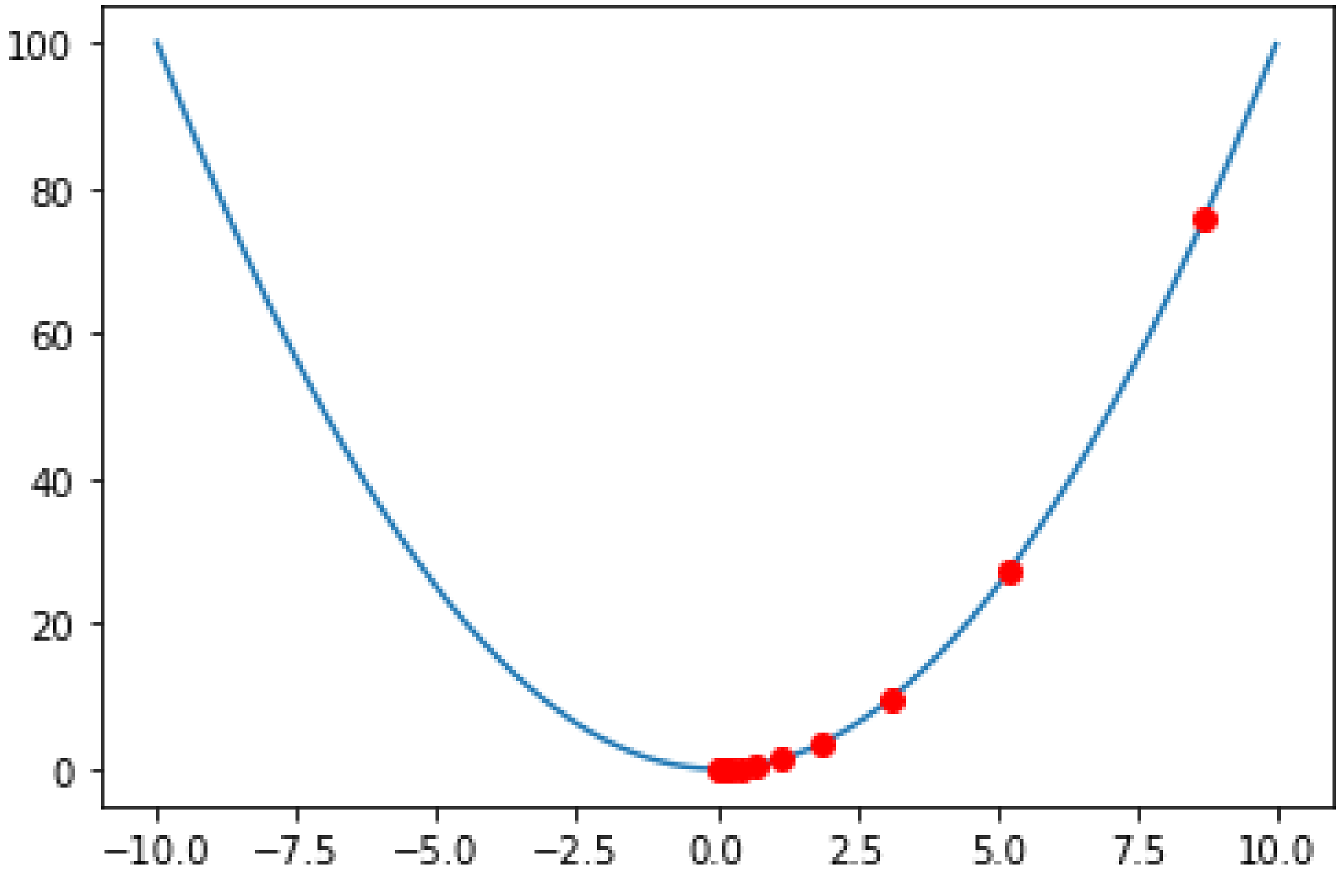
1D Gradient Descent

$$f(x) = x^2$$

$$f'(x) = 2x$$

$$x_0 = 8.7, \alpha = 0.2$$

$$x \leftarrow x - \alpha f'(x)$$



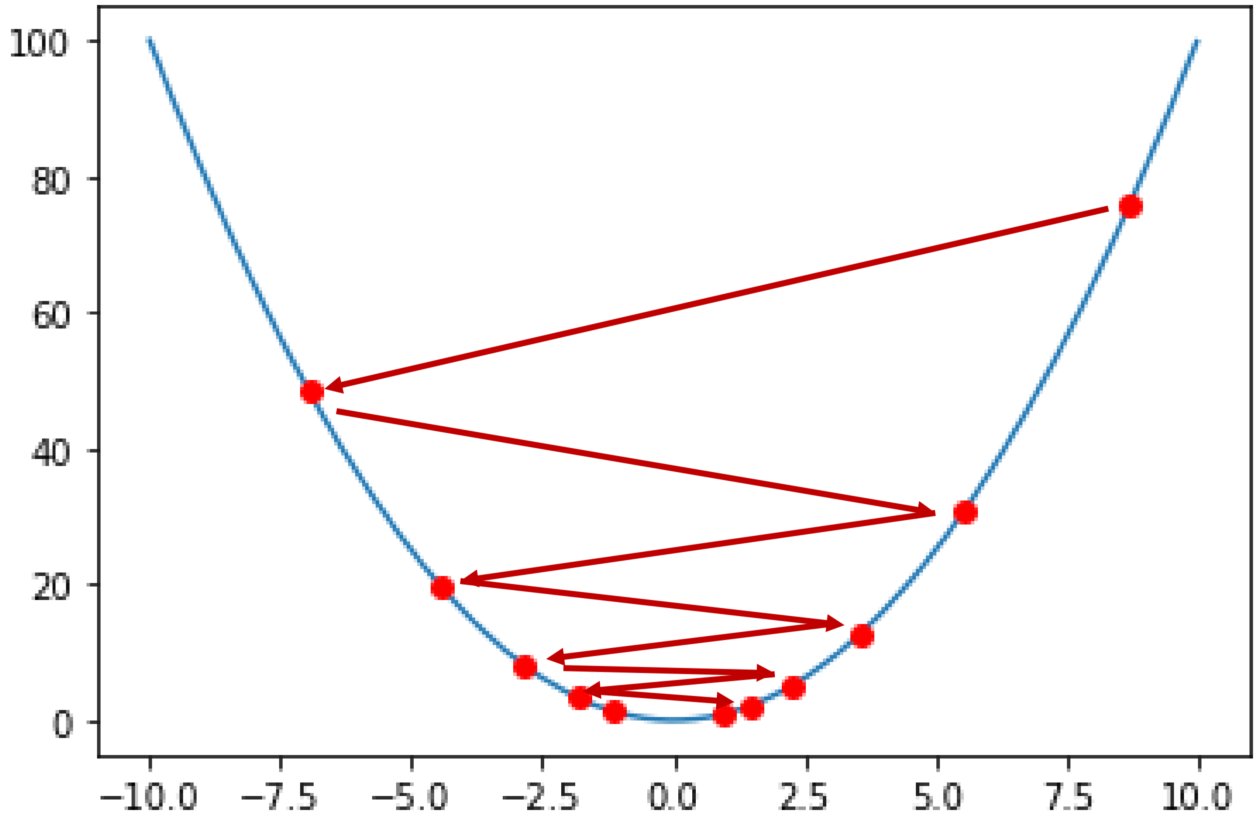
1D Gradient Descent

$$f(x) = x^2$$

$$f'(x) = 2x$$

$$x_0 = 8.7, \alpha = 0.9$$

$$x \leftarrow x - \alpha f'(x)$$



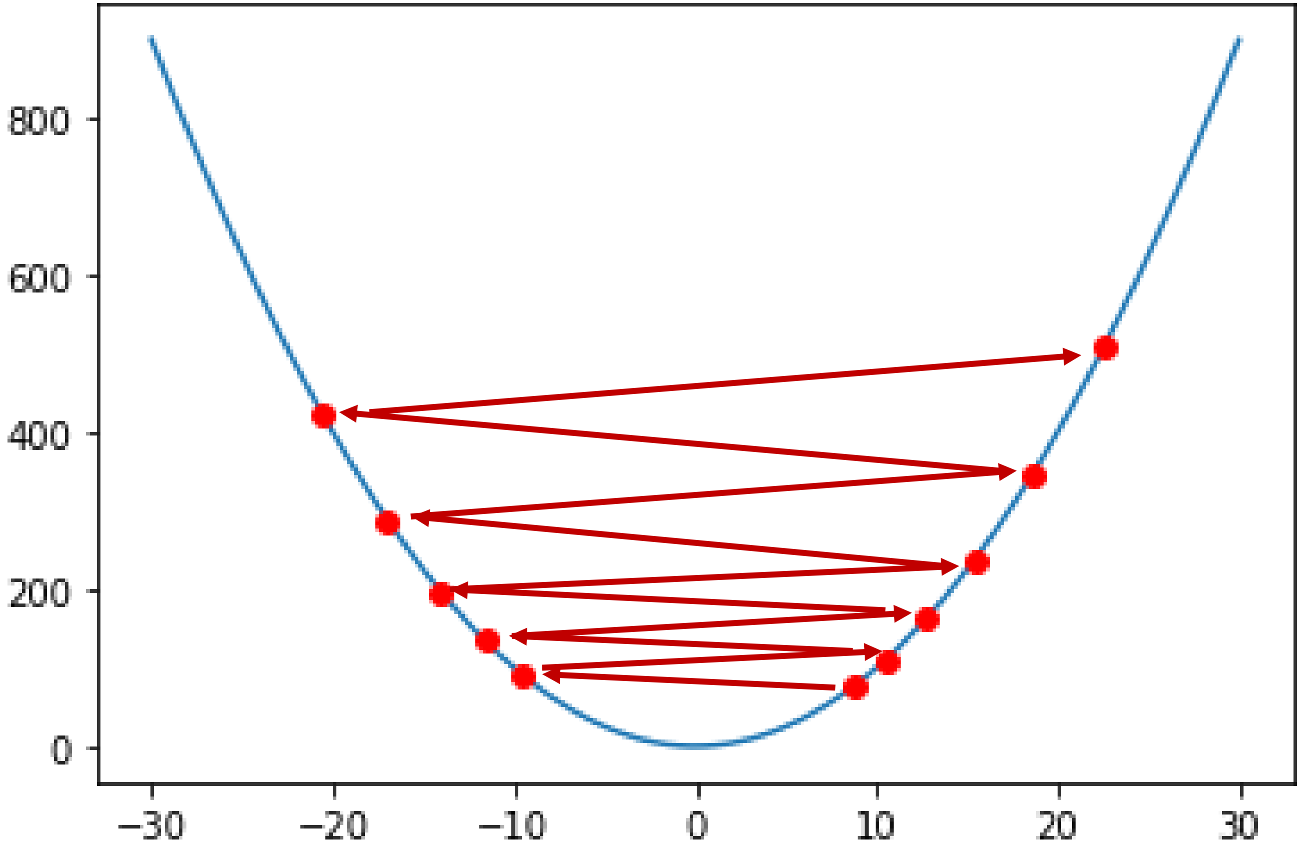
1D Gradient Descent

$$f(x) = x^2$$

$$f'(x) = 2x$$

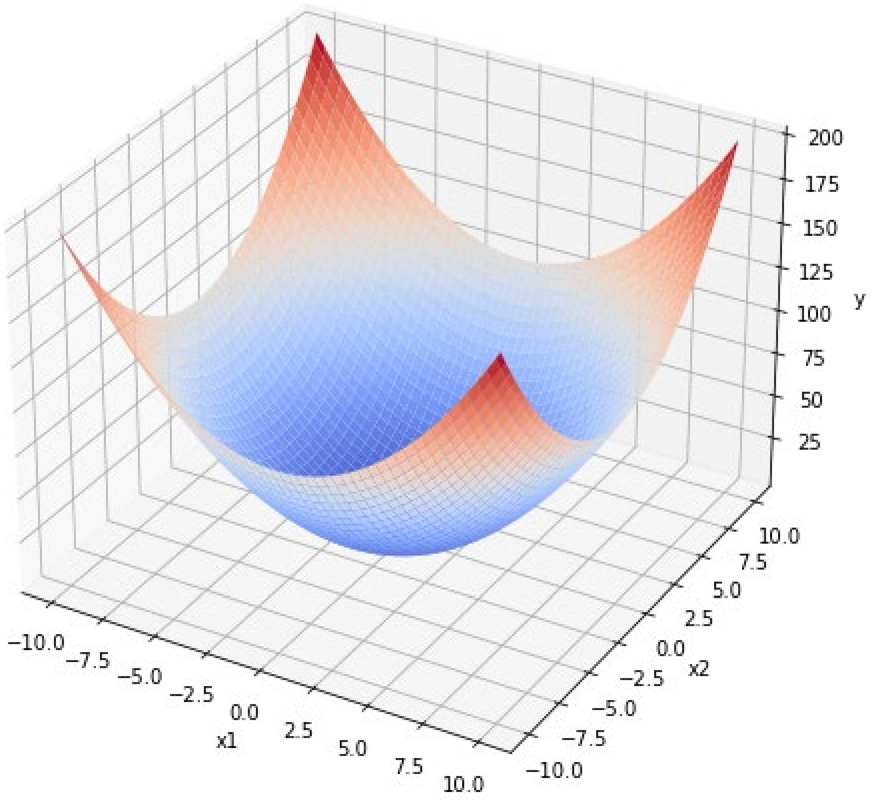
$$x_0 = 8.7, \alpha = 1.05$$

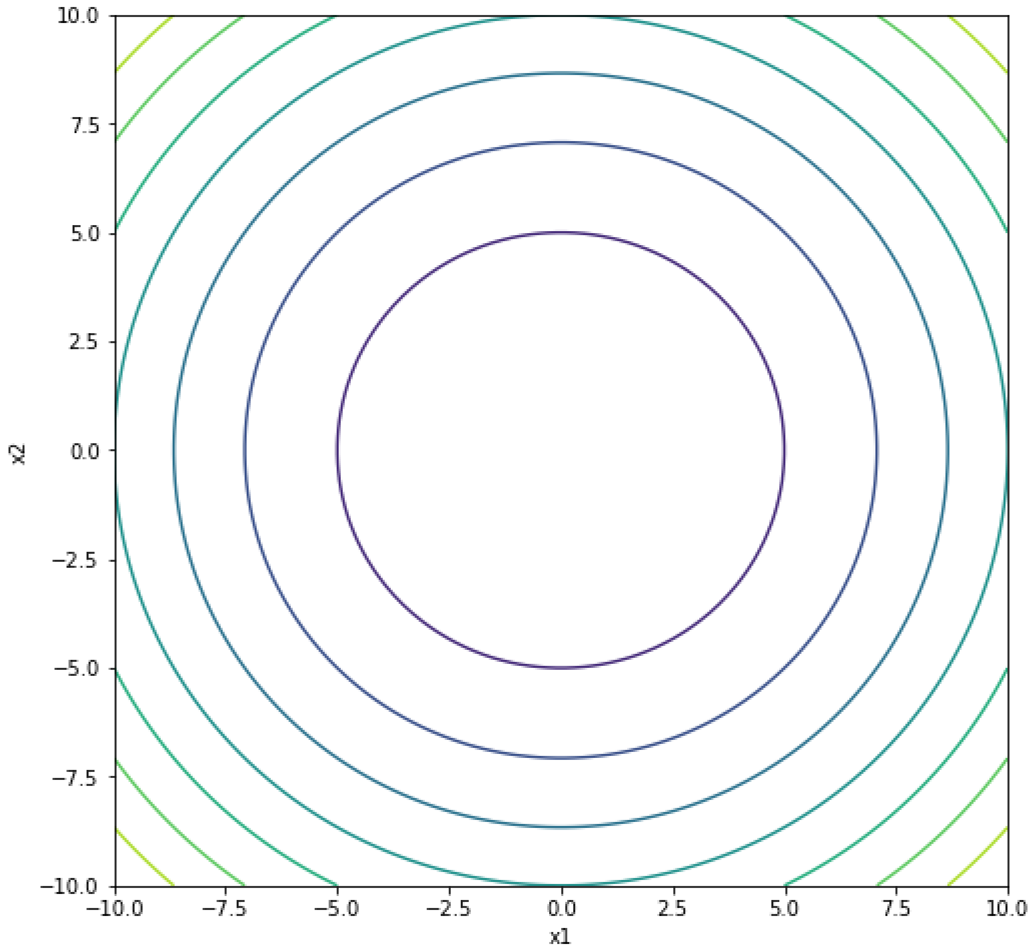
