4.4 행렬의 미분

지금까지는 스칼라값을 입력으로 받아 스칼라값을 출력하는 함수를 생각했다. 이제부터는 벡터나 행렬을 입력으로 받아서 벡터나 행렬을 출력하는 함수를 살펴본다.

여러개의 입력을 가지는 다변수 함수는 함수의 독립변수가 벡터인 경우로 볼 수 있다.

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

$$\text{vector } x \quad \rightarrow \quad \text{scalar } f$$

이를 확장하면 행렬을 입력으로 가지는 함수도 생각할 수 있다.

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

$$\text{matrix } x \quad \rightarrow \quad \text{scalar } f$$

반대로 벡터나 행렬을 출력하는 함수는 여러개의 함수를 합쳐놓은 것이다.

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

$$\text{scalar } x \quad \rightarrow \quad \text{vector } f$$

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

$$\text{scalar } x \quad \rightarrow \quad \text{matrix } f$$

벡터나 행렬을 입력받아 벡터나 행렬을 출력할 수도 있다.

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

$$\text{vector } x \quad \rightarrow \quad \text{vector } f$$

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

$$\text{vector } x \quad \rightarrow \quad \text{matrix } f$$

이러한 행렬을 입력이나 출력으로 가지는 함수를 미분하는 것을 **행렬미분(matrix differentiation)**이라고 한다. 사실 행렬미분은 정확하게는 미분이 아닌 편미분(partial derivative)이지만 여기에서는 편의상 미분이라고 쓰겠다. 또한 행렬미분에는 분자중심 표현법(Numerator-layout notation)과 분모중심 표현법(Denominator-layout notation) 두 가지가 있는데 여기에서는 분모중심 표현법으로 서술한다.

스칼라를 벡터로 미분하는 경우

데이터 분석에서는 함수의 출력변수가 스칼라이고 입력변수 x가 벡터인 다변수 함수를 사용하는 경우가 많다. 따라서 편미분도 $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots$ 등으로 여러 개가 존재한다.

이렇게 스칼라를 벡터로 미분하는 경우에는 결과를 열벡터로 표시한다. 이렇게 만들어진 벡터를 **그레디언트 벡터(gradient vector)**라고 하고 $\triangledown f$로 표기한다.

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

예제

다음과 같은 다변수 함수

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

에 대한 그레디언트 벡터를 구하면

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

가 된다.

연습 문제 4.4.1

다음 함수의 그레디언트 벡터를 구하라

(1)

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

(2)

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

연습 문제 4.4.2

그레디언트 벡터

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

에 대해서 x, y가 다음 위치일 때 그레디언트 벡터의 값을 구하고 평면상에 화살표로 나타내라.

(1) $x = 7, y = 1$

(2) $x = 2, y = 1$

2차원의 경우를 예로 들어 그레디언트 벡터를 표시하는 법을 알아보자. 2개의 입력변수를 가지는 2차원 함수 $f(x, y)$는 평면상에서 **컨투어(contour)플롯**으로 나타낼 수 있다. 그리고 입력 변수 x, y 위치에서의 그레디언트 벡터 $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$는 그 위치를 원점으로 하는 화살표로 표현할 수 있다. 그리고 그레디언트 벡터의 방향은 편미분 성분 $\frac{\partial f}{\partial x}$와 $\frac{\partial f}{\partial y}$의 부호에 의해 결정된다.

만약 어떤 위치 x, y에서 x가 증가할수록 f가 커지면 도함수 $\frac{\partial f}{\partial x}$은 양수이다. 반대로 y가 증가할수록 f가 작아지면 도함수 $\frac{\partial f}{\partial x}$은 음수이다. 벡터는 2차원 평면에서 화살표로 나타낼 수 있다. 가로 성분이 양수이고 세로 성분이 음수인 화살표는 우측 아래를 가리키는 화살이 될 것이다.

이렇게 컨투어 플롯 위에 그레디언트 벡터를 화살표로 나타낸 것을 플롯을 **퀴버(quiver)플롯**이라고 한다. 퀴버 플롯에서 화살표는 화살표 시작 지점의 그레디언트 벡터를 나타낸다.

