Machine Learning

2018/Fall

Review of Linear Algebra, Probability, Random Variable

Prof. Jewon Kang (jewonk@ewha.ac.kr)

This slide has many contents from Numerical Algorithms by J. Solomon, CRC press

Vector

- It is convenient in many applications to represent signals and system coefficients by vectors and matrices. Therefore, a brief overview of linear algebra is considered.
- A vector (denoted by a lowercase bold letter) is an array of real-valued or complex valued numbers or functions. We will assume column vectors. The *N*-dimensional vector is:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

: N-dim col vector

- The transpose of a vector is a row vector

$$\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}$$

- The Hermitian transpose is the complex conjugate of the transpose x

$$\mathbf{x}^H = \left(\mathbf{x}^T\right)^* = \begin{bmatrix} x_1^* & x_2^* & \cdots & x_N^* \end{bmatrix}$$

Vector Space

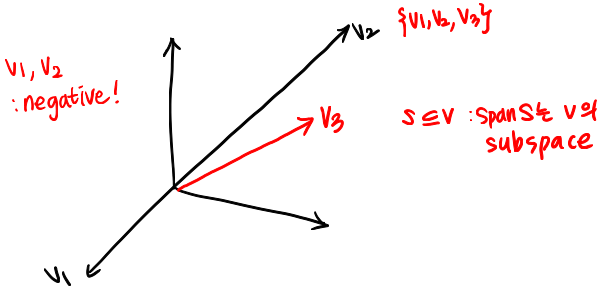
Definition 1.1 (Vector space over $\mathbb{R}$)**.** A *vector space* over $\mathbb{R}$ is a set $\mathcal{V}$ closed under addition and scalar multiplication satisfying the following axioms:

- *Additive commutativity and associativity:* For all $\vec{u}, \vec{v}, \vec{w} \in \mathcal{V}$, $\vec{v} + \vec{w} = \vec{w} + \vec{v}$ and $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$.
- *Distributivity:* For all $\vec{v}, \vec{w} \in \mathcal{V}$ and $a, b \in \mathbb{R}$, $a(\vec{v}+\vec{w}) = a\vec{v}+a\vec{w}$ and $(a+b)\vec{v} = a\vec{v}+b\vec{v}$.
- *Additive identity:* There exists $\vec{0} \in \mathcal{V}$ with $\vec{0} + \vec{v} = \vec{v}$ for all $\vec{v} \in \mathcal{V}$.
- *Additive inverse:* For all $\vec{v} \in \mathcal{V}$, there exists $\vec{w} \in \mathcal{V}$ with $\vec{v} + \vec{w} = \vec{0}$.
- *Multiplicative identity:* For all $\vec{v} \in \mathcal{V}$, $1 \cdot \vec{v} = \vec{v}$.
- *Multiplicative compatibility:* For all $\vec{v} \in \mathcal{V}$ and $a, b \in \mathbb{R}$, $(ab)\vec{v} = a(b\vec{v})$.

A member $\vec{v} \in \mathcal{V}$ is known as a *vector*; arrows will be used to indicate vector variables.

Span

≒ linear combination



Suppose we start with vectors $\vec{v}_1, \ldots, \vec{v}_k \in \mathcal{V}$ in vector space $\mathcal{V}$. By Definition 1.1, we have two ways to start with these vectors and construct new elements of $\mathcal{V}$: addition and scalar multiplication. *Span* describes all of the vectors you can reach via these two operations:

모든 공간을 표현가능.

Definition 1.2 (Span). The *span* of a set $S \subseteq \mathcal{V}$ of vectors is the set

$$\operatorname{span} S \equiv \{a_1\vec{v}_1 + \cdots + a_k\vec{v}_k : \vec{v}_i \in S \text{ and } a_i \in \mathbb{R} \text{ for all } i\}.$$

: Linear Combination $\vec{v}_i$: subset S의 원소

Figure 1.1(a-b) illustrates the span of two vectors. By definition, span S is a *subspace* of $\mathcal{V}$, that is, a subset of $\mathcal{V}$ that is itself a vector space. We provide a few examples:

$a_1\vec{v}_1 + a_2\vec{v}_2$
↳ 이렇게 표현할 수 있는 것이 Set S는 2차원평면 $\mathbb{R}^2$
∴ $S = V$ ($\mathbb{R}^2$ $\mathbb{R}^2$)

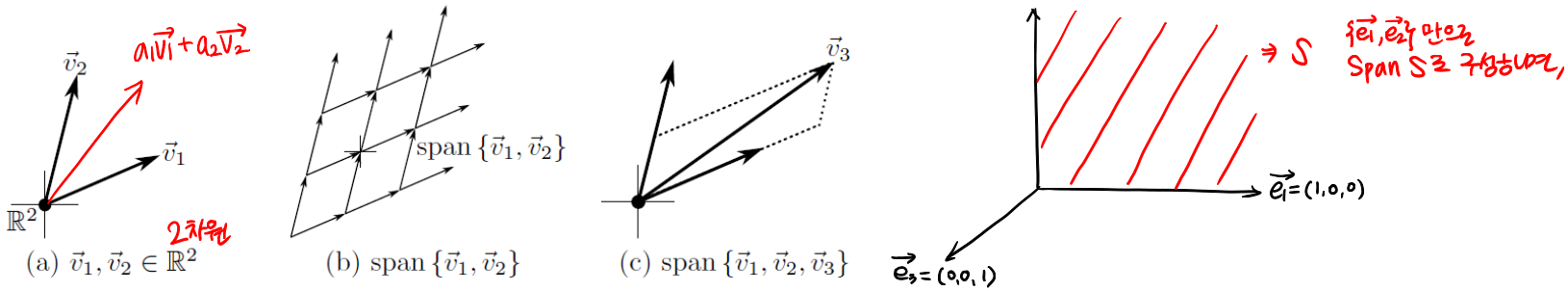


Figure 1.1 (a) Vectors $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$; (b) their span is the plane $\mathbb{R}^2$; (c) $\operatorname{span}\{\vec{v}_1, \vec{v}_2, \vec{v}_3\} = \operatorname{span}\{\vec{v}_1, \vec{v}_2\}$ because $\vec{v}_3$ is a linear combination of $\vec{v}_1$ and $\vec{v}_2$.

"칵테일 재료"

Example 1.3 (Mixology). The typical well at a cocktail bar contains at least four ingredients at the bartender's disposal: vodka, tequila, orange juice, and grenadine. Assuming we have this well, we can represent drinks as points in $\mathbb{R}^4$, with one element for each ingredient. For instance, a tequila sunrise can be represented using the point $(0, 1.5, 6, 0.75)$, representing amounts of vodka, tequila, orange juice, and grenadine (in ounces), respectively.

한 잔!!

4가지 재료!

The set of drinks that can be made with our well is contained in

보드카 테킬라 오렌지 grenadine

$$\text{span}\{(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)\}, \quad \mathbb{R}^4$$: unit vector =element vector

that is, all combinations of the four basic ingredients. A bartender looking to save time, however, might notice that many drinks have the same orange juice-to-grenadine ratio and mix the bottles. The new simplified well may be easier for pouring but can make fundamentally fewer drinks:

$$\text{span}\{(1,0,0,0), (0,1,0,0), (0,0,6,0.75)\}. \quad \mathbb{R}^3$$

차원이 줄어서 효율적이긴하지만, 고차원 공간은 모델의 표현 불가능

For example, this reduced well cannot fulfill orders for a screwdriver, which contains orange juice but not grenadine.

오렌지주스, grenadine을 고정적인 비율로 섞어서 가지고있으면 효율적!

Linear Independence

이러면 쓴이 필요가 없는 vec

$$v_k = a_1 v_1 + a_2 v_2 + \cdots + a_{k-1} v_{k-1} = \sum_{i=1}^{k-1} a_i v_i$$

: 이런식으로 한 벡터 v_k가 다른 vector들의 linear combination으로 표현 가능하면, $(v_1, v_2, \cdots, v_k)$는 linear dependent함

Definition 1.3 (Linear dependence). We provide three equivalent definitions. A set $S \subseteq \mathcal{V}$ of vectors is *linearly dependent* if: ↔ linear independence

한가지만 만족해도 linear dependence

1. One of the elements of S can be written as a linear combination of the other elements, or S contains zero. (S의 원소는 다른 원소로 표현가능하며, S는 zero를 포함한다.)
2. There exists a non-empty linear combination of elements $\vec{v}_k \in \mathcal{S}$ yielding $\sum_{k=1}^{m} c_k \vec{v}_k = 0$ where $c_k \neq 0$ for all k. (선형조합이 존재한다고 ??)
3. There exists $\vec{v} \in S$ ($\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_k$) such that $\operatorname{span} S = \operatorname{span} S \backslash \{\vec{v}\}$. That is, we can remove a vector from S without affecting its span. (임의의 벡터 $\vec{v}_1$를 빼도 span S가 유지된다.) 빼다!

