EE2211 Lecture 5: Linear Regression & Least Squares. $\hat{w} = (X^T X)^{-1} X^T y$

Agenda: i) Notation for sets & functions / linear & affine functions.
ii) Min & Argmin / Differentiation of functions (Grad).
iii) Linear Regression
Formulation / Objective (Loss) Function / Training / Prediction
iv) Predicting multiple outputs.

Reading: Lec_5.pdf Slides.

Office hours: Next Thursday 16 Feb 2pm
Zoom 98390985?4

Sets: A set is an unordered collection of objects.

Eg:
$S = \{1, 2, 3, 4, 5, 6\}$ possible outcomes of a dice toss.
$S' = \{6, 4, 5, 3, 2, 1\}$

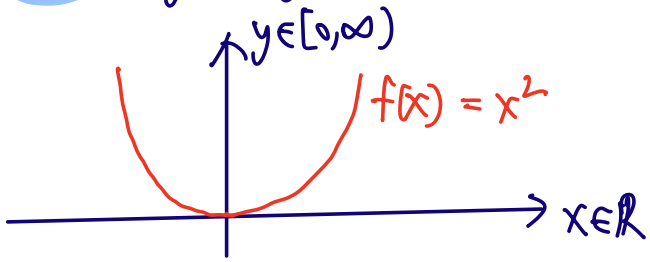
Eg: $S = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ ← set of all real numbers
$S = (a, b] = \{x \in \mathbb{R} : a < x \leq b\}$
} intervals of $\mathbb{R}$

Eg:
$\mathbb{R}$: set of real numbers
$\mathbb{R}^d$: set of d-dim real vectors ← column
$\mathbb{R}^{m \times d}$: set of all size $m \times d$ matrices. (m rows, d columns)

Functions: A function f is a map from a set X to another set Y.

$$f: X \to Y.$$

Eg: $f: \mathbb{R} \to [0,\infty)$ given by the recipe $f(x) = x^2$



Set of inputs X: domain

Set of possible outputs Y: codomain

Set $\{f(x): x \in X\}$: range or image.

$f: \{1,2,3,4\} \to \{1,2,3,4,5,6\}$ $\quad f(x) = x+2.$

domain: $\{1,2,3,4\}$ codomain: $\{1,2,3,4,5,6\}$ range or image: $\{3,4,5,6\}$

Range or image of f is $\{f(x): x \in X\} = \{3,4,5,6\}$

Linear functions: A function $f: \mathbb{R}^d \to \mathbb{R}$ is linear if

(i) [Homogeneity] For every vector $\underline{x} \in \mathbb{R}^d$ & scalar $a \in \mathbb{R}$

$$f(a\underline{x}) = a f(\underline{x}).$$

(ii) [Additivity] For every $\underline{x}, \underline{y} \in \mathbb{R}^d$,

$$f(\underline{x}+\underline{y}) = f(\underline{x}) + f(\underline{y}).$$

Question: A linear function $f: \mathbb{R}^d \to \mathbb{R}$ must pass through the origin $\underline{0} \in \mathbb{R}^d$, i.e.,

$$f(\underline{0}) = 0 \in \mathbb{R}$$

Question: If we have n vectors $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^d$ and n scalars $a_1, \dots, a_n \in \mathbb{R}$, and $f: \mathbb{R}^d \to \mathbb{R}$ is linear,

$$f\left(\sum_{i=1}^{n} a_i \underline{x}_i\right) = \sum_{i=1}^{n} a_i f(\underline{x}_i).$$

Question: $f: \mathbb{R} \to \mathbb{R}$ is specified by $f(x) = |x|$. Is f linear?

No: Take $x = x$, $a = -1$ $\quad f(ax) \overset{?}{=} a f(x)$

$$f(ax) = |ax| = |x| \quad , \quad a f(x) = -|x|$$

Question: Fix $\underline{a} \in \mathbb{R}^d$ ($\underline{a} = (a_1, \dots, a_d)$). Consider the function $f: \mathbb{R}^d \to \mathbb{R}$

$$f(\underline{x}) = \underline{a}^T \underline{x} = \sum_{i=1}^{d} a_i x_i \quad , \quad \underline{x} = (x_1, \dots, x_d)$$

f is called the inner product / dot product between $\underline{a}$ & $\underline{x}$.

