머신런닝 개요

Lecture 4: Mathematics for ML

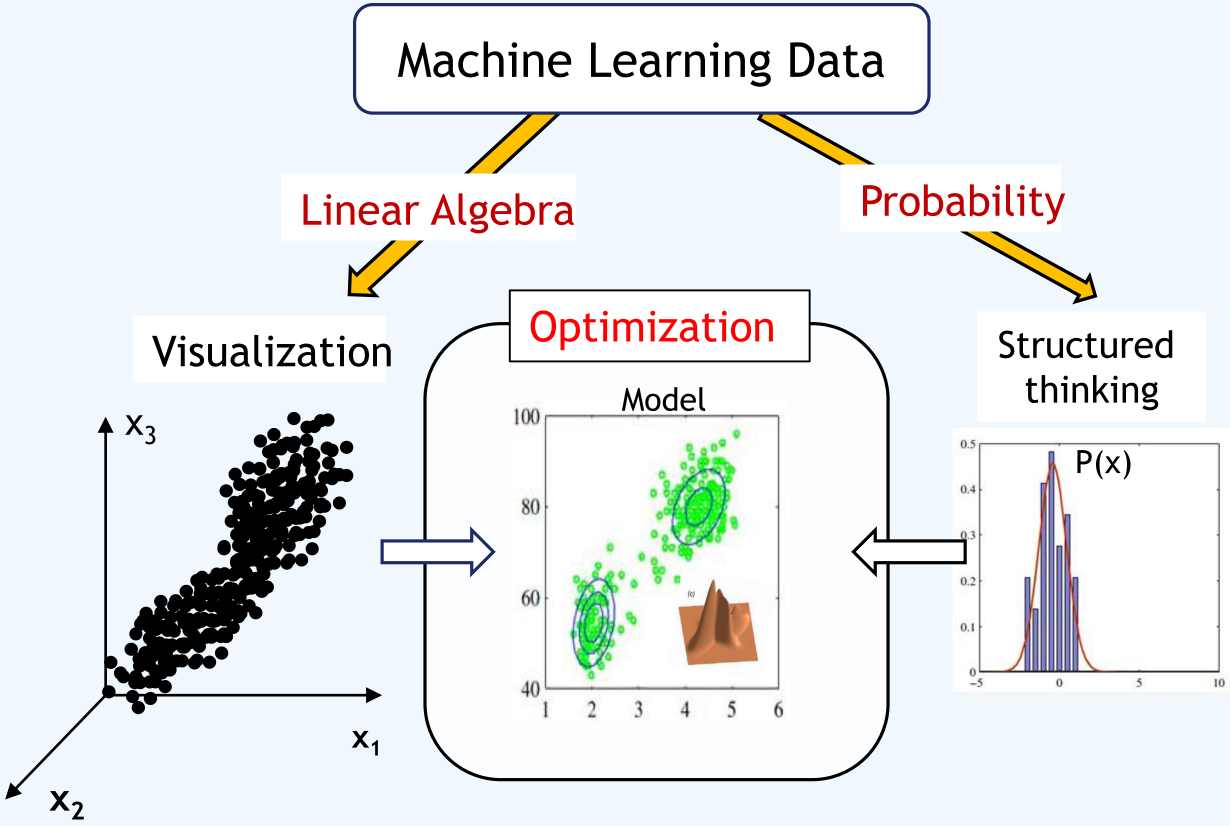
College of Information and Electronic Engineering

Kyung Hee Univeristy

Prof. Wonha Kim

(wonha@khu.ac.kr)

Math for Machine Learning



- 선형대수: 이 분야의 개념을 이용하면 학습 모델의 매개변수집합, 데이터, 선형연산의 결합 등을 행렬 또는 텐서로 간결하게 표현할 수 있다. 데이터를 분석하여 유용한 정보를 알아내거나 특징 공간을 변환하는 등의 과업을 수행하는 데 핵심 역할을 한다.
- 확률과 통계: 데이터에 포함된 불확실성을 표현하고 처리하는 데 활용한다. 베이즈 이론과 최대 우도 기법을 이용하여 확률 추론을 수행한다.
- 최적화: 목적함수를 최소화하는 최적해를 찾는 데 활용하며, 주로 미분을 활용한 방법을 사용한다. 수학자들이 개발한 최적화 방법을 기계 학습이라는 도메인에 어떻게 효율적으로 적용할지가 주요 관심사이다.

- Vector
  - 샘플을 특징 벡터로feature vector 표현
  - 예) Iris 데이터에서 꽃받침의 길이, 꽃받침의 너비, 꽃잎의 길이, 꽃잎의 너비라는 4개의 특징이 각각 5.1, 3.5, 1.4, 0.2인 샘플

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}$$

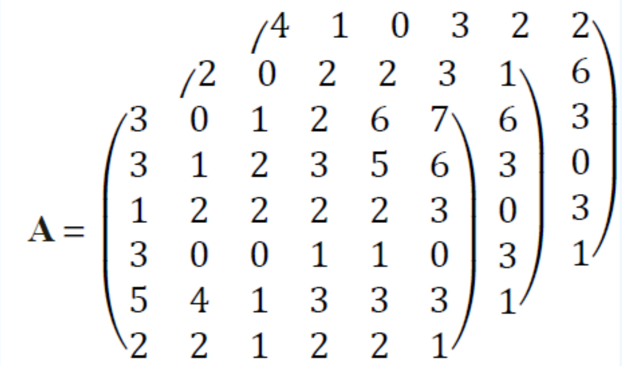
  - 여러 개의 특징 벡터는 첨자로 구분

$$\mathbf{x}_1 = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 4.9 \\ 3.0 \\ 1.4 \\ 0.2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 4.7 \\ 3.2 \\ 1.3 \\ 0.2 \end{pmatrix}, \cdots, \mathbf{x}_{150} = \begin{pmatrix} 5.9 \\ 3.0 \\ 5.1 \\ 1.8 \end{pmatrix}$$

- Matrix
  - vector의 배열

$$\mathbf{X} = \begin{pmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ \vdots & \vdots & \vdots & \vdots \\ 6.2 & 3.4 & 5.4 & 2.3 \\ 5.9 & 3.0 & 5.1 & 1.8 \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \\ x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{149,1} & x_{149,2} & x_{149,3} & x_{149,4} \\ x_{150,1} & x_{150,2} & x_{150,3} & x_{150,4} \end{pmatrix}$$

- Tensor
  - Matrix의 배열. 예) RGB color 영상



- **Transpose**

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix} \qquad \mathbf{A}^T = \begin{pmatrix} 3 & 0 \\ 4 & 5 \\ 1 & 2 \end{pmatrix}$$

– **Vector data들의 Matrix 표현**

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_{150}^T \end{pmatrix}$$

- **Special Matrix**

Identity Matrix: $\mathbf{I} \in \mathbb{R}^{n \times n}$

- $\mathbf{I} = \text{diag}(\mathbf{1})$
- $(\mathbf{I})_{ij} = I_{ij} = 1$ if $i = j$, 0 otherwise.
- $\mathbf{I}^T = \mathbf{I}$
- $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ (if $\mathbf{A}$ is square), $\quad \mathbf{Ix} = \mathbf{x}$
- $\mathbf{I} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, where only the i-th entries of $\mathbf{e}_i$ is 1 for all i.
- $\mathbf{e}_i^T \mathbf{e}_j = 1$ if $i = j$, 0 otherwise.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{I} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Diagonal Matrix: $\mathbf{D} \in \mathbb{R}^{n \times n}$

- $d_{ij} = 0$, if $i \neq j$
- $\mathbf{D} = \text{diag}(\mathbf{v}) \longrightarrow d_{ii} = v_i$
- $\mathbf{D}^T = \mathbf{D}$

$$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 8 \end{pmatrix}, \qquad \mathbf{D} = \begin{bmatrix} d_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{m} \end{bmatrix}$$

Symmetric Matrix: $\mathbf{S} \in \mathbb{R}^{n \times n}$

- $s_{ij} = s_{ji}$
- $\mathbf{S}^T = \mathbf{S}$

$$\begin{pmatrix} 1 & 2 & 11 \\ 2 & 21 & 5 \\ 11 & 5 & 1 \end{pmatrix}$$

Skew Symmetric(anti-symmetric) Matrix

- $s_{ij} = -s_{ji}$ and $s_{ii} = 0$
- $\mathbf{S}^T = -\mathbf{S}$

1. Linear Algebra : Matrix Operations

Addition

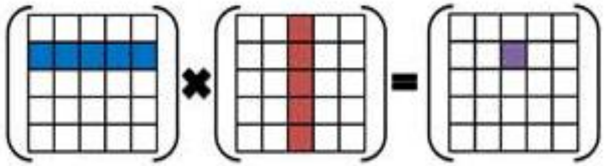
- $\mathbf{C} = \mathbf{A} + \mathbf{B}$ ► $c_{ij} = a_{ij} + b_{ij}$
- $\mathbf{D} = \mathbf{A} + k$ ► $d_{ij} = a_{ij} + k$
- $\mathbf{C} = \mathbf{A} + \mathbf{b}$ ► $c_{ij} = a_{ij} + b_j$ ► $\mathbf{c}_j = \mathbf{a}_j + b_j$ (*broadcasting*: convention in deep learning)

Multiplication

For $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{B} \in \mathbb{R}^{p \times n}$, $\mathbf{C} \in \mathbb{R}^{m \times n}$,

- $\mathbf{C} = \mathbf{AB}$ ► $c_{ij} = \sum_k a_{ik} b_{kj}$

  $(m \times p) \cdot (p \times n) = (m \times n)$
- $\mathbf{C} = k\mathbf{A}$ ► $c_{ij} = k a_{ij}$



$$\mathbf{A} = \begin{pmatrix} 3 & 4 & 1 \\ 0 & 5 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 0 & 5 \\ 4 & 5 & 1 \end{pmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} 14 & 5 & 24 \\ 13 & 10 & 27 \end{pmatrix}$$

Properties of basic operations

Provided that the dimension-matching for the matrix multiplication is satisfied,

- $\mathbf{AB} \neq \mathbf{BA}$ (quite often) – not commutative
- $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$ – associative
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ – distributive
- $\mathbf{A}^p = \mathbf{AAA} \cdots \mathbf{A}$ (p factors), $(\mathbf{A}^p)(\mathbf{A}^q) = \mathbf{A}^{p+q}$, $(\mathbf{A}^p)^q = \mathbf{A}^{pq}$
- $\mathbf{A}^0 = \mathbf{I}$ (identity matrix), $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ ($\mathbf{A}^{-1} \neq \mathbf{I} / \mathbf{A}$, there is no matrix division!)

- Definition of Vector Space V

Let V is a set of vectors defining

- Vector addition : $x + y \in V$ ($x, y \in V$, i.e, x, y are vector)
- Scalar multiplication : $\alpha x \in V$ ($x, y \in V$, $\alpha \in R$, i.e, α is a real number.)