If S is not linearly dependent, then we say it is *linearly independent.*

The concept of linear dependence provides an idea of "redundancy" in a set of vectors. In this sense, it is natural to ask how large a set we can construct before adding another vector cannot possibly increase the span. More specifically, suppose we have a linearly independent set $S \subseteq \mathcal{V}$, and now we choose an additional vector $\vec{v} \in \mathcal{V}$. Adding $\vec{v}$ to S has one of two possible outcomes:

1. The span of $S \cup \{\vec{v}\}$ is *larger* than the span of S.
2. Adding $\vec{v}$ to S has no effect on its span.

Dimension and Bases

Definition 1.4 (Dimension and basis). The *dimension* of $\mathcal{V}$ is the maximal size $|S|$ of a linearly independent set $S \subset \mathcal{V}$ such that span $S = \mathcal{V}$. Any set S satisfying this property is called a *basis* for $\mathcal{V}$.

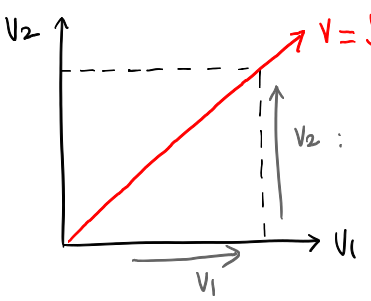
Example 1.5 ($\mathbb{R}^n$). The *standard basis* for $\mathbb{R}^n$ is the set of vectors of the form

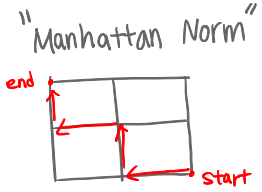
$$\vec{e}_k \equiv (\underbrace{0, \ldots, 0}_{k-1 \text{ elements}}, 1, \underbrace{0, \ldots, 0}_{n-k \text{ elements}}).$$

That is, $\vec{e}_k$ has all zeros except for a single one in the k-th position. These vectors are linearly independent and form a basis for $\mathbb{R}^n$; for example in $\mathbb{R}^3$ any vector (a, b, c) can be written as $a\vec{e}_1 + b\vec{e}_2 + c\vec{e}_3$. Thus, the dimension of $\mathbb{R}^n$ is n, as expected.

Example 1.6 (Polynomials). The set of monomials $\{1, x, x^2, x^3, \ldots\}$ is a linearly independent subset of $\mathbb{R}[x]$. It is infinitely large, and thus the dimension of $\mathbb{R}[x]$ is ∞.

Vector Norm $\|\vec{v_1}\|$





- Vector Norm = 벡터의 크기 = 벡터의 거리

If a norm p: $X \rightarrow R$ is given on a vector space X ← vector space X를 바꿔주는 함수
then the norm of a vector $x \in X$ denoted by $\|x\|_p = p(x)$ is

p-norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$ → p=1

1. $p(ax) = |a|\, p(x)$ a: 상수
2. $p(x+y) \le p(x) + p(y)$
3. if $p(x) = 0$, then x is the zero vector

↓ 이 조건을 만족해야 p-norm!

The Euclidean or L_2 norm: (유클리드한 distance)

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2}$$

The L_1 norm:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|$$

The L_∞ norm:

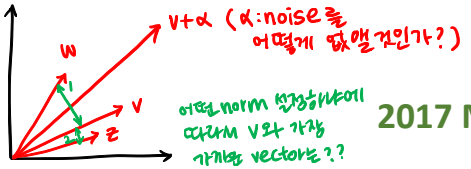
$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

절대값의 max

vector 차의

L_0 norm : x의 non-zero element 갯수

ex) $x = (1,0,0,1) \rightarrow L_0$ norm $= 2$



Cont.

- The squared norm represents the energy in the signal:

$$\|\mathbf{x}\|^2 = \sum_{n=0}^{N-1} |x_n|^2$$

희소성 ?

- The L0 norm represents the sparsity in the signal
- The norm can be used to measure the distance between two vectors:

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{\sum_{i=1}^{N} |x_i - y_i|^2}$$

- If the vector has a non-zero norm, it can be normalized as follows:

(벡터의 크기를 1로 바꾸고 싶을 때)

(unit vector)

$$\mathbf{v}_x = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$ (normalization !!)

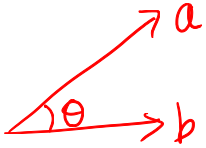
x의 norm

Inner Product Vector Space

- For two real vectors, the inner product becomes

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b} = \sum_{i=1}^{N} a_i b_i$$

- The inner product determines geometric quantities such as length and relative orientation of vectors



$$\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\| \|\mathbf{b}\| \cos\theta$$

where θ is the angle between two vectors. Therefore, two nonzero vectors **a** and **b** are **orthogonal** if their inner product is zero:

$$\langle \mathbf{a}, \mathbf{b} \rangle = 0$$

Two vectors that are orthogonal and have unit norm are called **orthonormal**.

synthesis : 합성 (요소들을 알고있을때)

어떤 vector를 다른 공간에 사영시킴 (변환)

One Dimensional Projection

Analysis or decomposition

- The inner product determines geometric quantities such as length and relative orientation of vectors

$u = u_1 e_1 + u_2 e_2 = a_1 u_{\perp v} + a_2 u_{|v}$: u가 어떻게 표현되는지 알기에 유용함

standard base

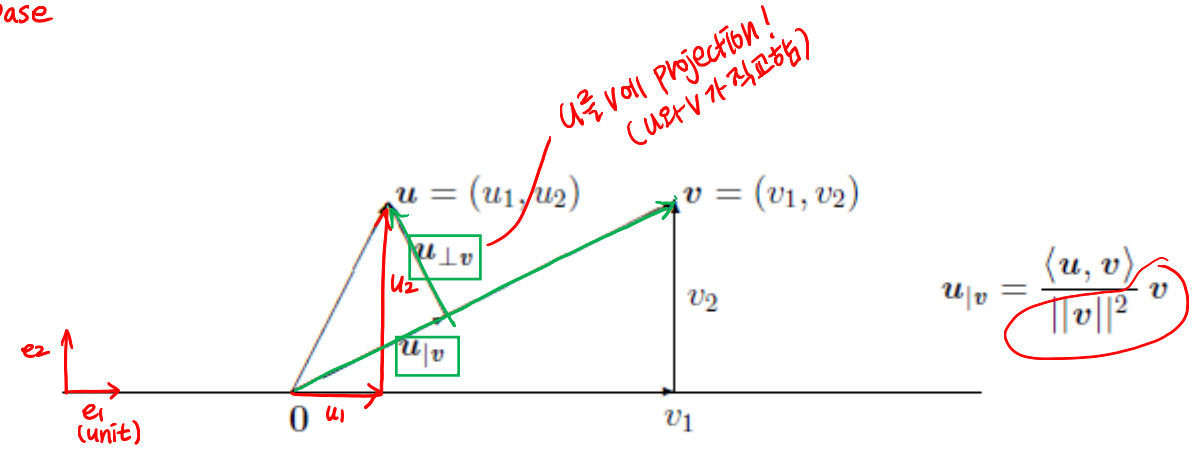


Figure 1: Two vectors, $\boldsymbol{u} = (u_1, u_2)$ and $\boldsymbol{v} = (v_1, v_2)$ in $\mathbb{R}^2$. Note that $||\boldsymbol{v}||^2 = \langle \boldsymbol{v}, \boldsymbol{v} \rangle = v_1^2 + v_2^2$ is the squared length of $\boldsymbol{v}$ (viewed as a directed line).

∴ u를 v로 projection!

Linearity 선형성

A function from one vector space to another that preserves linear structure is known as a *linear* function:

Definition 1.7 (Linearity). Suppose $\mathcal{V}$ and $\mathcal{V}'$ are vector spaces. Then, $\mathcal{L} : \mathcal{V} \to \mathcal{V}'$ is *linear* if it satisfies the following two criteria for all $\vec{v}, \vec{v}_1, \vec{v}_2 \in \mathcal{V}$ and $c \in \mathbb{R}$:

- $\mathcal{L}$ *preserves sums:* $\mathcal{L}[\vec{v}_1 + \vec{v}_2] = \mathcal{L}[\vec{v}_1] + \mathcal{L}[\vec{v}_2]$
- $\mathcal{L}$ *preserves scalar products:* $\mathcal{L}[c\vec{v}] = c\mathcal{L}[\vec{v}]$

★ 두 조건 만족할 때 "선형이다!"

