Math 7243-Machine Learning and Statistical Learning Theory – He Wang

Section 2. Linear Regression

1. Linear Regression
2. Least Squares
3. Matrix Calculus

House Price Example:

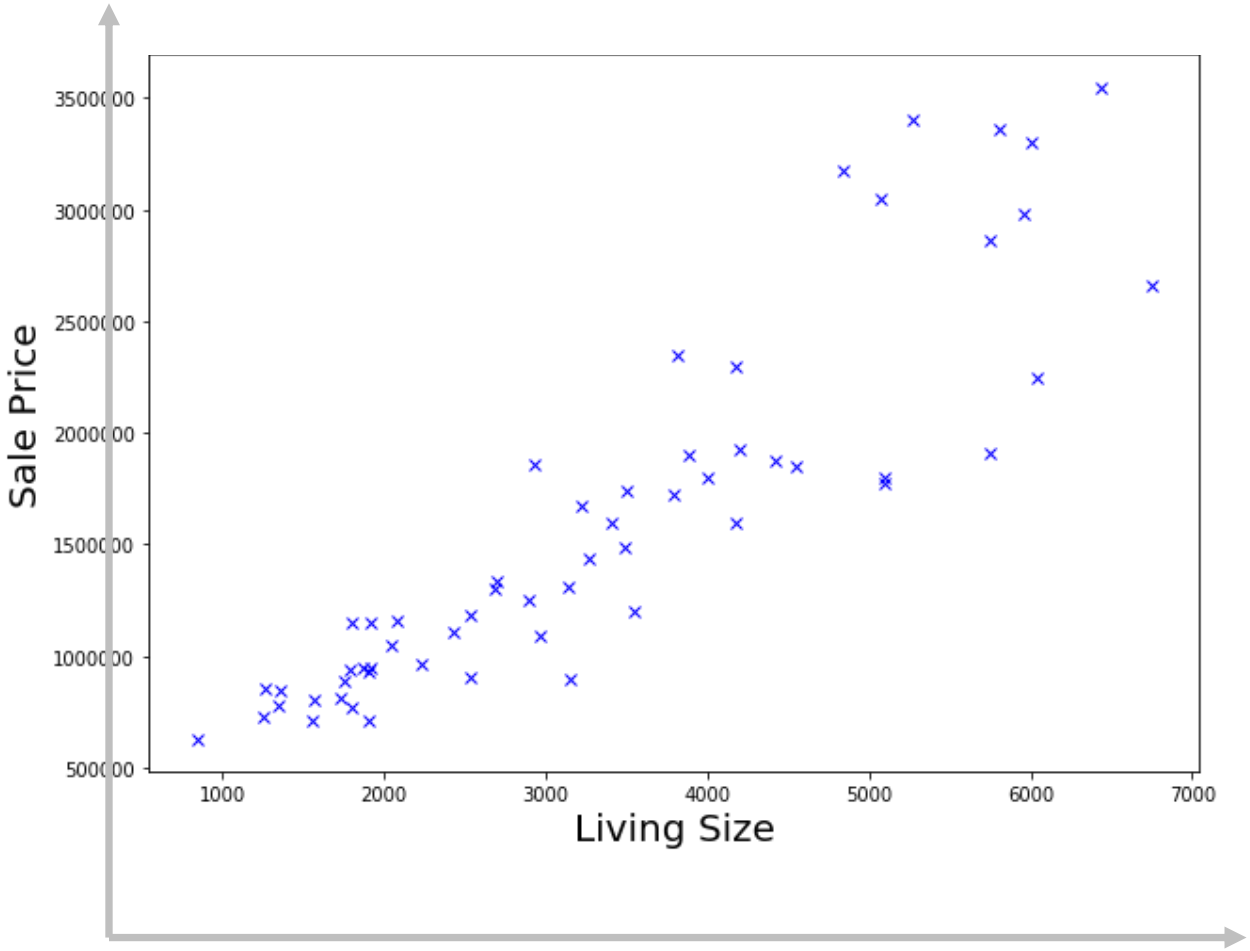
Consider the 59 single family residential houses sold in Newton, MA in Dec. 2020. (Data downloaded from www.redfin.com)

BEDS	BATHS	LOCATION	SQUARE_FEET	LOT_SIZE	YEAR_BUILT	PRICE
3	3	Newton	2969	15014	1967	1090000
3	2.5	Newton	1566	5582	1922	805000
4	2.5	Newton Corner	2532	6273	1953	905000
7	4.5	Newton Center	6748	26607	1902	2660000
4	4	West Newton	4200	20446	2007	1925000
4	2.5	Newton	2232	3966	1870	965000
2	1.5	Newton Corner	1344	5559	1851	775000
3	2.5	Newton	2898	12420	1943	1250000
2	2	West Newton	1729	4171	1953	815000
6	3	West Newton	3149	12616	1953	900000
5	3.5	West Newton	4000	12006	1912	1800000
4	3.5	West Newton	6430	30600	1920	3550000
4	1.5	Auburndale	1750	8222	1893	885000
2	2	Newton	840	5548	1955	630000
…	….	…	…	…	…	…

- Predict house price via living size (square_feet)