Show that the inner product is linear.

Question: Why does a linear f^n have to pass through $\underline{0} \in \mathbb{R}^d$?

Homo: $\forall a \in \mathbb{R}, \underline{x} \in \mathbb{R}^d$, $\quad f(a\underline{x}) = a f(\underline{x})$

Take $\underline{x} = \underline{x}$, $a = 0$ $\quad f(0\underline{x}) = 0 f(\underline{x}) = 0$

$$f(0\underline{x}) = f(\underline{0})$$

Affine Function: An affine function is a linear function plus a constant.

$f: \mathbb{R}^d \to \mathbb{R}$ is affine if $f(x) = g^T x + b$ for some $g \in \mathbb{R}^d$, $b \in \mathbb{R}$.

$g^T x$ — bias; b — offset

Eg: $f: \mathbb{R}^2 \to \mathbb{R}$ $\quad f(x) = f(x_1, x_2) = -x_1 + 3x_2 + 7$ is affine.

$= \begin{pmatrix} -1 \\ 3 \end{pmatrix}^T x + 7$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

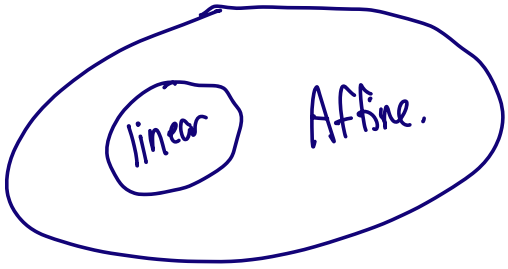
Question: Is a linear f^n affine? Yes. (Take $b = 0$)

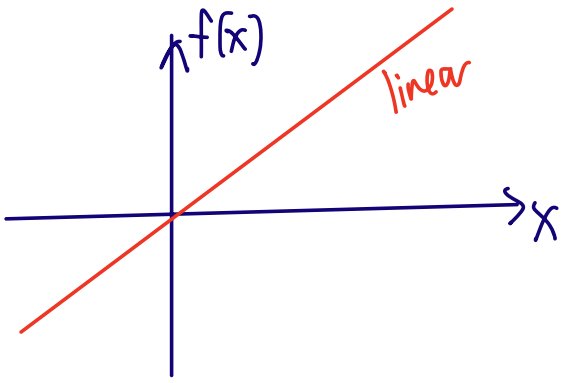
Is an affine f^n linear? Not necessarily.

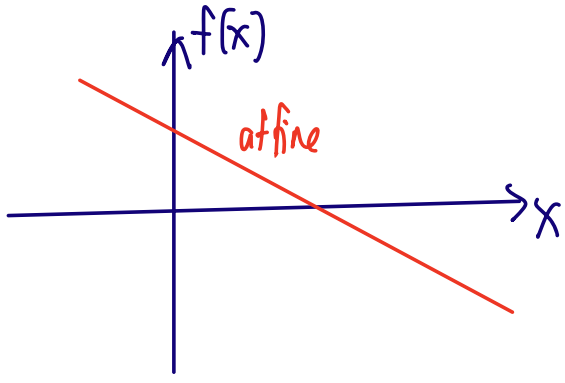
A linear f^n $f(0) = 0$.

Look at example $f(x_1, x_2) = -x_1 + 3x_2 + 7$. This function is not linear

$f(0,0) = 7 \neq 0$.







Min & Argmin

$f: [a,b] \to \mathbb{R}$.

