Least Squres Objective

$\hat{y}_i = w x_i$ linear regression

- Instead of "exact y_i", we evaluate the size of error in prediction
- Classic way is setting slope 'w' to minimize the sum of squared errors:

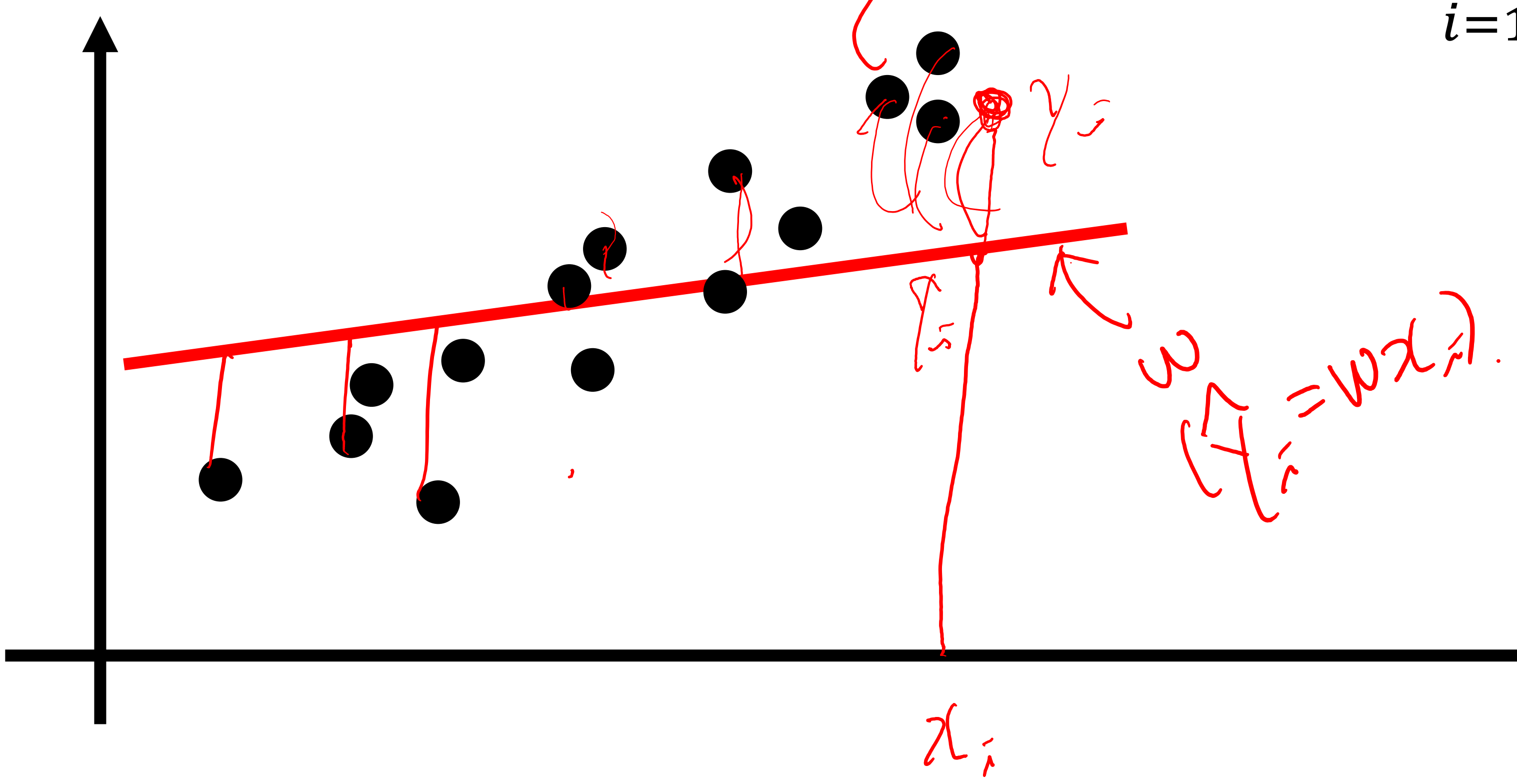
$$\arg\min_w f(w) = \underbrace{\sum_{i=1}^{n} (\underbrace{w x_i}_{\hat{y}_i} - y_i)^2}_{}$$

- A probabilistic interpretation is coming later in this course!
- But usually, it is done because it is easy to minimize.