We can write a particularly nice form for linear maps on $\mathbb{R}^n$. The vector $\vec{a} = (a_1, \ldots, a_n)$ is equal to the sum $\sum_k a_k \vec{e}_k$, where $\vec{e}_k$ is the k-th standard basis vector from Example 1.5. Then, if $\mathcal{L}$ is linear we can expand:

$$\begin{aligned}
\mathcal{L}[\vec{a}] &= \mathcal{L}\left[\sum_k a_k \vec{e}_k\right] \text{ for the standard basis } \vec{e}_k \\
&= \sum_k \mathcal{L}\left[a_k \vec{e}_k\right] \text{ by sum preservation} \\
&= \sum_k a_k \mathcal{L}\left[\vec{e}_k\right] \text{ by scalar product preservation.}
\end{aligned}$$

Matrix

data, signal 2 벡터 o

The expansion of linear maps above suggests a context in which it is useful to store multiple vectors in the same structure. More generally, say we have n vectors $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^m$. We can write each as a column vector:

$$\vec{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{m2} \end{pmatrix}, \cdots, \vec{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{mn} \end{pmatrix}.$$

Carrying these vectors around separately can be cumbersome notationally, so to simplify matters we combine them into a single $m \times n$ matrix:

col vector들을 옆에 붙인

$$\begin{pmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{pmatrix}.$$

★ col vector의 group !!

We will call the space of such matrices $\mathbb{R}^{m \times n}$.

Since we constructed matrices as convenient ways to store sets of vectors, we can use multiplication to express how they can be combined linearly. In particular, a matrix in $\mathbb{R}^{m\times n}$ can be multiplied by a column vector in $\mathbb{R}^n$ as follows:

Matrix V

$$\begin{pmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \equiv c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n.$$

weight

: 각 col vector의 linear combination

Expanding this sum yields the following explicit formula for matrix-vector products:

$$\begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} c_1v_{11} + c_2v_{12} + \cdots + c_nv_{1n} \\ c_1v_{21} + c_2v_{22} + \cdots + c_nv_{2n} \\ \vdots \\ c_1v_{m1} + c_2v_{m2} + \cdots + c_nv_{mn} \end{pmatrix}.$$

Example 1.11 (Identity matrix). We can store the standard basis for $\mathbb{R}^n$ in the $n \times n$ "identity matrix" $I_{n\times n}$ given by:

$$I_{n\times n} \equiv \begin{pmatrix} | & | & & | \\ \vec{e}_1 & \vec{e}_2 & \cdots & \vec{e}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

즉 각 col vector가 standard basis로 구성!

Transpose

$a_{ij} \rightarrow a_{ji}$

- If **A** is an *n* x *m* matrix, then the transpose $\mathbf{A}^T$ is the *m* x *n* matrix that is formed by interchanging the rows and columns of **A**. If the matrix is square, the transpose is formed by reflecting its elements with respect to the diagonal.
- A square matrix **A** is symmetric if

$$\mathbf{A} = \mathbf{A}^T \qquad a_{ij} = a_{ji}$$

  - For complex matrices, the Hermitian transpose is defined as

$$\mathbf{A}^H = \left(\mathbf{A}^*\right)^T = \left(\mathbf{A}^T\right)^*$$

element conjugate

1. $\left(\mathbf{A}+\mathbf{B}\right)^T = \mathbf{A}^T + \mathbf{B}^T$
2. $\left(\mathbf{A}^T\right)^T = \mathbf{A}$
3. $\left(\mathbf{AB}\right)^T = \mathbf{B}^T\mathbf{A}^T$

Example 1.16 (Residual norm)**.** Suppose we have a matrix A and two vectors $\vec{x}$ and $\vec{b}$. If we wish to know how well $A\vec{x}$ approximates $\vec{b}$, we might define a *residual* $\vec{r} \equiv \vec{b} - A\vec{x}$; this residual is zero exactly when $A\vec{x} = \vec{b}$. Otherwise, we can use the norm $\|\vec{r}\|_2$ as a proxy for the similarity of $A\vec{x}$ and $\vec{b}$. We can use the identities above to simplify:

residual norm

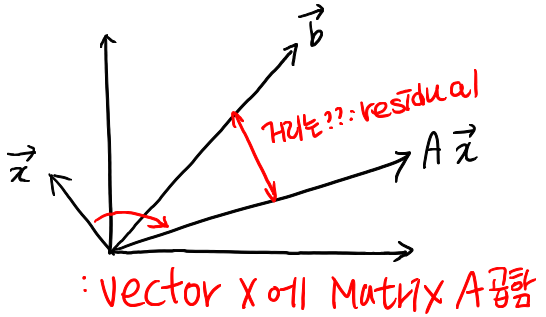
$$\|\vec{r}\|_2^2 = \|\vec{b} - A\vec{x}\|_2^2 \quad : \vec{b}\text{와 } A\vec{x}\text{의 거리}$$

$$= (\vec{b} - A\vec{x}) \cdot (\vec{b} - A\vec{x}) \quad \text{inner product}$$

$$= (\vec{b} - A\vec{x})^\top (\vec{b} - A\vec{x}) \text{ by our expression for the dot product above}$$

$$= (\vec{b}^\top - \vec{x}^\top A^\top)(\vec{b} - A\vec{x}) \text{ by properties of transposition}$$

$$= \vec{b}^\top\vec{b} - \vec{b}^\top A\vec{x} - \vec{x}^\top A^\top\vec{b} + \vec{x}^\top A^\top A\vec{x} \text{ after multiplication}$$



All four terms on the right-hand side are scalars, or equivalently 1×1 matrices. Scalars thought of as matrices enjoy one additional nice property $c^\top = c$, since there is nothing to transpose! Thus,

$$\vec{x}^\top A^\top \vec{b} = (\vec{x}^\top A^\top \vec{b})^\top = \vec{b}^\top A\vec{x}. \quad \therefore \text{constant !!}$$

This allows us to simplify even more:

$$\|\vec{r}\|_2^2 = \vec{b}^\top\vec{b} - 2\vec{b}^\top A\vec{x} + \vec{x}^\top A^\top A\vec{x}$$

$$= \|A\vec{x}\|_2^2 - 2\vec{b}^\top A\vec{x} + \|\vec{b}\|_2^2.$$

Rank

$C_m = \sum_{i=1}^{m-1} C_i$ (C_m이 다른 C_i들로 표현가능해면)

- Let **A** be an *n* x *m* matrix partitioned in a set of *m* column vectors

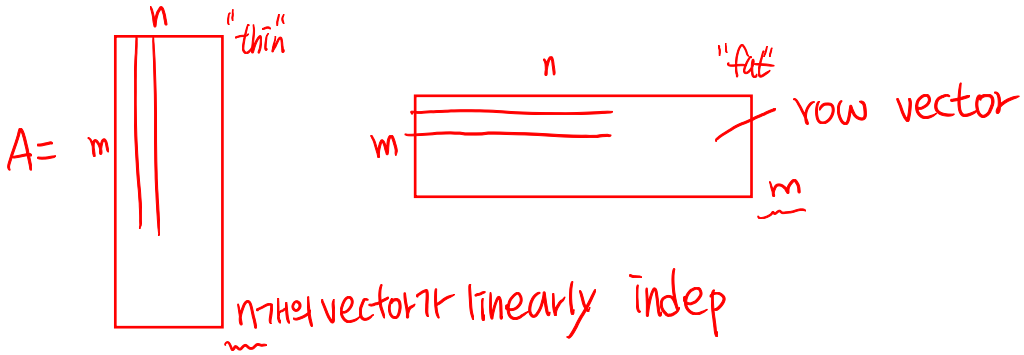
$$\mathbf{A} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_m]$$

- The **rank** of **A**, rank(**A**) is defined as the number of linearly independent columns in **A**, i.e., the number of linearly independent vectors in the set {**c**1, **c**2, … , **c***m*}. One of the properties of the rank is that

선형독립인 col vec수!

$$rank(\mathbf{A}) = rank(\mathbf{A}^H)$$

Cont.