예제

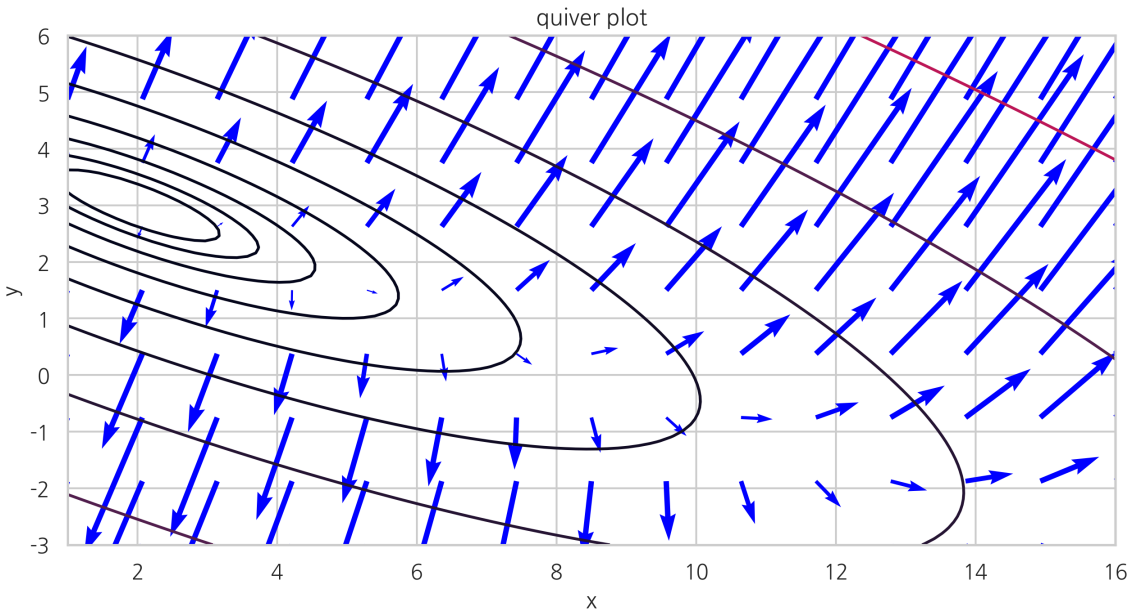
다음은 함수

$$2x^2 + 6xy + 7y^2 - 26x - 54y + 107$$

의 그레디언트 벡터를 표시한 퀴버플롯이다.

In [1]:

```
def f(x, y):
    return 2 * x**2 + 6 * x * y + 7 * y**2 - 26 * x - 54 * y + 107

xx = np.linspace(1, 16, 100)
yy = np.linspace(-3, 6, 90)
X, Y = np.meshgrid(xx, yy)
Z = f(X, Y)


def gx(x, y):
    return 4 * x + 6 * y - 26

def gy(x, y):
    return 6 * x + 14 * y - 54

xx2 = np.linspace(1, 16, 15)
yy2 = np.linspace(-3, 6, 9)
X2, Y2 = np.meshgrid(xx2, yy2)
GX = gx(X2, Y2)
GY = gy(X2, Y2)
plt.figure(figsize=(10, 5))
plt.contour(X, Y, Z, levels=np.logspace(0, 3, 10))
plt.quiver(X2, Y2, GX, GY, color='blue', scale=400, minshaft=2)

plt.xlabel('x')
plt.ylabel('y')
plt.title("퀴버 플롯(quiver plot)")
plt.show()
```



연습 문제 4.4.3

함수

$$2x^2 + 6xy + 7y^2 - 26x - 54y + 107$$

로 표현되는 지형을 상상하라. 이 지형의 (14, 4) 지점에 공을 두었다면 어떠한 경로로 공이 움직일지 경로를 그려라.

퀴버플롯에서 그레디언트 벡터는 다음과 같은 특징이 있다.

1. 그레디언트 벡터의 크기는 기울기를 의미한다. 즉 벡터의 크기가 클수록 함수 곡면의 기울기가 커진다.
2. 그레디언트 벡터의 방향은 함수 곡면의 기울기가 가장 큰 방향, 즉 단위 길이당 함수값(높이)이 가장 크게 증가하는 방향을 가리킨다.
3. 그레디언트 벡터의 방향은 등고선(isoline)의 방향과 직교한다.

어떤 점 x_0에서 다른 점 x로 이동하면서 함수 값이 얼마나 변하는지는 테일러 전개를 써서 근사할 수 있다.

$$f(x) - f(x_0) = \Delta f \approx \nabla f(x_0)^T (x - x_0)$$

변화의 방향 $x - x_0$가 그레디언트 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.

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

또는

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

같은 등고선 위의 다른 점 x_1를 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 전개로부터

$$\nabla f(x_0)^T (x_1 - x_0) = f(x_1) - f(x_0) = 0$$

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

행렬미분법칙

다변수 함수를 미분하여 그레디언트 벡터를 구할 때는 다음 두가지 법칙이 유용하게 쓰인다.

