4.4 행렬의 미분

지금까지는 스칼라 값을 입력으로 받아 스칼라값을 출력하는 함수를 생각했
이제부터는 벡터나 행렬을 입력으로 받아서 벡터나 행렬을 출력하는 함수를 살

"여러개의 입력을 갖는 다변수 함수" = "독립변수가 벡터인 함수"

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = f(x) = f(x_1, x_2)$$

vector $x \longrightarrow$ scalar f

이를 확장하면 행렬을 입력으로 가지는 함수도 생각할 수 있다.

$$f\left(\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}\right) = f(X) = f(x_{11}, \cdots x_{22})$$

matrix $X \longrightarrow$ scalar f

반대로 벡터나 행렬을 출력하는 함수는 여러개의 함수를 합쳐놓은 것이다.

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}$$

scalar $x \longrightarrow$ vector f

$$f(x) = \begin{bmatrix} f_{11}(x) & f_{12}(x) \\ f_{21}(x) & f_{22}(x) \end{bmatrix}$$

scalar $x \longrightarrow$ matrix f

벡터나 행렬을 입력받아 벡터나 행렬을 출력할 수는 있다.

$$f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}$$

vector $x \longrightarrow$ vector f

$$f(x) = \begin{bmatrix} f_{11}(x_1, x_2) & f_{12}(x_1, x_2) \\ f_{21}(x_1, x_2) & f_{22}(x_1, x_2) \end{bmatrix}$$

~~Matrix x~~

vector $x \longrightarrow$ matrix f

이러한 행렬을 입력이나 출력으로 가지는 함수를 미분 → 행렬 미분

- 엄밀히 말해서 partial differentiation, 편미분
- 행렬은 분자중심 표현법 (Numerator-layout notation)과
  분모중심 표현법 (Denominator-layout notation)
  두 가지가 있는데, 여기서는 분모중심 표현법으로 기술.

[스칼라를 벡터로 미분하는 경우.]

데이터분석에서는 함수의 출력변수가 스칼라이고 입력변수 ~~의~~ x가 벡터인 다변수 함수를 이용하는 경우가 많다. 따라서 편미분도 $\frac{\partial f}{\partial x_1}$, $\frac{\partial f}{\partial x_2}$, … 등 여러 개가 존재한다.

* 스칼라를 벡터로 미분하는 경우, 결과를 열벡터로 표시한다.
이렇게 만들어진 벡터를 그래디언트 벡터 (gradient vector) 라고 하며, ∇f로 표기한다.

$$\nabla f = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix}$$

예제

$$f(x,y) = 2x^2 + 6xy + 7y^2 - 26x - 54y + 107$$

에 대한 그래디언트 벡터를 구하면

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 4x + 6y - 26 \\ 6x + 14y - 54 \end{bmatrix}$$

연습문제 4.4.1

다음 함수의 그레디언트 벡터를 구하라.

(1) $f(x,y,z) = x+y+z$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(2) $f(x,y,z) = xyz$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} yz \\ xz \\ xy \end{bmatrix}$$

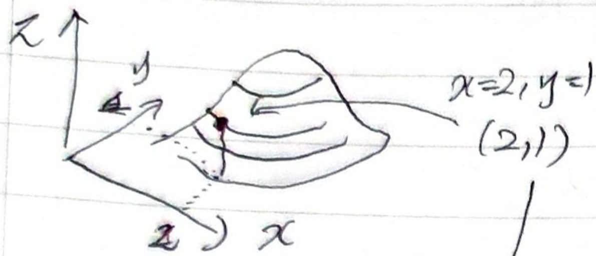
〈2차원에서의 그레디언트 벡터 표시〉 Quiver Plot

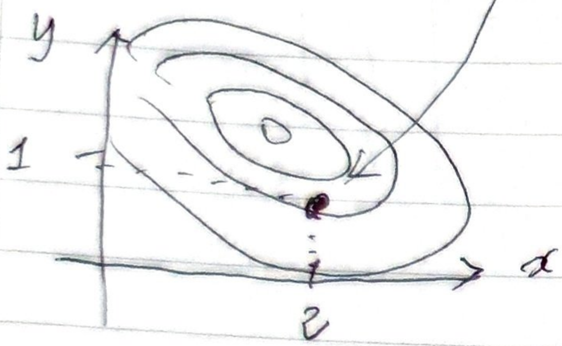
= Gradient vector on contour plot.