The function f has a local min at $c \in \mathbb{R}$ if

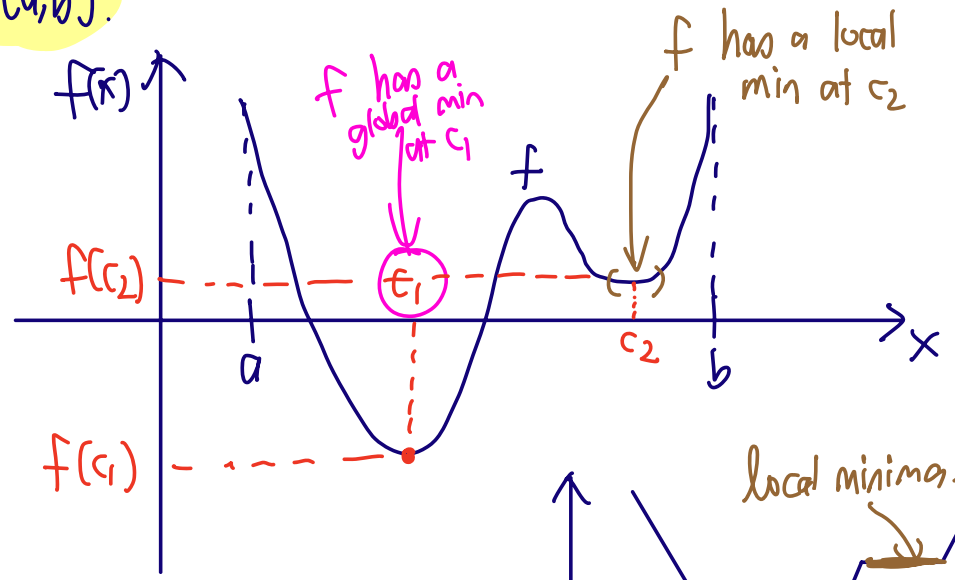
$$f(x) \geq f(c)$$

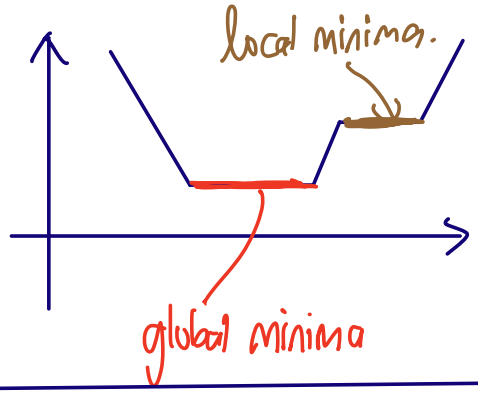
for all x in some neighborhood of c.

The function f has a global min at $c \in \mathbb{R}$ if

$$f(x) \geq f(c)$$

for all $x \in [a,b]$.





Minimum & argmin.

$f: X \to Y$, the $\min_{x \in X} f(x)$ returns the smallest value among all elements in the set $\{f(x): x \in X\}$

$f: X \to Y$, the argmin $x^* = \underset{x \in X}{\operatorname{argmin}} f(x)$ return the value $x \in X$ that minimizes $f(x)$.

$$f(x^*) = \min_{x \in X} f(x).$$

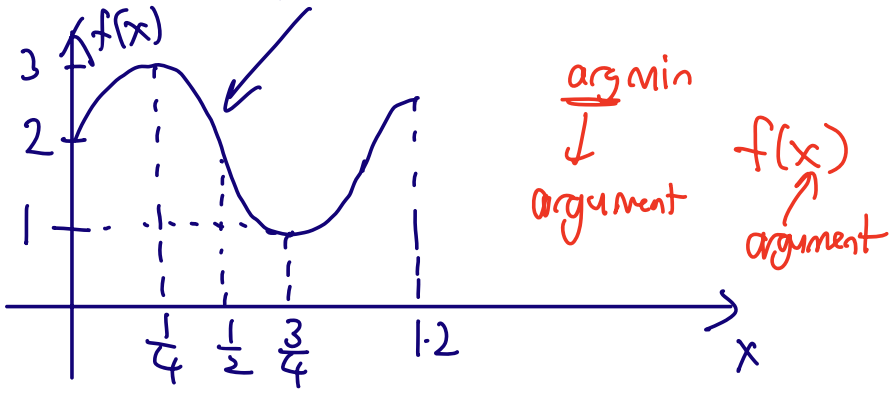
Ex: $X = \{0, 1\}$, $f: X \to \mathbb{R}$, $f(0) = 5$, $f(1) = 7$