- V is a vector space if the following 8 axioms are satisfied

1.(Community) : $x+y=y+x \in V$ 2.(Additive identity): $0 \in V$, $0+x=x$

3.(Associativity) : $(x+y)+z=x+(y+z)$ 4.(Additive inverse) : $-x \in V$, $-x+x=0$

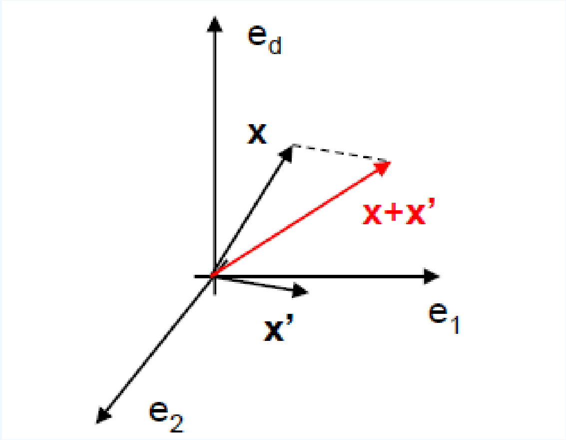
5.(Distributivity) : For a number α, $\alpha(x+y)= \alpha x+ \alpha y$

6.(Distributivity) : For number α, β, $(\alpha+ \beta)$ x = $\alpha x+ \beta$ y

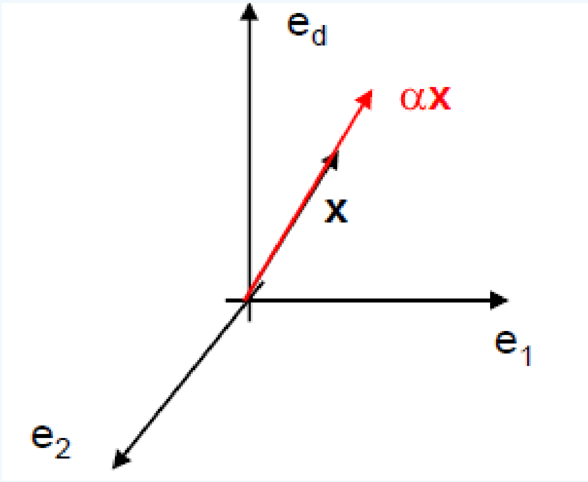
7.(Associativity) : $(\alpha\beta)$ x= $\alpha(\beta x)$

8. 1x=x

Vector Addition



Scalar multiplication



- Linear combination (spaces) : **선형결합으로 만들어지는 공간**

- For Y, $X_i \in$ V and $\alpha_i \in$R

$$Y = \alpha_0 X_0 + \alpha_1 X_1 + \ldots + \alpha_{(n-1)} X_{(n-1)} = \sum_{i=0}^{(n-1)} \alpha_{(n-1)} X_{(n-1)}$$

1 Linear Algebra : Contracting vector space

- Linear Independence

  - $\alpha_0 X_0 + \alpha_1 X_1 + \ldots + \alpha_{(n-1)} X_{(n\text{-}1)} = 0$ then $\alpha_0 = \alpha_1 = \ldots . \alpha_{(n-1)} = 0$

  - X_0 , X_1 …, $X_{(n\text{-}1)}$ are linearly independent .

  - No vectors can not be represented as linear combination of the remaining vectors.

- Spanning set $\{V_0 , V_1 \ldots, V_{(n\text{-}1)}\}$ of a vector space V

  - Every vector in V can be represented as a linear combinations of V_0 , V_1 …, $V_{(n\text{-}1)}$.

- Basis $\{V_0 , V_1 \ldots, V_{(n\text{-}1)}\}$ of a vector space V

  - V_0 , V_1 …, $V_{(n\text{-}1)}$ are linearly independent.

  - $\{V_0 , V_1 \ldots, V_{(n\text{-}1)}\}$ is a spanning set of V.

- Dimension of a vector space V

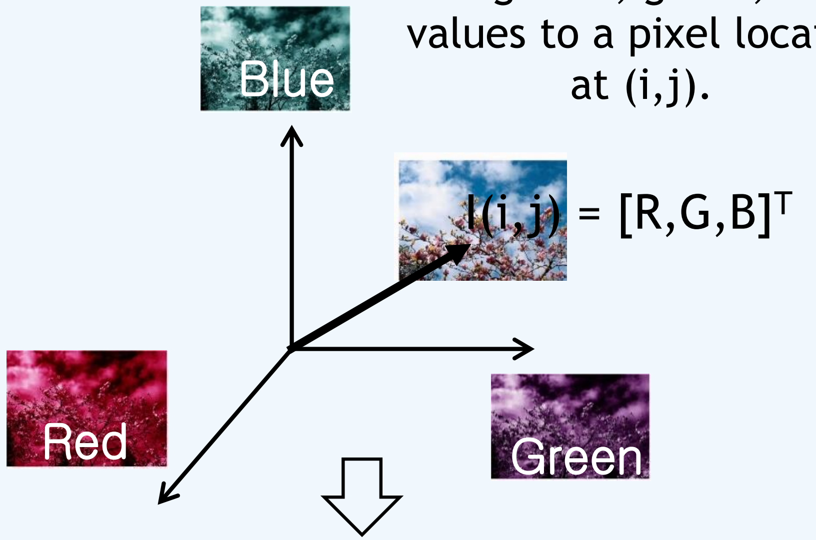
  -The number of basis vectors is same as the dimension of V.

  -Any set of linearly independent vectors span V or any n vectors spanning V are linearly independent.

1 Linear Algebra : Data in Vector Space-image data example

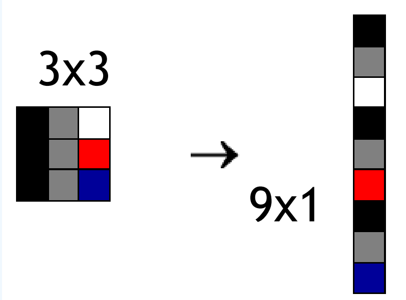
Image is the set of data point.
Image is the set of vectors at each location.

Assign red, green, blue values to a pixel located at (i,j).

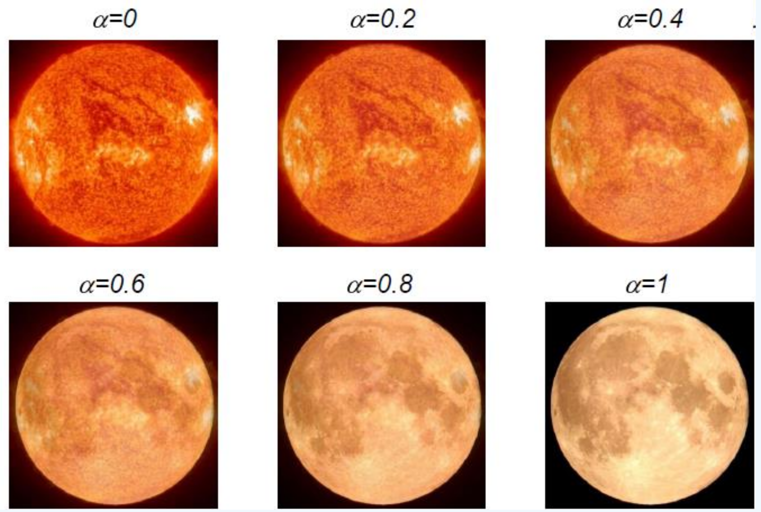


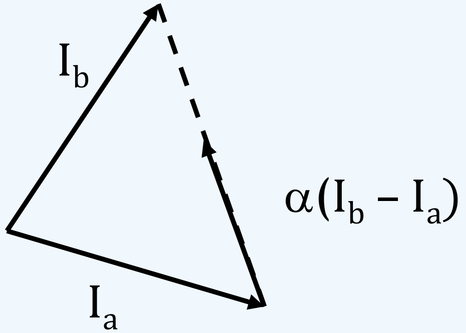
- Vector space representation is easier for understanding and analysis.
- we can manipulate image by manipulating their vector representation.

- A nxm image vector is transformed to a nmx1 vector.



- Example - Image morphing : $I(i,j) = (1-\alpha)\, I_a(i,j) + \alpha\, I_b(i,j)$

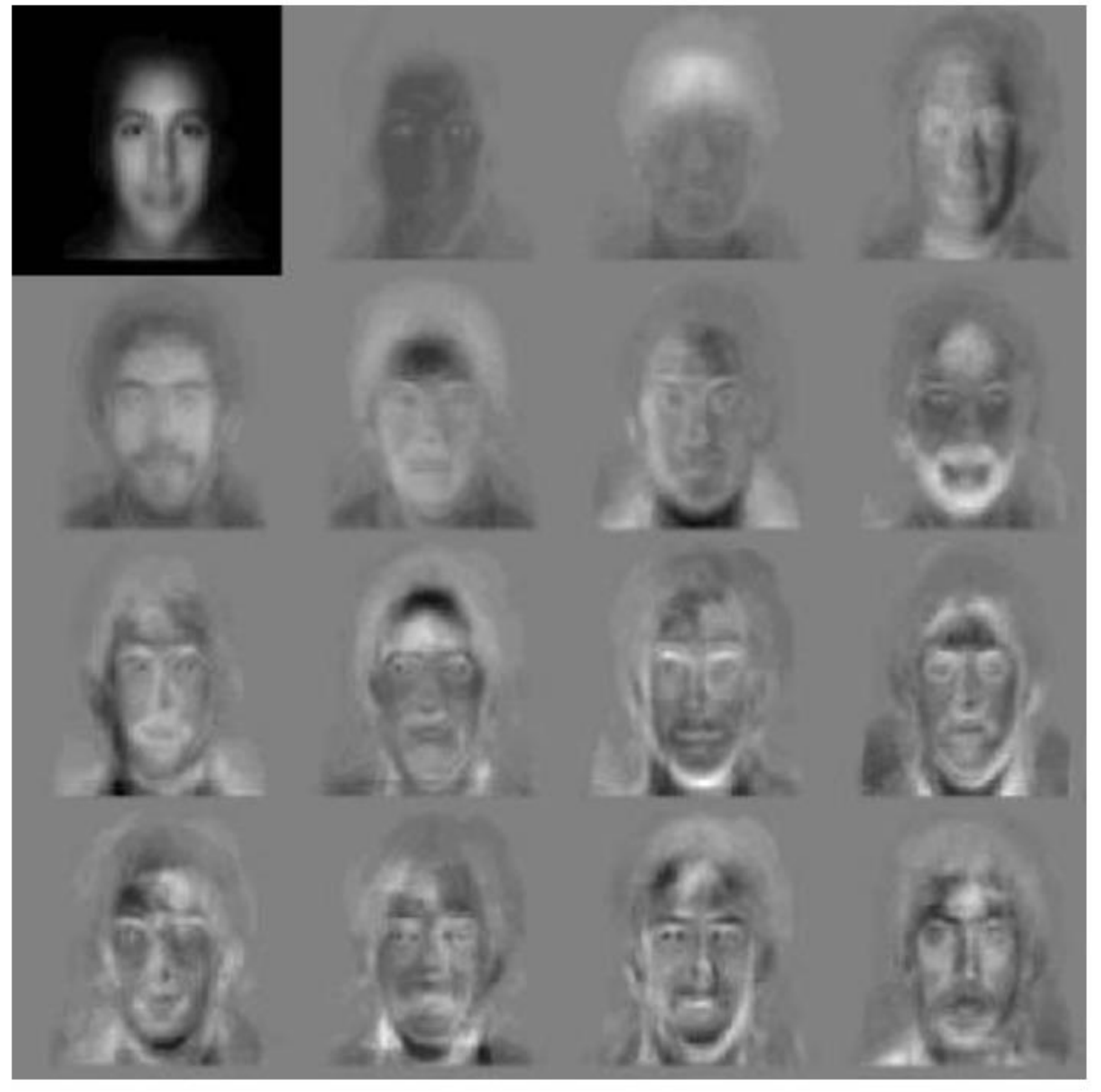




⇨ Possible because images are points in a vector space.

1 Linear Algebra : image space basis

- A basis of face image vector space.
- An face image can be made from a linearly combination of the basis image vectors.
- The first 16 basis image vector. There are more than 16.
- The vectors are orthonormal.



