Matrix Calculus

Derivative Definition

The simplest form of multivariable differentiation, vector differentiation generalizes the one-dimensional concept of a derivative to functions with vector-valued inputs or outputs.

We develop the concept of the gradient by generalizing the limit definition of the (single-variable) derivative, which is

$$f(x') = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

to functions where the input is a vector.

In the multivariable case, what $t \to 0$ means is less clear, as there are many directions in which one could approach a point in $\mathbb{R}^n$.

Given a vector $\vec{d}$ with the same dimension as $\vec{x}$, we could consider the limit

$$\nabla f(\vec{x})[\vec{d}] := \lim_{t \to 0} \frac{f(\vec{x} + t\vec{d}) - f(\vec{x})}{t}$$

which may be thought of as a function of both $\vec{x}$ and $\vec{d}$.

If we want a definition for the multidimensional derivative $\frac{df}{d\vec{x}}$ at a given point $\vec{x}$, it should not depend on $\vec{d}$.

Examples

- For any n-dimensional vector $\vec{x}$,

$$\frac{d\vec{x}}{d\vec{x}} = \mathbf{I}_n$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Proof: By definition, for any n-dimensional vector $\vec{d}$,

$$\begin{aligned}\frac{d\vec{x}}{d\vec{x}}\vec{d} &= \lim_{t\to 0}\frac{(\vec{x}+t\vec{d})-(\vec{x})}{t}\\ &= \vec{d}\end{aligned}$$

We note that $\frac{d\vec{x}}{d\vec{x}} = \mathbf{I}_n$ satisfies the limit definition.

- For any n-dimensional vector $\vec{x}$ and $n \times n$ constant matrix $\mathbf{A}$,

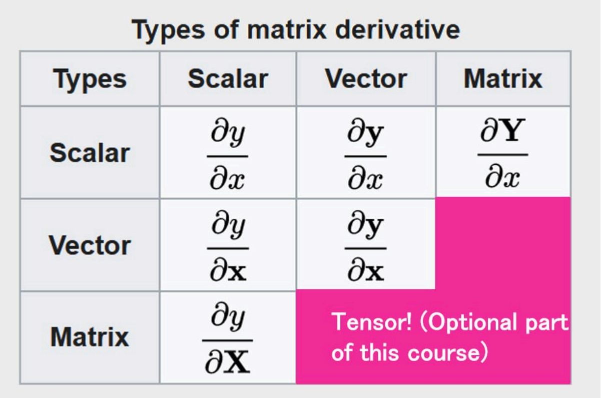
$$\frac{d\vec{x}^T\mathbf{A}\vec{x}}{d\vec{x}} = \vec{x}^T(\mathbf{A}+\mathbf{A}^T)$$

Proof: By definition, for any n-dimensional vector $\vec{d}$,

$$\begin{aligned}\frac{d\vec{x}^T\mathbf{A}\vec{x}}{d\vec{x}}\vec{d} &= \lim_{t\to 0}\frac{(\vec{x}+t\vec{d})^T\mathbf{A}(\vec{x}+t\vec{d})-\vec{x}^T\mathbf{A}\vec{x}}{t}\\ &= \lim_{t\to 0}\left(\vec{d}^T\mathbf{A}\vec{x}+\vec{x}^T\mathbf{A}\vec{d}+t\vec{d}^T\mathbf{A}\vec{d}\right)\\ &= \vec{d}^T\mathbf{A}\vec{x}+\vec{x}^T\mathbf{A}\vec{d}\\ &= \vec{x}^T\mathbf{A}^T\vec{d}+\vec{x}^T\mathbf{A}\vec{d}\\ &= \vec{x}^T(\mathbf{A}+\mathbf{A}^T)\vec{d}\end{aligned}$$

Notation

The notation that we will use may be different from other resources. For more information see this.



Types of matrix derivative

Types	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$	Tensor! (Optional part of this course)	

Vector-by-scalar

The derivative of a vector $y \in \mathbb{R}^m$, by a scalar x is written as:

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

Scalar-by-vector

The derivative of a scalar y, by a vector $x \in \mathbb{R}^n$ is written as:

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

Vector-by-vector

Each of the previous two cases can be considered as an application of the derivative of a vector with respect to a vector, using a vector of size one appropriately.

