MSBA Boot Camp 2024

Math 1: Calculus + Linear Algebra

About Me

Chinese Name: 张帷 (ZHANG, Wei)

Job Title: MSBA Programme Director

Rank: Associate Professor

Hometown: Chengdu

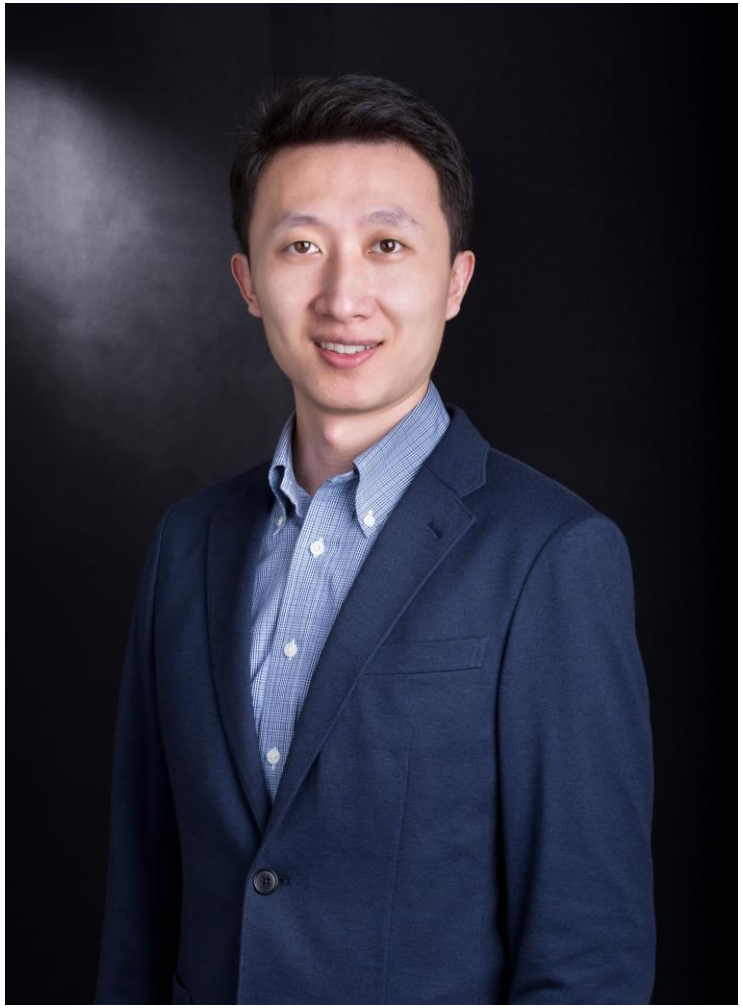
BA & MS at Tsinghua University

Ph.D. at UCLA

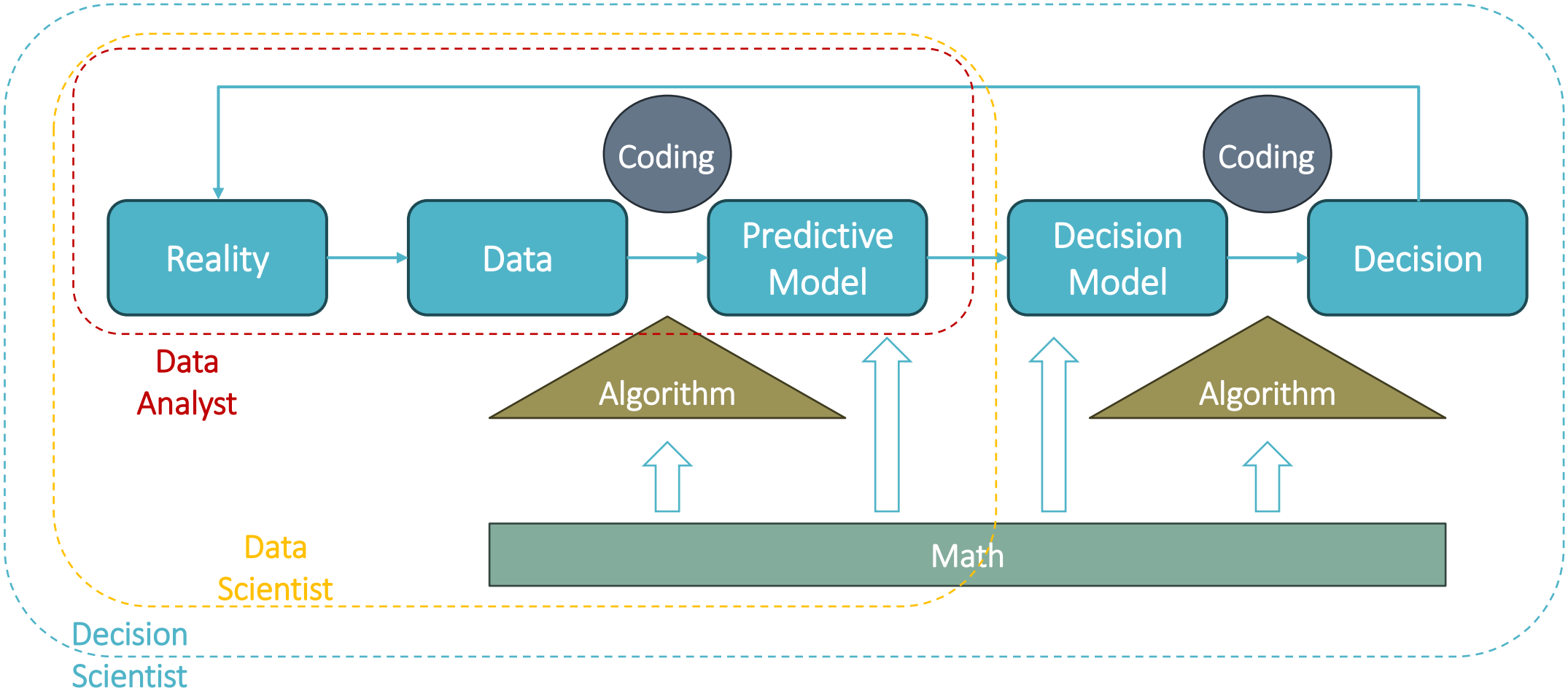
Office: K.K. Leung 814

Email: wzhang15@hku.hk

Tel: 3917 1685



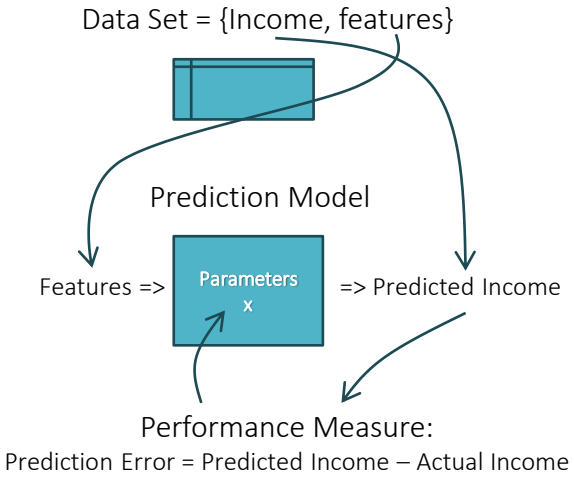
Analytics Process

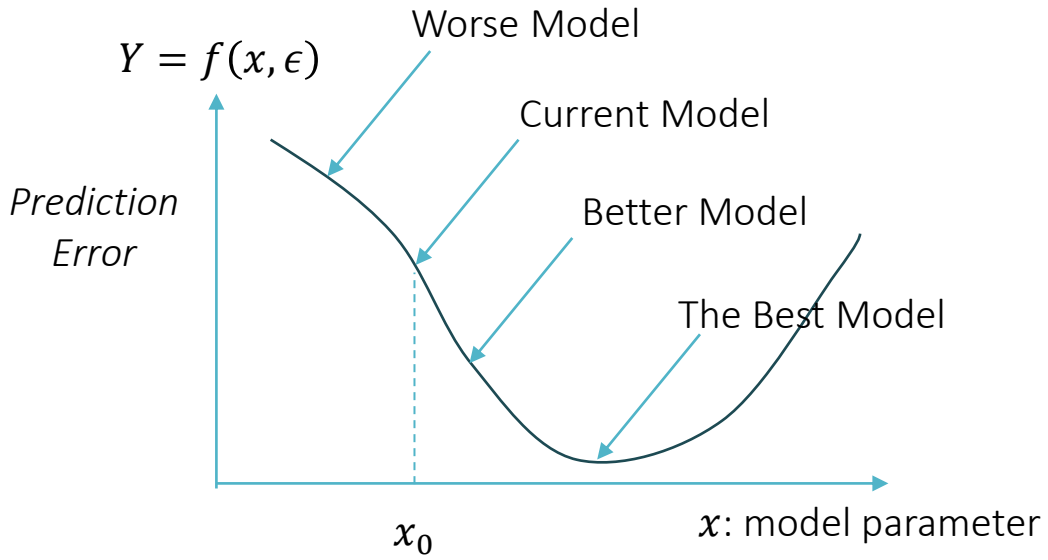


ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Why Calculus?

Calculus is the mathematical study of change: how the change of the inputs of a system affects the output, either in a marginal way (i.e., differential calculus) or a cumulative way (i.e., integral calculus).

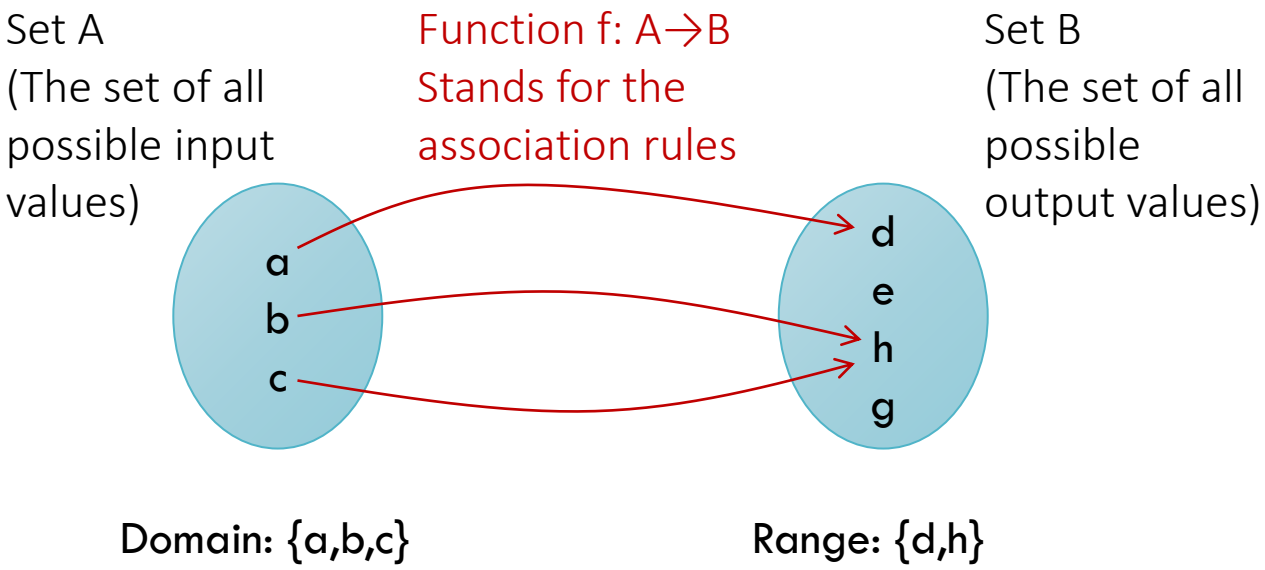




A Typical Machine Learning Process

Function

In short, a function is a relationship between some inputs and an output.