- The important property of the rank is

$$rank(\mathbf{A}) = rank(\mathbf{A}\mathbf{A}^H) = rank(\mathbf{A}^H\mathbf{A})$$

- Since the rank of a matrix is equal to the number of linearly independent rows and the number of linearly independent columns, then, if **A** is an *m* x *n* matrix:

$$rank(\mathbf{A}) \le \min(m, n)$$

row col

A: square, full rank !

- If **A** is an *m* x *n* matrix and $rank(\mathbf{A}) = \min(n,m)$, then **A** is of full rank. If **A** is a square matrix of full rank, then there exists a unique inverse matrix $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Cont.

where

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

=square matrix

is the identity matrix with ones along the main diagonal and zeros everywhere else. In this case, **A** is said to be invertible or nonsingular. If **A** is not of full rank, then it is noninvertible or singular and does not have an inverse.

=full rank라고함

If **A** and **B** are invertible, then full rank라면!

1. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
2. $\left(\mathbf{A}^{H}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{H}$
3. $(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\right)^{-1}\mathbf{DA}^{-1}$

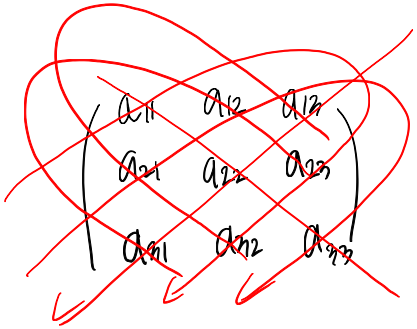
★Kalman Filter

;더 유용한수 있음

고차원의 inverse 구해야할!! 꿀팁!

각 component의 inverse로 표현가능

Determinant



- If **A** = a_{11} is a 1 x 1 matrix, the **determinant** is defined as det(**A**) = a11. For an *n* x *n* matrix, the determinant is defined recursively as

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det\left(\mathbf{A}_{ij}\right)$$

where $\mathbf{A}_{ij}$ is the (*n*-1) x (*n*-1) matrix formed by deleting the *ith* row and the *jth* column of **A**.

- **Property**: An *n* x *n* matrix **A** is invertible iff det(**A**) ≠ **0**

full rank

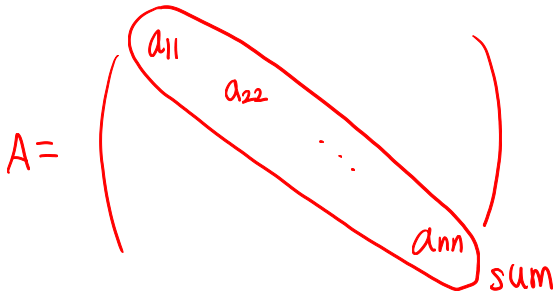
Cont.

For **A** and **B** being *n* x *n* matrices, if **A** is invertible, and a constant α, the following properties hold:

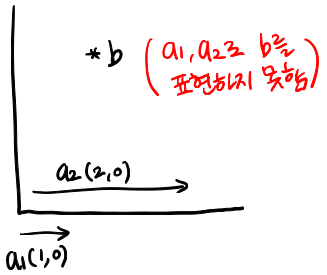
$$\begin{cases} 1.\ \det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B}) \\ 2.\ \det(\mathbf{A}^T) = \det(\mathbf{A}) \\ 3.\ \det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A}) \\ 4.\ \det(\mathbf{A}^{-1}) = \dfrac{1}{\det(\mathbf{A})} \end{cases}$$

For an *n* x *n* matrix **A**, the trace function is defined as

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$



Linear Equations



ex) $\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ (a_1 a_2 / A / b)

$Ax = b$

$\sum a_i x_i = b$

: Span $(a_1, a_2) \Rightarrow \mathbb{R}^1$ (직선으로 밖에 표현 못함)

if $b \in \mathbb{R}^1$ 이면, $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ 인 경우 표현가능!

- Consider the following set of *n* linear equations in the *m* unknowns x_i, $i = 1,2,\ldots,m$

unknown

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m = b_n \end{cases} \quad (n\text{개}) \rightarrow \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_m \\ | & | & & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

: b를 어떻게 만들까?

- These equations may be written in matrix form as follows

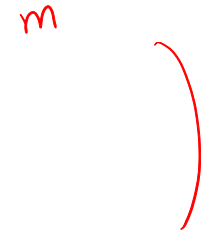
$$\mathbf{Ax} = \mathbf{b}$$

where **A** is an *n* x *m* matrix with entries a_{ij}, **x** is an *m*-dimensional vector of unknowns x_i, and **b** is an *n*-dimensional vector with elements b_i.

Cont.

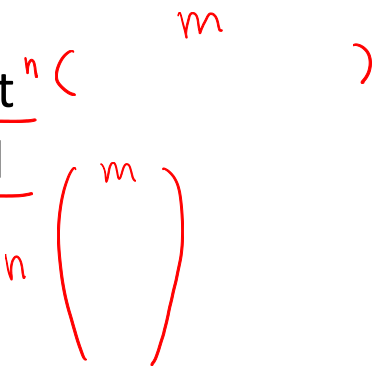
- A convenient way to view **Ax=b** is as an expansion of vector **b** in terms of a **linear combination** of the column vectors $\mathbf{a}_i$ of the matrix **A**:

$$\mathbf{b} = \sum_{i=1}^{m} x_i \mathbf{a}_i$$



Matrix 형태에 따라서 3가지로 표현가능 !

- Solving the equation above depends on a number of factors including the relative size of *m* and *n*, the rank of **A**, and the elements of **b**.
  - Square matrix (*m*=*n*)
  - Rectangular matrix (*n* < *m*): A is wide, fat
  - Rectangular matrix (*m* < *n*): A is thin, tall



Cont.(Square Matrix)

square matrix
Full rank

- If **A** is nonsingular, then the inverse $\mathbf{A}^{-1}$ exists and the solution is uniquely defined by

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

- However, if **A** is singular, then there may either be no solution (the equations are inconsistent) or many solutions: the columns of **A** are linearly dependent and there exist nonzero solutions to the homogeneous equation.

$$\mathbf{A}\mathbf{z} = \mathbf{0}$$

- In fact, there will be k = n – rank(A) linearly independent solutions to the homogeneous equations. Therefore, if there is at least one vector x_0 that solves **A**x=**b**, then any vector of the form

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1\mathbf{z}_1 + \ldots + \alpha_k\mathbf{z}_k$$

will also be a solution.

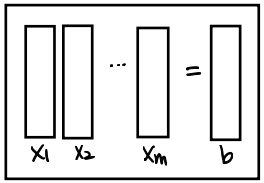
fat matrix 의 경우
unique한 solution X

$\text{rank}(A) \le \min(n, m) = n$

$$n\begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_m \\ | & | & & | \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Cont.(Rect. matrix $n < m$, Fat matrix)

ex) col vector가 100개 있는데, rank(A)=2면
두 식으로 b를 표현 가능함: many solution or 해 X → ex) x_1 x_2 b



m개 equation
↓ 새로,
n개만 있어도 됨

- In this situation, there are **fewer** equations than unknowns. Therefore, there can be many vectors satisfying the equations: the solution is **undetermined** or incompletely specified. A common approach to define a unique solution is to find the vector that satisfies the equations and has a minimum norm; i.e.

많은 x들 중에 x의 norm을 최소로하도록 하자
이 경우 unique한 solution존재

$$\text{①}\ \min\|\mathbf{x}\| \quad \textit{such that} \quad \text{②}\ \mathbf{Ax} = \mathbf{b} \quad \sum_i a_i x_i = b$$

- If rank(**A**) = *n* (the rows of **A** are linearly independent), then the *n* x *n* matrix $\mathbf{AA}^H$ is invertible and the minimum norm solution is

unique solution !

pseudo-inverse of **A**

$$\mathbf{x}_0 = \mathbf{A}^H\left(\mathbf{AA}^H\right)^{-1}\mathbf{b}$$

$$\mathbf{A}^+ = \mathbf{A}^H\left(\mathbf{AA}^H\right)^{-1}$$

ex) a_1 a_2 a_3 a_4 a_5

① $\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix}$ Many solution

$\begin{pmatrix}1\\2\end{pmatrix} = b$ 표현가능 a_1, a_2만으로 표현가능

② $\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}2\\0\end{pmatrix}\begin{pmatrix}3\\0\end{pmatrix}$ No solution

$\begin{pmatrix}1\\2\end{pmatrix} = b$

No Solution & Unique

Cont.(Rect. matrix n > m, Thin matrix)

$$\underset{n}{\overset{m}{\begin{pmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{pmatrix}}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

- In this situation, there are **more** equations than unknowns and, in general, no solution exists. The equations are inconsistent and the solution is overdetermined.