The derivative of a vector $y \in \mathbb{R}^m$, by a vector $x \in \mathbb{R}^n$ is written as:

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Matrix-by-scalar

The derivative of a matrix $Y \in \mathbb{R}^{m \times n}$ by a scalar x is given by:

$$\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial Y_{11}}{\partial x} & \frac{\partial Y_{12}}{\partial x} & \cdots & \frac{\partial Y_{1n}}{\partial x} \\ \frac{\partial Y_{21}}{\partial x} & \frac{\partial Y_{22}}{\partial x} & \cdots & \frac{\partial Y_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_{m1}}{\partial x} & \frac{\partial Y_{m2}}{\partial x} & \cdots & \frac{\partial Y_{mn}}{\partial x} \end{bmatrix}$$

Scalar-by-matrix

The derivative of a scalar y by a matrix $X \in \mathbb{R}^{m \times n}$ is given by:

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \frac{\partial y}{\partial X_{21}} & \cdots & \frac{\partial y}{\partial X_{m1}} \\ \frac{\partial y}{\partial X_{12}} & \frac{\partial y}{\partial X_{22}} & \cdots & \frac{\partial y}{\partial X_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial X_{1n}} & \frac{\partial y}{\partial X_{2n}} & \cdots & \frac{\partial y}{\partial X_{mn}} \end{bmatrix}$$

Gradient

For notational conventions that we use, the gradient of $f(x) : \mathbb{R}^n \to \mathbb{R}$ is the derivative (some resources use the transpose of the derivative).

$$\nabla f(x) = \frac{\partial f(x)}{\partial x}$$

Note that the size of $\nabla_x f(x)$ is always the same as the size of x, but transposed. So if $x \in \mathbb{R}^n$, then we have

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Hessian

Suppose that $f(x) : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the Hessian matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$$

Rules

We will present the product rule and chain rule based on our notational conventions.

Chain rule

We want to generalize the chain rule for single-valued functions, $\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x}$, to multi-valued functions.

1. Matrix-scalar and scalar-matrix

By the definition, we can easily conclude that for scalars x, u and matrix Y, we have:

$$\frac{\partial Y}{\partial x} = \frac{\partial Y}{\partial u} \frac{\partial u}{\partial x}$$

, and also for matrix X and scalars y, u:

$$\frac{\partial y}{\partial X} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial X}$$

2. Vector-vector: $\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial x}$

For vectors $x \in \mathbb{R}^n, y \in \mathbb{R}^m, u \in \mathbb{R}^l$ we have:

$$(\frac{\partial y}{\partial x})_{ij} = \frac{\partial y_i}{\partial x_j}$$

If y_i is a function of vector $u = \begin{bmatrix} u_1 \\ \vdots \\ u_l \end{bmatrix}$, we can write

$$\frac{\partial y_i}{\partial x_j} = \sum_{k=1}^{l} \frac{\partial y_i}{\partial u_k} \frac{\partial u_k}{\partial x_j} = \sum_{k=1}^{l} (\frac{\partial y}{\partial u})_{ik} (\frac{\partial u}{\partial x})_{kj} = \left(\frac{\partial y}{\partial u} \frac{\partial u}{\partial x}\right)_{ij}$$

So we can conclude that $\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x}$.

Product Rule

For single-valued functions $f(x), g(x) : \mathbb{R} \to \mathbb{R}$ we know that

$$\big(f(x) + g(x)\big)' = f'(x)g(x) + f(x)g'(x)$$

Let's generalize this to multi-valued functions.