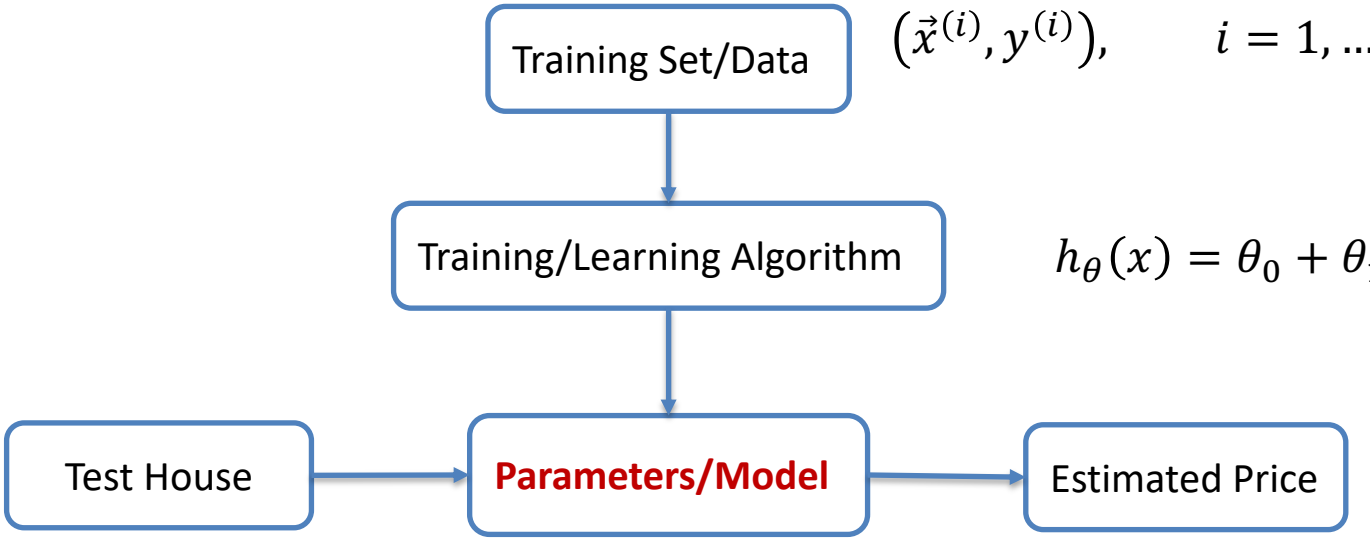
Input: a dataset that contains n samples. $\left(\vec{x}^{(i)}, y^{(i)}\right), \quad i = 1, \ldots n$

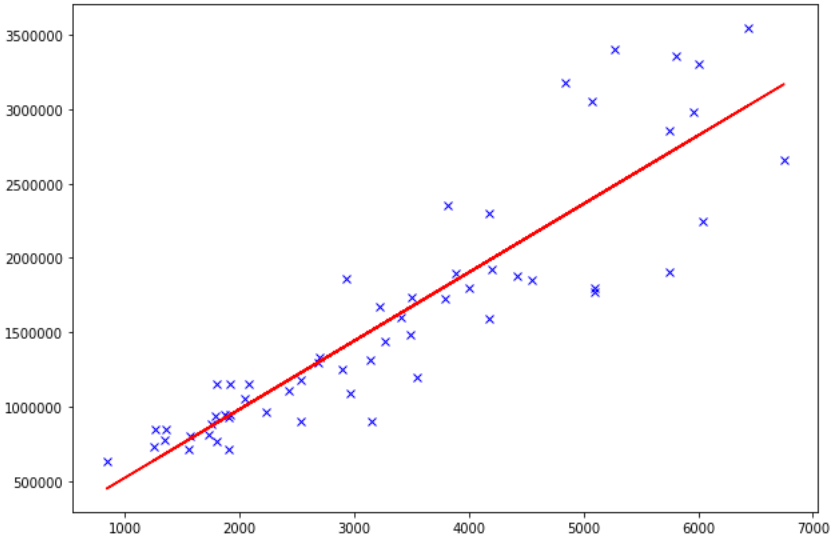
Task: if a house has x square feet, predict its price?



SQUARE_FEET	PRICE
2969	1090000
1566	805000
2532	905000
6748	2660000
4200	1925000
2232	965000
1344	775000
2898	1250000
1729	815000
3149	900000
4000	1800000
6430	3550000
1750	885000
840	630000
…	…

- Predict house price.