$$x \leftarrow x - \alpha f'(x)$$



2D Gradient Descent

$$f(x_1, x_2) = x_1^2 + x_2^2$$



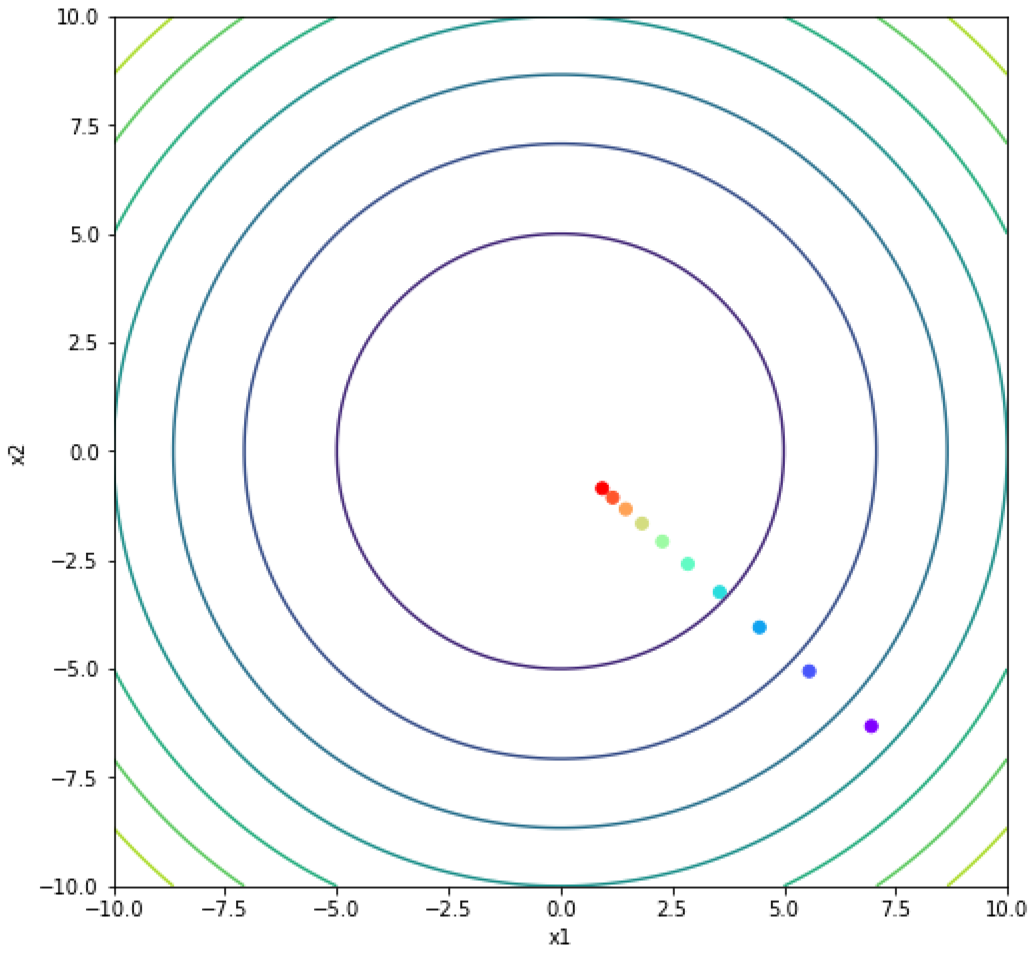


2D Gradient Descent

$$f(x_1, x_2) = x_1^2 + x_2^2$$

$$x_0 = [8.7, -7.9], \alpha = 0.1$$

$$x \leftarrow x - \alpha \nabla f(x)$$

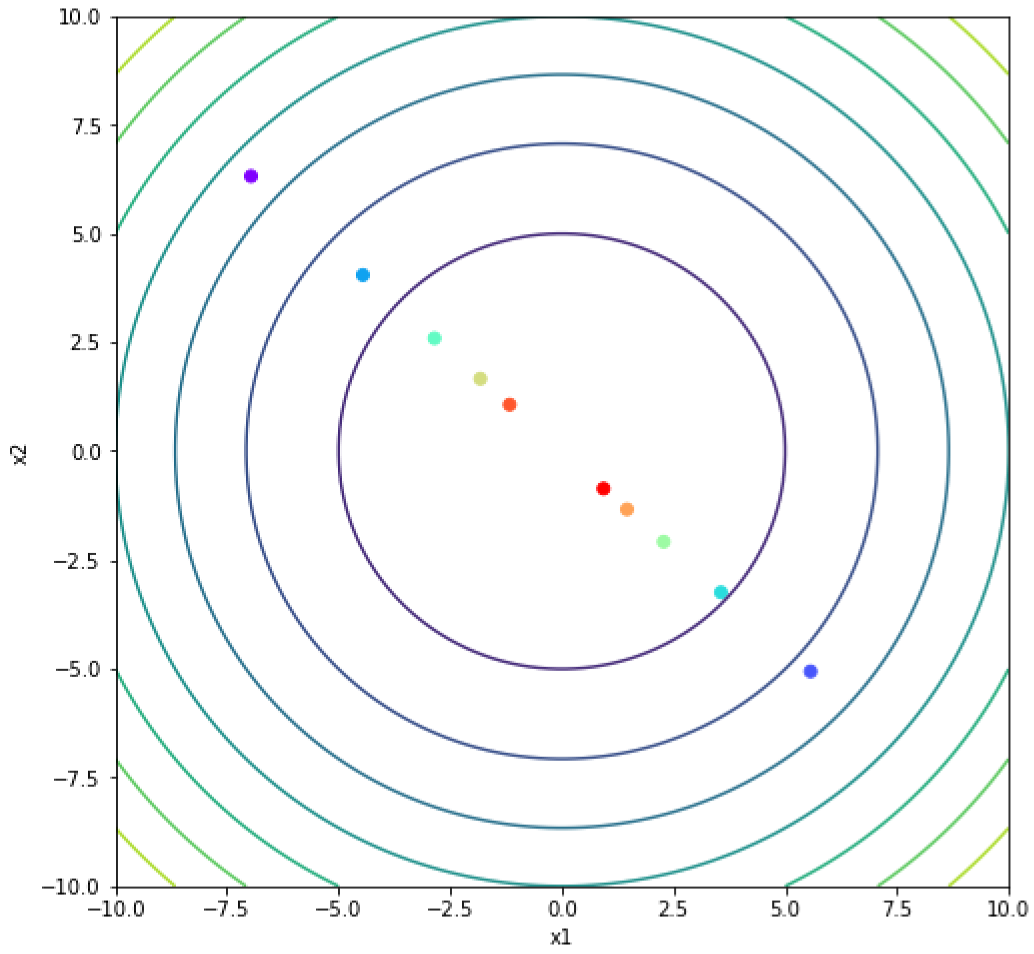


2D Gradient Descent

$$f(x_1, x_2) = x_1^2 + x_2^2$$

$$x_0 = [8.7, -7.9], \alpha = 0.9$$

$$x \leftarrow x - \alpha \nabla f(x)$$

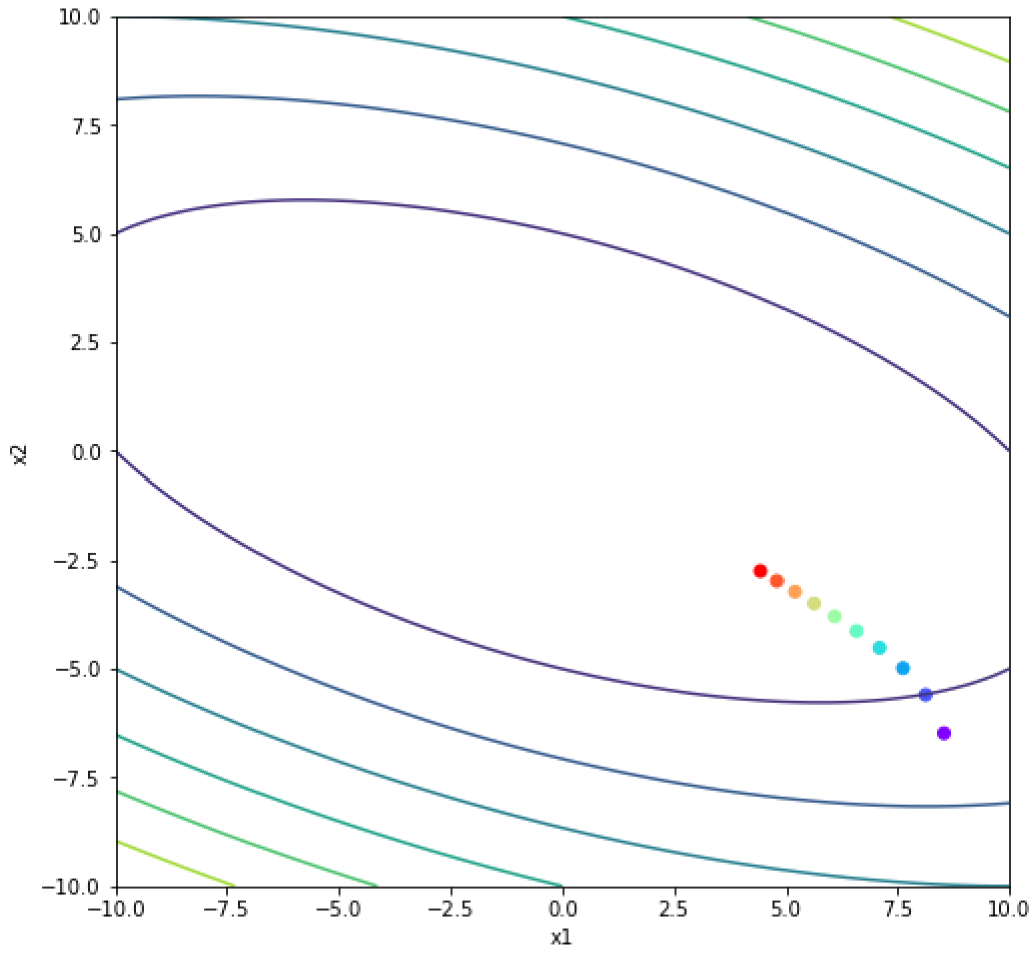