- Inner Product : <x,y> for x,y ∈ V

  i) <x,x> ≥ 0 , ii) <x,x> $= 0$ iff x=0, iii) <x,y>=<y,x>

- Dot product : <X,Y>=$\sum_{i=0}^{(n-1)} x_i\, y_i$
- Norm (for a certain inner product) : $\|X\|^2$ = <X,X>

  -L^p nom : $\|X\|$p $\equiv \left(\sum_{i=0}^{(n-1)} |x_i|^p\right)^{1/P}$

  - L^2 norm (Euclidean norm) : $\|X\|^2 \equiv \left(\sum_{i=0}^{(n-1)} |x_i|^2\right)^{1/2}$

  - L^1 norm : $\|X\|^1 \equiv \sum_{i=0}^{(n-1)} |x_i|$

  - L^∞ norm (max norm) : $\|X\|^\infty \equiv$ max $(|x_i|)$

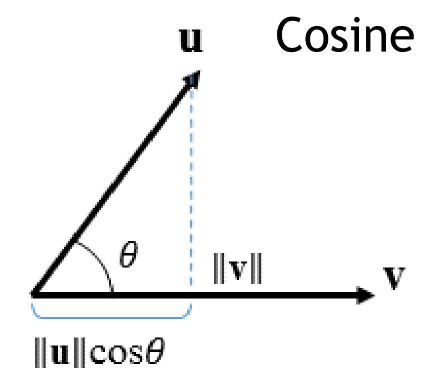
- Distance (Metric): D(X,Y) =$\|X - Y\|$

  - Euclidean Distance : $\|X - Y\| \equiv \left(\sum_{i=0}^{(n-1)} |x_i - y_i|^2\right)^{1/2}$

  - D(X,Y) $\leq$ D(X,Z) + D(Z,Y), D(X,Y)=D(Y,X)

- Dot product : <u,v>=$\sum_{i=0}^{(n-1)} u_i\, v_i$

  - $\mathbf{u}^T\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta,\quad \dfrac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \cos\theta$



Cosine Similarity

- $\mathbf{u}^T\mathbf{v} = 0 \rightarrow \mathbf{u} \perp \mathbf{v} \rightarrow (\mathbf{u}^T\mathbf{v} > 0 \Rightarrow \theta < 90^o,\ \mathbf{u}^T\mathbf{v} < 0 \Rightarrow \theta > 90^o)$
- Schwarz Inequality: $|\mathbf{u}^T\mathbf{v}| < \|\mathbf{u}\|\|\mathbf{v}\|$
- Triangle Inequality: $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$

1 Linear Algebra : Matrix-Vector operations

- A real $m \times n$ Matrix A ($\in \mathrm{R}^{mxn}$) is a linear operator mapping an n-dimension vector X ($\in \mathrm{R}^{n}$) to an M-dimension vector y ($\in \mathrm{R}^{m}$).

m x 1 vector   *n x 1 vector*   *m x n matrix*

$$Y = A \cdot X \Longleftrightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

- $$\begin{bmatrix} \vdots \\ y_i \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots & & \vdots \\ a_{i1} & \cdots & a_{in} \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \vdots \\ \sum_{j=1}^{n} a_{ij} x_j \\ \vdots \end{bmatrix}$$

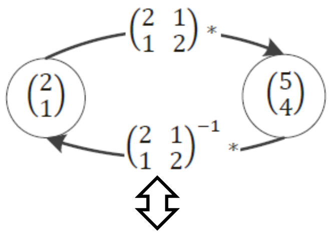
- y_i is the inner product of i^{th} row of A and X.
- Y is the projection of X on the space spanned by the rows of A.
- Y is coordinate of X in the ***row space*** of A.
- Maps R^n to R^n . Basis change.

- $$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \vdots \\ \sum_{j=1}^{n} a_{ij} x_j \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} x_1 + \cdots + \begin{bmatrix} | \\ a_n \\ | \end{bmatrix} x_n$$

-x_i is the i^{th} component of y in the space spanned by the columns of A.
- Y is a linear combination of the columns of A.
- X is coordinate of Y in the ***column space*** of A.
- Maps R^n to R^m. Dimension change.

- Rank of a matrix A ($\in \mathrm{R}^{mxn}$) is the number of linearly independent columns of A. Rank(A) $\leq$ min(m,n).

- **Inverse Matrix (for a square matrix)**
  - If A is not invertible, it is called singular matrix.
  - $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \qquad \mathbf{I}^{-1} = \mathbf{I}$
  - $(c\mathbf{A})^{-1} = \frac{1}{c}\mathbf{A}^{-1}$
  - $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \qquad (\mathbf{A}^{T})^{-1} = (\mathbf{A}^{-1})^{T} = \mathbf{A}^{-T}$



$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1} * \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$$

- **Orthogonality**
  - $\mathbf{u} \perp \mathbf{v}$ iff $\mathbf{u} \neq \mathbf{0}, \mathbf{v} \neq \mathbf{0}, \mathbf{u}^{T}\mathbf{v} = 0$
  - Orthonormal vectors = orthogonality + unit norm
- **Orthogonal matrix: $\mathbf{Q}$**
  - Rows are mutually orthonormal, and columns are mutually orthonormal
  - $\mathbf{Q}^{T}\mathbf{Q} = \mathbf{I} \implies \mathbf{Q}^{T} = \mathbf{Q}^{-1}$
- **Determinants(for a square matrix): det(A) or |A|**
  - $\det(\mathbf{A}) = 0$ if $\mathbf{A}$ has no inverse(singular)
  - $\det(\mathbf{A}^{-1}) = \dfrac{1}{\det(\mathbf{A})}$
  - $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \mathbf{A}^{-1} = \dfrac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \dfrac{1}{\det(\mathbf{A})}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

    $\det(\mathbf{A}) = ad - bc$

$$\det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

- **Homogeneous systems: $\mathbf{Ax} = \mathbf{0}$**
  - Homogeneous systems are always consistent
  - If A is $m \times n$, then $\mathbf{Ax} = \mathbf{0}$ has a nontrivial solution if $n > m$.
- **Solution and Singularity**
  - $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution $\mathbf{0}$ $\Leftrightarrow$ $\mathbf{A}$ is nonsingular
  - $\mathbf{Ax} = \mathbf{b}$ has a unique solution $\Leftrightarrow$ $\mathbf{A}$ is nonsingular
- **Solution of $\mathbf{Ax} = \mathbf{b}$, (where $\mathbf{A}$ is square matrix)**

  If A is invertible, i.e., nonsingular

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Definition

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, a scalar $\lambda \in \mathbb{C}$ is is said to be an **eigenvalue** or a **characteristic value** of $\mathbf{A}$ if there exists a nonzero vector $\mathbf{x}$ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

The vector $\mathbf{x}$ is said to be an **eigenvector** or a **characteristic vector** belonging to λ.

Following statements are equivalent:

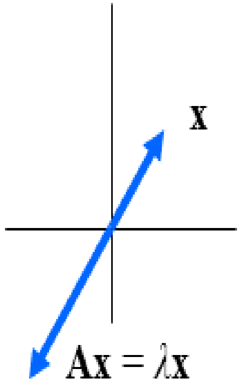
(a) λ is an eigenvalue of $\mathbf{A}$.

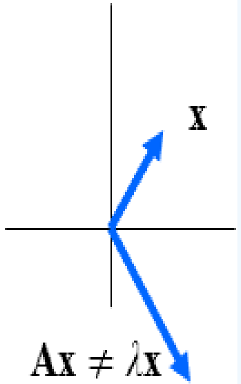
(b) $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ has a nontrivial solution.

(c) $\mathrm{Nul}(\mathbf{A} - \lambda\mathbf{I}) \neq \{\mathbf{0}\}$

(d) $(\mathbf{A} - \lambda\mathbf{I})$ is singular.

(e) $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$





Properties

- $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$
- $\det(\mathbf{A}) = |\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\Updownarrow$$

$$\lambda_1 = 3, \lambda_2 = 1$$

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

- **정수 3717은 특성이 보이지 않지만, 3*3*7*59로 소인수 분해하면 특성이 보이듯이, 행렬도 분해하면 여러모로 유용함**

[그림 2-12]의 반지름이 1인 원 위에 있는 4개의 벡터 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$가 $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$에 의해 어떻게 변환되는지 살펴보자. 변환 후의 벡터를 각각 $\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3, \mathbf{x}'_4$로 표기한다.

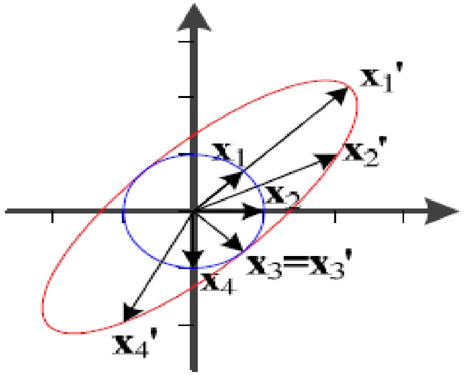
$$\mathbf{x}'_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 3/\sqrt{2} \\ 3/\sqrt{2} \end{pmatrix}$$

$$\mathbf{x}'_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\mathbf{x}'_3 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

$$\mathbf{x}'_4 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$$

눈 여겨 볼 점은 $\mathbf{A}$의 고유 벡터 $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$과 방향이 같은 $\mathbf{x}_1$과 $\mathbf{x}_3$이다. 이들은 변환 때문에 길이가 달라지더라도 방향은 그대로 유지한다. 식 (2.20)을 충실히 따르고 있다. 이때 길이의 변화는 고윳값 λ에 따른다. 즉, $\mathbf{x}_1$은 3배만큼, $\mathbf{x}_3$은 1배만큼 길이가 변한다. 나머지 $\mathbf{x}_2$와 $\mathbf{x}_4$는 길이와 방향이 모두 변한다. 파란 원 위에 있는 모든 점을 변환하면 빨간색의 타원이 된다. 파란 원 위에 존재하는 무수히 많은 점(벡터) 중에 방향이 바뀌지 않는 것은 고유 벡터에 해당하는 $\mathbf{x}_1$과 $\mathbf{x}_3$뿐이다.



- **Quadratic form**

  For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$ ($\mathbf{A}$ is symmetric)

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i(\mathbf{A}\mathbf{x})_i = \sum_{i=1}^{n} x_i\left(\sum_{j=1}^{n} a_{ij}x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$$

- **Positive definite**

  For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$,

  $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 2x_2^2 > 0$$

A= $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$: Positive definite

- **Positive semi-definite**

  For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$,

  $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \geq 0$

- **Indefinite**
  - Neither positive semi-definite nor negative semi-definite, i.e.,
  - If there exist $\mathbf{x}_1$, $\mathbf{x}_2$ such that $\mathbf{x}_1^{\mathrm{T}}\mathbf{A}\mathbf{x}_1 > 0$ and $\mathbf{x}_2^{\mathrm{T}}\mathbf{A}\mathbf{x}_2 < 0$.

- **Properties**
  - $\mathbf{A}$ is positive definite $\Leftrightarrow$ $-\mathbf{A}$ is negative definite
  - $\mathbf{A}$ is positive definite $\Leftrightarrow$ $\mathbf{A}$ is full rank & invertible

Diagonalization

- Let $\mathbf{X}$ be a matrix of n independent eigenvector of $\mathbf{A}$, the diagonal matrix $\Lambda = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ is the eigenvalue matrix of $\mathbf{A}$.
- $\mathbf{A}$ is said to be diagonalizable if it is similar to a diagonal matrix.
- $\mathbf{A}$ is diagonalizable if and only if $\mathbf{A}$ has n linearly independent eigenvectors

$$\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Eigendecomposition

- If $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \Lambda$, then $\mathbf{X}(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}) = \mathbf{X}\Lambda \Rightarrow \mathbf{A}\mathbf{X} = \mathbf{X}\Lambda \Rightarrow \mathbf{A}\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}\Lambda\mathbf{X}^{-1}$

  $\Rightarrow \mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1}$

  where X is a matrix of n independent eigenvector of $\mathbf{A}$, and Λ is a diagonal eigenvalue matrix of $\mathbf{A}$
- Power: $\mathbf{A}^2 = (\mathbf{X}^{-1}\Lambda\mathbf{X})(\mathbf{X}^{-1}\Lambda\mathbf{X}) = \mathbf{X}^{-1}\Lambda^2\mathbf{X}$, $\mathbf{A}^3 = (\mathbf{X}^{-1}\Lambda^2\mathbf{X})(\mathbf{X}^{-1}\Lambda\mathbf{X}) = \mathbf{X}^{-1}\Lambda^3\mathbf{X}, \ldots$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}$$

Independence of eigenvectors

- Let the eigenvalues of $\mathbf{A}$, $\lambda_1, \lambda_2, \ldots, \lambda_n$, are all different, then the corresponding n eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent and $\mathbf{A}$ is diagonalizable.