$\min_{x \in X} f(x) = 5$. $\max_{x \in X} f(x) = 7$

$\operatorname{argmin}_{x \in X} f(x) = 0$ $\operatorname{argmax}_{x \in X} f(x) = 1$.

Eg: $f: [0, 1.2] \to \mathbb{R}$ $f(x) = 2 + \sin(2\pi x)$

$|\sin \theta| \leq 1$



argmin → argument

f(x) → argument

argument that min

$\underset{x \in X}{\operatorname{argmin}} f(x) = \frac{3}{4}$ $\min_{x \in X} f(x) = 1$.

Derivatives or gradient

$f: \mathbb{R}^d \to \mathbb{R}$ $f(x) = y$

d-dim.

Gradient vector / derivative at $x \in \mathbb{R}^d$ is the vector

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \in \mathbb{R}^d$$

$$f(x_1, x_2) = 2x_1^2 + 5x_1x_2 + 3x_2^2$$

$d=2$

$$\frac{\partial f}{\partial x_1} = 4x_1 + 5x_2 \qquad \frac{\partial f}{\partial x_2} = 5x_1 + 6x_2$$

$$\nabla f(\underline{x}) = \begin{bmatrix} 4x_1+5x_2 \\ 5x_1+6x_2 \end{bmatrix}$$

Example: Fix $\underline{a} \in \mathbb{R}^d$, $f(\underline{x}) = \underline{a}^T\underline{x} = \sum_{i=1}^{d} a_i x_i$

$$\nabla f(\underline{x}) = \underline{a}$$

Example: Fix $A \in \mathbb{R}^{d\times d}$, $f(\underline{x}) = \underline{x}^T A \underline{x} = \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j A_{ij}$

$\underline{x}^T$: $1\times d$, A: $d\times d$, $\underline{x}$: $d\times 1$, result: 1×1

$$\nabla f(\underline{x}) = (A + A^T)\underline{x}$$

Most of the time (in this class) A is symmetric ($A = A^T$)

$$\Rightarrow \nabla f(\underline{x}) = 2A\underline{x}.$$

Linear Regression is linear approach for modelling the r/s between a scalar response $y \in \mathbb{R}$ and explanatory variables (attributes, features) $\underline{x} \in \mathbb{R}^d$.

Dataset $\{(\underline{x}_i, y_i) : 1 \le i \le m\}$, $\underline{x}_i \in \mathbb{R}^d$ (training vector)

$y_i \in \mathbb{R}$ (target).

Without offset

Form the design matrix

$$X = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,d} \end{bmatrix} \in \mathbb{R}^{m \times d}$$

target vector

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{m \times 1}$$

Goal: Find a solution (or approx solution) $\underline{w} \in \mathbb{R}^d$ to the linear system $X\underline{w} = y$

With offset: Design a function/model/regressor $f_{\underline{w},b}$ linear comb. of features in $\underline{x}$ **plus possibly a constant.**

$$f_{\underline{w},b}(\underline{x}) = \underline{w}^T\underline{x} + b$$

offset: does not have to be constrained to pass thru origin.

$\underline{w} \in \mathbb{R}^d$: d-dim unknown weight vector

$b \in \mathbb{R}$: 1-dim offset, bias scalar unknown.