2D Gradient Descent

$$f(x_1, x_2) = 0.5x_1^2 + 2x_2^2 + x_1x_2$$

$$x_0 = [8.7, -7.9], \alpha = 0.2$$

$$x \leftarrow x - \alpha \nabla f(x)$$



Steepest Descent

- *‘The negative gradient is the direction of steepest descent’*

Directional Derivatives

- The gradient vector is a vector of partial derivatives
- It represents the instantaneous rates of change of the function f w.r.t one of its variables
  - $\frac{\partial f}{\partial x_i}$: how much f changes as x_i change while fixing other components at any given point
- Directional derivative is about how much f changes as all components change together at any given point

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

(gradient)

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

(partial derivative)

$$D_u f(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + u_1 h, y_0 + u_2 h) - f(x_0, y_0)}{h}$$

(directional derivative)

$u = [u_1, u_2], \|u\| = 1$ (unit vector)

Directional Derivatives (Chain Rules)

- Given a function $f(x, y)$ then the rate of change with respect to t along a curve $x(t), y(t)$ is

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$$

- The line through (x_0, y_0) in the direction $u = u_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is

$$x(t) = u_1 t + x_0, \qquad y(t) = u_2 y + y_0$$

$$D_u f(x, y) = \frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2 = \nabla f(x, y) \cdot u$$

Directional Derivatives

$$a \cdot b = \|a\|\|b\|\cos(\theta)$$

$$D_u f(x_0, y_0) = \nabla f(x_0, y_0) \cdot u = \|\nabla f(x_0, y_0)\|\|u\|\cos(\theta)$$

- When $\theta = 0, \cos(\theta) = 1,$ $D_u f$ is maximized, u is the direction of steepest ascent

$$u = \frac{\nabla f(x_0, y_0)}{\|\nabla f(x_0, y_0)\|}$$

- When $\theta = \pi, \cos(\theta) = -1,$ $D_u f$ is minimized, u is the direction of steepest descent

$$u = -\frac{\nabla f(x_0, y_0)}{\|\nabla f(x_0, y_0)\|}$$

Gradient Descent in Linear Regression (1D)

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(wx^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial L(w)}{\partial w} = \sum_{i=1}^{N}\left(wx^{(i)} - y^{(i)}\right)x^{(i)}$$

$$w := w - \alpha\left(\sum_{i=1}^{N}\left(wx^{(i)} - y^{(i)}\right)x^{(i)}\right)$$

(descent) (step-size) (gradient)

Gradient Descent in Linear Regression (N-D)

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial L}{\partial w_k} = \sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)x_k^{(i)}$$

$$w_k := w_k - \alpha\left(\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)x_k^{(i)}\right)$$

