Matrix calculus

Ka Yu Wong

Two natrual challenges naturally arises in calculating the derivative of a quantity with respect to a matrix. First, if the quantity is not a scalar, the result cannot be represented by a matrx. Second, the variable matrix has shape information (i.e. number of rows and columns) that we want to retain in the result. The use of differentials solves both problems elegantly by packing the derivative of a scalar with respect to a matrix (which is a matrix) into a differential form that captures all the derivative information.
Documents I found online are often incomplete, which motivates me to write a rigorous summary of the technique.

1 Differential form

Definition 1.1. $\mathbb{R}^n$ is the n-dimensional Euclidean space. $e_i \in \mathbb{R}^n$ is the i-th standard basis vector of $\mathbb{R}^n$.

Definition 1.2. $(\mathbb{R}^n)^*$ is the dual space of $\mathbb{R}^n$, that is, the set of linear functions from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 1.3. $x_i : \mathbb{R}^n \to \mathbb{R}$ is the i-th coordinate function, that is, $x_i(e_j) = \delta_{ij}$.

Proposition 1.4. $(\mathbb{R}^n)^*$ is an n-dimensional vector space. $(x_i)_{i=1}^n$ is a basis of $(\mathbb{R}^n)^*$.

Definition 1.5. A 1-form on $\mathbb{R}^n$ is a function $\omega : \mathbb{R}^n \to (\mathbb{R}^n)^*$.

Remark 1.6. A function $f : \mathbb{R}^n \to \mathbb{R}$ and an 1-form w on $\mathbb{R}^n$ has a natural product as a

1-form: $f\omega : \mathbb{R}^n \to (\mathbb{R}^n)^*, f\omega(p) = f(p)\omega(p)$.

Definition 1.7. $\mathrm{d}x_i : \mathbb{R}^n \to (\mathbb{R}^n)^*$ is the constant 1-form that sends all $p \in \mathbb{R}^n$ to x_i. That is, $\mathrm{d}x_i(p) = x_i$ for all $p \in \mathbb{R}^n$.

Definition 1.8. Let $f : \mathbb{R}^n \to \mathbb{R}$ be smooth. Define $\mathrm{d}f : \mathbb{R}^n \to (\mathbb{R}^n)^*$,

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \mathrm{d}x_i.$$

Proposition 1.9. For every 1-form ω on $\mathbb{R}^n$, there exist unique $f_i : \mathbb{R}^n \to \mathbb{R}$ such that

$$\omega = \sum_{i=1}^{n} f_i \, \mathrm{d}x_i.$$

Proof. This follows directly from that for each $p \in \mathbb{R}^n$, $\omega(p)$ has a unique representation by $(x_i)_{i=1}^n$. □

This uniqueness is the key to why we can use differentials to fully capture the derivative.

2 Matrix calculus

Now we consider functions whose variables and values are matrices. We preserve those information in our definitions of differentials. Let $f, g : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$, $h : \mathbb{R}^{m\times n} \to \mathbb{R}^{q\times r}$.

Definition 2.1. Define $X : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as the identity map:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix}.$$

Denote f also as $[f]$:

$$f = [f] = \begin{bmatrix} f_{11} & \dots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \dots & f_{pq} \end{bmatrix}.$$

Define $[f]^T : \mathbb{R}^{m\times n} \to \mathbb{R}^{q\times p}, [f]^T(X) = f(X)^T$. Define $[f][h] : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times r}, [f][h](X) = f(X)h(X)$, the matrix product of images. If $p = q$ and $f(X)$ is invertible for all $X \in \mathbb{R}^{m\times n}$,

then define $[f]^{-1} : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times p}, [f]^{-1}(X) = f(X)^{-1}$. If $p = q$, define $|f| : \mathbb{R}^{m\times n} \to \mathbb{R}$, $|f|(X) = |f(X)|$, the determinant of its value.