If function f is defined like this, then we say that f(a)=d and f(b)=f(c)=h.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

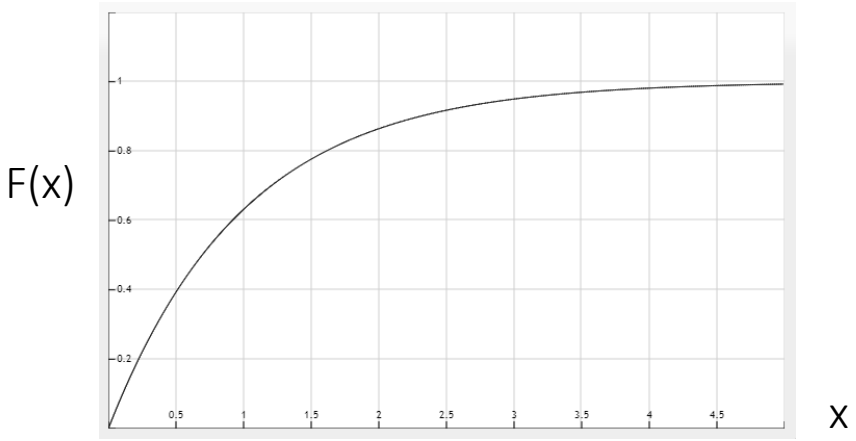
Function

For example, profit is a function of revenue and cost:

profit = f(revenue, cost) = revenue – cost

The cumulative probability function F of a random variable X maps a particular value x to a probability value between 0 and 1. For example, if X follows Exponential distribution with a mean of 1, then

Cumulative Probability Pr(X≤x) = F(x) = 1 – exp(-x)



ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

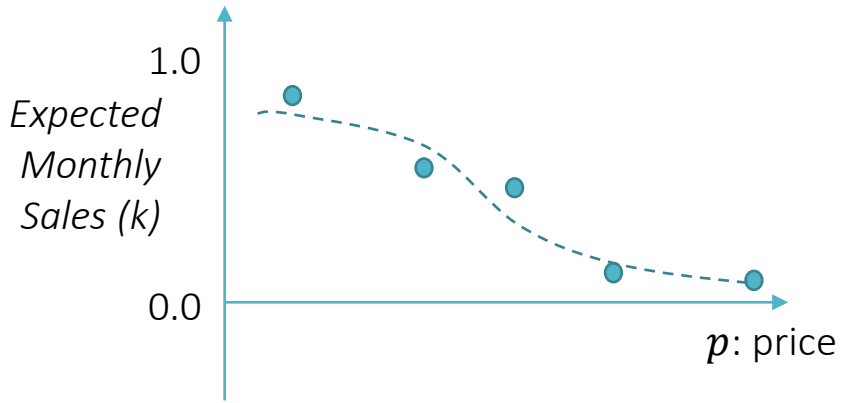
Function

Sometimes there is a relationship between X and Y such that Y is uniquely determined given X holding everything else constant, but we do not know the functional form or the association rule.

Statistical learning is the process of learning the functional form between X and Y with a certain amount and scope of data.

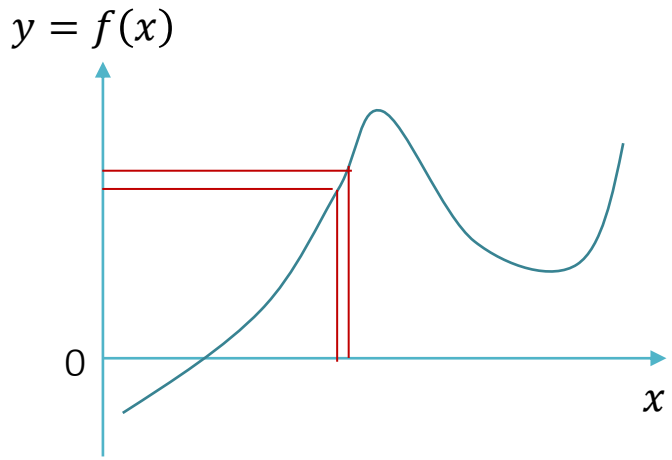
For example, the expected sales number of a product next month is a function of the price (and of course some other factors)! The total squared error is a function of the parameters.

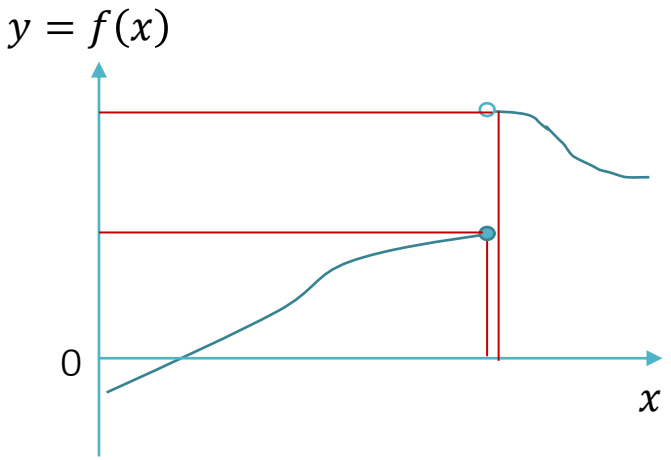
$$S = f(p, \epsilon) = a + bp + cp^2 + dp^3 + \epsilon$$



ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Continuity





A function is continuous if for a small enough change in x, the change in y is also small enough.