1. Matrix-scalar: $\frac{\partial AB}{\partial \alpha}$

If $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times l}$ be matrices which elements are functions of scalar α.

$$\begin{aligned}
\frac{\partial (AB)_{ij}}{\partial \alpha} &= \sum_{k=1}^{n} \frac{\partial (A_{ik}B_{kj})}{\partial \alpha} \\
&= \sum_{k=1}^{n} \left[A_{ik}\frac{\partial B_{kj}}{\partial \alpha} + \frac{\partial A_{ik}}{\partial \alpha}B_{kj}\right] \\
&= (A\frac{\partial B}{\partial \alpha})_{ij} + (\frac{\partial A}{\partial \alpha}B)_{ij} \\
\Rightarrow \frac{\partial (AB)}{\partial \alpha} &= (A\frac{\partial B}{\partial \alpha}) + (\frac{\partial A}{\partial \alpha}B)
\end{aligned}$$

2. Scalar-vector: $\frac{\partial y^T z}{\partial x}$

If $y, z \in \mathbb{R}^n$ be vectors which elements are functions of $x \in \mathbb{R}^m$.

$$\frac{\partial y^T z}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial y_i z_i}{\partial x_k} = \sum_{i=1}^{n} \left[y_i \frac{\partial z_i}{\partial x_k} + \frac{\partial y_i}{\partial x_k} z_i\right] = y^T(\frac{\partial z}{\partial x_k}) + z^T(\frac{\partial y}{\partial x_k})$$

$$\Rightarrow \frac{\partial y^T z}{\partial x} = y^T(\frac{\partial z}{\partial x}) + z^T(\frac{\partial y}{\partial x})$$

For example we know that for a matrix $A \in \mathbb{R}^{n\times n}$ that is not a function of $x \in \mathbb{R}^n$, the derivative $\frac{\partial Ax}{\partial x}$ equals A. Because

$$(\frac{\partial (Ax)}{\partial x})_{ij} = \frac{\partial (Ax)_i}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{k=1}^{n} A_{ik}x_k = A_{ij}$$

So we can compute the derivative below by chain rule:

$$\frac{\partial(x^TAx)}{\partial x}=x^T\frac{\partial(Ax)}{\partial x}+(Ax)^T\frac{\partial x}{\partial x}$$

We know that $\frac{\partial x}{\partial x}=I$ so

$$\frac{\partial(x^TAx)}{\partial x}=x^TA+(Ax)^T=x^T(A+A^T)$$

Examples

Gradients

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x(f(x)+g(x))=\nabla_xf(x)+\nabla_xg(x)$.
- For $t\in\mathbb{R}$, $\nabla_x(tf(x))=t\nabla_xf(x)$.

1. Linear functions ($f(x)=b^Tx$)

For $x\in\mathbb{R}^n$, let $f(x)=b^Tx$. So:

$$f(x)=\sum_{i=1}^{n}b_ix_i\Rightarrow\frac{\partial f(x)}{\partial x_k}=b_k\Rightarrow\nabla_xf(x)=b^T$$

This should be compared to the analogous situation in single variable calculus, where $\frac{\partial}{\partial x}ax=a$.

Clearly $\nabla_x^2f(x)=0$.

2. Quadratic functions ($f(x)=x^TAx$)

Now consider $f(x)=x^TAx$ for $A\in\mathbb{S}^n$. So:

$$\begin{aligned}
f(x) &= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_ix_j \\
\Rightarrow \frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k}\Big[\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_ix_j + \sum_{i\neq k} A_{ik}x_ix_k + \sum_{j\neq k} A_{kj}x_kx_j + A_{kk}x_k^2\Big] \\
&= \sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k \\
&= \sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j \\
&= (A^Tx)_k + (Ax)_k \\
\Rightarrow \nabla_x f(x) &= ((A + A^T)x)^T = x^T(A^T + A)
\end{aligned}$$

To compute Hessian we have:

$$\begin{gathered}
\frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\partial f(x)}{\partial x_k}\Big[\sum_{i=1}^{n}(A_{il} + A_{li})x_i\Big] = (A_{kl} + A_{lk}) \\
\Rightarrow \nabla_x^2 f(x) = A^T + A
\end{gathered}$$

If A is symmetric, then $\nabla_x f(x) = 2x^TA$ and $\nabla_x^2 f(x) = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\frac{\partial^2}{\partial x^2}ax^2 = 2a$).