- Properties of Eigendecomposition

  $\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}$

  - $\det(\mathbf{A}) = \det(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}) = \det(\mathbf{X})\det(\boldsymbol{\Lambda})\det(\mathbf{X})^{-1} = \det(\mathbf{A}) = \lambda_1\lambda_2\cdots\lambda_n$
  - $\mathbf{A}^{-1} = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})^{-1} = \mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^{-1} = \mathbf{X}\mathrm{diag}(1/\lambda_1,\ 1/\lambda_2,\ \ldots,\ 1/\lambda_n)\mathbf{X}^{-1}$
  - $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}) = \mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{X}^{-1}\mathbf{X}) = \mathrm{Tr}(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$
  - $\mathbf{A}^k = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})^k = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})\cdots(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}) = \mathbf{X}\boldsymbol{\Lambda}^k\mathbf{X}^{-1} = \mathbf{X}\mathrm{diag}(\lambda_1^k,\ \lambda_2^k,\ \ldots,\ \lambda_n^k)\mathbf{X}^{-1}$
  - $e^{\mathbf{A}} = \mathbf{X}e^{\boldsymbol{\Lambda}}\mathbf{X}^{-1}$, where $e^{\boldsymbol{\Lambda}} = \mathrm{diag}(e^{\lambda_1},\ e^{\lambda_2},\ \ldots,\ e^{\lambda_n})$

- Properties of Eigenvalues/Eigenvectors for Symmetric Matrices

  Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and symmetric

  - All the eigenvalues of $\mathbf{A}$ are real
  - All eigenvectors with different eigenvalues of $\mathbf{A}$ are mutually orthogonal
  - The eigenvectors of $\mathbf{A}$ are orthonormal

    $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ (instead of $\mathbf{X}$)
  - $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ (since $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$)
  - $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\mathbf{x} = \mathbf{y}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2$

    Since $y_i^2 > 0$, $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ if all $\lambda_i > 0$

    Therefore, if all $\lambda_i > 0 \implies \mathbf{A}$ is positive definite

- $n*m$ 행렬 A의 Singular Value Decomposition

$$\mathbf{A} = \mathbf{U\Sigma V^T}$$

– 왼쪽 특이행렬 U는 AA^T의 고유 벡터를 열에 배치한 $n*n$ 행렬

– 오른쪽 특이행렬 V는 A^TA의 고유 벡터를 열에 배치한 $m*m$ 행렬

- Σ는 AA^T의 고유값의 제곱근을 대각선에 배치한 $n*m$ 대각행렬

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 2 \\ 3 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} -0.1914 & -0.2412 & 0.1195 & -0.9439 \\ -0.5144 & 0.6990 & -0.4781 & -0.1348 \\ -0.6946 & -0.6226 & -0.2390 & 0.2697 \\ -0.4651 & 0.2560 & 0.8367 & 0.1348 \end{pmatrix}$$

$$\begin{pmatrix} 3.7837 & 0 & 0 \\ 0 & 2.7719 & 0 \\ 0 & 0 & 1.4142 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.7242 & -0.4555 & -0.5177 \\ -0.6685 & 0.2797 & 0.6891 \\ 0.1690 & -0.8452 & 0.5071 \end{pmatrix}$$

2 Probability : Definition

- Probability : Mathematics for dealing with processes or experiments that are non-deterministic.
- Outcome: Possible case.
- (Random) Experiment : Taking values in a set of outcomes.
- (Random) Event : Set of outcomes.
- Probability of an event : A real number between 0 and 1 expressing the chance that the event will occur when a random experiment is performed.

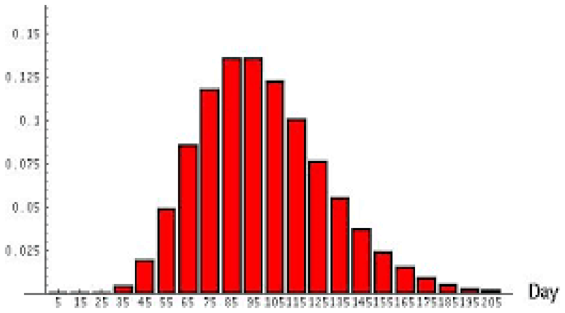
- Sample space U : Set of experimental outcomes that must stratify the following properties.
  - Collectively Exhaustive : When experiment is performed, one of these outcomes must occur.
    $\Rightarrow$ There is no possible event to which a probability cannot be assigned
  - Mutually Exclusive : Only one outcomes happens and no other can occur.
    => Simplifying the calculation of the probability of event.

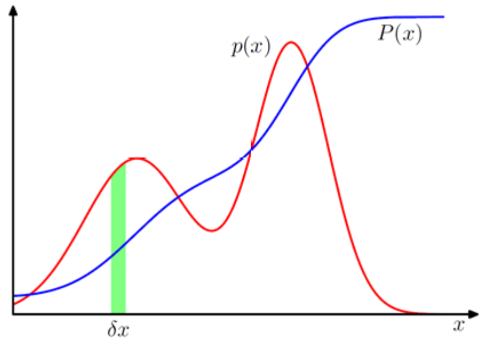
- Probability measure
  - $P(\phi) = 0$, $P(A) \geq 0$, P(U) =P(Sample space) =1
  - If $A \subseteq B$, $P(A) \leq P(B)$
  - $P(A \cup B) = P(A)+P(B)-P(A \cap B)$
  - If $A \cap B = \emptyset$, then $P(A \cup B) = P(A)+P(B)$
  - $P(A^c) = 1- P(A)$

2 Probability : Random Variables (RV)

- Random variable (RV) x (RV could be a scalar or vector).
  - Function that assigns a real value to each outcome.
  - That is, $X(a_i) = x$ where a_i ($\in$ U) is an outcome and x($\in$ R) is a real value assinged to ai.
  - In notation, the random variable is subscript and the value is the argument.
    For example, $P_X(x_1,x_2) = 1/36 \Rightarrow Prob[X=(x_1,x_2)]=1/36$.

Discrete RV	Continuous RV
• Take a value of finite or at most countable set of outcomes. • Probability assignments are given by a probability mass function (pmf). • $0 \le P_X(x=k) \le 1$ and $\sum_k P_X(k)=1$.	• Take arbitrary values in a real interval for scalar RV, in an area for a vector RV. • Probability assignments are given by a probability density function (pdf). - $0 \le P_X(x)$ and $\int_{-\infty}^{\infty} P_X(x)\,dx = 1$, - $P_x(X=a) = \int_{-a}^{a} P_X(x)\,dx = 0$ - $P_x(a \le X \le b) = \int_{a}^{b} P_X(x)\,dx$ • Cumulative Distribution Function (CDF) - $P(z) = P_x(x \le z) = \int_{-\infty}^{z} P_X(x)\,dx$





- Multiple RVs: RV is a collection of random data, that is, a vector.
  - Multiple RVs : A RV data $X=[X_1, X_2, \ldots, X_n]^T$
  - Medical examination : X=[temperature, blood pressure, weight, age] ^{T}.
  - pdf for multiple RVs is Joint probability distribution : $P_X(x_1,x_2,x_3, \ldots.x_n)$.

- Marginalization :
  - Delete the irrelevant or unnecessary RVs from joint pdf.
  - EX: Application to K-Univ. X=[Language score, weight, Math score, height]= [x1, x2, x3, x4].

  => 'weight' and 'height' are not relevant to possibility of admission.

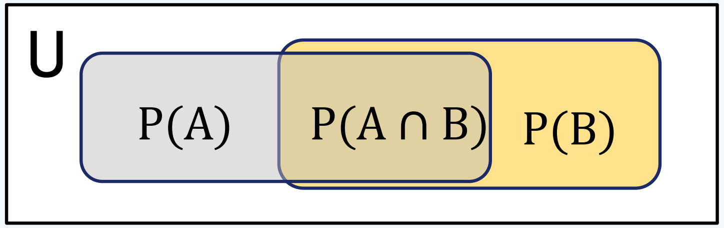
  $\Rightarrow$ $P_{x_1x_3}(x_1,x_3) = \sum_{x_2}\sum_{x_4} P_{x_1x_2x_3x_4}(x_1,x_2,x_3,x_4)$ for discrete RV

  $\Rightarrow$ $P_{x_1x_3}(x_1,x_3) = \iint_{x_2x_4} P_{x_1x_2x_3x_4}(x_1,x_2,x_3,x_4)$ for continuous RV

2 Probability : Conditional Probability

- Conditional Probability: P(A|B)
  - Conditioned on (or Given) the probabilities of an event B, measure of the probability of an event A.
  - "The conditional probability of A given B" or "The probability of A under the condition B"
  - Values of some variables are known (or given), while other variables are unknown.

- Calculation of P(A|B)
  - $P(A|B) = P(A \cap B)/P(B)$.
  - $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$



- Example : Applying to an University
  - For judging the possibility of admission, define new variable with state $Y \in \{\text{Accept}, \text{Not Accept}\}$.
  - Decide the probability that an application is accepted for given (known) scores of Math (x3=93).

  Namely, decide $P_{Y|x_3}(Y=\text{Accept} \mid x_3=93) = P_{Y,x_3}(Y=\text{Accept} , x_3=93) / P_{x_3}(x_3=93)$

- Independence
  - Random variable x_1 and x_2 are independent if $P_{x_1 x_2}(x1,x2) = P_{x_1}(x1)\, P_{x_2}(x2)$
  - In conditional prob. : $P_{x_1|x_2}(x_1|x_2) = P_{x_1 x_2}(x_1,x_2) / P_{x_2}(x_2) = P_{x_1}(x_1)$.
  - Knowing x_2 does not change the estimating probability concerned with x_1.

2 Probability : Rule of Chain

- Chain rule of probability:
  - $P_{x_1x_2}(x_1,x_2) = P_{x_2|x_1}(x_2|x_2)\ P_{x_1}(x_1)$
  - $P_{x_1x_2,\dots,x_n}(x_1,\dots, x_n) = P_{x_1|x_2,\dots x_n}(x_1|x_2,\dots, x_n) \times P_{x_2|x_3,\dots x_n}(x_2|x_3,\dots, x_n) \dots \mathrm{X}\ P_{x_{n\_1}|xn}(x_{n-1}|x_n)\ P_{x_n}(x_n)$

- Marginal probability with conditional prob. :
  - $P(A) = \sum_{b\in B} P_X(A,B=b) = \sum_{b\in B} P_X(A|B=b)\ P(B=b)$, $\quad P_y(y) = \int P_{yx}(y,x)\,dx = \int P_{y|x}(y|x)\ P_x(x)dx$
  - Combining the chain rule with the marginalization makes difficult problem simpler.
    - $\Rightarrow$ $P_x(x)$ is computed from data observation or historgram
    - $\Rightarrow$ $P_{y|x}(y|x)$ is easier to be estimated because size of observed data is much reduced.
    - $\Rightarrow$ Bayesian probability => Likelihood Estimation

EX : Break down a hard question (EX: Prob. of Accept and math score 93) into two easier questions.
- $\Rightarrow$ It is well know that $P_{Y,x_3}(Y=\text{Accept}|x_3=93)$ is closed to 1.
- $\Rightarrow$ Even though computing $P_{x_3}(x_3=93)$ is hard, easier than $P_{Y,x_3}(Y=\text{Accept}, x_3=93)$.
- $\Rightarrow$ In order to compute $P_{x_3}(x_3=93)$, gather a number of applicants and make an histogram of math scores.