Continuity is important for optimization. Why? Derivative or linear search cannot be done.

Examples of continuous functions: $f(x) = x^2 - x$, $f(x) = e^{1+x}$, and $f(x) = |x|$.

Examples of discontinuous functions: $f(x) = (1 - x)^{-1}$.

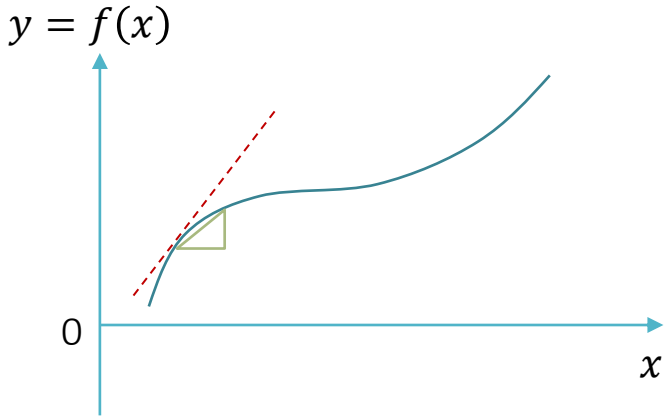
ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Differentiation

The derivative of a function $f(x)$ with respect to x is written as $f'(x)$. It is the ratio of the small change in f over the small change in x. It measures how fast f changes with x.

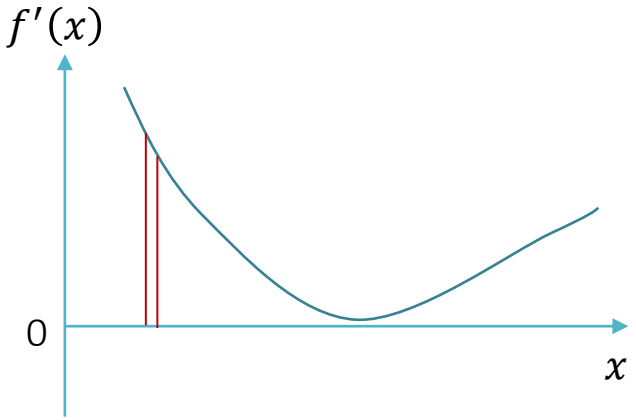
Mathematically, $f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x}$.

If $f'(a) = 0$, it means f does not change with x when $x = a$.



The derivative $f'(x)$ is also a function. The integral $\int_{x_1}^{x_2} f'(x)dx$ measures the change of function value $f(x)$ from x_1 to x_2.

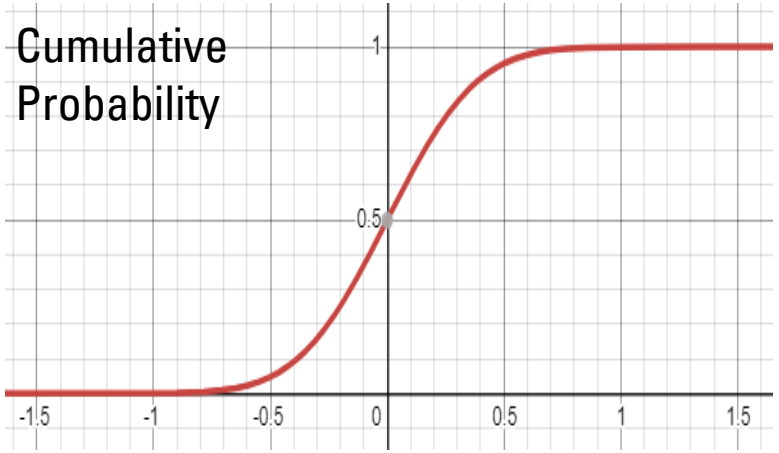
Example: $f(x) = 1 + x - x^2$. Check the rate of change.

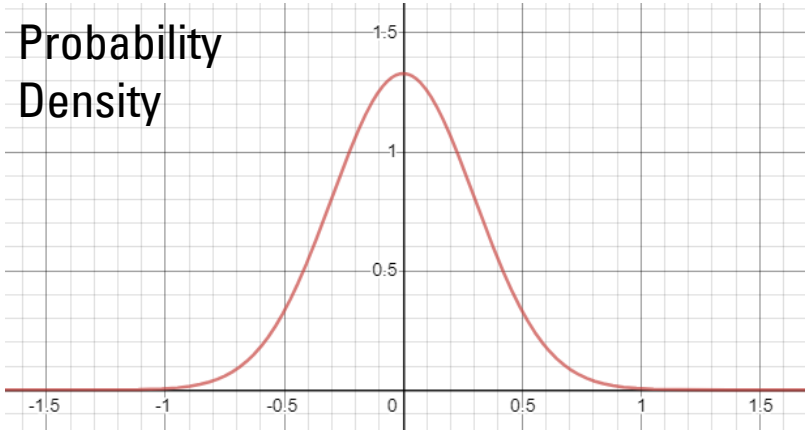


ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Probability Density

For continuous random variables, the probability density function is the derivative of the cumulative probability. It measures how likely the random variable takes a value around a point. Why?