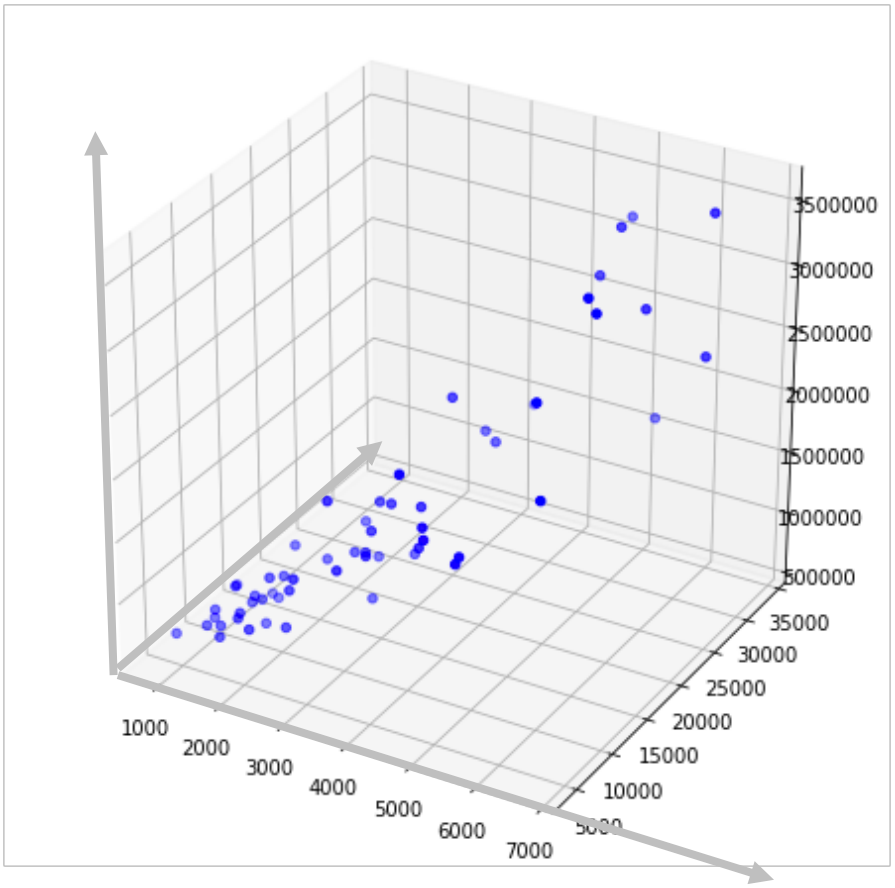
$h_\theta(x) = \theta_0 + \theta_1 x_1$

➢ **Predict house price.**

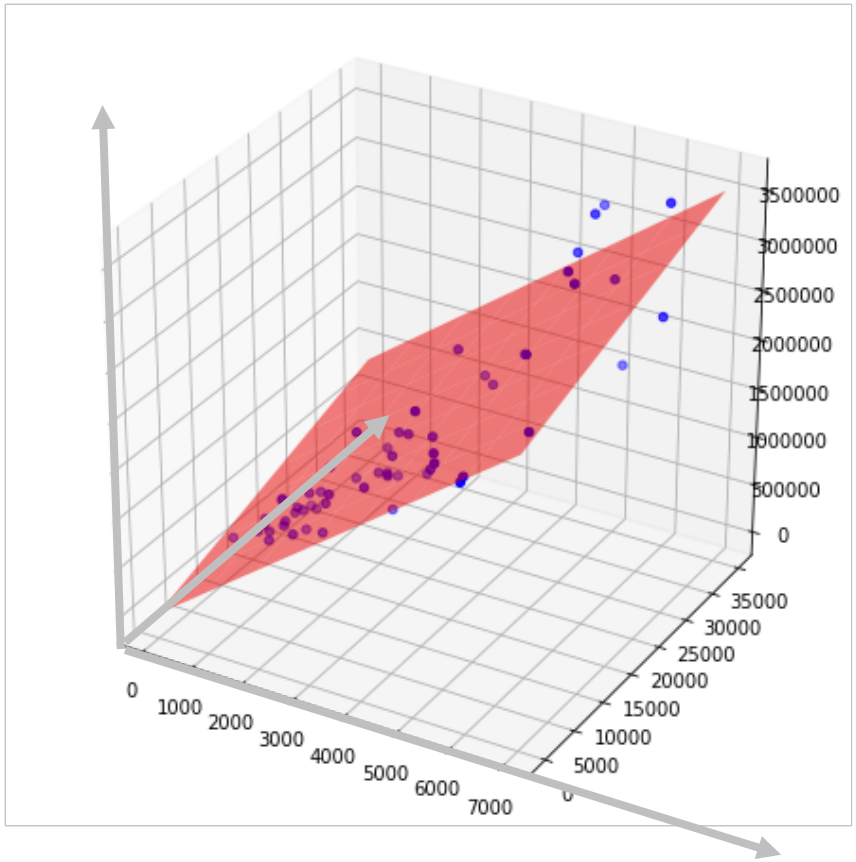
Input: a dataset that contains n samples $\left(\vec{x}^{(i)}, y^{(i)}\right),\ i = 1, \ldots n$

Task: if a house has x_1 (ft^2) living size and x_2 (ft^2) lot size, predict its price?

Model: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$



SQUARE_FEET	LOT_SIZE	PRICE
2969	15014	1090000
1566	5582	805000
2532	6273	905000
6748	26607	2660000
4200	20446	1925000
2232	3966	965000
1344	5559	775000
2898	12420	1250000
1729	4171	815000
3149	12616	900000
4000	12006	1800000
6430	30600	3550000
1750	8222	885000
840	5548	630000
…	…	…



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

➢ **Linear Regression** (Parametric Method)

Input: a dataset that contains n samples.

$$D = \left\{\left(\vec{x}^{(i)}, y^{(i)}\right), \qquad i = 1, \dots n\right\}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix} \xrightarrow{\quad h \quad} y$$

SQUARE_FEET	LOT_SIZE	BEDS	BATHS	PRICE
2969	15014	3	3	1090000
1566	5582	3	2.5	805000
2532	6273	4	2.5	905000
6748	26607	7	4.5	2660000
4200	20446	4	4	1925000
2232	3966	4	2.5	965000
1344	5559	2	1.5	775000
2898	12420	3	2.5	1250000
1729	4171	2	2	815000
3149	12616	6	3	900000