2 Probability :Momentum (Statistics)

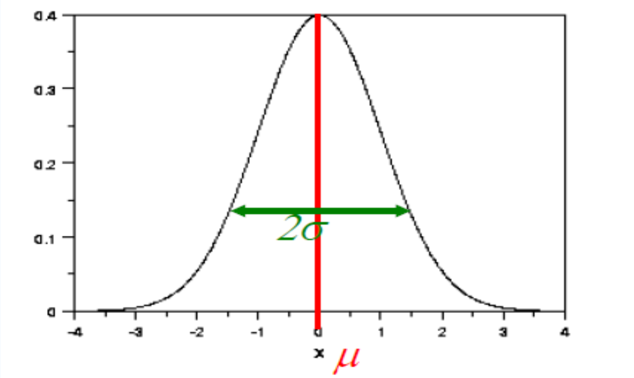
- Moments are important properties of random variables.

	Discrete	continuous	
Mean : μ = E[X]	$\mu = \sum_{k} k\, P_X(k)$	$\mu = \int x\, P(x)\, dx$	E[(X+Y)] = E[X]+ E[Y],
Variance: σ^2 = Var(X) = E[(x-μ)^2]	$\sigma^2 = \sum_{k} (k-\mu)^2\, P_X(k)$	$\sigma^2 = \int (x-\mu)^2 P(x)\, dx$	• Var(X+Y) $\neq$ Var(X) + Var(Y) • E[(x-μ)^2] = E[x^2]-μ^2

- nth order (non-central) moment : E[X^n]
- nth order (central) moment : E[$(X-\mu)^n$]

- Nice distribution are well specified by a very few moments. Ex: Gaussian by mean variance

Gaussian distribution



2 Probability :Covariance

- Covariance for scalar RVs x, y :

$$\mathrm{cov}[x,y] = \mathbb{E}[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

- Covariance matrix for random vector RVs (multiple RVs) **x**:Kx1, **y**:Lx1
  - $\mathrm{cov}[\mathbf{x},\mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\{\mathbf{x}-\mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}}-\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}] = \mathbb{E}_{x,y}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]$

$$= \begin{pmatrix} \mathrm{cov}[x_1,y_1] & \mathrm{cov}[x_1,y_2] & \cdots & \mathrm{cov}[x_1,y_L] \\ \mathrm{cov}[x_2,y_1] & \mathrm{cov}[x_2,y_2] & \cdots & \mathrm{cov}[x_2,y_L] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}[x_K,y_1] & \mathrm{cov}[x_K,y_2] & \cdots & \mathrm{cov}[x_K,y_L] \end{pmatrix}$$

  - $\mathrm{cov}[\mathbf{x},\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}] \triangleq \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}]$: symmetric & Positive definite

$$= \begin{pmatrix} \mathrm{var}[x_1] & \mathrm{cov}[x_1,x_2] & \cdots & \mathrm{cov}[x_1,x_D] \\ \mathrm{cov}[x_2,x_1] & \mathrm{var}[x_2] & \cdots & \mathrm{cov}[x_2,x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}[x_D,x_1] & \mathrm{cov}[x_D,x_2] & \cdots & \mathrm{var}[x_D] \end{pmatrix} \Leftrightarrow$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad \boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

예제 2-7

Iris 데이터베이스의 샘플 중 8개만 가지고 공분산 행렬을 계산하자.

$$\mathbb{X} = \left\{\mathbf{x}_1 = \begin{pmatrix}5.1\\3.5\\1.4\\0.2\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}4.9\\3.0\\1.4\\0.2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}4.7\\3.2\\1.3\\0.2\end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix}4.6\\3.1\\1.5\\0.2\end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix}5.0\\3.6\\1.4\\0.2\end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix}5.4\\3.9\\1.7\\0.4\end{pmatrix}, \mathbf{x}_7 = \begin{pmatrix}4.6\\3.4\\1.4\\0.3\end{pmatrix}, \mathbf{x}_8 = \begin{pmatrix}5.0\\3.4\\1.5\\0.2\end{pmatrix}\right\}$$

먼저 평균벡터를 구하면 $\boldsymbol{\mu} = (4.9125,\ 3.3875,\ 1.45,\ 0.2375\ )^{\mathrm{T}}$이다. 첫 번째 샘플 $\mathbf{x}_1$을 식 (2.39)에 적용하면 다음과 같다.

$$(\mathbf{x}_1 - \boldsymbol{\mu})(\mathbf{x}_1 - \boldsymbol{\mu})^{\mathrm{T}} = \begin{pmatrix}0.1875\\0.1125\\-0.05\\-0.0375\end{pmatrix}(0.1875 \quad 0.1125 \quad -0.05 \quad -0.0375)$$

$$= \begin{pmatrix}0.0325 & 0.0211 & -0.0094 & -0.0070\\0.0211 & 0.0127 & -0.0056 & -0.0042\\-0.0094 & -0.0056 & 0.0025 & 0.0019\\-0.0070 & -0.0042 & 0.0019 & 0.0014\end{pmatrix}$$

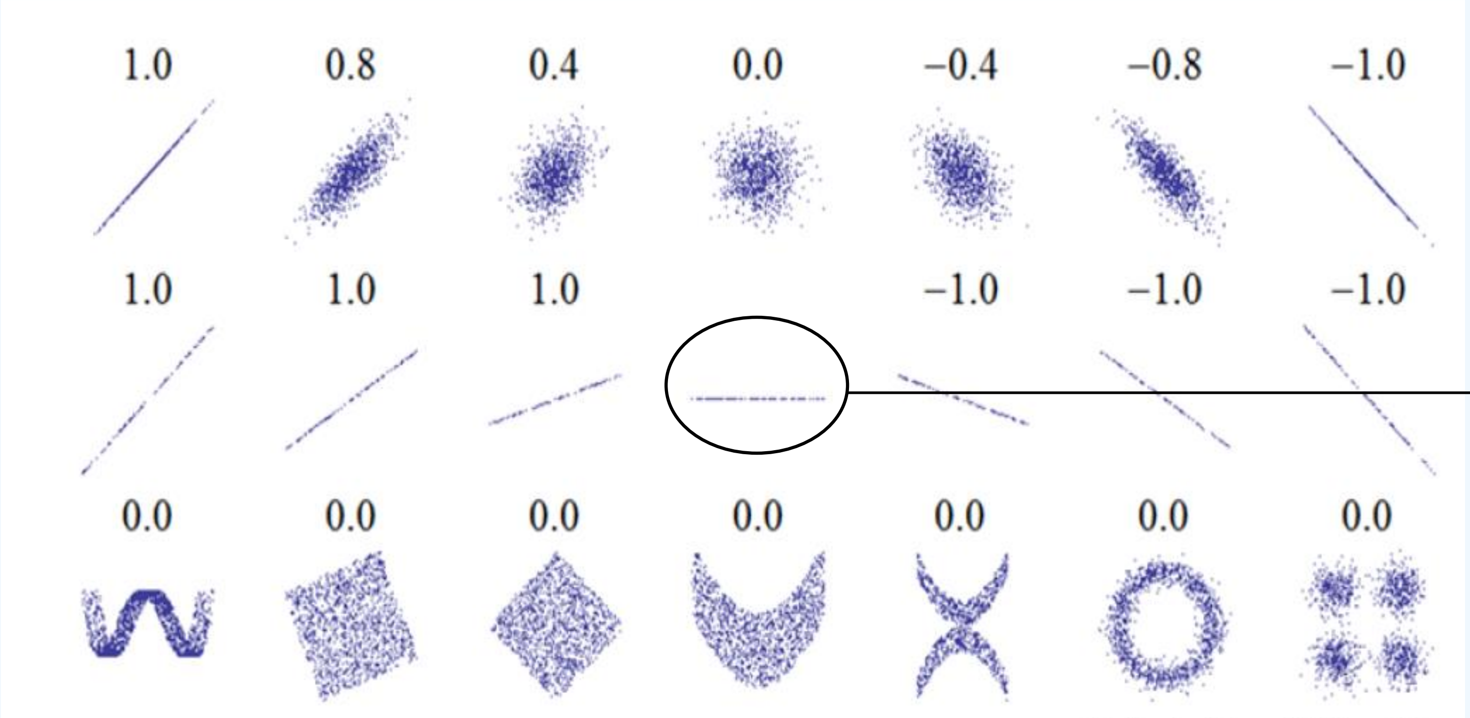
나머지 7개 샘플도 같은 계산을 한 다음, 결과를 모두 더하고 8로 나누면 다음과 같은 공분산 행렬을 얻는다.

$$\boldsymbol{\Sigma} = \begin{pmatrix}0.0661 & 0.0527 & 0.0181 & 0.0083\\0.0527 & 0.0736 & 0.0181 & 0.0130\\0.0181 & 0.0181 & 0.0125 & 0.0056\\0.0083 & 0.0130 & 0.0056 & 0.0048\end{pmatrix}$$

- Correlation coefficient for scalar RVs, *x*,*y* :

$$\operatorname{corr}[x,y] = \frac{\operatorname{cov}[x,y]}{\sqrt{\operatorname{var}[x]\operatorname{var}[y]}}$$

- Sets of (x,y) points, with corr[x,y] for each set.



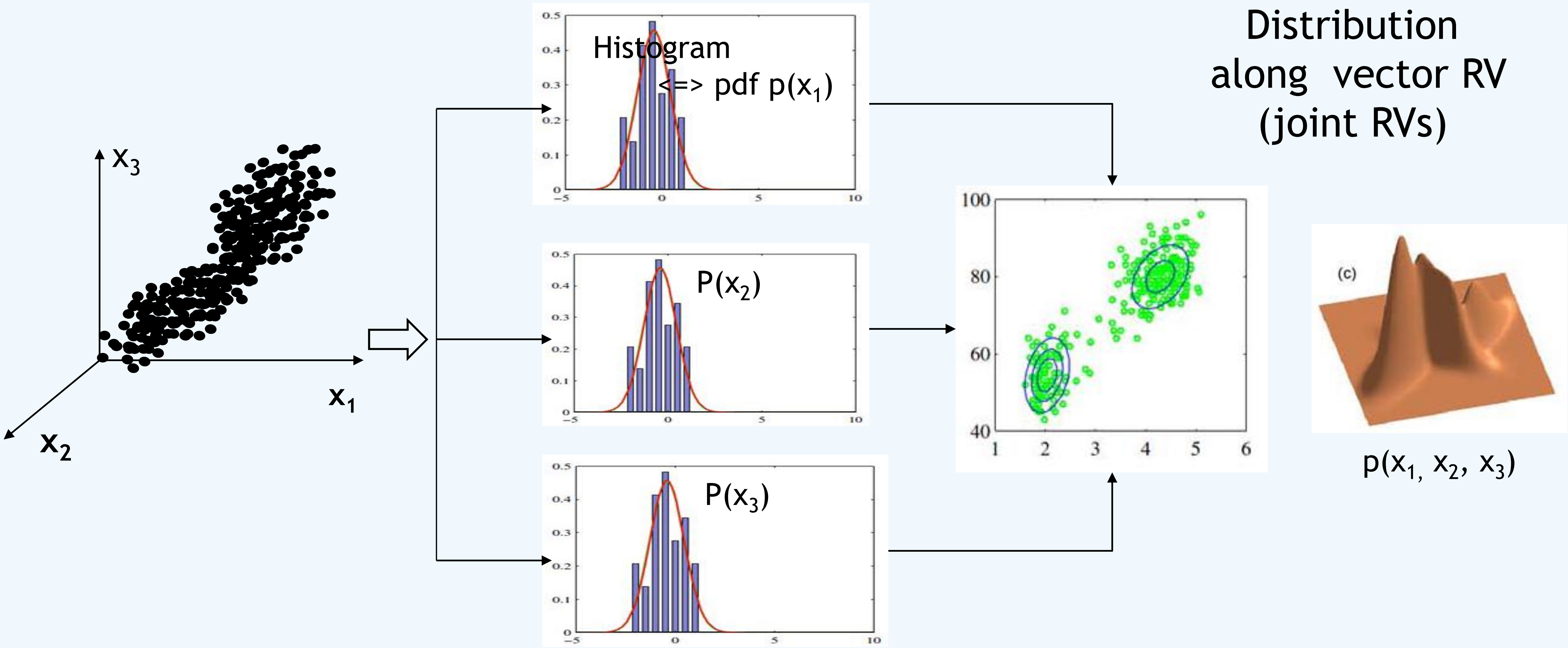
[Adopted from Wikipedia]

var[y]=0 $\Rightarrow$ Corr is undefined.