ex) $A = \left[\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right] \Rightarrow b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$

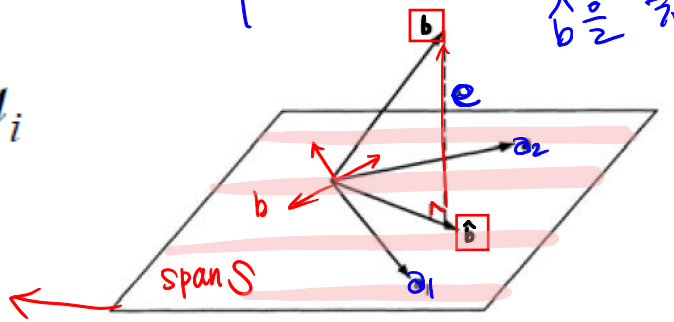
rank(A) = 2

if $b_2 = 0$ 해 O
$b_2 \neq 0$ 해 X

- Since an arbitrary vector **b** cannot be represented as a linear combination of the columns of **A,** the goal is to find the coefficients *xi* producing the best approximation to **b**:

{ span S 가 어떤곳에서 b가 정의될때, 거리최소화되는 $\hat{b}$을 찾자! }

$$\hat{\mathbf{b}} = \sum_{i=1}^{m} x_i a_i$$



if) b가 $\vec{a_1}, \vec{a_2}$ 공간인 span S에 존재한다면, b를 표현가능함

3 equations in 2 unknowns

Cont.(Rect. matrix n > m, Thin matrix)

- The approach commonly used in this situation is to find the least squares solution; i.e., the vector **x** minimizing the norm of the error

$$\|\mathbf{e}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$$

- The least squares solution has the property that the error

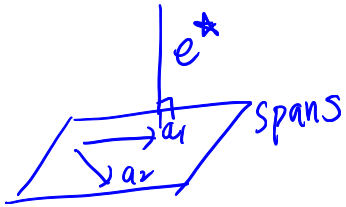
$$\mathbf{e} = \mathbf{b} - \mathbf{A}\mathbf{x}$$

min ||e|| Find !!!

- is orthogonal to the column vectors of **A**. This orthogonality implies that

$$\mathbf{A}^H\mathbf{e} = 0$$

$$\mathbf{A}^H\mathbf{A}\mathbf{X} = \mathbf{A}^H\mathbf{b}$$



$a_1 \perp e \Rightarrow \langle a_1, e\rangle = 0 : a_1'e = 0$

$\langle a_2, e\rangle = 0 : a_2'e = 0$

A의 열벡터와 e는 직교

Cont.(Rect. matrix n > m, Thin matrix)

A : full rank ⟶ $(A^H A)^{-1}$ 존재!

- If the columns of **A** are linearly independent (**A** has full rank), the matrix $\mathbf{A}^H\mathbf{A}$ is invertible and the least square solution is

$$\mathbf{x}_0 = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{b}$$

$$\mathbf{x}_0 = \mathbf{A}^+\mathbf{b}$$

$$\mathbf{A}^+ = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H$$: pseudo inverse !

is the pseudo-inverse of the matrix **A** for the overdetermined problem. Furthermore, the best approximation of **b** is given by the projection of the vector onto the subspace spanned by the vectors **a***i*

2차원공간의 projection vector ←

$$\hat{\mathbf{b}} = \mathbf{A}\mathbf{x}_0 = \mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{b}$$

3차원공간 vector

P (Projection Matrix)

3차원공간 vector를 2차원공간 vector로 projection 시킴!

Cont.(Rect. matrix n > m, Thin matrix)

즉, (col vec space / Target b의 dim) 에 따라서

↓ solution

{ unique / many → min ||e|| / No

or $$\hat{\mathbf{b}} = \mathbf{P}_A \mathbf{b}$$

where $$\mathbf{P}_A = \mathbf{A}\left(\mathbf{A}^H \mathbf{A}\right)^{-1} \mathbf{A}^H$$

is called the projection matrix. Finally, using the orthogonality condition, the minimum mean square error is

distance : $$\boxed{\min \|\mathbf{e}\|^2 = \mathbf{b}^H \mathbf{e} = \mathbf{b}^H \mathbf{b} - \mathbf{b}^H \mathbf{A}\mathbf{x}_0}$$

$e = b - Ax$

Special Form of Matrix

- Diagonal Matrix - a square matrix having all entries equal to zero except, possibly, for those along the main diagonal:

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

The diagonal matrix may be written as

$$\mathbf{A} = \operatorname{diag}\{a_{11} \quad a_{22} \quad \cdots \quad a_{nn}\}$$

Cont.

- Identity matrix: a diagonal matrix with ones along a diagonal

$$\mathbf{I} = \mathrm{diag}\{1 \;\; 1 \;\; \ldots \;\; 1\} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

대각 1

- Block diagonal matrix – a diagonal matrix whose entries along the diagonal are replaced with matrices

matrix !!

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & 0 & \cdots & 0 \\ 0 & \mathbf{A}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_{m} \end{bmatrix}$$

↳ Block of diagonal !!

Cont.

- Upper triangular matrix – a square matrix where all entries below the diagonal are zero, i.e.

$$a_{ij} = 0 \quad for \quad i > j$$

e.g

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}$$

- Lower triangular matrix – a square matrix where all entries above the diagonal are zero, i.e.

$$a_{ij} = 0 \quad for \quad i < j$$

Cont.

- Toeplitz matrix – a square *n* x *n* matrix, in which all entries along each of the diagonals have the same value

$$a_{ij} = a_{i+1,j+1} \qquad \forall\, i < n,\ j < n$$

- Hankel matrix – a square *n* x *n* matrix, in which all entries along each of the cross-diagonals have the same value

$$a_{ij} = a_{i+1,j-1} \qquad \forall\, i < n,\ j \le n$$

Toeplitz matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 1 & 3 & 5 \\ 4 & 2 & 1 & 3 \\ 6 & 4 & 2 & 1 \end{bmatrix}$$

Hankel matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 3 & 5 & 7 & 4 \\ 5 & 7 & 4 & 2 \\ 7 & 4 & 2 & 1 \end{bmatrix}$$

Orthogonal Matrix

- Orthogonal matrix – a real *n* x *n* matrix with orthogonal columns (and rows):

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$$

각 col vector orthogonal

$$\mathbf{a}_i^T \mathbf{a}_j = \begin{cases} 1 & for\ i = j \\ 0 & for\ i \neq j \end{cases}$$

then **A** is orthogonal. We observe that if **A** is orthogonal, then

$$\mathbf{A}^T \mathbf{A} = \mathbf{I}$$

Therefore, the inverse of an orthogonal matrix is equal to its transpose

$$\mathbf{A}^{-1} = \mathbf{A}^T$$

Unitary Matrix

- Unitary matrix – a complex *n* x *n* matrix with orthogonal columns (rows):

$$\mathbf{a}_i^H \mathbf{a}_j = \begin{cases} 1 & for\ i = j \\ 0 & for\ i \neq j \end{cases}$$

then

$$\mathbf{A}^H \mathbf{A} = \mathbf{I}$$

The inverse of a unitary matrix equals to its Hermitian transpose

$$\mathbf{A}^{-1} = \mathbf{A}^H$$

Quadratic Form of Matrix

- The quadratic form of a real symmetric *n* x *n* matrix **A** is the scalar defined by

$$Q_A(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

- where **x** is a vector of *n* real variables. Observe that the quadratic form is a quadratic function in the *n* variables *x1*, *x2*, …, *xn*. For example, the quadratic form of

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

x' A x

$$(x_1 \quad x_2)\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

is

$$Q_A(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} = 3x_1^2 + 2x_1x_2 + 2x_2^2$$

Positive Definite Matrix

- If the quadratic form of a matrix **A** is positive for all nonzero vectors **x**,

$$Q_A(\mathbf{x}) > 0$$

then **A** is said to be positive definite and we write **A** > 0. For example, if

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

$$Q_A(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 3x_2^2$$

is positive definite since $Q_A(\mathbf{x}) > 0$ for all $\mathbf{x} \neq 0$.

Cont.

If the quadratic form of a matrix **A** is non negative for all nonzero vectors **x**,

$$Q_A(\mathbf{x}) \geq 0$$

then **A** is said to be positive semidefinite.

If the quadratic form of a matrix **A** is negative for all nonzero vectors **x**,

$$Q_A(\mathbf{x}) < 0$$

then **A** is said to be negative definite.