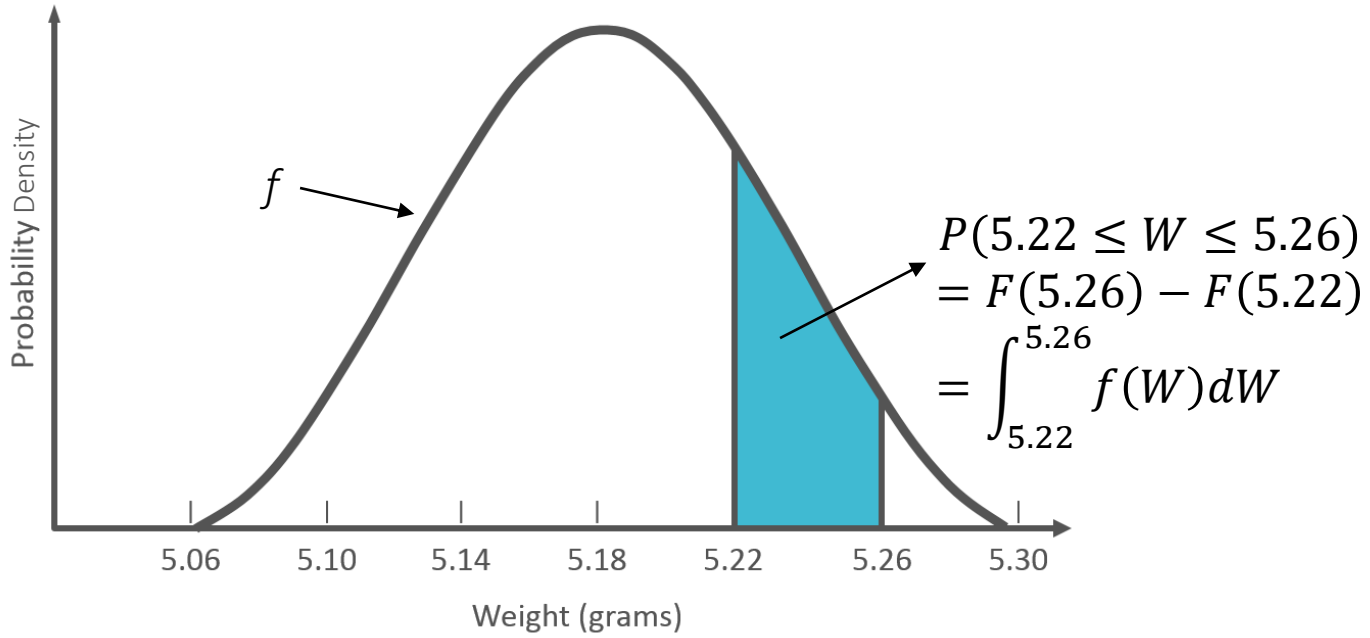


Probability density is not probability!

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Probability Density

W = the weight of a particular machined part



ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

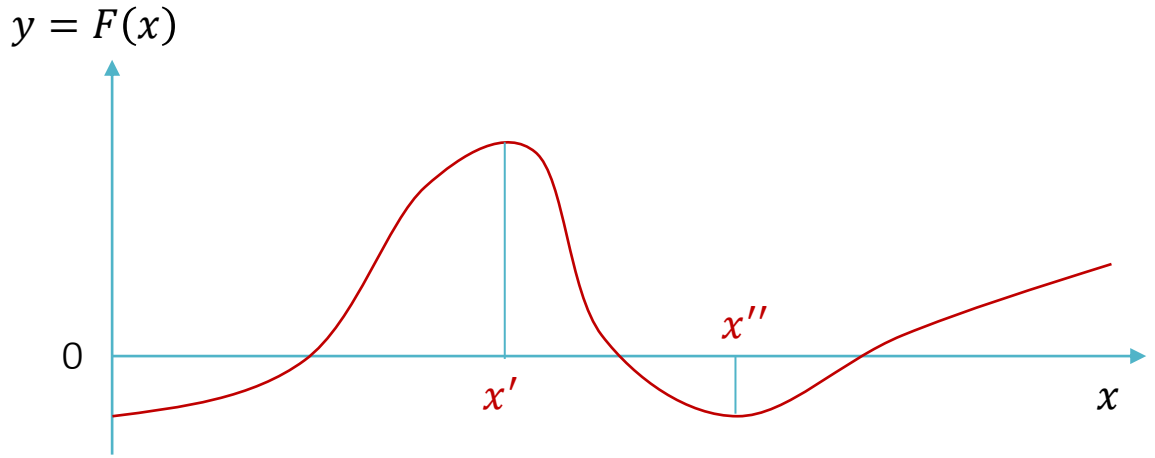
Optimization

How to teach a machine to find the maximizer of $F(x)$ given x in $[x', x'']$ without knowing the shape of the curve?

Monotonic?

U-shaped?

Inverted-U?



Check the sign of $F'(x)$! When the sign changes from positive to negative, we are at a local maximum.