행렬미분법칙 1: 선형 모형

선형 모형을 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)

$$\frac{\partial(w^Tx)}{\partial x} = \begin{bmatrix} \frac{\partial(w^Tx)}{\partial x_1} \\ \frac{\partial(w^Tx)}{\partial x_2} \\ \vdots \\ \frac{\partial(w^Tx)}{\partial x_N} \end{bmatrix} = \begin{bmatrix} \frac{\partial(w_1x_1 + \cancel{w_2x_2} + \cdots + \cancel{w_Nx_N})}{\partial x_1} \\ \frac{\partial(\cancel{w_1x_1} + w_2x_2 + \cdots + \cancel{w_Nx_N})}{\partial x_2} \\ \vdots \\ \frac{\partial(\cancel{w_1x_1} + \cancel{w_2x_2} + \cdots + w_Nx_N)}{\partial x_N} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = w$$

행렬미분법칙 2: 이차 형식

이차 형식을 미분하면 행렬과 벡터의 곱으로 나타난다.

$$f(x) = x^TAx$$

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

(증명)

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

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

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

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

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

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

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

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

위의 두 가지 경우는 1차 다항식과 2차 다항식에 대한 스칼라 미분과 비슷하다. 두 경우를 비교해 보면 이 공식을 외우는데 도움이 된다.

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

벡터를 스칼라로 미분하는 경우

벡터

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

를 스칼라 x로 미분하는 경우에는 결과를 행 벡터로 표시한다.

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_2}{\partial x} & \cdots & \frac{\partial f_M}{\partial x} \end{bmatrix}$$

벡터를 벡터로 미분하는 경우

벡터 x를 입력받아 벡터를 출력하는 함수 $f(x)$를 생각하자.

벡터를 벡터로 미분하면 미분을 당하는 벡터의 원소가 여러개($i = 1, \ldots, N$)이고 미분을 하는 벡터의 원소도 여러개($j = 1, \ldots, M$)이므로 미분의 결과로 나온 도함수는 2차원 배열 즉, 행렬이 된다.

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_2}{\partial x} & \cdots & \frac{\partial f_N}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_M} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_M} & \frac{\partial f_2}{\partial x_M} & \cdots & \frac{\partial f_N}{\partial x_M} \end{bmatrix}$$

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)

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

$$\frac{\partial(Ax)}{\partial x} = \begin{bmatrix} \frac{\partial(Ax)}{\partial x_1} \\ \frac{\partial(Ax)}{\partial x_2} \\ \cdots \\ \frac{\partial(Ax)}{\partial x_M} \end{bmatrix} = \begin{bmatrix} \frac{\partial(c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_1} \\ \frac{\partial(c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_2} \\ \cdots \\ \frac{\partial(c_1x_1 + c_2x_2 + \cdots + c_Mx_M)^T}{\partial x_M} \end{bmatrix} = \begin{bmatrix} {c_1}^T \\ {c_2}^T \\ \cdots \\ {c_M}^T \end{bmatrix} = A^T$$

함수의 출력변수와 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 벡터변수로 미분해서 생기는 행렬의 **전치행렬**이다. 따라서 행/열의 방향이 다르다는 점에 유의한다.

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

연습 문제 4.4.4

다음 함수의 자코비안 행렬을 구하라

$$f(x) = \begin{bmatrix} \sum_{i=1}^{3} x_i \\ \prod_{i=1}^{3} x_i \end{bmatrix}$$

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

헤시안 행렬은 그레디언트 벡터의 자코비안 행렬의 전치 행렬로 정의한다.

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

풀어쓰면 다음과 같다.

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

즉,

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

함수가 연속이고 미분가능한 함수라면 헤시안 행렬은 대칭행렬이 된다.

연습 문제 4.4.5

다음 함수의 헤시안 행렬을 구하라

$$f(x) = \sum_{i=1}^{3} x_i^2$$

스칼라를 행렬로 미분

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

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial x_{1,1}} & \frac{\partial f}{\partial x_{1,2}} & \cdots & \frac{\partial f}{\partial x_{1,N}} \\ \frac{\partial f}{\partial x_{2,1}} & \frac{\partial f}{\partial x_{2,2}} & \cdots & \frac{\partial f}{\partial x_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{M,1}} & \frac{\partial f}{\partial x_{M,2}} & \cdots & \frac{\partial f}{\partial x_{M,N}} \end{bmatrix}$$

행렬미분법칙 4: 행렬 곱의 대각성분

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

$$f(X) = \text{tr}(WX)$$

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

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

(증명)

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

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

행렬미분법칙 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

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