Definition 2.2. Define $\mathrm{d}f$ as a function from $\mathbb{R}^{m\times n}$ to a $p \times q$ matrix of 1-forms:

$$\mathrm{d}f = \begin{bmatrix} \mathrm{d}f_{11} & \dots & \mathrm{d}f_{1q} \\ \vdots & \ddots & \vdots \\ \mathrm{d}f_{p1} & \dots & \mathrm{d}f_{pq} \end{bmatrix}.$$

Specifically, $\mathrm{d}X$ is a function from $\mathbb{R}^{m\times n}$ to a $m \times n$ matrix of 1-forms:

$$\mathrm{d}X = \begin{bmatrix} \mathrm{d}x_{11} & \dots & \mathrm{d}x_{1n} \\ \vdots & \ddots & \vdots \\ \mathrm{d}x_{m1} & \dots & \mathrm{d}x_{mn} \end{bmatrix}.$$

Definition 2.3. When $p = q = 1$, that is, $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, define the gradient $\frac{\partial f}{\partial X} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$:

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \dots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \dots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}.$$

Proposition 2.4. Some basic properties. Let A, B be constant matrices.

(a) $\mathrm{d}A = 0$

(b) $\mathrm{d}(A[f]B) = A(\mathrm{d}f)B$

(c) $\mathrm{d}(f + g) = \mathrm{d}f + \mathrm{d}g$

(d) $\mathrm{d}(f^T) = (\mathrm{d}f)^T$

(e) $\mathrm{d}(\mathrm{tr}(f)) = \mathrm{tr}(\mathrm{d}f)$

(f) $\mathrm{d}([f][h]) = (\mathrm{d}f)h + f(\mathrm{d}h)$

(g) $\mathrm{d}[f]^{-1} = -[f]^{-1}(\mathrm{d}f)[f]^{-1}$

Proof. We only prove (f) and (g).

Proof for (f). If f, h are both scalars, then $\mathrm{d}(fh) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial fg}{\partial x_{ij}}\,\mathrm{d}x_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n}(\frac{\partial f}{\partial x_{ij}}h + f\frac{\partial h}{\partial x_{ij}})\,\mathrm{d}x_{ij} = (\mathrm{d}f)h + f(\mathrm{d}h)$. In general, let $1 \le k \le p, 1 \le l \le r$. Then we have $\mathrm{d}([f][h])_{kl} = \mathrm{d}(\sum_{u=1}^{q} f_{ku}h_{ul}) = \sum_{u=1}^{q}\mathrm{d}(f_{ku}h_{ul}) = \sum_{u=1}^{q}(\mathrm{d}f_{ku})h_{ul} + f_{ku}(\mathrm{d}h_{ul}) = ((\mathrm{d}f)h)_{kl} + (f(\mathrm{d}h))_{kl}$

Proof for (g). $0 = \mathrm{d}(I_p) = \mathrm{d}([f][f]^{-1}) = (\mathrm{d}f)[f]^{-1} + [f]\,\mathrm{d}[f]^{-1}$. □

We are ready to state the main result.

Theorem 2.5. Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $g : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$. Then $g = \frac{\partial f}{\partial X}$ if and only if $\mathrm{d}f = \mathrm{tr}(g^T\,\mathrm{d}X)$.

Proof. If $A, B \in \mathbb{R}^{m\times n}$, then $\mathrm{tr}(A^TB) = \sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}B_{ij}$, the sum of elementwise product. This is called the Hilbert–Schmidt inner product. Then $\mathrm{tr}(g^T\,\mathrm{d}X) = \sum_{i=1}^{m}\sum_{j=1}^{n}g_{ij}\,\mathrm{d}x_{ij}$. By definition, $\mathrm{d}f = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial f}{\partial x_{ij}}\,\mathrm{d}x_{ij}$. The theorem then follows from Proposition 1.9. □

This theorem is useful when we try to find the derivative of a scalar with respect to a matrix. We can write out the $\mathrm{d}f$ and infer the form of the derivative.

Example 2.6. Let $a, b \in \mathbb{R}^m$ be constant. Compute the derivative of $a^T\Sigma^{-1}b$ with respect to Σ. We write out the differential form:

$$\begin{aligned}
\mathrm{d}(a^T\Sigma^{-1}b) &= \mathrm{d}(\mathrm{tr}(a^T\Sigma^{-1}b)) \\
&= \mathrm{tr}(\mathrm{d}(a^T\Sigma^{-1}b)) \\
&= \mathrm{tr}(a^T(\mathrm{d}\Sigma^{-1})b) \\
&= \mathrm{tr}(a^T(-\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1})b) \\
&= \mathrm{tr}(-\Sigma^{-1}ba^T\Sigma^{-1}\,\mathrm{d}\Sigma)
\end{aligned}$$

Hence $\frac{\partial a^T\Sigma^{-1}b}{\partial\Sigma} = (-\Sigma^{-1}ba^T\Sigma^{-1})^T = -\Sigma^{-T}ab^T\Sigma^{-T}$.