If the quadratic form of a matrix **A** is non positive for all nonzero vectors **x**,

$$Q_A(\mathbf{x}) \leq 0$$

then **A** is said to be negative semidefinite.

A matrix **A** that is none of the above is called indefinite.

대비 Review 해줌 (나중에 대변)

Eigenvalues and Eigenvectors

Let **A** be an *n* x *n* matrix and consider the following set of linear equations:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

where λ is a constant. Equivalently:

$$\left(\mathbf{A} - \lambda\mathbf{I}\right)\mathbf{v} = 0$$

In order for a nonzero vector **v** to be a solution to this equation, it is necessary for the matrix **A** - λ**I** to be singular. Therefore, the determinant of **A** - λ**I** must be zero:

$$p(\lambda) = \det\left(\mathbf{A} - \lambda\mathbf{I}\right) = 0$$

where $p(\lambda)$ is an n^{th} order polynomial in λ. This polynomial is called the characteristic polynomial of the matrix **A** and its *n* roots λ_i for $i = 1, 2, \ldots, n$ are called the eigenvalues of **A**.

Cont.

For each eigenvalue λ_i for $i = 1, 2, \ldots, n$ the matrix $\mathbf{A} - \lambda\mathbf{I}$ will be singular and there will be at least one nonzero vector $\mathbf{v}_i$ that solves

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

These vectors $\mathbf{v}_i$ are called the eigenvectors of $\mathbf{A}$. For any eigenvector $\mathbf{v}_i$, $\alpha\mathbf{v}_i$ will also be an eigenvector for any constant α. Therefore, eigenvectors are often normalized to have unit norm.

Property 1: The nonzero eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n$ corresponding to the distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are linearly independent.

If $\mathbf{A}$ is an $n \times n$ singular matrix, then there are nonzero solutions to the homogeneous equation

$$\mathbf{A}\mathbf{v}_i = 0$$

and it follows that $\lambda = 0$ is an eigenvalue of $\mathbf{A}$.

There are n – rank($\mathbf{A}$) linearly independent solutions Therefore, $\mathbf{A}$ will have rank($\mathbf{A}$) nonzero eigenvalues and n – rank($\mathbf{A}$) eigenvalues that are equal zero.

Cont.

Property 2: The eigenvalues of a Hermitian matrix are real.

Property 3: A Hermitian matrix is positive definite **A** > 0 iff eigenvalues of **A** are positive: $\lambda_k > 0$.

The determinant of a matrix is related to its eigenvalues as

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

Property 4: The eigenvectors of a Hermitian matrix corresponding to distinct eigenvalues are orthogonal; i.e.

$$if\ \lambda_i \neq \lambda_j,\ then\ \left\langle \mathbf{v}_i, \mathbf{v}_j \right\rangle = 0$$

Spectral Theorem: Any Hermitian matrix **A** can be decomposed as

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^H + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^H + \ldots + \lambda_n \mathbf{v}_n \mathbf{v}_n^H$$

where λ_i are the eigenvalues of **A** and $\mathbf{v}_i$ are a set of orthonormal eigenvectors.

Cont.

Property 5: Let **B** be an *n* x *n* matrix with eigenvalues λ_i and let **A** be a matrix that is related to **B** as follows

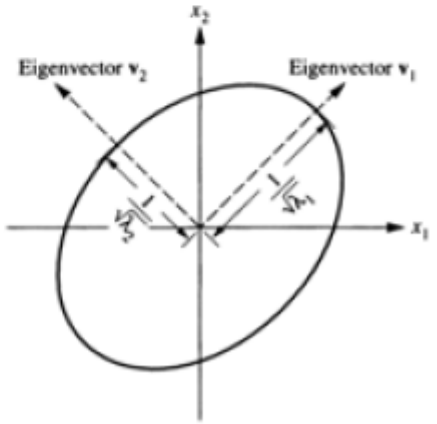
$$\mathbf{A} = \mathbf{B} + \alpha\mathbf{I}$$

Then **A** and **B** have the same eigenvectors and the eigenvalues of **A** are λ_i + α.

Property 6: For a symmetric positive definite matrix **A**, the equation

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 1$$

defines an ellipse in *n* dimensions whose axes are in the direction of the eigenvectors $\mathbf{v}_j$ of **A** with the half-length of these axes equal to $1/\sqrt{\lambda_j}$



Probability

- Given a sample space Ω, a function P defined on the subsets of Ω is a probability measure if the following four axioms are satisfied:

1. $P(A) \geq 0$
2. $P(\emptyset) = 0.$
3. $P(\Omega) = 1.$
4. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $A_1, A_2, \ldots$ are events that are mutually exclusive or pairwise disjoint.

The probability measure ***P***: ***F*** -> [0, 1] is a function on ***F*** that assigns to
an event ***A*** **to** a number in [0,1], such that above axioms are satisfied.

Cumulative distribution function

CDF

- Important measures of probability are a cumulative distribution function (cdf) and a probability density/mass function (pdf, pmf).

cdf (continuous): $F_x(x') \equiv P\{x \le x'\}$

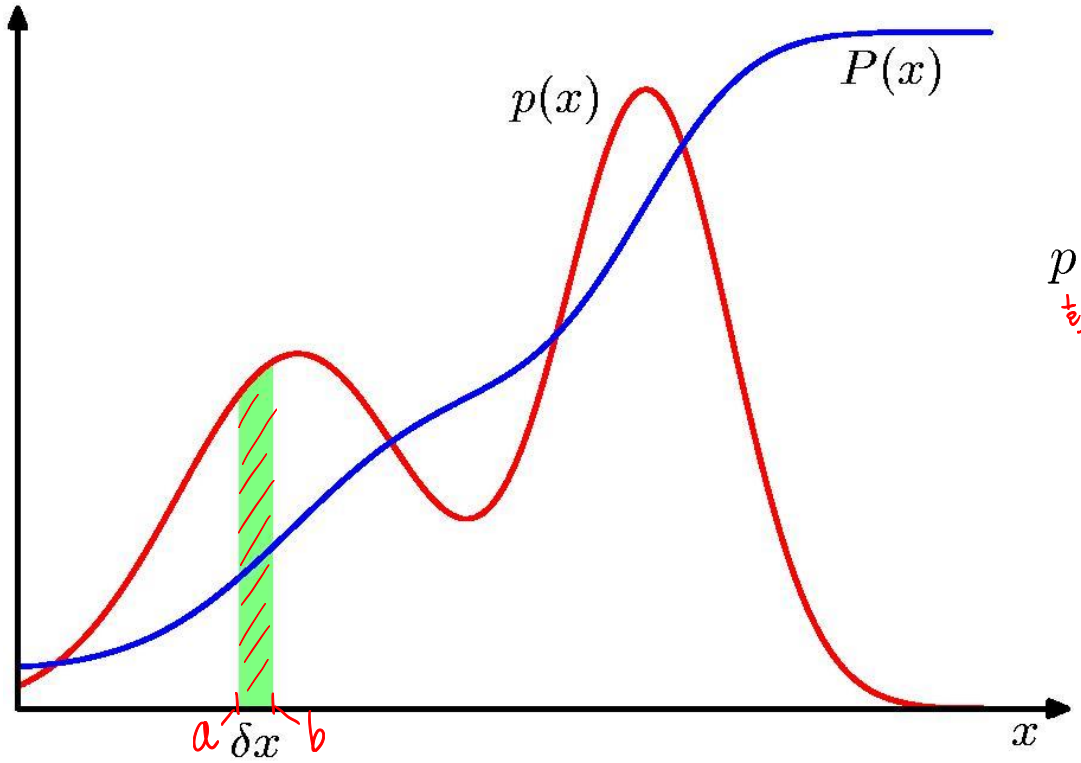
~~cdf~~ cmf (discrete): $F_x(x') \equiv \sum_k P\{x = k\}\, u_{x-k}$

u_{x-k}: a step function

Properties of cdf:

1) $F_x(-\infty) = 0$ – $P\{x < -\infty\} = 0$
2) $F_x(\infty) = 1$ – *true event*
3) $F_x(x') = \lim_{\varepsilon \to 0} = F_x(x' + \varepsilon) = F_x(x'^{+})$ – *continuous on the right*
4) *If* $x'_2 > x'_1 \Rightarrow F_x(x'_2) \ge F_x(x'_1)$ – *cdf is nondecreasing*

Probability Densities



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

확률!