Assumption: linear model

$$h_\theta(\vec{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d$$

Data give us $h_\theta(\vec{x}^{(i)}) = y^{(i)}$ for $i = 1, \ldots, n$

Data and linear assumption implies $(\vec{x}^{(i)})^T \vec{\theta} = y^{(i)}$ for $i = 1, \ldots, n$

Matrix Notation:

$$\boldsymbol{X}\vec{\theta} = \vec{y}$$

Data Matrix:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & & x_d^{(2)} \\ & \vdots & & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}^{(1)T} \\ \vec{x}^{(2)T} \\ \vdots \\ \vec{x}^{(n)T} \end{bmatrix}$$

Target vector:

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

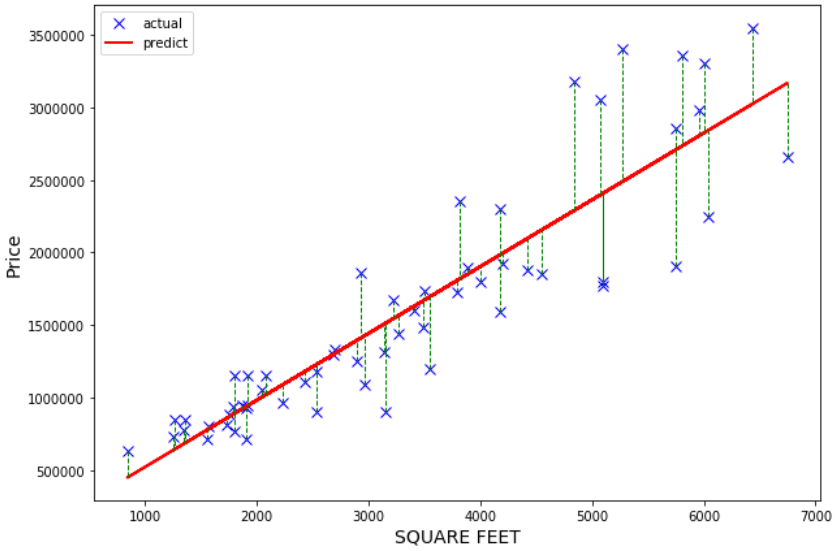
Parameter vector:

$$\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

- **Evaluate the model:**

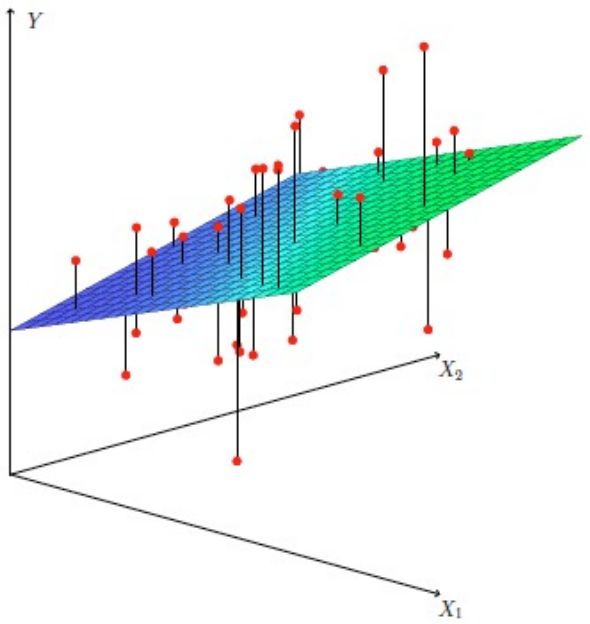
Prediction Vector:

$$h(\boldsymbol{X}) := \begin{bmatrix} h\left(\vec{x}^{(1)}\right) \\ h\left(\vec{x}^{(2)}\right) \\ \vdots \\ h\left(\vec{x}^{(n)}\right) \end{bmatrix}$$



Difference Vector

$$h(\boldsymbol{X}) - \vec{y} = \begin{bmatrix} h\left(\vec{x}^{(1)}\right) - y^{(1)} \\ h\left(\vec{x}^{(2)}\right) - y^{(2)} \\ \vdots \\ h\left(\vec{x}^{(n)}\right) - y^{(n)} \end{bmatrix}$$



Cost/Loss Functions

- **Mean** Absolute Error

$$L(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left|h_\theta(\vec{x}^{(i)}) - y^{(i)}\right|$$

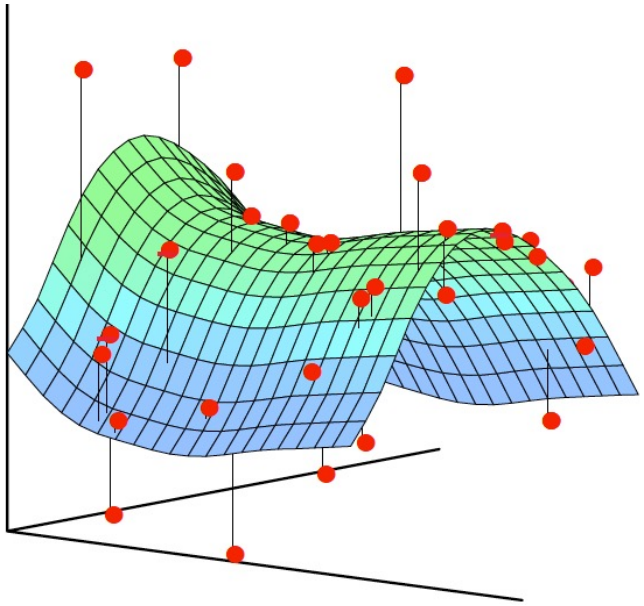
- **Mean** Residual Sum of Squares

$$L(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(\vec{x}^{(i)}) - y^{(i)}\right)^2$$