Proposition 2.7. Recall that X is the identity map, and $|X| : \mathbb{R}^{m\times m} \to \mathbb{R}$ sends a matrix to its determinant. We have $\mathrm{d}|X| = |X|\mathrm{tr}([X]^{-1}\,\mathrm{d}X)$.

Proof. Let $1 \le i \le m, 1 \le j \le n$. The Laplace expansion along the i-th row gives $\frac{\partial |X|}{\partial x_{ij}} = (-1)^{i+j}|X_{-i,j}|$, where $X_{-i,j}$ is the matrix with i-th row and j-th column removed. This is precisely the i,j-th entry of the classical adjoint of X, so $\frac{\partial |X|}{\partial X} = |X|([X]^{-1})^T$. Then by Theorem 2.5, $\mathrm{d}|X| = \mathrm{tr}((|X|([X]^{-1})^T)^T\,\mathrm{d}X) = |X|\mathrm{tr}([X]^{-1}\,\mathrm{d}X)$. □

Now we present the chain rule.

Theorem 2.8. Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}, g : \mathbb{R}^{p\times q} \to \mathbb{R}$. Then

$$\mathrm{d}(g \circ f) = \mathrm{tr}((\frac{\partial g}{\partial f})^T\,\mathrm{d}f).$$

Proof.

$$\begin{aligned}
(\frac{\partial g}{\partial f})^T\,\mathrm{d}f &= \sum_{k=1}^{p}\sum_{l=1}^{q}\frac{\partial g}{\partial f_{kl}}\,\mathrm{d}f_{kl} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{q}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial g}{\partial f_{kl}}\frac{\partial f_{kl}}{\partial x_{ij}}\,\mathrm{d}x_{ij} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}(\sum_{k=1}^{p}\sum_{l=1}^{q}\frac{\partial g}{\partial f_{kl}}\frac{\partial f_{kl}}{\partial x_{ij}})\,\mathrm{d}x_{ij} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial g\circ f}{\partial x_{ij}}\,\mathrm{d}x_{ij} \\
&= (\frac{\partial g\circ f}{\partial x})^T\,\mathrm{d}X
\end{aligned}$$

□

Example 2.9. $\mathrm{d}\ln|X| = \mathrm{tr}(\frac{1}{|X|}\,\mathrm{d}|X|) = \mathrm{tr}(\frac{1}{|X|}|X|X^{-1}\,\mathrm{d}X) = \mathrm{tr}(X^{-1}\,\mathrm{d}X)$. Hence $\frac{\partial \ln|X|}{\partial X} = X^{-T}$.

Example 2.10. Compute $\frac{\partial |X^T X|}{\partial X}$.

$$\begin{aligned}
\mathrm{d}|X^T X| &= \mathrm{tr}(|X^T X|(X^T X)^{-1}\,\mathrm{d}(X^T X)) \\
&= \mathrm{tr}(|X^T X|(X^T X)^{-1}(\mathrm{d}X^T)X) + \mathrm{tr}(|X^T X|(X^T X)^{-1}X^T\,\mathrm{d}X) \\
&= \mathrm{tr}((|X^T X|(X^T X)^{-1}(\mathrm{d}X^T)X)^T) + \mathrm{tr}(|X^T X|(X^T X)^{-1}X^T\,\mathrm{d}X) \\
&= \mathrm{tr}(X^T\,\mathrm{d}X(|X^T X|(X^T X)^{-1}) + \mathrm{tr}(|X^T X|(X^T X)^{-1}X^T\,\mathrm{d}X) \\
&= 2|X^T X|\mathrm{tr}((X(X^T X)^{-1})^T\,\mathrm{d}X)
\end{aligned}$$

Hence $\frac{\partial |X^T X|}{\partial X} = 2|X^T X|X(X^T X)^{-1}$.

References

[1] Kaare Brandt Petersen, Michael Syskind Pedersen, *The Matrix Cookbook*. `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

[2] Pili HU, *Matrix Calculus: Derivation and Simple Application*. `https://project.hupili.net/tutorial/hu2012-matrix-calculus/hu2012matrix-calculus.pdf`