Least Squres Objective

- Example – Large f(w)

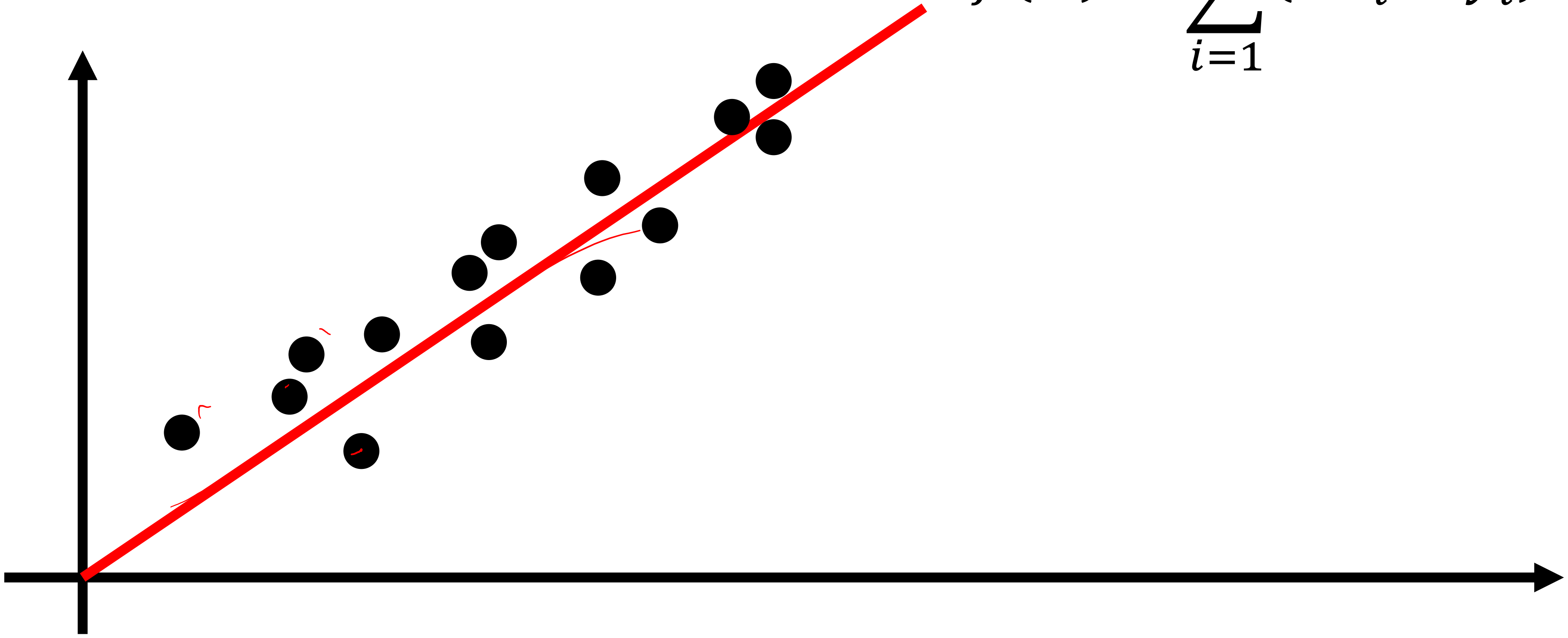
$$f(w) = \sum_{i=1}^{n} (wx_i - y_i)^2$$



Least Squres Objective

- Example – Small f(w)

$$f(w) = \sum_{i=1}^{n} (wx_i - y_i)^2$$



Finding Least Squares Solution

- Not change the solution!
  - Multiply ‘f’ by any positive constant
  - Add some constants to ‘f’

$$f'(w) = C_1 \sum_{i=1}^{n} (wx_i - y_i)^2 + C_2$$

- Finding ‘w’ that minimizes sum of squared errors:

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{n}\left[w^2x_i^2 - 2wx_iy_i + y_i^2\right]$$

$$= \frac{w^2}{2}\sum_{i=1}^{n} x_i^2 - w\sum_{i=1}^{n} x_iy_i + \frac{1}{2}\sum_{i=1}^{n} y_i^2 = \frac{w^2}{2}a - wb + c$$

$$f'(w) = wa - b = 0$$

$$\boxed{w = \frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}}$$

Multiple Dimension Linear Function

w_d

- A simple way is with a d-dimensional linear model
  - $\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + \cdots + x_d x_{id}$
  - In words, our model is that the output is a weighted sum of the inputs
- We can re-write this in summation notation:
  - $\hat{y}_i = \sum_{j=1}^{d} w_j x_{ij}$
- We can also re-write this in vector notation: (inner product)
  - $\hat{y}_i = \mathbf{w}^{\mathrm{T}} \mathbf{x}_i$
  - In this course, a vector is a column vector

Least Squares in d-Dimensions

- The linear least squares model in d-dimensions minimizes:

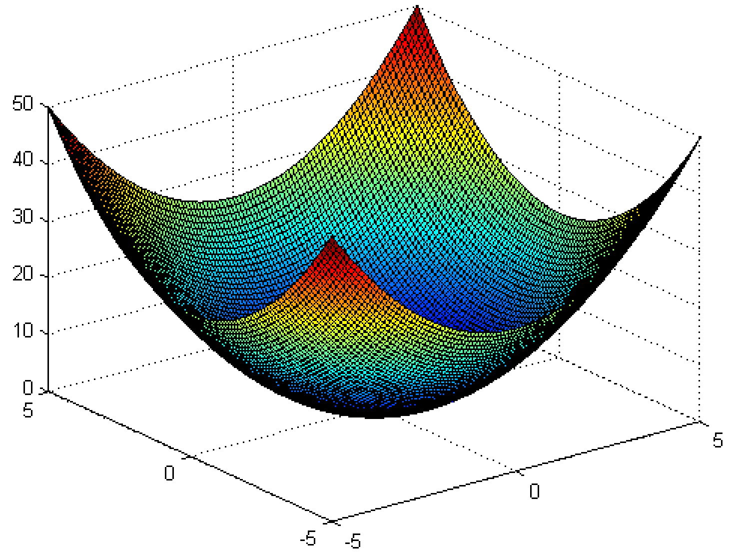
$$f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$$

- Least Squares Partial Derivatives for 1 sample

$$f(w_1, w_2, \dots, w_d) = \frac{1}{2}\left(\sum_{j=1}^{d} w_j x_{ij}\right)^2 - \left(\sum_{j=1}^{d} w_j x_{ij}\right) y_i + \frac{1}{2} y_i^2$$

$$\frac{\partial}{\partial w_k} f(w_1, w_2, \dots, w_d) = \left(\sum_{j=1}^{d} w_j x_{ij}\right) x_{ik} - y_i x_{ik} = (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i) x_{ik} = 0.$$

Least Squares in d-Dimensions



- Least Squares Partial Derivatives for all samples

$$\frac{\partial}{\partial w_k} f(w_1, w_2, \ldots, w_d) = \sum_{i=1}^{n} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i - y_i\right) x_{ik}$$

- Unfortunately, the partial derivative for w_j depends on all $\{w_1, w_2, \ldots, w_d\}$
  - Thus, we can't just set equal to 0 and solve for w_j
  - **We need to find 'w' where the gradient vector equals the zero vector!**

$$\nabla f(w_1, w_2, \ldots, w_d) = \left[\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \frac{\partial f}{\partial w_3}, \ldots, \frac{\partial f}{\partial w_d}\right]^T = 0$$

Linear and Quadratic Gradients

$A = A^T$.

$$f(w) = aw^2 \quad \Rightarrow \quad \frac{\partial}{\partial w} f(w) = 2aw$$

$$g(w) = bw \quad \Rightarrow \quad \frac{\partial}{\partial w} g(w) = b$$

$$\mathrm{h}(w) = c \quad \Rightarrow \quad \frac{\partial}{\partial w} h(w) = 0$$

$$f(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \quad \Rightarrow \quad \nabla f(\mathbf{w}) = \mathbf{A}\mathbf{w}$$

If A is symmetric

$$g(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{b} \quad \Rightarrow \quad \nabla g(\mathbf{w}) = \mathbf{b}$$

$$\mathrm{h}(\mathbf{w}) = \mathbf{c} \quad \Rightarrow \quad \nabla h(\mathbf{w}) = 0$$

Linear and Quadratic Gradients

$\frac{1}{2}(Xw-y)^T(Xw-y)$

- We can re-write the d-dimensional quadratic:
  - $f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 = \boxed{\frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2} = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boxed{\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y}}$
  - $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{b} + \mathbf{c}$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \left(\begin{bmatrix} \hat{y}_1 \\ y_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right)$$

- Thus, the gradient is given by:
  - $\nabla f(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{X}^{\mathrm{T}}\mathbf{y}$ $= 0$

- **Normal equations:**
  - $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$ (in the form of $\mathbf{A}\mathbf{x} = \mathbf{b}$)
  - When $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible, $\boxed{\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$

$(X^TX)w = X^Ty$.

$A = X^TX$,
$A^T = (X^TX)^T = X^TX$.

Pseudo-inverse.

Inverse Matrix & Pseudo-inverse Matrix

- For $m \times n$ system A,
  - $Ax = b$
  - Let's find x that minimizes the energy of $||Ax - b||^2$
- The derivative of $||Ax - b||^2$ becomes
  - $Ax - b = 0$
  - $Ax = b$
  - But, we can't estimate the inverse of A because A is not square!

$w = (\underbrace{X^T X}_{\boxed{100 \times 100}})^{-1} X^T y$

$X \in \mathbb{R}^{n \times d}$

Inverse Matrix & Pseudo-inverse Matrix

- The derivative of $||Ax - b||^2$ becomes
  - $Ax - b = 0$
  - $Ax = b$
  - But, we can’t estimate the inverse of A because A is not square!

- Then, let’s make it square matrix
  - $A^T Ax = A^T b$
  - When the columns of A are linearly independent, $A^T A$ is invertible. Thus,
  - $x = (A^T A)^{-1} A^T b \equiv A^+ b$