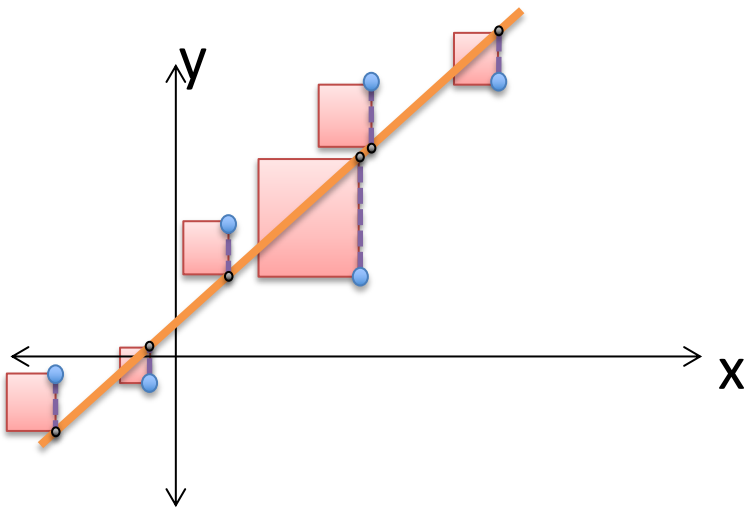
- **Residual Sum of Squares (RSS):**

$$RSS(\vec{\theta}) := \sum_{i=1}^{n}\left(h_\theta(\vec{x}^{(i)}) - y^{(i)}\right)^2$$

$$= \|h_\theta(\boldsymbol{X}) - \vec{y}\|^2$$



Picture Interpretation of RSS



➢ Review: Inner Product, Norm, Metric on vector spaces V

Let V be a real vector space. For example, V is a subspace of $\mathbb{R}^n$.

Definition (Inner Product). An **inner product** on V is a binary function

$$\langle -, - \rangle : V \times V \longrightarrow \mathbb{R}$$

such that for vectors $\vec{u}, \vec{v}, \vec{w} \in V$ and a scalar $c \in \mathbb{R}$, the following hold:

(1.) $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$

(2.) $\langle \vec{u} + \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle + \langle \vec{v}, \vec{w} \rangle$

(3.) $\langle c\vec{u}, \vec{v} \rangle = c\langle \vec{v}, \vec{u} \rangle$

(4.) $\langle \vec{u}, \vec{u} \rangle \geq 0$

(5.) $\langle \vec{u}, \vec{u} \rangle = 0$ if and only if $\vec{u} = \vec{0}$

We call V an **inner product space** with inner product $\langle -, - \rangle$.

Example: Dot product on $\mathbb{R}^n$.

Example: Weighted dot product on $\mathbb{R}^n$.

$$\langle \vec{u}, \vec{v} \rangle_W := \vec{u}^T W \vec{v}$$

Here, W is a positive-definite symmetric matrix

Definition (Norm). Let V be a real vector space. A **norm** on V is a function

$$\|-\|: V \longrightarrow \mathbb{R}$$

such that for vectors $\vec{u}, \vec{v} \in V$ and a scalar $c \in \mathbb{R}$, the following hold:

(1.) $\|\vec{u}\| \geq 0$

(2.) $\|\vec{u}\| = 0$ if and only if $\vec{u} = \vec{0}$

(3.) $\|c\vec{u}\| = c\,\|\vec{u}\|$

(4.) The triangle inequality $\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$

We call V an normed space with norm $\|-\|$.

Example: l_2-norm induced by dot product.

Example: l_p-norm on $\mathbb{R}^n$

Example: l_p-cost function $L(\vec{\theta}) = \|h_\theta(\boldsymbol{X}) - \vec{y}\|_p$

Definition (Metric). Let S be a **set**. A **metric**(distance) on S is a binary function

$$d: S \times S \longrightarrow \mathbb{R}$$

such that for vectors $\vec{u}, \vec{v}, \vec{w} \in S$ and a scalar $c \in \mathbb{R}$, the following hold:

(1.) $d(\vec{u}, \vec{v}) = d(\vec{v}, \vec{u})$

(2.) $d(\vec{u}, \vec{v}) = 0$ if and only if $\vec{u} = \vec{v}$

(3.) $d(\vec{u}, \vec{w}) \leq d(\vec{u}, \vec{v}) + d(\vec{v}, \vec{w})$

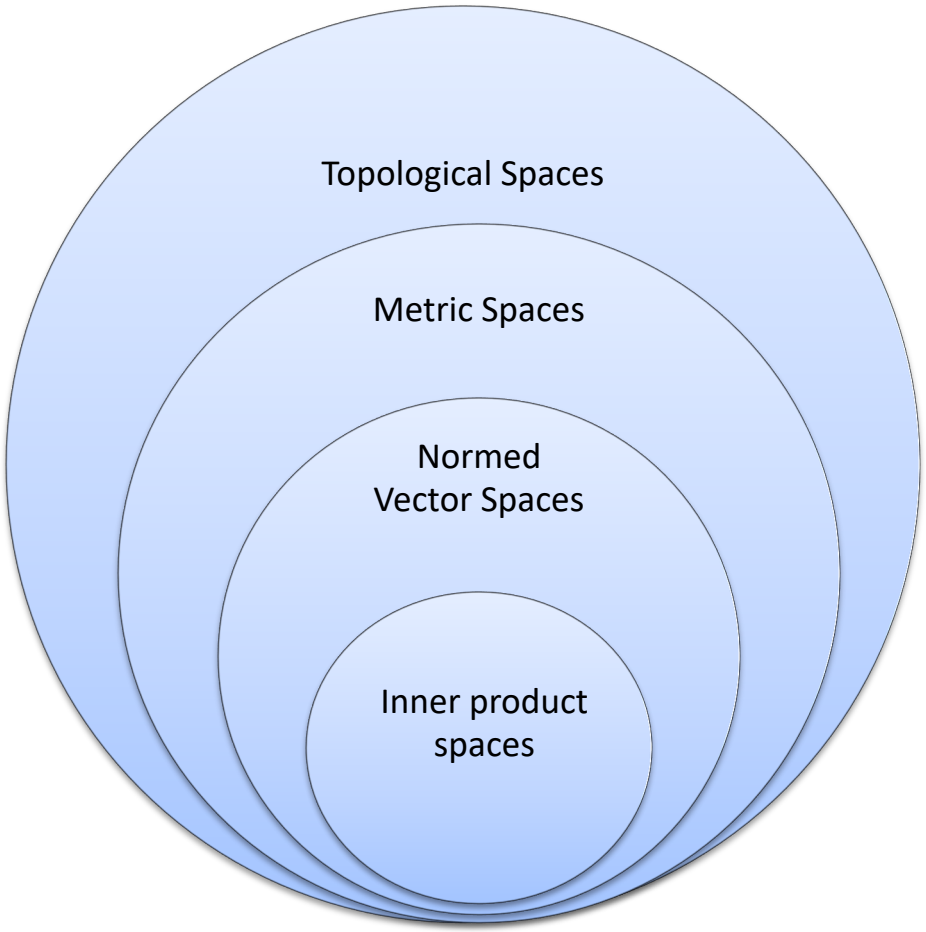
We call S a metric space metric function d.

Examples:

1. If S is a vector space, metric is equivalent to norm.
2. The **discrete metric on S**, where $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ otherwise.
3. The **positive real numbers** with distance function $d(x, y) = |\log(y/x)|$ is a metric space.

Given a distance function d on the label space $\mathcal{C}^n$

- Cost/Loss Function: $L(\vec{\theta}) = d(h_\theta(\boldsymbol{X}), \vec{y})$


Topological Spaces
Metric Spaces
Normed
Vector Spaces
Inner product
spaces

➢ Minimize Cost/Loss Functions

- Given labeled **Data** $\mathcal{D} = \left\{\left(\vec{x}^{(i)}, y^{(i)}\right),\ i = 1, \dots n\right\}$
- Assumption(Model): $h_\theta(-)$
- Cost/Loss Function: $L\left(\vec{\theta}\right) = d(h_\theta(\boldsymbol{X}), \vec{y})$ for some distance d on the label space.

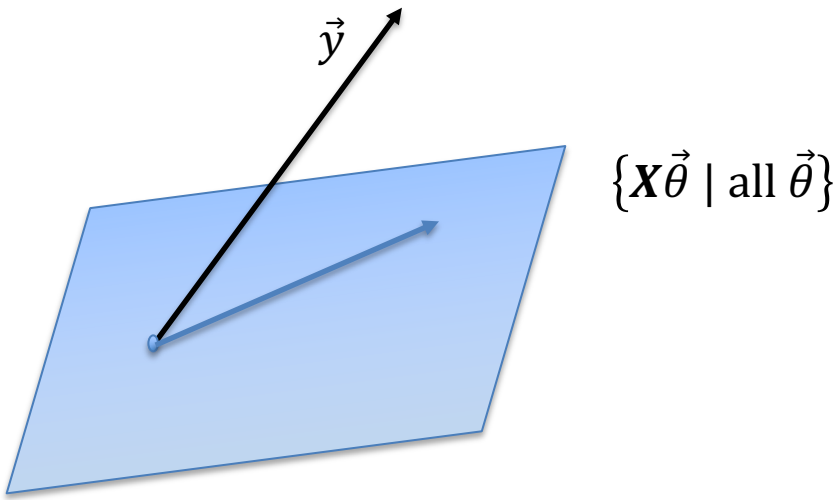
Goal: Find $\vec{\boldsymbol{\theta}}$ to minimize the cost $\boldsymbol{L\left(\vec{\theta}\right)}$

Equivalently, find $\underset{\vec{\theta}}{\mathbf{argmin}}\, \boldsymbol{L(\vec{\theta})} = \{\vec{\theta}$ such that $\boldsymbol{L\left(\vec{\theta}\right)}$ is minimized$\}$

➢ Minimize Cost/Loss Functions (linear regression)

- Given labeled **Data** $\mathcal{D} = \left\{\left(\vec{x}^{(i)}, y^{(i)}\right),\ i = 1, \ldots n\right\}$, where $y^{(i)} \in \mathbb{R}$
- Assumption(**Linear** Model): $h_\theta(\vec{x}\,) = \vec{x}^T\vec{\theta}$
- Cost/Loss Function: $L\left(\vec{\theta}\right) = d(h_\theta(\boldsymbol{X}), \vec{y})$ for some distance d on $\mathbb{R}^n$.
- **Find** $\underset{\vec{\theta}}{\operatorname{argmin}}\, L(\vec{\theta}) = \{\vec{\theta} \text{ such that } L\left(\vec{\theta}\right) \text{ is minimized}\}$