- 2개의 입력변수를 가지는 2차원 함수 $f(x,y)$
  - 평면상 컨투어(contour) 플롯으로 나타낼 수 있음.
- 입력변수 x, y 위치에서의 그레디언트 벡터 $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$는 그 위치를 원점으로 하는 화살표로 표현할 수 있다.
  - 그레디언트 벡터의 방향은 편미분 성분 $\frac{\partial f}{\partial x}$와 $\frac{\partial f}{\partial y}$의 부호에 의해 결정

* 어떤 위치 x, y에서 x 증가할수록 f 커지면 $\frac{\partial f}{\partial x} > 0$, 반대의 경우라면 $\frac{\partial f}{\partial x} < 0$. 벡터는 2차원 평면에서 화살표로 나타낼 수 있으며, 가로 성분이 양수이고 세로 성분이 음수인 화살표는 우측 아래를 가리킬 것.

$z = f(x, y)$





Gradient vector — arrow

어떤 지점을 찍어주면

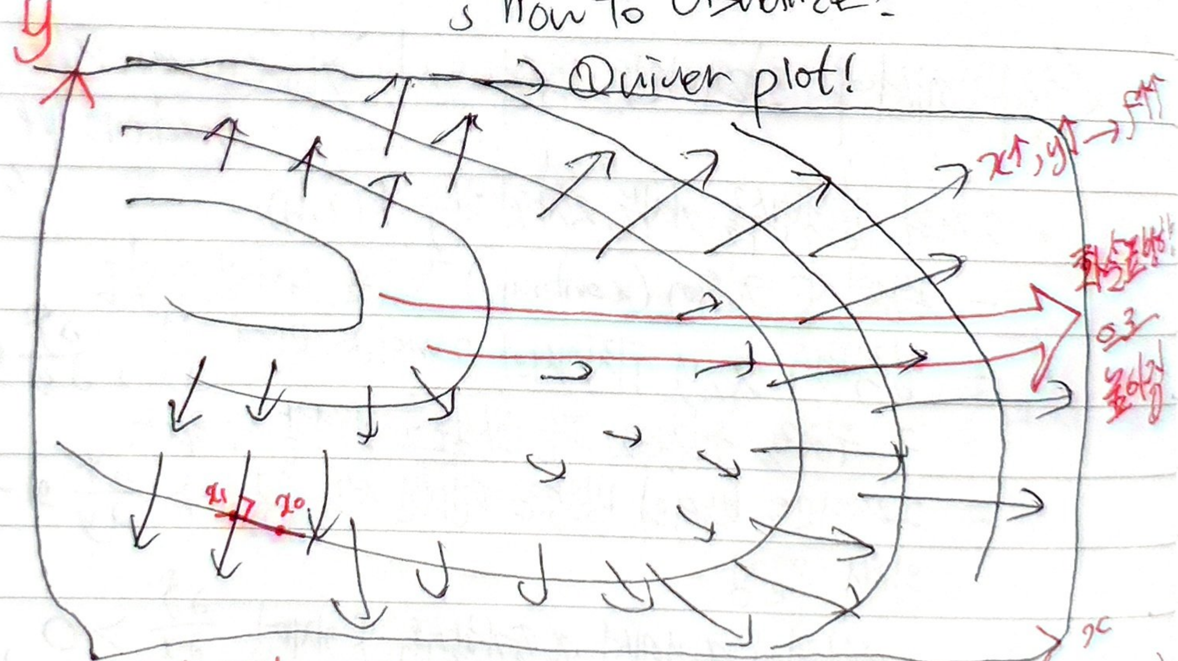
거기서부터 출발하는 arrow vector

; how to visualize?

→ Quiver plot!

화살표의 의미

1 step 이동할 때 가장 경사가 가파른 방향!!



방향의 의미; x 증가 → gradient rightward (x↑ → f↑)

y 증가 → gradient upward (y↑ → f↑)

{케라스에서 그래디언트 벡터가 갖는 특징}

1. 그래디언트 벡터 크기는 기울기를 의미한다.
   — 즉, 벡터의 크기가 클수록 함수 곡면의 기울기가 커진다.
2. 그래디언트 벡터의 방향은 함수곡면의 기울기가 가장 큰 방향,
   즉 단위 길이당 함수값 (높이)이 가장 크게 증가하는 방향을 가리킨다.
3. 그래디언트 벡터의 방향은 등고선 (iso line) 방향과 직교한다.