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

CDF

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Conditional Probability

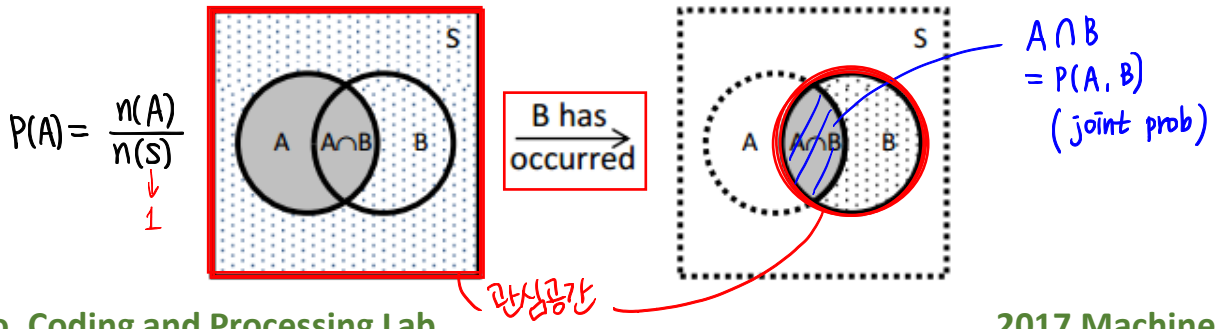
ex) P(A) : 사람이 죽을 확률
P(B) : 암에 걸릴 확률
↓
P(A|B)에 관심!! (암에 걸렸을 때, 죽을 확률)

- If A and B are two events, the probability of event A when we already know that event B has occurred is

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

joint prob

- Interpretation
  - The new evidence "B has occurred" has the following effects
    - The original sample space S (the square) becomes B (the rightmost circle)
    - The event A becomes A ∩ B : 전체공간이 S → B로 바뀌기도
  - P[B] simply re-normalizes the probability of events that occur jointly with B



Joint/Marginal Probability

★ joint prob 알면 → 다 아는것!

- Define the probability of the joint event *A* and *B* as follows:

$p(x_1, \cdots, x_N)$

$$P(A,B) = P(A \mid B)P(B)$$

called the product rule. Given a joint distribution on two events *P*(*A*,*B*), we define the *marginal* distribution as follows (sum rule):

- $P(A) = \sum_{b \in B} P(A,B) = \sum_{b} P(A \mid B = b)P(B = b)$ (conditional)

- $P(B) = \sum_{a \in A} P(A,B) = \sum_{a} P(B \mid A=a)\,P(A=a)$

일반화)

Product and Sum Rules

- *H* denotes some background assumptions.
- Product rule

condition

$$P(A, B \mid H) = P(A \mid B, H) P(B \mid H)$$

$$= \frac{P(A,B,H)}{P(B,H)} \cdot \frac{P(B,H)}{P(H)} = \frac{P(A,B,H)}{P(H)} = P(A,B \mid H)$$

- Sum rule

$$P(A \mid H) = \sum_{b} P(A, B = b \mid H) = \sum_{b} P(A \mid B = b, H) P(B = b \mid H)$$

Total Probability

- Let $B_1, B_2 \ldots B_N$ be a partition of S (mutually exclusive in S)
- Any event A can be represented as

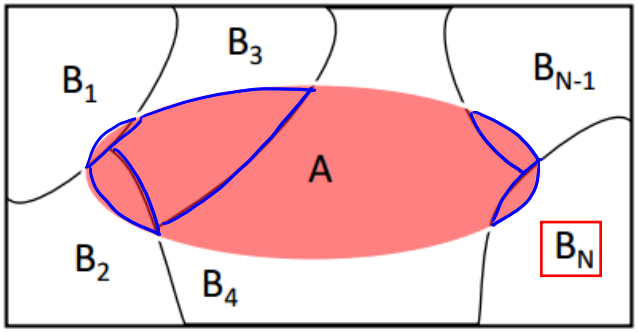
$$A = A \cap S = A \cap (B_1 \cup B_2 \ldots B_N) = (A \cap B_1) \cup (A \cap B_2) \ldots (A \cap B_N)$$

  - Since $B_1, B_2 \ldots B_N$ are mutually exclusive, then

$$P[A] = P[A \cap B_1] + P[A \cap B_2] + \cdots + P[A \cap B_N]$$

  - Therefore,

$$P[A] = P[A|B_1]P[B_1] + \cdots P[A|B_N]P[B_N] = \sum_{k=1}^{N} P[A|B_k]P[B_k]$$



Independence and Conditional Independence

강한 가정 " 독립이다! "

- We say *A* and *B* are unconditionally independent or marginally independent, denoted $A \perp B$, if we can represent the joint as the product of the two marginals

$$A \perp B \overset{\text{iff}}{\Longleftrightarrow} P(A, B) = p(A)p(B)$$

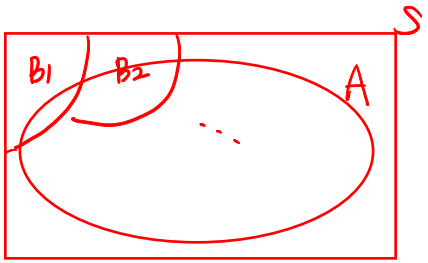
약한 가정

- *A* and *B* are conditionally independent (CI) given *C* iff the conditional joint can be written as a product of conditional marginals:

$$A \perp B \mid C \overset{\text{iff}}{\Longleftrightarrow} p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

Machine Learning 에서는 가장 중요함

Bayes Theorem



- Assume $B_1, B_2 \ldots B_N$ is a partition of S (mutually exclusive)
- Suppose that event A occurs, then what is the probability of B_j?
- Using the definition of conditional probability and the Theorem of total probability we obtain

$$P[B_j|A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A|B_j]P[B_j]}{\sum_{k=1}^{N} P[A|B_k]P[B_k]}$$

전확률공식

known as Bayes Theorem or Bayes Rule, and is (one of) the most useful relations in probability and statistics.

Bayes Theorem and Statistical Pattern Recognition

ex) $P(B_j|A) = \begin{pmatrix} 0.1 \\ 0.7 \\ 0.2 \end{pmatrix}$ 사과 / 바나나 / 배 — B_j에 대한 확률 : 바나나구나!

- When used for pattern classification, BT is generally expressed as

$$P[B_j|A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A|B_j]P[B_j]}{\sum_{k=1}^{N} P[A|B_k]P[B_k]}$$

mutually exclusive!

where B_j is the j-th class and A is the feature/observation/evidence vector.

ex) 과일 : 바나나 / ex) 노란색

A typical decision rule is to choose class B_j with **highest** P [B_j |A]

– Each term in the Bayes Theorem has a special name

A발생!!

- P [B_j] : prior probability (of class B_j) 아무것도 모를때의 Assumption, hypothesis — 일종의 선험적 = net prior
- P [B_j |A] : posterior probability (of class B_j given the observation A) : 더 정확해짐
- P [A|B_j] : likelihood (probability of observation A given class B_j)
- P [A]: normalization constant (does not affect the decision)

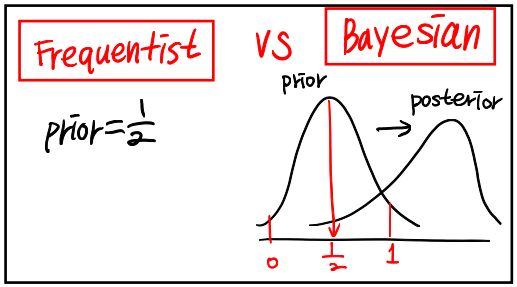
Maximum A Posteriori

ex) 앞면 ⑩ $P(B)=\frac{1}{2}$ (Prior : Assumption)

관측 { Ⓐ Ⓐ ⋮ → Posterior $P(⑩|A)=\frac{1}{1.5}$

관측 : 앞면이 많이 나왔을 확률이 더 높아짐

(어떤 관측이 나왔을 때 posterior dist.)



- Based on Bayes rule, we can compute the maximum a posteriori hypothesis for the data:

$$B_{MAP} = \arg\max_{B_j \in B} P(B_j \mid A)$$

$$= \operatorname{argmax}_{B_j \in B} \frac{P(A \mid B_j)P(B_j)}{P(A)}$$

$$= \operatorname{argmax}_{B_j \in B} P(A \mid B_j)P(B_j)$$

prior × maximum Likelihood = maximum posterior

P(A)는 normalization 상수 — P(B)와 관계 X

A에 관한 Prob

ex) 과일이 바나나 일 때, 노란색일 확률!

→ if) prior : uniform 인 경우, $= \operatorname{argmax}_{B_j}$ [Likelihood] 문제와 같아진다.