$f_{\underline{w},b}$: function f is parameterized by $\underline{w}, b$.

$$f_{\underline{w},b}(\underline{x}) = \begin{bmatrix} b \\ \underline{w} \end{bmatrix}^T \begin{bmatrix} 1 \\ \underline{x} \end{bmatrix} = b + \underline{w}^T\underline{x}$$

unknown

$\begin{bmatrix} b \\ \underline{w} \end{bmatrix} \in \mathbb{R}^{d+1}$, $\begin{bmatrix} 1 \\ \underline{x} \end{bmatrix} \in \mathbb{R}^{d+1}$

$$\underline{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \qquad \begin{bmatrix} b \\ \underline{w} \end{bmatrix} = \begin{pmatrix} b \\ w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \qquad \begin{bmatrix} 1 \\ \underline{x} \end{bmatrix} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

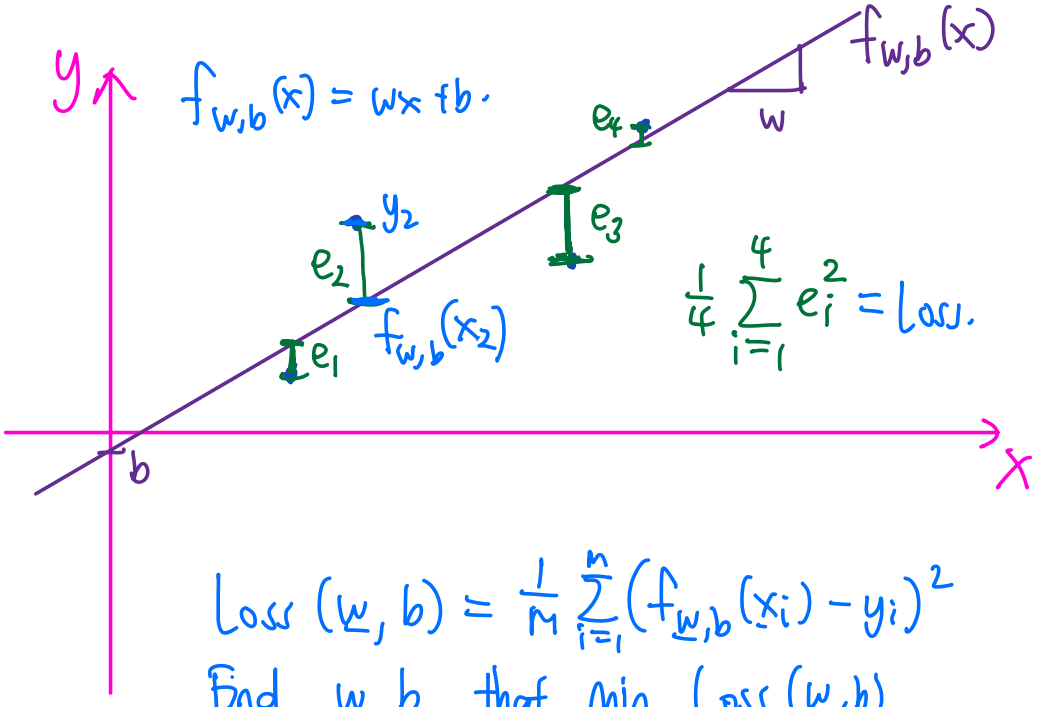
Concatenation.

Minimize the error e_i bet. the prediction $f_{\underline{w},b}(\underline{x}_i)$ and target y_i.

$$e_i = f_{\underline{w},b}(\underline{x}_i) - y_i$$

$$\text{Loss}(\underline{w}, b) = \frac{1}{m}\sum_{i=1}^{m} e_i^2 = \frac{1}{m}\sum_{i=1}^{m}\underbrace{(f_{\underline{w},b}(\underline{x}_i) - y_i)^2}_{\text{per-sample loss.}}$$

The squared loss (l_2 loss) is differentiable.



$$\text{Loss}(\underline{w}, b) = \frac{1}{n}\sum_{i=1}^{n}(f_{\underline{w},b}(\underline{x}_i) - y_i)^2$$

Find $\underline{w}, b$ that min Loss $(\underline{w}, b)$.