어떤 점 x_0에서 다른 점 x로 이동하면서 함수값이 얼마나 변하는지는 테일러 근사를 써서 근사할 수 있다.

$$\underbrace{f(x) - f(x_0)}_{x \to x_0 \text{ 이동하여}} = \underbrace{\Delta f}_{\text{변한 함수값}} \approx \underbrace{\nabla f(x_0)^T (x - x_0)}_{\text{taylor expansion approx.}}$$

변화의 방향 $x - x_0$가 그래디언트 벡터와 같은 방향일 때, Δf가 가장 커지는 것을 알 수 있다.

등고선은 $f(x)$의 값이 일정한 x의 집합이므로 다음과 같은 방정식으로 표현할 수 있다.

$$f(x) = f(x_0)$$

or

$$f(x) - f(x_0) = 0$$

같은 등고선 위의 다른점 x_1을 향해 움직이는 등고선 방향의 움직임은 $x_1 - x_0$이고, x_0, x_1 모두 같은 등고선 위의 점이므로 $f(x_0) = f(x_1)$이다. 따라서 테일러 전개에서 $\nabla f(x_0)^T (x_1 - x_0) = f(x_1) - f(x_0) = 0$

등고선 방향 $(x_1 - x_0)$과 $\nabla f(x_0)$이 직교함을 알 수 있다.

행렬미분법칙

수많은 matrix calculus 법칙이 존재하지만,
중요 수준을 커버하는 것은 5개 법칙으로 대부분 커버 가능

행렬미분법칙 1: 선형모형

선형모형을 미분하면 그레디언트 벡터는 가중치 벡터이다.

$$f(x) = w^T x$$

$$\nabla f = \frac{\partial w^T x}{\partial x} = \frac{\partial x^T w}{\partial x} = w$$

(증명)

$$\frac{\partial (w^T x)}{\partial x} = \begin{bmatrix} \frac{\partial (w^T x)}{\partial x_1} \\ \frac{\partial (w^T x)}{\partial x_2} \\ \vdots \\ \frac{\partial (w^T x)}{\partial x_N} \end{bmatrix} = \begin{bmatrix} \frac{\partial (w_1 x_1 + \cancel{w_2 x_2} + \cdots \cancel{w_N x_N})}{\partial x_1} \\ \frac{\partial (\cancel{w_1 x_1} + w_2 x_2 + \cdots \cancel{w_N x_N})}{\partial x_2} \\ \vdots \\ \frac{\partial (\cancel{w_1 x_1} + \cancel{w_2 x_2} + \cdots + w_N x_N)}{\partial x_N} \end{bmatrix}$$

$$= \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = w$$

행렬미분법칙 2 : 이차형식

이차 형식을 미분하면 행렬과 벡터의 곱으로 ~~나타난~~ 나타난다.

$$f(x) = x^T A x$$

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

(증명)

$$\frac{\partial(x^T A x)}{\partial x} = \begin{bmatrix} \frac{\partial(x^T A x)}{\partial x_1} \\ \frac{\partial(x^T A x)}{\partial x_2} \\ \vdots \\ \frac{\partial(x^T A x)}{\partial x_N} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial\left(\sum_{i=1}^N \sum_{j=1}^N a_{ij} x_i x_j\right)}{\partial x_1} \\ \frac{\partial\left(\sum_{i=1}^N \sum_{j=1}^N a_{ij} x_i x_j\right)}{\partial x_2} \\ \vdots \\ \frac{\partial\left(\sum_{i=1}^N \sum_{j=1}^N a_{ij} x_i x_j\right)}{\partial x_N} \end{bmatrix}$$

1열, 1행 제외 나머지 = 0.

2열, 2행 제외 나머지 = 0.

$$= \begin{bmatrix} \frac{\partial \begin{pmatrix} a_{11}x_1x_1 + a_{12}x_1x_2 \cdots a_{1N}x_1x_N \\ a_{21}x_2x_1 + \cancel{a_{22}x_2x_2} \cdots \cancel{a_{2N}x_2x_N} \\ \vdots \\ a_{N1}x_Nx_1 + \cancel{a_{N2}x_Nx_2} \cdots \cancel{a_{NN}x_Nx_N} \end{pmatrix}}{\partial x_1} \\ \frac{\partial \begin{pmatrix} a_{11}x_1x_1 + a_{12}x_1x_2 \cdots a_{1N}x_1x_N \\ a_{21}x_2x_1 + a_{22}x_2x_2 \cdots a_{2N}x_2x_N \\ \vdots \\ a_{N1}x_Nx_1 + a_{N2}x_Nx_2 \cdots a_{NN}x_Nx_N \end{pmatrix}}{\partial x_2} \\ \vdots \end{bmatrix}$$