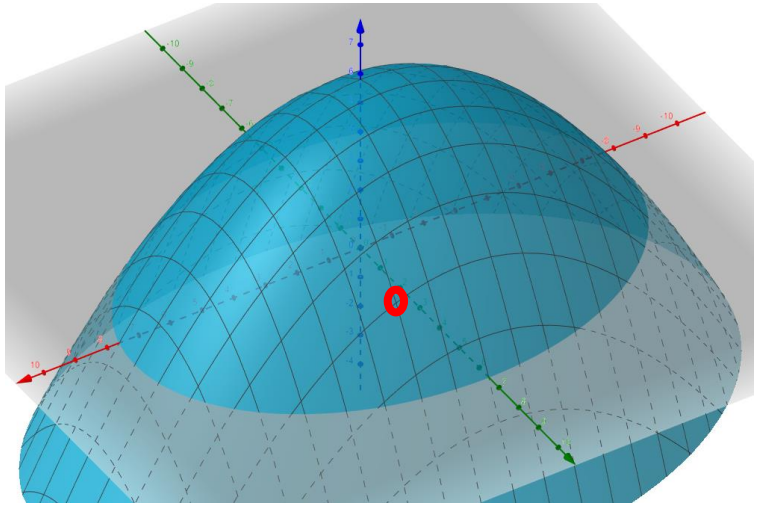
What is the maximizer of f(x) =3x - e^{2x} without boundaries? What if the domain of x is [1, 2]?

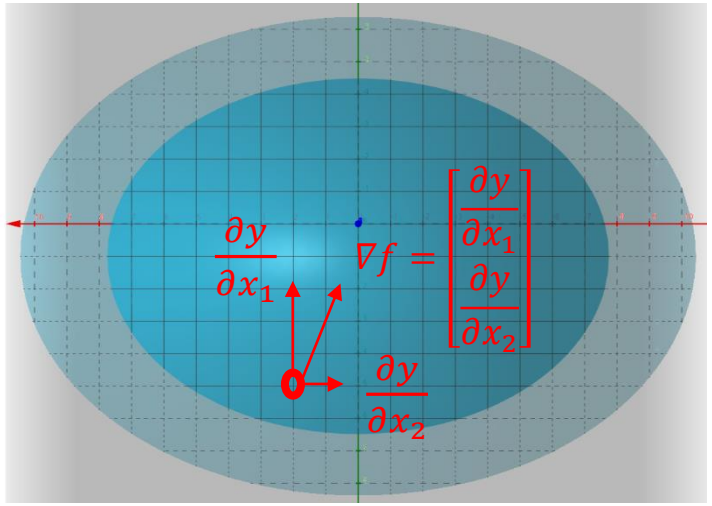
ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Partial Differentiation

If y is a function of two variables, e.g., $y = f(x_1, x_2)$, then how y changes with one variable is called a partial derivative. For example, we are interested in how productivity depends on a worker's gender and age, respectively and jointly.

The partial derivative $\partial y/\partial x_1$ is just the usual derivative with respect to x_1 when x_2 is treated as a constant. The vector of partial derivatives given (x_1, x_2) is called the gradient of f at (x_1, x_2):





The gradient is always pointing to the direction to which the function value can be increased the most.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Partial Differentiation

Compute the partial derivatives of the following function:

$$y = 2z - x \cdot z + x^2$$

$$y = \log(x); \frac{dy}{dx} = \frac{1}{x}$$

$$y = \exp(x); \frac{dy}{dx} = \exp(x)$$

$$y = x^c; \frac{dy}{dx} = cx^{c-1}$$

$$y = c; \frac{dy}{dx} = 0$$

$$y = f(x), u = g(x); \frac{d(y+u)}{dx} = \frac{dy}{dx} + \frac{du}{dx}$$

$$y = f(x), u = g(x); \frac{d(y \cdot u)}{dx} = u \cdot \frac{dy}{dx} + y \cdot \frac{du}{dx}$$

$$y = f(x), u = g(x); \frac{d(y/u)}{dx} = \left(u \cdot \frac{dy}{dx} - y \cdot \frac{du}{dx}\right) / u^2$$

$$y = f(u), u = g(x); \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Gradient Descent/Ascent

Minimize: $f(x,z) = 8x - z + 3x^2 + 5z^2 + 2xz$.

Gradient: $\frac{\partial f}{\partial x} = 8 + 6x + 2z$, $\frac{\partial f}{\partial z} = -1 + 10z + 2x$.

Step size = 0.1

Starting point: (0,0) => gradient = (+8,-1)

Next point: (-0.8,+0.1) => gradient = (+3.4,-1.6)

Next point: …

…

Finally, the points converge to (-1.4643, 0.3929).

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Why Linear Algebra?

Linear algebra is deeply related to geometry.

It gives geometric interpretation to data of multiple dimensions.

ID	Age	Income	Spending
A	36	99	50
B	20	70	60
C	45	120	100

U.D.	A	B	C
A	-	-	-
B	34.6	-	-
C	55.0	68.7	-

Cos	A	B	C
A	-	-	-
B	0.969	-	-
C	0.976	0.998	-

Question: how similar are these individuals?

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Vector

A vector is a set of numbers that represent a direction or a point in a high dimensional space.

For example, a two-dimensional vector $v = (1, 2)'$ or $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is a point in a 2D space with coordinate (1,2) or the direction pointing from the origin to the point (1,2).

The *sum* of two vectors is the diagonal vector of the parallelogram spanned by the two vectors.

Vector

The *inner product* of two equal-dimension vectors ($v \cdot w$) is defined as the sum of the products of the corresponding components of the two vectors.

The inner product of a vector with itself is the **squared** *Euclidean norm* of the vector: $\|v\|^2$.

The inner product of two equal-dimension vectors ($v \cdot w$) is equal to the product of the projected norm in either direction: $v \cdot w = \|v\| \cdot \|w\| \cdot \cos\theta$, where θ is the degree of the angle formed by the two vectors. If $v \cdot w = 0$, then we say v and w are perpendicular.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Similarity

In data analysis (especially text mining), cosine similarity is used to measure the similarity of two documents.

We define a sequence of key words indexed by 1, 2, …, N and count the number of each key word in a document. Denote as $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ the vector of counts for document i.