  Minimize the cost is the same as minimize the distance from $\vec{y}$ to $\boldsymbol{X}\vec{\theta}$



- If we the norm/distance is induced by an **inner product**, then the solution of $\underset{\vec{\theta}}{\operatorname{argmin}} L(\vec{\theta})$ is

$$\boldsymbol{X}\,\vec{\theta} = \mathrm{Proj}_{\mathrm{im}(\mathbf{X})}\vec{y}$$

- If we the norm is induced by **dot product**, the cost function is the residual sum of squares:

$$L(\vec{\theta}) = RSS(\vec{\theta}) \coloneqq \|h_\theta(\boldsymbol{X}) - \vec{y}\|^2 = \left\|\boldsymbol{X}\vec{\theta} - \vec{y}\right\|^2 = \sum_{i=1}^{n}\left(\left(\vec{x}^{(i)}\right)^T\vec{\theta} - y^{(i)}\right)^2$$

then the solution of $\underset{\vec{\theta}}{\operatorname{argmin}} L(\vec{\theta})$ is obtained by solving

$$\boldsymbol{X^T X}\vec{\theta} = \boldsymbol{X^T}\vec{y}$$

This is called the **normal equation** of $\boldsymbol{X}\vec{\theta} = \vec{y}$

Lemma: Rank $X^T X$ = Rank X

- If rank $X = d + 1$, then $X^T X$ is invertible.

In this case, the solution for the normal equation is

$$\boxed{\vec{\theta} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T} \vec{y}}$$

The matrix $H = X(X^T X)^{-1} X^T$ is the projection matrix.

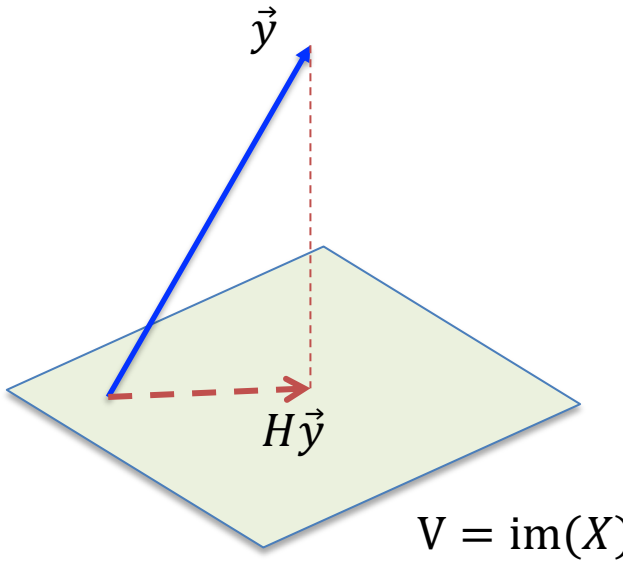
$$\hat{y} = X\vec{\theta} = H\vec{y} = \text{Proj}_{\text{V}}\, \vec{y}$$

H is symmetric and idempotent.

Eigenvalues of H are either 1 or 0.

Trace(H)= rank(H)

$$\|\vec{y}\|^2 = \left\|X\vec{\theta}\right\|^2 + \left\|\vec{y} - X\vec{\theta}\right\|^2$$



- If rank $X = d + 1$ and $X = QR$ where Q is an orthogonal matrix and R is an upper triangular matrix, then the solution for the normal equation is

$$\vec{\theta} = R^{-1}Q^T\vec{y}$$

- Suppose has the singular value decomposition $X = UDV$

Remarks:

- If we the norm $\|\vec{u}\|_W := \langle \vec{u}, \vec{u} \rangle_W$ is induced by **weighted inner product** $\langle \vec{u}, \vec{v} \rangle_W := \vec{u}^T W \vec{v}$, the cost function is the:

$$L(\vec{\theta}) = \|\boldsymbol{X}\vec{\theta} - \vec{y}\|_W^{\ 2}$$

The solution of $\underset{\vec{\theta}}{\operatorname{argmin}} L(\vec{\theta})$ is by solving the weighted normal equation

$$\boldsymbol{X^T W X \vec{\theta} = X^T W \vec{y}}$$

- If the norm is not induced by inner product (e.g., the l_p-norm), then finding $\underset{\vec{\theta}}{\operatorname{argmin}} L(\vec{\theta})$ is a hard optimization question. We will need to use gradient descent or Newton's method to minimize the cost $L(\vec{\theta})$)

Matrix Calculus

$$\boldsymbol{RSS}(\vec{\boldsymbol{\theta}}) = \|\boldsymbol{X}\vec{\theta} - \vec{y}\| = (\boldsymbol{X}\vec{\theta} - \vec{y})^T(\boldsymbol{X}\vec{\theta} - \vec{y})$$

$$= (\vec{\theta}^T\boldsymbol{X}^{\boldsymbol{T}} - \vec{y}^T)(\boldsymbol{X}\vec{\theta} - \vec{y})$$

$$= \vec{\theta}^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\vec{\theta} - \vec{\theta}^T\boldsymbol{X}^{\boldsymbol{T}}\vec{y} - \vec{y}^T\boldsymbol{X}\vec{\theta} + \vec{y}^T\vec{y}$$

$$= \vec{\theta}^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\vec{\theta} - 2\vec{y}^T\boldsymbol{X}\vec{\theta} + \vec{y}^T\vec{y}$$

Quadratic Linear Constant

To minimize $RSS(\vec{\theta})$, we need to find critical points.