하나씩 떼서 정리.

$$= \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + \cdots a_{1N}x_N \\ a_{21}x_2 + 0 + \cdots 0 \\ \vdots \\ a_{N1}x_N + 0 + \cdots 0 \\ 0 + a_{12}x_2 + \cdots + 0 + \\ a_{21}x_1 + 2a_{22}x_2 + \cdots a_{2N}x_N + 0 \\ \vdots \\ 0 + a_{N2}x_N + \cdots + 0 \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \left(\sum_{i=1}^{N} a_{1i}x_i + \sum_{i=1}^{N} a_{i1}x_i\right) \\ \left(\sum_{i=1}^{N} a_{2i}x_i + \sum_{i=1}^{N} a_{i2}x_i\right) \\ \vdots \\ \left(\sum_{i=1}^{N} a_{Ni}x_i + \sum_{i=1}^{N} a_{iN}x_i\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{N} a_{1i}x_i \\ \sum_{i=1}^{N} a_{2i}x_i \\ \vdots \\ \sum_{i=1}^{N} a_{Ni}x_i \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{N} a_{i1}x_i \\ \sum_{i=1}^{N} a_{i2}x_i \\ \vdots \\ \sum_{i=1}^{N} a_{iN}x_i \end{bmatrix}$$

$$= Ax + A^T x = (A + A^T)x$$

*대칭행렬 A라면 $(A+A^T)x = 2Ax$

스칼라 미분 for x	벡터/행렬 미분 for x
$ax \rightarrow a$	$w^T x \rightarrow w$
$ax^2 \rightarrow 2ax$	$x^T A x \rightarrow (A+A^T)x$

* 벡터를 스칼라로 미분하는 경우

벡터 $f(x) = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{bmatrix}$

을 스칼라 x로 미분하는 경우, 결과를 행 벡터로 표시한다.

(gradient vector와 혼동하지 않도록!)

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f_1}{\partial x} \; \frac{\partial f_2}{\partial x} \; \cdots \; \frac{\partial f_M}{\partial x} \right]$$

"<u>벡터를 벡터로 미분하는 경우</u>"

미분을 당하는 벡터의 원소가 여러개 ($i=1,2,\cdots N$)이고,
미분을 하는 벡터 원소도 여러개 ($j=1,2,\cdots M$)이므로,
미분 결과로 나온 도함수는 $N \times M$ 행렬이 된다.

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f_1}{\partial x} \; \frac{\partial f_2}{\partial x} \; \cdots \; \frac{\partial f_N}{\partial x} \right] = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_N} & \frac{\partial f_2}{\partial x_N} & \cdots & \frac{\partial f_N}{\partial x_N} \end{bmatrix}$$

행렬미분법칙 3 : 행렬과 벡터의 곱의 미분

행렬 A와 벡터 x의 곱 Ax를 벡터 x로 미분하면 행렬 A^T가 된다.

$$f(x) = Ax$$

$$\nabla f(x) = \frac{\partial (Ax)}{\partial x} = A^T$$

증명)

$$Ax = c_1x_1 + c_2x_2 + \cdots + c_Mx_M$$

$$\frac{\partial (Ax)}{\partial x} = \begin{bmatrix} \frac{\partial (Ax)}{\partial x_1} \\ \frac{\partial (Ax)}{\partial x_2} \\ \vdots \\ \frac{\partial (Ax)}{\partial x_M} \end{bmatrix} = \begin{bmatrix} \frac{\partial (c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_1} \\ \frac{\partial (c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_2} \\ \vdots \\ \frac{\partial (c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_M} \end{bmatrix}$$

$$= \begin{bmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_M^T \end{bmatrix} = A^T$$

함수의 출력변수과 입력변수가 모두 벡터(다차원) 데이터인 경우에는 입력변수 각각과 출력변수 각각의 조합에 대해 모두 미분이 존재한다. 따라서 도함수는 행렬 형태가 된다. 이렇게 만들어진 도함수의 행렬을 자코비안 행렬(Jacobian matrix)이라고 한다. 자코비안 행렬은 벡터함수를 벡터함수로 미분해서 생기는 행렬의 전치행렬이다. 따라서 행/열의 방향이 다르다는 점에 유의한다.

$$Jf(x) = J = \left(\frac{\partial f}{\partial x}\right)^T = \begin{bmatrix} \left(\frac{\partial f_1}{\partial x}\right)^T \\ \vdots \\ \left(\frac{\partial f_M}{\partial x}\right)^T \end{bmatrix} = \begin{bmatrix} \nabla f_1^T \\ \vdots \\ \nabla f_M^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix}$$

$$J \approx \nabla$$

자코비안을 구해라 ≈ 그레디언트 구해라!