The cosine similarity between documents i and j is defined as: $\cos(\theta_{ij}) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|} = \frac{\sum_{n=1}^{N} x_{in}x_{jn}}{\sqrt{\sum_{n=1}^{N} x_{in}^2 \sum_{n=1}^{N} x_{jn}^2}}$.

$\cos(\theta_{ij})$ ranges from -1 to 1. The closer is it to 1, the greater is the similarity.

If the scale (e.g., the length of document) does not matter, we can normalize $\boldsymbol{x}_i$ to be the vector of word frequencies. However, cosine similarity is not appropriate if the scale matters.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Matrix

An $m \times n$ (m-by-n) matrix A

$$A = [a_{ij}]_{m\times n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

A square matrix has the same number of rows and columns.

A matrix with only one row (column) is called a row (column) vector.

The identity matrix I is a square matrix defined as

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Matrix Algebra

The sum of two equal-dimension matrices: $A + B = [a_{ij} + b_{ij}]_{m \times n}$.

Let A be an $m \times n$ matrix and B be an $n \times p$ matrix.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \text{ and } B = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix}$$

The matrix multiplication is defined as $AB = \left[\sum_{k=1}^{n} a_{ik} b_{kj}\right]_{m \times p}$.

For a square matrix A, if there exists a square matrix B such that $AB = BA = I$, then A is invertible and the inverse of A is $A^{-1} = B$.

Properties of matrix multiplication: (1) (AB)C = A(BC). (2) (A + B)C = AC + BC. (3) A(B + C) = AB + AC. (4) AB ≠ BA. (5) AI = IA = A. (6) (AB)' = B'A'.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Linear Transformation

The product of an n-by-n matrix **A** and an n-dimensional vector **v** is still an n-dimensional vector **w**.

The new vector **w** is called the linear transformation of **v**, and matrix **A** is called the linear map.

Consider the example of

$$\begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{5}{\sqrt{10}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}.$$

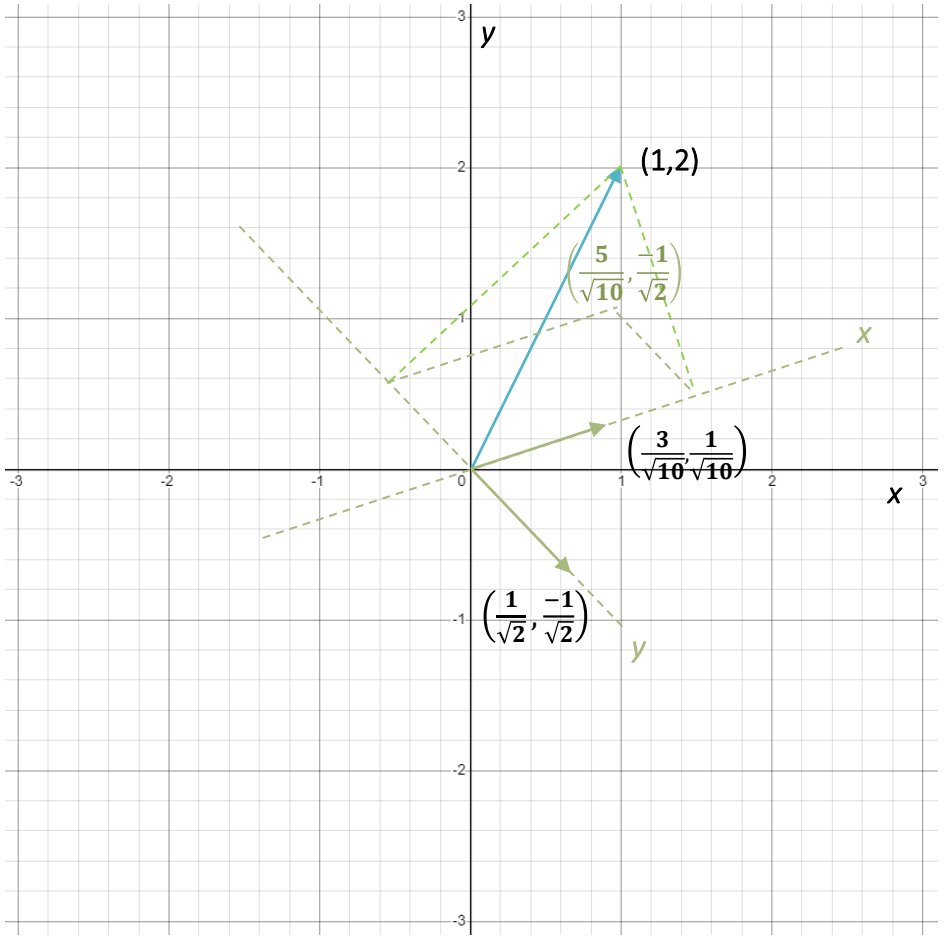
The input vector in the original 2D space is projected into a transformed 2D space formed by two new axes given by the two row vectors of the linear map: $\begin{bmatrix} \frac{3}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \end{bmatrix}$ and $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$.