3. Least Squares ($||Ax - b||_2^2$)

Suppose we are given matrix $A \in \mathbb{R}^{m\times n}$ (for simplicity we assume A is full rank) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$. In this situation we will not be able to find a vector $x \in \mathbb{R}^n$, such that $Ax = b$, so instead we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.

$$\begin{aligned}
||Ax - b||_2^2 &= (Ax - b)^T(Ax - b) \\
&= x^TA^TAx - 2b^TAx + b^Tb
\end{aligned}$$

Taking the gradient with respect to x we have, and using the properties we derived in the previous section

$$\nabla_x ||Ax - b||_2^2 = 2x^TA^TA - 2b^TA$$

Setting this last expression equal to zero and solving for x gives the normal equations

$$x = (A^T A)^{-1} A^T b$$

4. Determinant ($|A|$ and $\log|A|$)

Recall from our discussion of determinants that

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\backslash i, \backslash j}| \quad (\text{for any} j \in 1, \ldots, n)$$

so

$$\frac{\partial}{\partial A_{kl}} |A| = \frac{\partial}{\partial A_{kl}} \sum_{i=1}^{n} (-1)^{i+l} A_{il} |A_{\backslash i, \backslash l}| = (-1)^{k+l} |A_{\backslash k, \backslash l}| = (\text{adj}(A))_{lk}$$

From this it immediately follows from the properties of the adjugate matrix

$$\nabla_A |A| = \text{adj}(A) = |A| A^{-1}$$

Now let's consider the function $f : \mathbb{S}^n \to \mathbb{R}$, $f(A) = \log|A|$. Note that we have to restrict the domain of f to be the positive definite matrices, since this ensures that $|A| > 0$, so that the log of $|A|$ is a real number. In this case we can use the chain rule to see that

$$\frac{\partial \log|A|}{\partial A_{ij}} = \frac{\partial \log|A|}{\partial |A|} \frac{\partial |A|}{\partial A_{ij}} = \frac{1}{|A|} \frac{\partial |A|}{\partial A_{ij}}$$

From this it should be obvious that

$$\nabla_A \log|A| = A^{-1}$$

Note the similarity to the single-valued case, where $\frac{\partial}{\partial x} \log x = 1/x$.

5. $f(A) = y^T A x$

For matrix $A \in \mathbb{R}^{m \times n}$:

$$\begin{aligned} \frac{\partial f(A)}{\partial A_{kl}} &= \frac{\partial}{\partial A_{kl}} \sum_{i=1}^{m} \sum_{i=1}^{n} A_{ij} y_i x_j \\ &= y_k x_l = (xy^T)_{lk} \\ \Rightarrow \nabla_A f(A) &= xy^T \end{aligned}$$

Trace

According to trace definition, for a square matrix $A \in \mathbb{S}^n$:

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

We can conclude that:

- $\partial \mathrm{tr}(X)/\partial X = I$.
- $\partial \mathrm{tr}(U+V)/\partial X = \partial \mathrm{tr}(U)/\partial X + \partial \mathrm{tr}(V)/\partial X$.
- $\partial \mathrm{tr}(\alpha U)/\partial X = \alpha \partial \mathrm{tr}(U)/\partial X$. a is not a function of X.

1. $\mathrm{tr}(AX)$

For $A \in \mathbb{R}^{n\times m}$ and $X \in \mathbb{R}^{m\times n}$.

$$(AX)_{kl} = \sum_{i=1}^{m} A_{ki}X_{il}$$
$$\mathrm{tr}(AX) = \sum_{k=1}^{n}\sum_{i=1}^{m} A_{ki}X_{ik} = \mathrm{tr}(XA)$$
$$\frac{\partial}{\partial X_{ij}}\mathrm{tr}(AX) = A_{ji} \Rightarrow \frac{\partial}{\partial X}\mathrm{tr}(AX) = A$$

Based on this result we can solve part 5 of previous section, differently:

$$f(A) = y^T Ax = \mathrm{tr}(y^T Ax) = \mathrm{tr}(xy^T A) \Rightarrow \frac{\partial f(A)}{\partial A} = xy^T$$

Generally, we have the differential form:

$$\partial f = \mathrm{tr}(A\partial X) \Rightarrow \frac{\partial f}{\partial X} = A$$

2. $\mathrm{tr}(X^T AX)$

The product rule applys to the differential form, and this is the way to derive many of the identities involving the trace function, combined with the fact that the trace function allows transposing and cyclic permutation.