- Note when the prior is uniform, this shrinks to a maximum likelihood hypothesis

즉, likelihood !!

Mean, Variance, Moment, and Covariance

pdf

$$\text{Mean } E[x] = \int_{-\infty}^{\infty} x f(x) dx$$

$$\text{Conditional Mean } E[x \mid M] = \int_{-\infty}^{\infty} x f(x \mid M) dx$$

$$\text{Variance } \sigma^2 = \int_{-\infty}^{\infty} (x - E[x])^2 f(x) dx$$

$$\text{Moment } m_n = E[x^n] = \int_{-\infty}^{\infty} x^n f(x) dx$$

$$\text{Covariance } C_{xy} = E[(x - E[x])(y - E[y])]$$

$$= E[xy] - E[x]E[y]$$

Rxy

Random Vector and Covariance

- A random vector $x \in R^n$ is a collection of *n* random variables. The probability density function of the random vector *x* is defined by the joint density function,

$$P(x) = P(x_1, \ldots, x_n)$$

- A complete set of second moments is given by the correlation matrix

Autocorrelation

$$\mathbf{R_x} = E\{\mathbf{x}\mathbf{x}^H\} = E\{\mathbf{x}\mathbf{x}^{*T}\} = \begin{bmatrix} E\{|x_1|^2\} & E\{x_1x_2^*\} & \cdots & E\{x_1x_N^*\} \\ E\{x_2x_1^*\} & E\{|x_2|^2\} & \cdots & E\{x_2x_N^*\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_Nx_1^*\} & E\{x_Nx_2^*\} & \cdots & E\{|x_N|^2\} \end{bmatrix}$$

Rxy

Autocovariance

$$\begin{aligned} \mathbf{C_x} &= E\{(\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^H\} \\ &= E\{\mathbf{x}\mathbf{x}^H\} - \mathbf{m_x}E\{x^H\} - E\{\mathbf{x}\}\mathbf{m_x}^H + \mathbf{m_x}\mathbf{m_x}^H \\ &= \mathbf{R_x} - \mathbf{m_x}\mathbf{m_x}^H \end{aligned}$$

Various Distributions

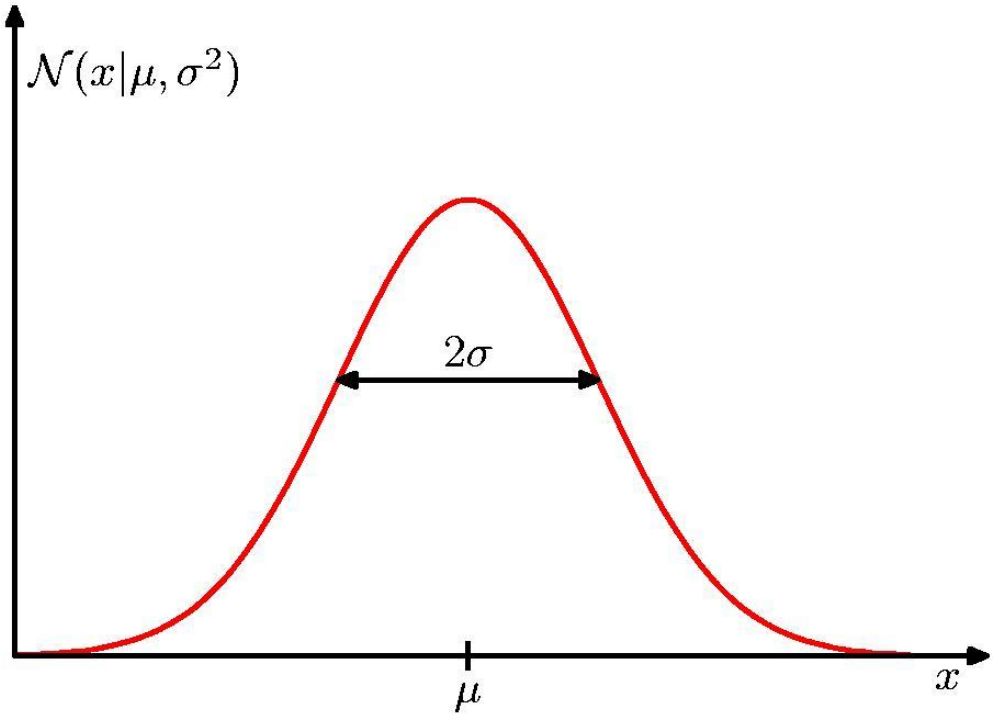
Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x / x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

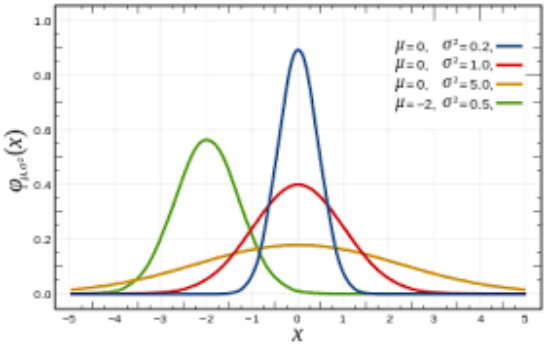
The (Univariate) Gaussian Distribution

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

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\, \mathrm{d}x = 1$$





The Multivariate Gaussian Distribution

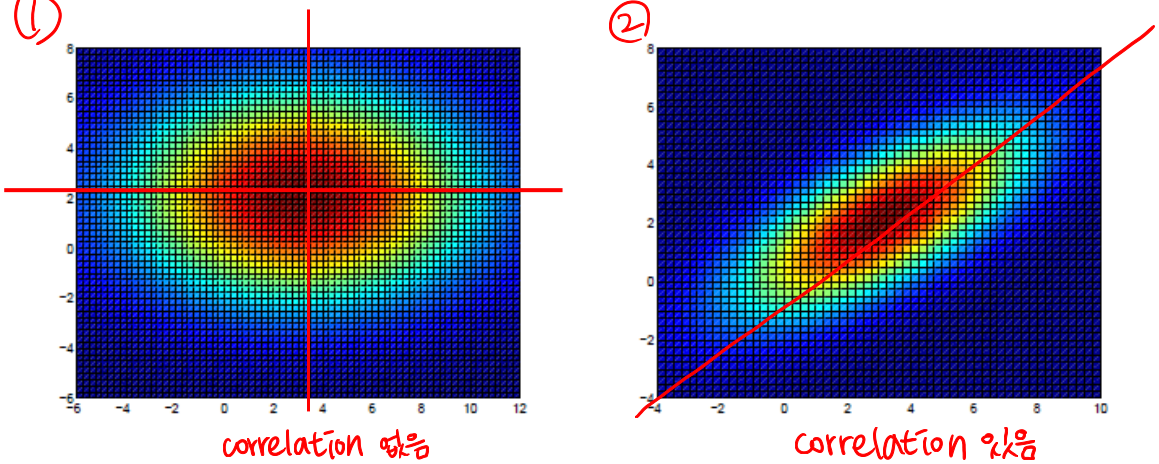
random vector — 각 element가 R.V

- A vector-valued random variable X ={X1 · · · Xn} is said to have a multivariate normal (or Gaussian) distribution with mean μ ∈ R^n and covariance matrix Σ ∈ $R^{n \times n}$ if its probability density function is given by

vector

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

① ②



correlation 없음 / correlation 있음

$$\text{mean } \mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

① $\Sigma = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$ diagonal : correlation = 0

② $\Sigma = \begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}$ correlation 있으면 ≠0

The diagonal covariance matrix case

- Assuming, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ correlation = 0

- The multivariate Gaussian density has a form:

$$p(x;\mu,\Sigma) = \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1-\mu_1) \\ \frac{1}{\sigma_2^2}(x_2-\mu_2) \end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right).$$

indep !!!

univariate Gaussian의 곱과 같다는 뜻

(The product of two independent Gaussian densities.)

Why Gaussian Distribution

- Completely characterized by mean and covariance.
- Central limit theorem. (Suppose $\{X_1, \ldots, X_n\}$ are independent and identically distributed random variables, each with mean μ and finite variance σ^2. The sum $X_1 + \ldots + X_n$ has mean $n\mu$ and variance $n\sigma^2$.

$$\bar{X}_n = S_n = \frac{X_1 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

- The distribution is again normal after a nonsingular linear transform.
- Affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b$, $Y \sim N(a\mu+b, a^2\sigma^2)$
- Marginal density is also normal and conditional density is also normal.
- There exists a linear transform which diagonalizes the covariance matrix (whitening, data sphering).