■ 다변수 함수의 2차 도함수는 그레디언트 벡터를 입력변수 벡터로 미분한 것으로, 헤시안 행렬(Hessian matrix)이라고 한다.

헤시안 행렬은 그레디언트 벡터의 자코비안 행렬의 전치행렬로 정의한다.

$$Hf(x) = H = J(\nabla f(x))^T$$

벡터 x → 스칼라 f

↓

∇f 그레디언트 (벡터)

↓

$\nabla^2 f$, $\nabla\nabla f$ 헤시안 (행렬)

$$\boxed{H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}}$$

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \frac{\partial^2 f}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix}$$

" 함수가 연속이고 미분 가능한 함수라면,
헤시안 행렬은 대칭행렬이 된다."

스칼라를 행렬로 미분

출력변수 f가 스칼라값이고 입력변수 X가 행렬인 경우에는
도함수 행렬의 모양이 입력 변수 X와 같다.

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1N}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{M1}} & \frac{\partial f}{\partial x_{M2}} & \cdots & \frac{\partial f}{\partial x_{MN}} \end{bmatrix}$$

행렬미분법칙 4: 행렬곱의 대각성분

두 정방행렬을 곱해서 만들어진 행렬의 대각성분 (trace)는 스칼라이다.
이 스칼라를 뒤의 행렬로 미분하면 앞 행렬의 전치행렬이 나온다.

$$f(X) = tr(WX)$$

$$W \in R^{N \times N}, X \in R^{N \times N}$$

$$\frac{\partial f}{\partial X} = \frac{\partial tr(WX)}{\partial X} = W^T$$

(증명)

$$tr(WX) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ji} x_{ij}$$

$$\frac{\partial tr(WX)}{\boxed{\partial x_{ij}}} = w_{ji}$$

// 대응하는 항목은 $w_{ji} x_{ij}$ 단 하나밖에 없음 //

$$\frac{\partial tr(WX)}{\partial x_{ij}} = \frac{\partial \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ji} x_{ij}}{\partial x_{ij}} = \frac{\partial (w_{ji} x_{ij} + \cdots)}{\partial x_{ij}}$$

$$= w_{ji}$$

행렬 미분법칙 5 : 행렬식의 로그

행렬식 (determinant)은 스칼라값이고, 이 값의 로그 값도 스칼라이다. 이 값을 원래의 행렬로 미분하면 원래 행렬의 역행렬의 전치 행렬이 된다.

$$f(X) = \log |X|$$

$$\frac{\partial f}{\partial X} = \frac{\partial \log |X|}{\partial X} = (X^{-1})^T$$

(증명)

행렬식의 정의에서

$$\frac{\partial}{\partial x_{i,j}} |X| = C_{i,j}$$

행렬식과 역행렬의 관계에서

$$\frac{\partial}{\partial X} |X| = C = |X| (X^{-1})^T$$

로그 함수 공식에 대입하면

$$\frac{d}{dx} \log f(x) = \frac{f'(x)}{f(x)} = \frac{|X| (X^{-1})^T}{|X|} = (X^{-1})^T$$

정리

$w^T x \Rightarrow (w^T)^T = w$

$x^T A x \Rightarrow (A + A^T) x$

$Ax \Rightarrow A^T$

$\mathrm{tr}(Ax) \Rightarrow A^T$

$\log |X| \Rightarrow (X^{-1})^T$

확률론, 최적화를 위한 미적분

done!

functional calculus // calculus of variations

변분법 : input function이 변할 때, functional F 출력이 어떻게 달라지나

데이터 분석에서는 함수와 더불어 다양한 범함수가 사용된다.
이 절에서는 범함수의 개념과 범함수의 미분에 해당하는 변분법을 공부한다.

* 범함수 (functional)
; 함수를 입력받아서 실수를 출력

ex 기댓값 계산, 엔트로피 계산

$$\text{function } y(x) \rightarrow \boxed{\text{Functional } F} \rightarrow \text{real number}$$

표기 알파벳 대문자,
입력변수를 대괄호 (square bracket)

$$F[y(x)]$$

● 범함수 값의 계산 — 정적분

ex. 확률변수 X의 기댓값, 엔트로피 (= 확률밀도함수의 정적분 값)

$$E[p(x)] = \int_{-\infty}^{\infty} xp(x)dx$$

$$H[p(x)] = \int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$