2. Probability : Distribution

Distribution
along a scalar RV

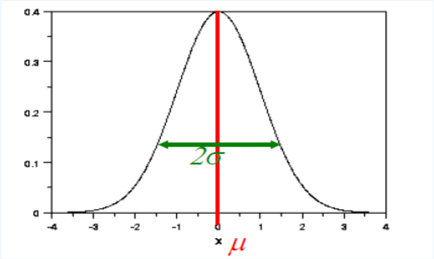
Distribution
along vector RV
(joint RVs)



$p(x_1, x_2, x_3)$

- Univariate Gaussian Distribution

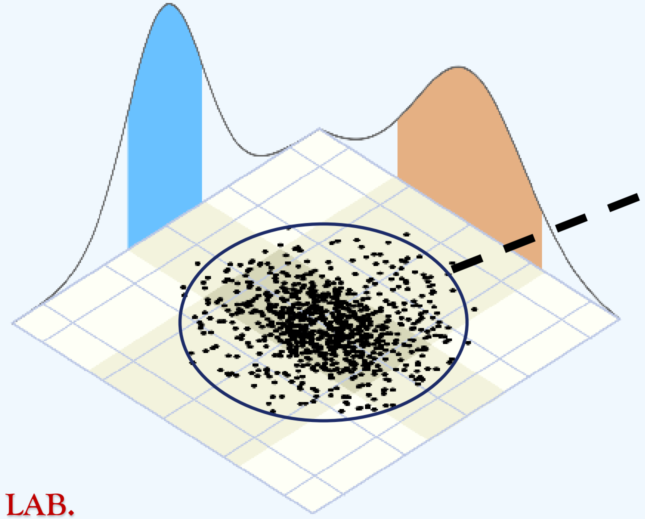
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

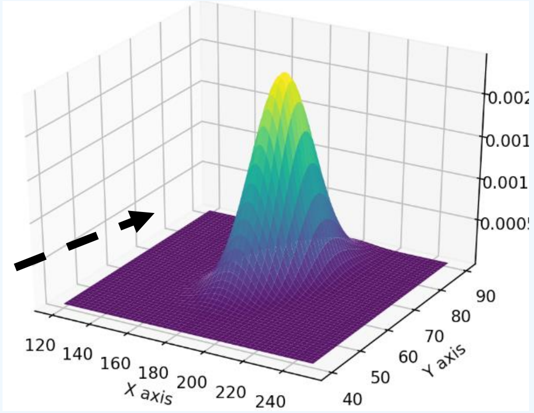


- Multivariate Gaussian Distribution for $x=(x_1,x_2,\ldots, x_D)$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

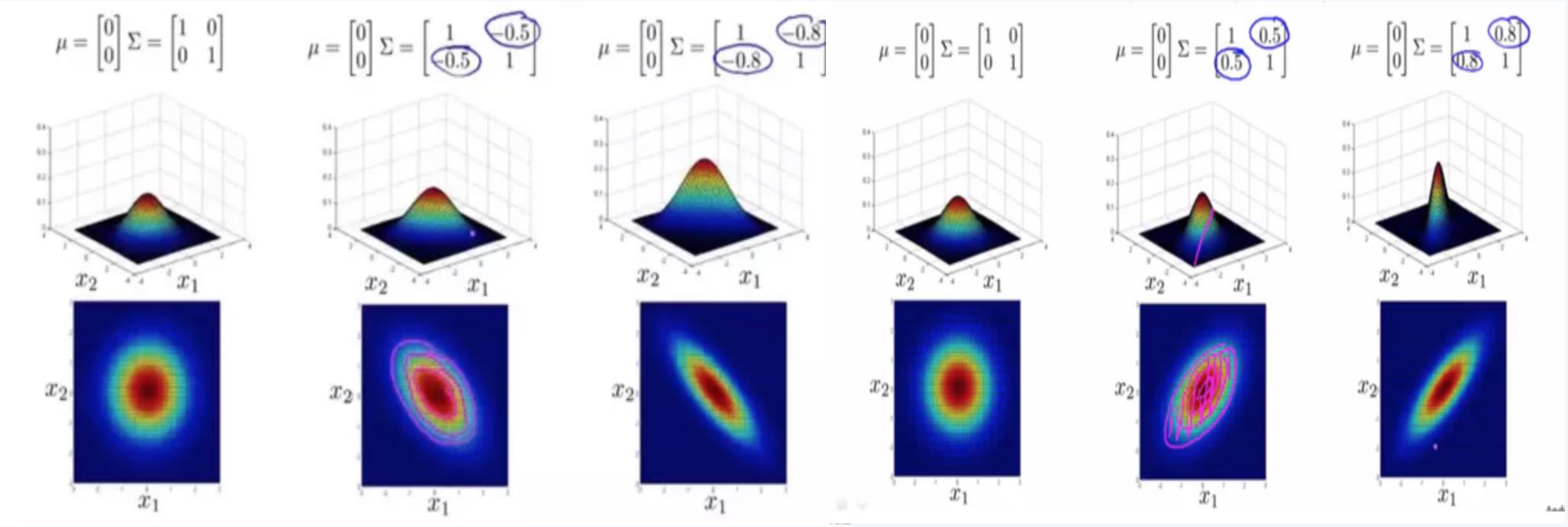
where $\boldsymbol{\Sigma} : \mathrm{cov}[\mathbf{x}]$





2. Probability : Multivariate Gaussian Distri.

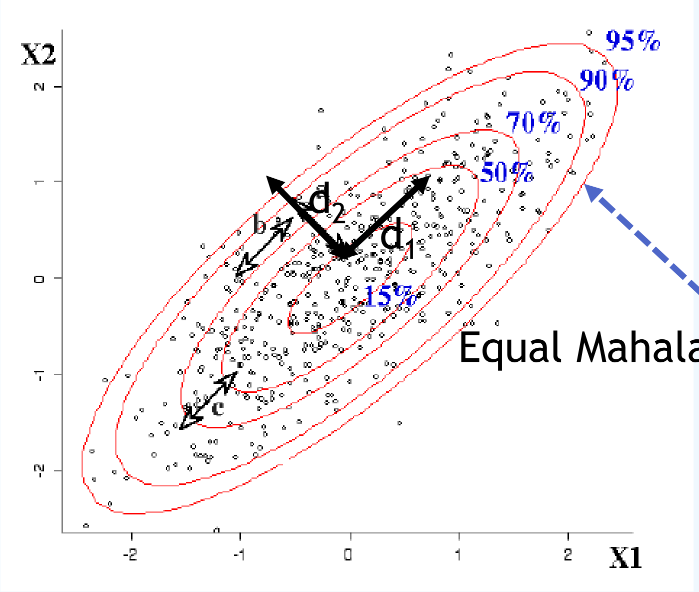
- Multivariate Gaussian Distribution with different $\sum$ (=cov[x]).



- N-dimensional data vector : $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$
- Mahalanobis Distance : $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})}$

- Measure of correlation between variables.

In Euclidean distance, $d_1=d_2$.
In Mahalonobis distance, $d_1< d_2$.

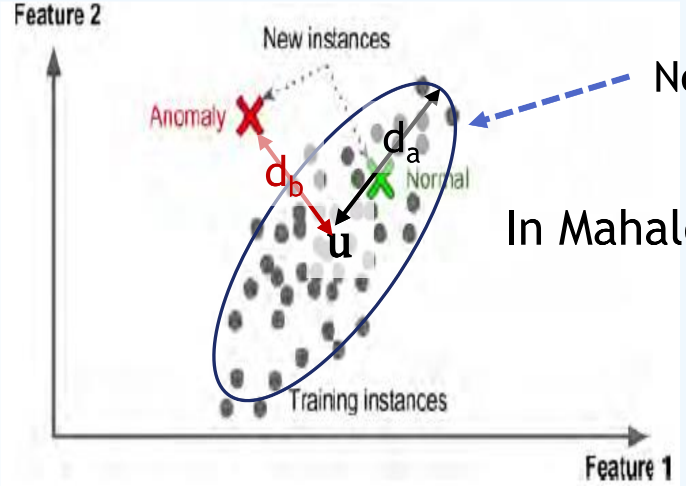


Equal Mahalanobis distance

- If we employ the center of a distribution, the distance may detect the outliers.

$$d(\mathbf{x}) = \sqrt{(\mathbf{x}-\mathbf{u})^T \Sigma^{-1} (\mathbf{x}-\mathbf{u})}$$

where $\mathbf{u} = [u_1, u_2, \; u_N]^T$ is mean vector.



Normal data boundary

In Mahalonobis distance : $d_a << d_b$

(* In Euclian distance, $d_a > d_b$)

- Bernoulli distribution :
  - Given a binary random variable $x \in \{0, 1\}$ (e.g. tossing a coin)
  - 성공(x=1) 확률 p, 실패(x=0) 확률이 1-p인 분포

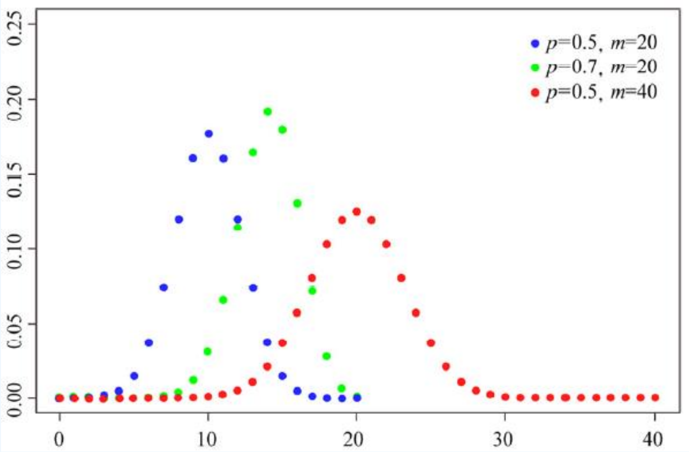
$$Ber(x;p) = p^x(1-p)^{1-x} = \begin{cases} p, & x = 1\text{일 때} \\ 1-p, & x = 0\text{일 때} \end{cases}$$

E(x) =p, var[x] =$p(1-p)$

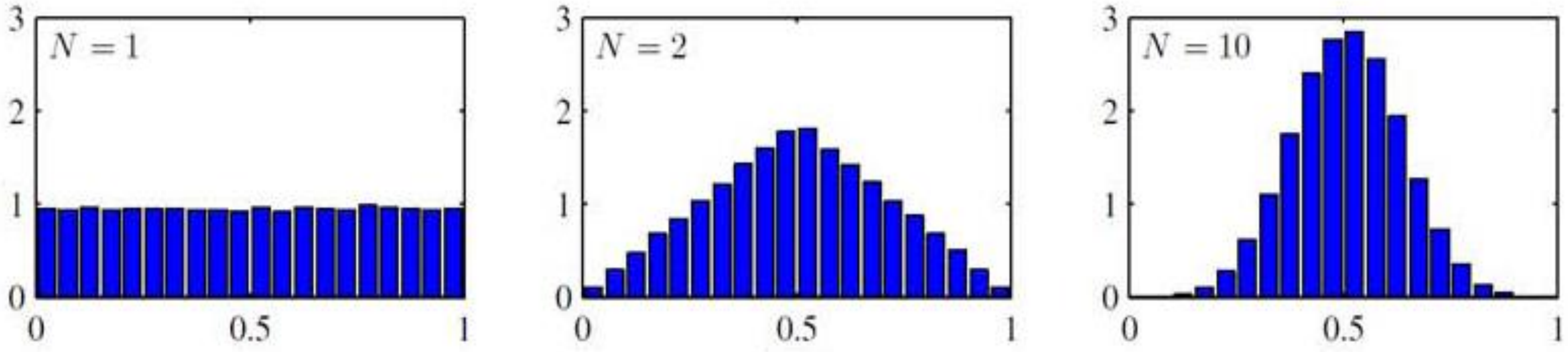
- Bernoulli distribution :
  - **성공 확률이 p인 베르누이 실험을 m번 수행할 때 성공할 횟수의 확률분포**

$$B(x;m,p) = C_m^x p^x(1-p)^{m-x} = \frac{m!}{x!\,(m-x)!}p^x(1-p)^{m-x}$$

E(x) =mp, var[x] =$mp(1-p)$



- Sum of multiple random variables
  - Maximum entropy condition for continuous variable $\Rightarrow$ Gaussian
  - N random variables $x_1, \ldots, x_N$, each of x_i has $\mathrm{Unif}(x \mid 0, 1)$. Considering the distribution of the mean $(x_1 + \cdots + x_N)/N$. For large N, this distribution tends to a Gaussian $\Longrightarrow$ **central limit theorem**