(descent) (step-size) (gradient)

Gradient Descent in Linear Regression (N-D)

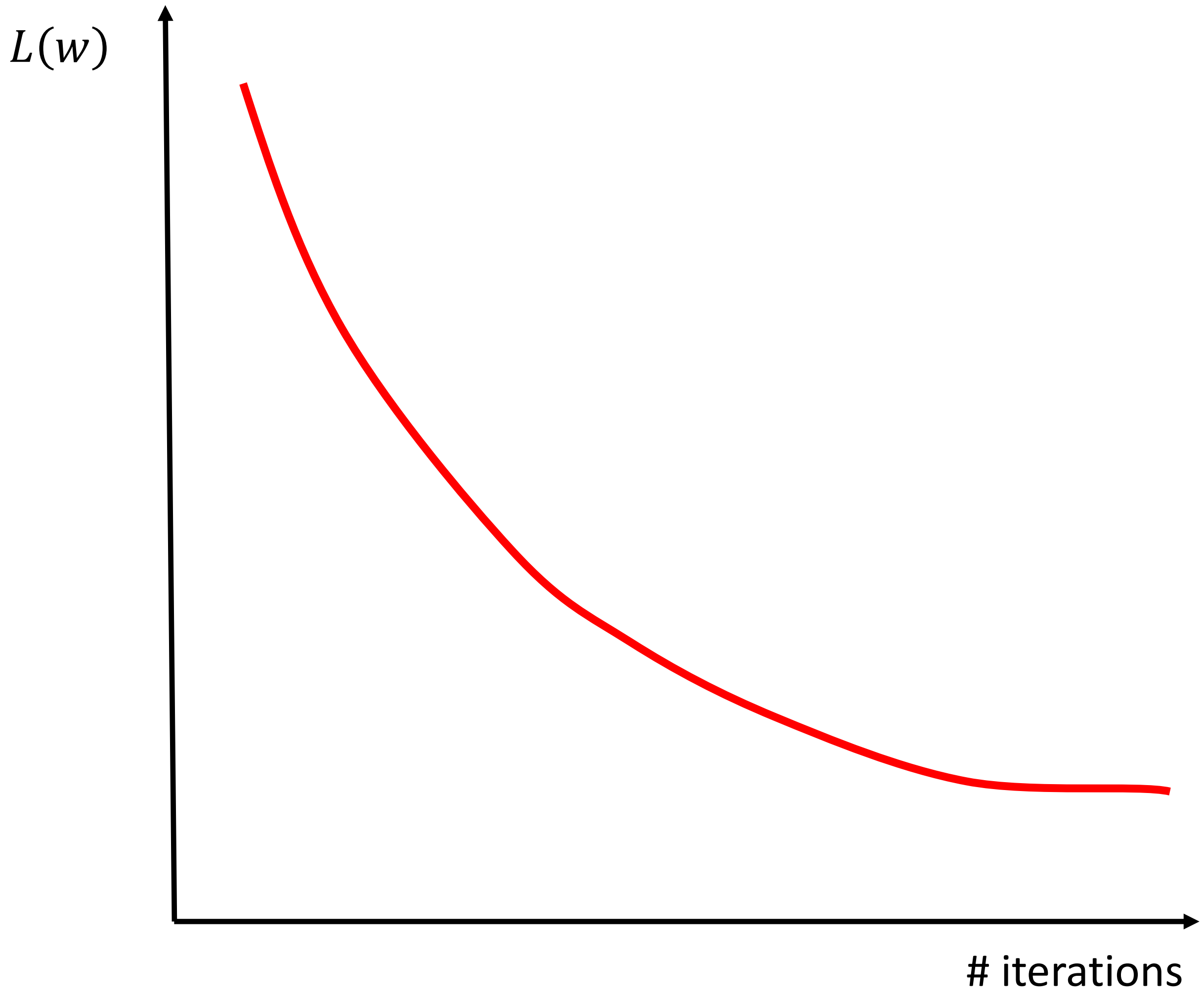
$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)x^{(i)} = X^\top(Xw - Y)$$

$$w \coloneqq w - \alpha\left(X^\top(Xw - Y)\right)$$

(descent) (step-size) (gradient)

Gradient Descent in Practice



- $L(w)$ should decrease every iteration
- If $L(w)$ decreases by very small amount, then it's considered as convergence