➢ Matrix Calculus

Definitions. (Gradient/Partial derivative)

(1) If $f: \mathbb{R}^n \to \mathbb{R}$, the **gradient** of f is defined to be

$$\frac{\partial f}{\partial \vec{x}} := \nabla_{\vec{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

(2) If $f: \mathbb{R}^{m \times n} \to \mathbb{R}$, the **gradient** of f is defined to be

$$\frac{\partial f}{\partial X} := \nabla_X f := \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}$$

(3) If $F: \mathbb{R}^n \to \mathbb{R}^m$, the **derivative** of F is defined to be

$$\frac{\partial F}{\partial \vec{x}} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Remark: The above notation is called **denominator** layout notation, which is a generalization of the gradient.

There is another **numerator layout** convention for the derivative of F, which is the transpose of the denominator layout

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Theorem(Linear Property)

Let f and g be functions and c be a real number.

(1) $\nabla(f+g) = \nabla f + \nabla g$

(2) $\nabla(cf) = c\nabla f$

Here ∇ can be $\nabla_{\vec{x}}$, or ∇_X, or $\frac{\partial}{\partial \vec{x}}$

Proposition: If $f(\vec{x}) = \vec{b}^T\vec{x}$, then $\nabla f = \vec{b}$.

Proposition: If $F(\vec{x}) = A\vec{x}$, then $\frac{\partial F}{\partial \vec{x}} = A^T$

Proposition: (quadratic function):

If $f(\vec{x}) = \vec{x}^T A\vec{x}$, then $\nabla f = (A^T + A)\vec{x}$

Proof: $f(\vec{x}) = \vec{x}^T A\vec{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} a_{i1}x_i + \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{i=1}^{n} a_{in}x_i + \sum_{j=1}^{n} a_{nj}x_j + \end{bmatrix} = A^{\mathrm{T}}\vec{x} + A\vec{x}.$$

Example: $J(\vec{\theta}) = RSS(\vec{\theta}) = \vec{\theta}^T \boldsymbol{X}^T \boldsymbol{X} \vec{\theta} - 2\vec{y}^T \boldsymbol{X} \vec{\theta} + \vec{y}^T \vec{y}$

$\nabla_{\vec{\theta}} J =$

Theorem: (Product Rule) (denominator layout)

Suppose $F: \mathbb{R}^n \to \mathbb{R}^m$ and $G: \mathbb{R}^n \to \mathbb{R}^m$ and $H = G^T F$. Then,

$$\frac{\partial H}{\partial \vec{z}} = \frac{\partial G}{\partial \vec{z}} F + \frac{\partial F}{\partial \vec{z}} G$$

Definition: The **Hessian matrix** of $f: \mathbb{R}^n \to \mathbb{R}$ is defined by

$$H(f) := \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Proposition: If $f(\vec{x}) = \vec{b}^T \vec{x}$, then $H(f) = 0$.

Proposition: If $f(\vec{x}) = \vec{x}^T A \vec{x}$, for a symmetric matrix A, then $\mathrm{H}(f) = 2A$

Example: $J(\vec{\theta}) = RSS(\vec{\theta}) = \vec{\theta}^T \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} \vec{\theta} - 2\vec{y}^T \boldsymbol{X} \vec{\theta} + \vec{y}^T \vec{y}$

Theorem: (Second Derivative Test)

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is smooth, then a **critical point** $\vec{a} \in \mathbb{R}^n$ (i.e., $\nabla f(\vec{a}) = \vec{0}$,) is

(1) a local minimum if $H(f(\vec{a}))$ is positive definite;

(2) a local maximum if $H(f(\vec{a}))$ is negative definite;

(3) a saddle point if $H(f(\vec{a}))$ contains positive and negative eigenvalues;

(4) there is no conclusion for the other cases.

Examples:

$$f(x,y) = x^2 + y^2$$

$$f(x,y) = -x^2 - y^2$$

$$f(x,y) = x^2 - y^2$$

$f(x,y) = -x^4$

$f(x,y) = x^4$

$f(x,y) = x^2 + y^3$

Example: $f(x,y) = (x+y)(xy + xy^2)$

https://www.geogebra.org/3d/bjsj7erx

Definition: A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **convex** if

$$f(\lambda\,\vec{u} + (1-\lambda)\vec{v}\,) \leq \lambda\, f(\vec{u}) + (1-\lambda) f(\vec{v}) \qquad \text{for any } 0 \leq \lambda \leq 1$$

Definition: A set C is **convex** if and only if

$$\vec{u}, \vec{v} \in C \implies \lambda\vec{u} + (1-\lambda)\vec{v} \in C \quad \text{for any } 0 \leq \lambda \leq 1$$

Theorem: A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is smooth and convex if and only if

$$f(\vec{u}) - f(\vec{v}) \geq \nabla f \cdot (\vec{u} - \vec{v}) \qquad \text{for all } \vec{u}, \vec{v}$$

- Any local minimum of a convex function is also a global minimum.
- If its Hessian $H(f(\vec{x}))$ is everywhere positive semi-definite, then f is convex.