- Proof of Bayesian Rule

$$p(B \mid A)p(A) = p(A \cap B)$$
$$p(A \mid B)p(B) = p(A \cap B)$$
$$\Rightarrow p(A \mid B)p(B) = p(B \mid A)p(A) \Rightarrow p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} = \frac{p(B \mid A)p(A)}{\sum_{\alpha \in A} p(B \mid A)p(A)}$$

- $P(w|X) = \frac{P(X|w)P(w)}{P(X)} \Longleftrightarrow \text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$

  - Prior P(w) : Presumption, i.e, hypothesis. In a parametric model, parameters to be set.
  - Evidence P(X) : Determined from observed data.
  - Likelihood P(X|w) (L(X|w)): Under w, the possibility of X. The function of w. Decided by counting frequencies.
  - Posterior P(w|X) : (Required) inferencing w from observation X.

- A measure of Information–by Shannon

1. Information contained in events should be defined in terms of some measure of the uncertainty of the event.

2. Less probable events should contain more information.

3. Info. of uncorrelated (independent) events should equal to the sum of info. of each event.

$P(\alpha)$: the prob. of a event α.

$I(\alpha)$: the information measure of the event α.

$$\boxed{I(\alpha) = -\log_2 P(\alpha)} \quad \text{(bits)}$$

* If $P(\alpha \cap \beta) = P(\alpha) * P(\beta)$, that is, event α, β are independent.

$$\mathbf{I(\alpha \cap \beta)} = -\log P(\alpha \cap \beta) = -\log (P(\alpha) * P(\beta))$$

$$= -\log P(\alpha) - \log P(\beta) = \mathbf{I(\alpha)} + \mathbf{I(\beta)}$$

- 메시지가 지닌 정보를 수량화
  - “고비 사막에 눈이 왔다” 와 “대관령에 눈이 왔다” 라는 두 메시지 중 어느 것이 더 많은 정보를 가지나?
  - 정보이론의 기본 원리 → 확률이 작을수록 많은 정보
  - 정보를 정량화 할 필요가 있음.

2 Probability : Information Theory

- **Entropy for discrete case or RV**

Discrete Memorlyess Source (DMS)

- An event set U ={ $A_1, A_2, A_3, \ldots, A_N$} with probability $P(A_k)$.
- U is independent of time.

Entropy (the average amount of information) for DMS

$$H(U) = \sum_{k=1}^{N} P(A_k) \bullet I(A_k) = -\sum_{k=1}^{N} P(A_k) \bullet \log P(A_k) \text{ (bits/symbol)}$$

U ={ 0, 1, 2, 3}

- Symbol stream 1 : [0 0 1 2 0 3 2 1]

$P(0) = 3/8,\ P(1) = 2/8$
$P(2) = 2/8,\ P(3) = 1/8$

$H(U) =$
$- \{ (3/8) \log_2(3/8) + (2/8) \log_2(2/8)$
$+ (2/8) \log_2(2/8) + (1/8) \log_2(1/8)\}$
$= 1.7$ bits/symbol

- Symbol stream 2 : [0 2 1 2 0 3 1 3]

$P(0) = 2/8,\ P(1) = 2/8$
$P(2) = 2/8,\ P(3) = 2/8$

$H(U) =$
$- \{ (2/8) \log_2(2/8) + (2/8) \log_2(2/8)$
$+ (2/8) \log_2(2/8) + (2/8) \log_2(2/8)\}$
$= 2$ bits/symbol

- **Entropy for continuous RV.**

$$H(x) = -\int_{\mathbb{R}} P(x) \log_2 P(x)\, dx \ (bits)$$

- Symbol stream 2의 Entropy가 높음
  - Event들의 확률이 동일함.
  - 어떤 Event가 발생될지 예측이 어려움
  - 불확실성이 높음.

- Event들의 확률이 동일할 때 최대 entropy가 됨
- 특정 event의 확률이 높을 수록 entropy가 작음

- **Cross Entropy**

$$H(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,} P(e_i)\log_2 Q(e_i)$$

$$\begin{aligned} H(P,Q) &= -\sum_{x} P(x)\log_2 Q(x) \\ &= -\sum_{x} P(x)\log_2 P(x) + \sum_{x} P(x)\log_2 P(x) - \sum_{x} P(x)\log_2 Q(x) \\ &= H(P) + \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} \end{aligned}$$

$= P$의 엔트로피 $+$ P와 Q 간의 KL 다이버전스

-> pdf

Kullback-Leibler (KL) Divergence

- **KL Divergence**

$$KL(P \parallel Q) = \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)}$$

: 오류의 중요도에 관계없이 동일한 penalty를 주는 MSE의 단점을 보완

MSE: 오류의 중요도에 관계없이 동일한 penalty를 주는것

- **Cross Entropy**

[그림 2-21]과 같이 정상적인 주사위와 찌그러진 주사위가 있는데, 정상적인 주사위의 확률분포는 *P*, 찌그러진 주사위의 확률분포는 *Q*를 따르며, *P*와 *Q*가 다음과 같이 분포한다고 가정하자.

$$P(1)=\frac{1}{6},\ P(2)=\frac{1}{6},\ P(3)=\frac{1}{6},\ P(4)=\frac{1}{6},\ P(5)=\frac{1}{6},\ P(6)=\frac{1}{6}$$
$$Q(1)=\frac{3}{12},\ Q(2)=\frac{1}{12},\ Q(3)=\frac{1}{12},\ Q(4)=\frac{1}{12},\ Q(5)=\frac{3}{12},\ Q(6)=\frac{3}{12}$$

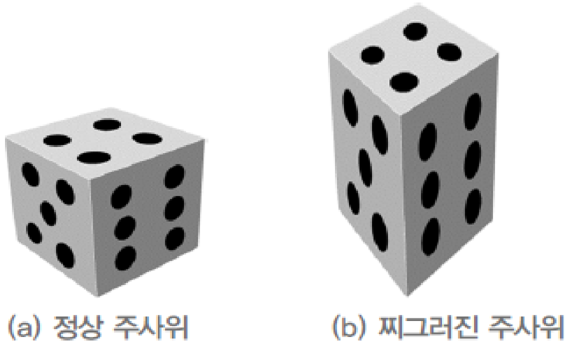


그림 2-21 **확률분포가 다른 두 주사위**

확률분포 *P*와 *Q* 사이의 교차 엔트로피와 KL 다이버전스는 다음과 같다.

$$H(P,Q)=-\left(\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{3}{12}\right)=2.7925$$
$$KL(P\parallel Q)=\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2\frac{2}{3}=0.2075$$

[예제 2-8]에서 *P*의 엔트로피 *H*(*P*)는 2.585이었다. 따라서 식 (2.49)가 성립함을 알 수 있다.

3. Optimization : Optimization task of ML

- **Optimization for Machine Learning**

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} J(\Theta)$$ ⇨ Find the optimal solution $\hat{\Theta}$ that minimizes the cost function $J(\Theta)$.

- **기계 학습의 최적화는 식이 아니고 훈련집합이 주어지고, 훈련집합에 따라 정해지는 목적함수의 최저점을 찾아야 함.**
  - **식이 아닌 데이터로 미분하는 과정 필요**
  - **주로 Stochastic Gradient Descent (SDG)에 근거한 오류 역 전파 (Backward propagation) 알고리즘을 사용**
  - **Data를 이용한 미분 방법**

Exhaustive search

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$
출력 : 최적해 $\hat{\Theta}$

```
가능한 해를 모두 생성하여 집합 S에 저장한다.
min을 충분히 큰 값으로 초기화한다.
for (S에 속하는 각 점 Θ_current에 대해)
    if(J(Θ_current)< min) min=J(Θ_current), Θ_best = Θ_current
Θ̂ = Θ_best
```

Random search

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$
출력 : 최적해 $\hat{\Theta}$

```
min을 충분히 큰 값으로 초기화한다.
repeat
    무작위로 해를 하나 생성하고 Θ_current라 한다.
    if(J(Θ_current)<min) min=J(Θ_current), Θ_best = Θ_current
until(멈춤 조건)
Θ̂ = Θ_best
```

Gradient Decent

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$
출력 : 최적해 $\hat{\Theta}$

```
난수를 생성하여 초기해 Θ을 설정한다.
repeat
    J(Θ)가 작아지는 방향 dΘ를 구한다.
    Θ = Θ + dΘ
until(멈춤 조건)
Θ̂ = Θ
```

* **ML에서 주로 사용하는 방법**

3. Optimization : Gradient (1/2)

- Gradient of $f(\mathbf{x})$: ∇f, $\frac{\partial f}{\partial \mathbf{x}}$, $\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)^{\mathrm{T}}$

EX: $f(\mathbf{x}) = f(x_1, x_2) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$

$$\nabla f = f'(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)^{\mathrm{T}} = (2x_1^5 - 8.4x_1^3 + 8x_1 + x_2, 16x_2^3 - 8x_2 + x_1)^{\mathrm{T}}$$

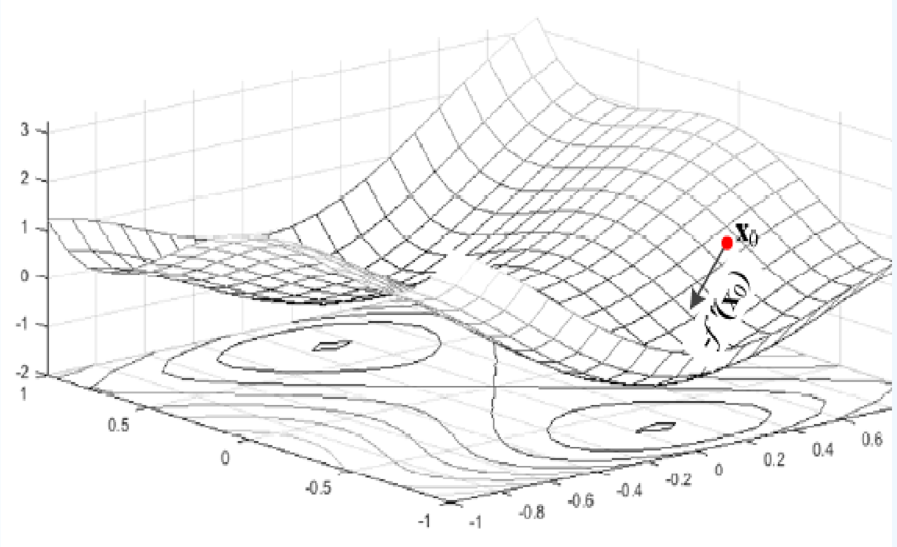


그림 2-25 그레이디언트는 최저점으로 가는 방향을 알려 줌

초기점 $\mathbf{x}_0=(-0.5,0.5)^{\mathrm{T}}$라고 하자. $\mathbf{x}_0$에서의 그레이디언트는 $f'(\mathbf{x}_0) = (-2.5125, -2.5)^{\mathrm{T}}$ 즉, $\nabla f|_{\mathbf{x}_0} = (-2.5125, -2.5)^{\mathrm{T}}$이다. [그림 2-25]는 $\mathbf{x}_0$에서 그레이디언트를 화살표로 표시하고 있어, $-f'(\mathbf{x}_0)$은 최저점의 방향을 제대로 가리키는 것을 확인할 수 있다. 하지만 얼마만큼 이동하여 다음 점 $\mathbf{x}_1$로 옮겨갈지에 대한 방안은 아직 없다. 2.3.3절에서 공부하는 경사 하강법은 이에 대한 답을 제공한다.