$$\begin{aligned}\partial \text{tr}(X^TAX) &= \text{tr}(\partial(X^TAX)) \\ &= \text{tr}(\partial(X^T)AX) + \text{tr}(X^T\partial(AX)) \\ &= \text{tr}((X^TA^T\partial(X^T)^T)^T) + \text{tr}(X^TA\partial(X)) \\ \Rightarrow \frac{\partial \text{tr}(X^TAX)}{\partial X} &= X^T(A+A^T)\end{aligned}$$

3. $y^TA^{-1}x$

$$y^TA^{-1}x = \partial \text{tr}(y^TA^{-1}x) = \text{tr}(xy^T\partial(A^{-1}))$$

To compute $\partial(A^{-1})$, we use the definition of inverse of matrix:

$$\begin{aligned}A^{-1}A &= I \\ \partial(A^{-1}A) &= \partial I = 0 \\ A^{-1}\partial(A) + \partial(A^{-1})A &= 0 \\ &\Rightarrow \partial(A^{-1}) = -A^{-1}\partial(A)A^{-1}\end{aligned}$$

Now, we continue:

$$\begin{aligned}\text{tr}(xy^T\partial(A^{-1})) &= \text{tr}(-xy^TA^{-1}\partial(A)A^{-1}) \\ &= -\text{tr}(A^{-1}xy^TA^{-1}\partial(A)) \Rightarrow \frac{\partial \text{tr}(y^TA^{-1}x)}{\partial X} &= -A^{-1}xy^TA^{-1}\end{aligned}$$

Some resources may use a different definition for scalar-by-matrix derivative. See this for its reason.

4. $\text{tr}(BA^{-1})$

$$\begin{aligned}\partial \text{tr}(BA^{-1}) &= \text{tr}(B\partial(A^{-1})) = \text{tr}(-BA^{-1}\partial(A)A^{-1}) \\ &= -\text{tr}(A^{-1}BA^{-1}\partial(A)) \\ \Rightarrow \frac{\partial \text{tr}(BA^{-1})}{\partial A} &= -A^{-1}BA^{-1}\end{aligned}$$

5. Jacobi's formula

In matrix calculus, Jacobi's formula expresses the derivative of the determinant of a matrix A in terms of the adjugate of A and the derivative of A.

$$\frac{d}{dt}\det A(t) = \operatorname{tr}\Big(\operatorname{adj}\big(A(t)\big)\frac{dA(t)}{dt}\Big) = \det A(t)\operatorname{tr}\Big(A(t)^{-1}\frac{dA(t)}{dt}\Big)$$

To proof this, we know that $|A|$ is a function of $A_{11}, A_{12}, \ldots, A_{1n}, A_{21}, A_{22}, \ldots, A2n, \ldots, A_{nn}$, so we can write

$$\frac{\partial |A|}{\partial t} = \sum_{ij} \frac{\partial |A|}{\partial A_{ij}} \frac{\partial A_{ij}}{\partial t}$$

We already know that $\frac{\partial |A|}{\partial A_{ij}} = (\operatorname{adj}(A))_{ji}$. So

$$\frac{\partial |A|}{\partial t} = \sum_{ij} (\operatorname{adj}(A))_{ji} (\frac{\partial A}{\partial t})_{ij}$$

By the definition of the trace, we can rewirte the above equation

$$\frac{\partial |A|}{\partial t} = \operatorname{tr}\Big(\operatorname{adj}(A)\frac{\partial A}{\partial t}\Big)$$

By the definition of the Adjugate matrix, $\operatorname{adj}(A) = |A|A^{-1}$, so

$$\frac{\partial |A|}{\partial t} = \operatorname{tr}\Big(|A|A^{-1}\frac{\partial A}{\partial t}\Big) = |A|\operatorname{tr}\Big(A^{-1}\frac{\partial A}{\partial t}\Big)$$

Refrences

1. Linear Algebra Review - CS229 Stanford
2. Matrix Calculus - Wikipedia
3. Math StackExchange
4. CE282: Linear Algebra - Hamid R. Rabiee & Maryam Ramezani