Define $\bar{w} \in \mathbb{R}^{d+1}$, $\bar{x}_i \in \mathbb{R}^{d+1}$ (i^{th} training sample)

$$\bar{w} = \begin{bmatrix} b \\ \underline{w} \end{bmatrix} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \qquad \bar{x}_i = \begin{bmatrix} 1 \\ \underline{x}_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,d} \end{bmatrix}$$

Objective (Loss) in Linear Regression

Want to find $\bar{w}^* = \begin{bmatrix} b^* \\ \underline{w}^* \end{bmatrix} \in \mathbb{R}^{d+1}$ that min

$$\bar{w}^* = \underset{\bar{w} = \begin{bmatrix} b \\ \underline{w} \end{bmatrix}}{\operatorname{argmin}} \; \text{Loss}(\underline{w}, b)$$

← is a constant

$$\text{Loss}(\underline{w}, b) = \frac{1}{m} \sum_{i=1}^{m} e_i^2 = \cancel{\frac{1}{m}} \sum_{i=1}^{m} (f_{\underline{w},b}(\underline{x}_i) - y_i)^2$$

$$f_{\underline{w},b}(\underline{x}_i) - y_i = \underbrace{\begin{bmatrix} 1 \\ \underline{x}_i \end{bmatrix}^T}_{\bar{x}_i^T} \underbrace{\begin{bmatrix} b \\ \underline{w} \end{bmatrix}}_{\bar{w}} - y_i = \underbrace{\underline{x}_i^T \underline{w} + b}_{\text{affine function in } \underline{x}_i} - y_i$$

So that

$$\sum_{i=1}^{m} (f_{\underline{w},b}(\underline{x}_i) - y_i)^2 = \sum_{i=1}^{m} (\underline{x}_i^T \underline{w} + b - y_i)^2$$

$$= (X\bar{w} - y)^T (X\bar{w} - y)$$

X: design matrix $m \times (d+1)$

unknown $\begin{bmatrix} b \\ \underline{w} \end{bmatrix} = \bar{w}$

$$X = \begin{bmatrix} 1 & \underline{x}_1^T \\ 1 & \underline{x}_2^T \\ \vdots & \\ 1 & \underline{x}_m^T \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ 1 & x_{2,1} & \cdots & x_{2,d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{m,1} & \cdots & x_{m,d} \end{bmatrix} \in \mathbb{R}^{m \times (d+1)}$$

$(X^TX)^T = X^TX$

Objective: $J(\underline{w}) = (X\underline{w} - \underline{y})^T (X\underline{w} - \underline{y})$

$\underline{a}^T\underline{b} = \underline{b}^T\underline{a}$

$$= \underline{w}^T X^T X \underline{w} - \underline{w}^T X^T \underline{y} - \underline{y}^T X \underline{w} + \underline{y}^T \underline{y}.$$

$$= \underline{w}^T (X^TX) \underline{w} - 2\underline{w}^T (X^T\underline{y}) + \underline{y}^T\underline{y}$$

($\underline{y}^T\underline{y}$: const wrt $\underline{w}$)

$$\nabla_{\underline{w}} J(\underline{w}) = 2(X^TX)\underline{w} - 2X^T\underline{y} = \underline{0}$$

$$(X^TX)\underline{w} = X^T\underline{y} \Rightarrow \underline{w}^* = (X^TX)^{-1}X^T\underline{y}.$$

Are we allowed to do this?

if X^TX is invertible

$\Updownarrow$

X has full column rank.

$\underline{w}^* = (X^TX)^{-1}X^T\underline{y}$ is the least squares solution.

Given a dataset $\{(x_i, y_i)\}$.

Form the design matrix & target vector

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ 1 & x_{2,1} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m,1} & \cdots & x_{m,d} \end{bmatrix} \in \mathbb{R}^{m \times (d+1)} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$$

Find the least squares solution.

$$\underline{w}^* = \begin{bmatrix} b^* \\ \underline{w}^* \end{bmatrix} = (X^T X)^{-1} X^T y$$

Training / Learning

if X has full col. rank.