3. Optimization : Gradients (2/2)

- **Chain Rule :** Derivative of $f(x) = g(h(x))$

$$f'(x) = g'\big(h(x)\big)h'(x)$$
$$f'(x) = g'\big(h(i(x))\big)h'(i(x))i'(x)$$

Ex: $f(x) = 3(2x^2 - 1)^2 - 2(2x^2 - 1) + 5$

Let $h(x) = 2x^2 - 1$

$$f'(x) = \underbrace{(3 * 2(2x^2 - 1) - 2)}_{g'(h(x))}\underbrace{(2 * 2x)}_{h'(x)} = 48x^3 - 32x$$

- **Jacobian Matrix:** derivative matrix of $\mathbf{f}: \mathbb{R}^d \mapsto \mathbb{R}^m$

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_d} \end{pmatrix}$$

$\mathbf{f}: \mathbb{R}^2 \mapsto \mathbb{R}^3$ 의 $\mathbf{f}(\mathbf{x}) = (2x_1 + x_2^2, -x_1^2 + 3x_2, 4x_1x_2)^{\mathrm{T}}$

$$\mathbf{J} = \begin{pmatrix} 2 & 2x_2 \\ -2x_1 & 3 \\ 4x_2 & 4x_1 \end{pmatrix} \qquad \mathbf{J}|_{(2,1)^{\mathrm{T}}} = \begin{pmatrix} 2 & 2 \\ -4 & 3 \\ 4 & 8 \end{pmatrix}$$

- **Hessian Matrix:** Second derivative matrix

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 x_1} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n x_n} \end{pmatrix}$$

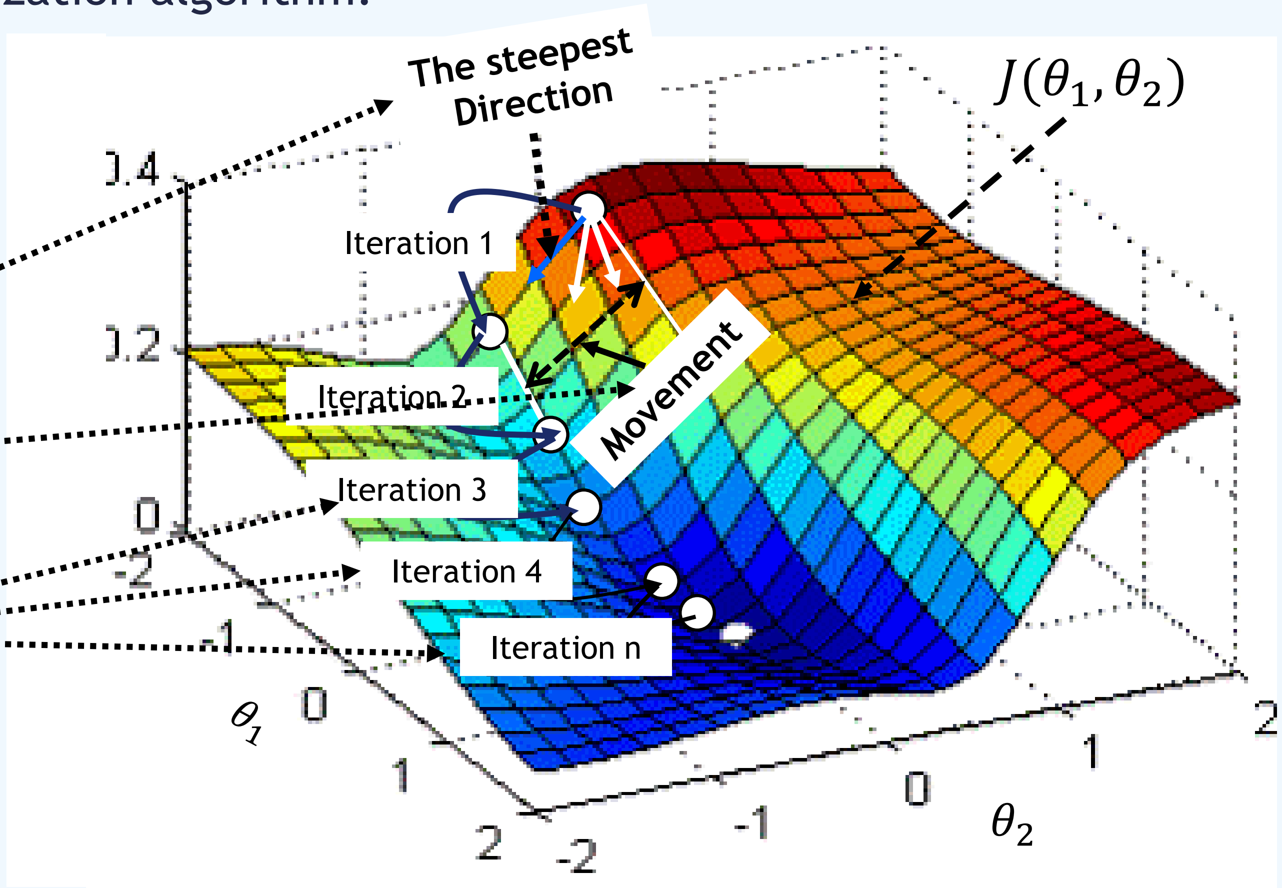
$$f(\mathbf{x}) = f(x_1, x_2)$$
$$= \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$$

$$\mathbf{H} = \begin{pmatrix} 10x_1^4 - 25.2x_1^2 + 8 & 1 \\ 1 & 48x_2^2 - 8 \end{pmatrix}$$

$$\mathbf{H}|_{(0,1)^{\mathrm{T}}} = \begin{pmatrix} 8 & 1 \\ 1 & 40 \end{pmatrix}$$

- Gradient Descent
  - Goal : *Numerically* determine the minimum of a smooth function $J(\theta_1, \theta_2, \ldots, \theta_M)$.
  - Gradient descent is a first-order iterative optimization algorithm.
  - Procedure :
    - ➢ Decide the direction that diminishes the function at most. (Steepest direction)
    - ➢ Decide proper movement rate.

      Movement = rate * slope
    - ➢ Iterate until the function values saturate or reaches at the minimum.



4. NN Learning : Gradient Descent (SDG) (2/3)

From Taylor series

- $L = J(\boldsymbol{\Theta}) = J(\theta_1, \theta_2)$
- $L + \Delta L = J(\theta_1 + \Delta\theta_1, \theta_2 + \Delta\theta_2) \approx J(\theta_1, \theta_2) + \dfrac{\partial J(\theta_1, \theta_2)}{\partial \theta_1}\Delta\theta_1 + \dfrac{\partial J(\theta_1, \theta_2)}{\partial \theta_2}\Delta\theta_2$
- $\Delta L \approx \dfrac{\partial J(\theta_1,\theta_2)}{\partial \theta_1}\Delta\theta_1 + \dfrac{\partial J(\theta_1,\theta_2)}{\partial \theta_2}\Delta\theta_2 = \left(\dfrac{\partial J(\theta_1,\theta_2)}{\partial \theta_1}, \dfrac{\partial J(\theta_1,\theta_2)}{\partial \theta_2}\right) \cdot (\Delta\theta_1, \Delta\theta_2) = \Delta J(\boldsymbol{\Theta}) \cdot \Delta\boldsymbol{\Theta}$

To decrease ΔL maximally, $\Delta\boldsymbol{\Theta} = -\eta \Delta J(\boldsymbol{\Theta})$ (Learning rate: $\eta > 0$)

Take steps proportional to the negative of the gradient of the function at the current point.

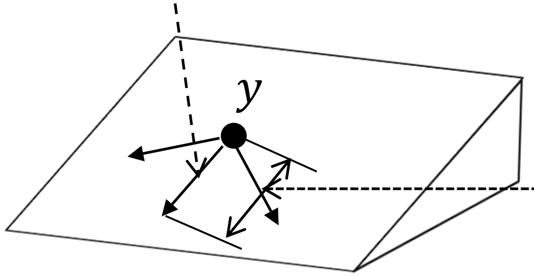
$$(\Delta\theta_1, \Delta\theta_2) = -\eta\left(\frac{\partial J(\theta_1, \theta_2)}{\partial \theta_1}, \frac{\partial J(\theta_1, \theta_2)}{\partial \theta_2}\right)$$

The steepest direction:
Direction of $\Delta J(\boldsymbol{\Theta})$

Movement:

$|\Delta\boldsymbol{\Theta}| = \eta |\Delta J(\boldsymbol{\Theta})|$

= rate * The slope of the steepest direction



Weight Update

New **Old**

$$\theta_1^{n+1} = \theta_1^n + \Delta\theta_1 = \theta_1^n - \eta \frac{\partial J(\theta_1, \theta_2, \ldots, \theta_M)}{\partial \theta_1}$$

$$\theta_2^{n+1} = \theta_2^n + \Delta\theta_2 = \theta_2^n - \eta \frac{\partial J(\theta_1, \theta_2, \ldots, \theta_M)}{\partial \theta_2}$$

$$\vdots$$

$$\theta_M^{n+1} = \theta_M^n + \Delta\theta_M = \theta_M^n - \eta \frac{\partial J(\theta_1, \theta_2, \ldots, \theta_M)}{\partial \theta_M}$$

- Gradient Descent Method

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, J(\boldsymbol{\Theta}) \iff \boldsymbol{\Theta} = \boldsymbol{\Theta} - \rho \bullet \nabla J = \boldsymbol{\Theta} - \rho \bullet \frac{\partial J}{\partial \boldsymbol{\Theta}}$$

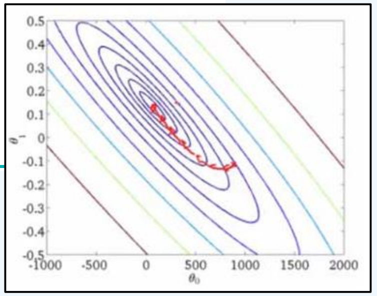
Training set : $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \rightarrow \mathbb{Y} = \{y_1, y_2, \cdots, y_n\}$

Stochastic Gradient Descent(SDG)

알고리즘 2-4 배치 경사 하강 알고리즘(BGD)

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$, 학습률 ρ
출력 : 최적해 $\hat{\boldsymbol{\Theta}}$

```
난수를 생성하여 초기해 Θ를 설정한다.
repeat
    X에 있는 샘플의 그레이디언트 ∇_1, ∇_2, ⋯, ∇_n을 계산한다.
    ∇_total = (1/n) Σ_{i=1,n} ∇_i   // 그레이디언트 평균을 계산
    Θ = Θ − ρ∇_total
until(멈춤 조건)
Θ̂ = Θ
```



알고리즘 2-5 스토케스틱 경사 하강 알고리즘(SGD)

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$, 학습률 ρ
출력 : 최적해 $\hat{\boldsymbol{\Theta}}$

```
난수를 생성하여 초기해 Θ를 설정한다.
repeat
  X의 샘플의 순서를 섞는다.
  for (i=1 to n)
        i번째 샘플에 대한 그레이디언트 ∇_i를 계산한다.
        Θ = Θ − ρ∇_i
until(멈춤 조건)
Θ̂ = Θ
```