Because the Euclidean norms (lengths) of the two vectors are both 1, there is no scaling.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Geometric Interpretations

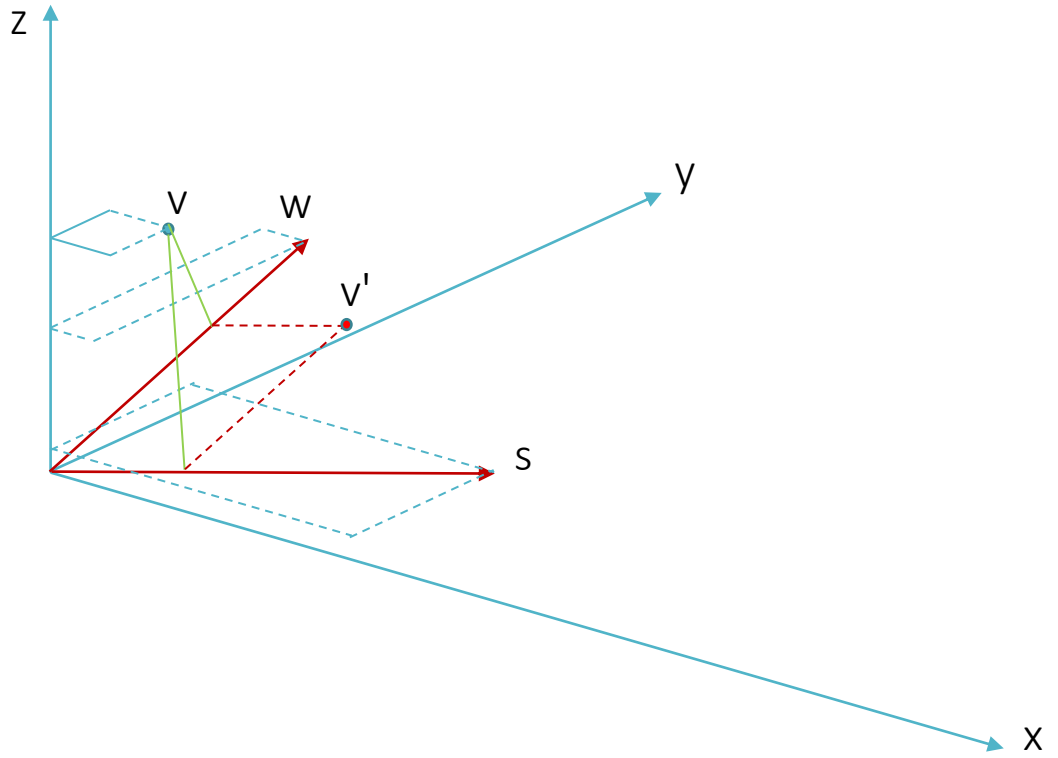
$$\begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{5}{\sqrt{10}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$



ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Dimension Reduction

For a high dimensional vector v, we can reduce its dimensionality and condense its information by projecting it into a lower dimensional space defined by a chosen linear map.



Invertible Matrix & Linear Dependency

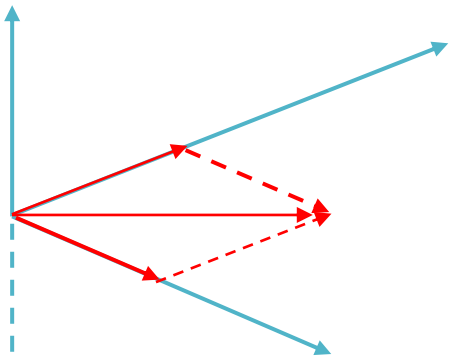
If 2-by-2 matrices A and B are each other's inverse matrix, then we have

$$AB = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which means that (1) $\begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix}$ and $[a_{21} \quad a_{22}]$ are perpendicular to each other and that (2) $\begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix}$ and $[a_{11} \quad a_{12}]$ are perpendicular to each other. Similar results hold for n-by-n matrices.

Observation: if a square matrix is invertible, then we can find a linear map in which the row vectors are perpendicular to all but one different column of the matrix.

Corollary: If the column or row vectors of a square matrix are linearly dependent, then the matrix is not invertible and vice versa.



(Suppose the last column can be generated by some other columns and the inverse matrix exists. In the inverse matrix, the last row must be perpendicular to all but the last column. Because the last column is the in the plane formed by other columns, the last row of the inverse matrix must be perpendicular to the last column. A contradiction.)

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Invertible Matrix & Linear Dependency

Are the following matrices invertible?

(1) $\begin{bmatrix} 1 & 3 \\ 2.5 & 7.5 \end{bmatrix}$

(2) $\begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & -2 & 1 \end{bmatrix}$

(3) $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 1 \end{bmatrix}$ (Hint: Try to solve the set of three linear equations with two unknowns.)

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Quadratic Functions

A quadratic function $f(x_1, \dots, x_n) = a_{11}x_1x_1 + a_{12}x_1x_2 + a_{21}x_2x_1 + \cdots + a_{nn}x_nx_n$ can be written as

$$f(x_1, \dots, x_n) = [x_1 \dots x_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x'Ax$$

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

$$= [x_1\ x_2] \begin{bmatrix} 1 & 1.5 \\ 1.5 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [x_1\ x_2] \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

For sum of squared values, $f(x_1, \dots, x_n) = x_1^2 + \cdots + x_n^2 = x'x$.

ZHANG, WEI - THE UNIVERSITY OF HONG KONG - WZHANG15@HKU.HK

Data Representation and Manipulation

If we have a data set containing information of 100 individuals regarding their ID number, gender, age, and income, then we can write the data set as a 100-by-4 matrix $X = [\ x_{ij}\ ]_{100\times4}$, where x_{ij} represents the value of the i-th individual's **j**-th attribute. For example, x_{13} represents the age of the first individual (or the first data point). If we assign a score y_i to the i-th individual according to the formula:

$y_i = b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4}$,

then we can describe the 100 scores as a column vector y = Xb, where $b=[\ b_1\ b_2\ b_3\ b_4\ ]'$ is a 4-by-1 column vector.

The sum of squared y values: $\sum_{i=1}^{100} y_i^2 = y'y = b'X'Xb$.

Define **1** as a 100-by-1 column vector with all 1s. Then the mean of y can be written as **1**'Xb/100.