$x_{new} \in \mathbb{R}^d$

Prediction:

$$\hat{y}_{new} = \begin{bmatrix} 1 \\ \underline{x}_{new} \end{bmatrix}^T \begin{bmatrix} b^* \\ \underline{w}^* \end{bmatrix}$$

$$= b^* + \underline{x}_{new}^T \underline{w}^*.$$

Prediction / Testing

Dataset $x_1 = -7,\ x_2 = -5,\ x_3 = 1,\ x_4 = 5$ $\quad d = 1$

$y_1 = -6\ \ y_2 = -4,\ y_3 = -1,\ y_4 = 4$ $\quad m = 4.$

$$X = \begin{bmatrix} 1 & -7 \\ 1 & -5 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \qquad y = \begin{bmatrix} -6 \\ -4 \\ -1 \\ 4 \end{bmatrix}$$

$X\bar{w} = y$:

$\cap$

$\mathbb{R}^2$

$\text{rank}(X) = 2$

$\text{rank}(\tilde{X}) = \text{rank}([X\ y]) = 3$

$\text{rank}(X) < \text{rank}(\tilde{X}) \Rightarrow$ no solution.

$$\bar{w}^* = (X^T X)^{-1} X^T y = \begin{bmatrix} -0.5879 \\ 0.7747 \end{bmatrix}.$$

$x_{new} = 2$ (Predict its y value)

$$\hat{y}_{new} = \begin{bmatrix} 1 \\ x_{new} \end{bmatrix}^T \bar{w}^* = 1(-0.5879) + 2 \times 0.7747$$
$$= -2.1374 .$$

To get the design matrix from a dataset $\{(\underline{x}_i, y_i)\}$
$\underline{x}_i \in \mathbb{R}^d$

$$\underline{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \underline{x}_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \underline{x}_3 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}. \qquad d=2. \quad m=3.$$

$$X = \begin{bmatrix} 1 & \underline{x}_1^T \\ 1 & \underline{x}_2^T \\ 1 & \underline{x}_3^T \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 4 & 7 \end{bmatrix} \in \mathbb{R}^{m \times (d+1)}$$

$\underline{x}_1^T = [1 \quad 2]$ $\underline{x}_2^T = [2 \quad 3]$ $\underline{x}_3^T = [4 \quad 7]$

Converted the set of training vectors $\underline{x}_i$, $1 \le i \le m$ into a design matrix that can be used to learn an affine function

$$(X\bar{w} - y)^T (X\bar{w} - y) = \sum_{i=1}^{m} (X\bar{w} - y)_i^2$$

$(\underline{a})_i = i^{th}$ comp. of vector $\underline{a}$.

$$= \sum_{i=1}^{m} (\underline{x}_i^T \underline{w} + b - y_i)^2$$

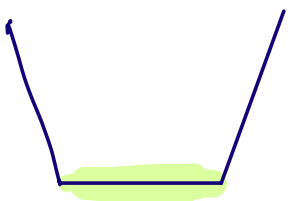
X : full column rank $\Rightarrow$ $\bar{w}^* = (X^TX)^{-1}X^Ty$ is a unique global min for the loss function

TRUE.

X: not full col. rank $\Rightarrow$ $(X^TX)\bar{w}^* = X^Ty$

SVD : singular value decomp. of X^TX

$\Rightarrow$ Multiple global min



$f(\underline{x}) = A\underline{x}$, $f: \mathbb{R}^d \to \mathbb{R}^m$ $A \in \mathbb{R}^{m \times d}$.

Check $\forall a, b \in \mathbb{R}$, $\underline{x}, \underline{y} \in \mathbb{R}$ $f(a\underline{x} + b\underline{y}) = af(\underline{x}) + bf(\underline{y})$.

LHS $= A(a\underline{x} + b\underline{y}) = A(a\underline{x}) + A(b\underline{y})$ (additivity)

$= aA\underline{x} + bA\underline{y}$

RHS $= aA\underline{x} + bA\underline{y}$ //