Matrix Calculus

Nanyi, UIBE

October 19, 2022

1 Differentiation of a function

Definition 1.1 (Derivatives)**.** suppose f is a map from $(\mathbb{R}, d_{\mathbb{R}})$ to (V, d_V). Then the derivative of f at x_0 is define by

$$f'(x_0) := \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

A direct consequence of this definition is $f'(x_0) \in V$.

The definition of differentiation in calculus does not capture the essence of differentiation. In fact, differentiation operator is a map from $B(V, W) \to B(V, W)$(given x_0), where V and W are normed vector spaces. Now, suppose $x_0, v \in V$.

Definition 1.2 (Differentiation operator)**.** Suppose V and W are normed vector spaces, x_0 and v are vectors in V. If there exists a linear map S from V to W(an element of $B(V, W)$) such that

$$\lim_{v \to 0} \frac{\|T(x_0 + v) - T(x_0) - Sv\|}{\|v\|} = 0$$

Then we say T is differentiable at x_0 and S is the differentiation of T at x_0. For historical reasons, the differentiaion S is usually written as $dT(x_0)$.

Theorem 1.1 (Jacobian Matrix)**.** Suppose f is a map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the matrix of differentiation of f at $x_0(\mathrm{d}f(x_0))$ with respect to the standard orthonormal basis is

$$\mathcal{M}(\mathrm{d}f(x_0)) = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x_0) & \frac{\partial}{\partial x_2} f_1(x_0) & \cdots \frac{\partial}{\partial x_n} f_1(x_0) \\ \frac{\partial}{\partial x_1} f_2(x_0) & \frac{\partial}{\partial x_2} f_2(x_0) & \cdots \frac{\partial}{\partial x_n} f_2(x_0) \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x_0) & \frac{\partial}{\partial x_2} f_m(x_0) & \cdots \frac{\partial}{\partial x_n} f_m(x_0) \end{bmatrix}$$

$\mathcal{M}(\mathrm{d}f(x_0))$ is also known as the Jacobian Matrix of $\mathbf{f}$.

Theorem 1.2 (Chain rules)**.** Suppose $f : U \to V$, $g : V \to W$. If f is differentiable at x_0 and g is differentiable at $f(x_0)$, then $g \circ f$ is differentiable at x_0 and

$$(\mathrm{d}(g \circ f))(x_0) = (\mathrm{d}g)(f(x_0)) \circ df(x_0).$$

Taking matrice of the equality above, we obtain

$$\mathcal{M}((\mathrm{d}(g \circ f))(x_0)) = \mathcal{M}((\mathrm{d}g)(f(x_0)) \circ df(x_0)) = \mathcal{M}((\mathrm{d}g)(f(x_0)))\mathcal{M}(\mathrm{d}f(x_0)).$$

Definition 1.3. Suppose $x \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$. Then $df(x_0)$ and $d\varphi(x_0)$ is defined by:

$$\mathrm{d}f(x_0) := \mathbb{R} \to \mathbb{R}, h \mapsto f'(x_0)h.$$

$$\mathrm{d}\varphi = \mathrm{d}\varphi(x_0) := \mathbb{R} \to \mathbb{R}, h \mapsto h.$$

For historical reasons, the notation $\mathrm{d}x$ is widely used in many textbooks. However, In this notes, we prefer to use $\mathrm{d}\varphi$ since this notation is well-defined.

Definition 1.4. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x}_0 \in \mathbb{R}^n$. Then $\mathrm{d}f(\mathbf{x}_0)$ is defined by

$$\mathrm{d}f(\mathbf{x}_0) := \mathbb{R}^n \to \mathbb{R}, h \mapsto \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\mathbf{x}_0) h_i,$$

and

$$\mathrm{d}\varphi_k(\mathbf{x}_0) = \mathrm{d}x_k = \mathrm{d}\varphi_k := \mathbb{R}^n \to \mathbb{R}, h \mapsto h_k.$$

Remark 1.

$$\mathrm{d}f(\mathbf{x}_0) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\mathbf{x}_0) \mathrm{d}\varphi_k.$$

Definition 1.5. Suppose $v_1, \cdots, v_n$ is a basis in V. And the dual space of V is the collection of all bounded linear maps to from V to $\mathbb{F}$, denoted by V'. A basis $\varphi_1, \cdots, \varphi_n$ of V is called a dual basis if

$$\varphi_i(\nu_k) = \delta_{ik}.$$

In this situation, suppose $h = \sum_{i=1}^{n} h_i e_i$. we have that

$$\mathrm{d}x_i(h) = h_k.$$

You should note that the definition of $\mathrm{d}x_i$ is consistent with the dual basis.

Example 1.1. Find the differentiation of $f(t, s) = t^2 + s^3$ at (t, s).

$$\mathrm{d}f(t,s) = 2t\mathrm{d}t + 3s^2\mathrm{d}s$$

where $\mathcal{M}(\mathrm{d}t) = \begin{bmatrix}1, 0\end{bmatrix}$ and $\mathcal{M}(\mathrm{d}s) = \begin{bmatrix}0, 1\end{bmatrix}$. In other words, $\mathrm{d}t = \mathrm{d}\varphi_1$, $\mathrm{d}s = \mathrm{d}\varphi_2$.

Definition 1.6. Suppose $f : \mathbb{R}^{m,n} \to \mathbb{R}$. Then $\mathrm{d}f(\mathbf{X}_0)$ is defined by

$$\mathrm{d}f(\mathbf{X}_0) : \mathbb{R}^{m,n} \to \mathbb{R},\ H \mapsto \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial}{\partial x_{ij}} f(\mathbf{X}_0) H(i,j)$$

$$\mathrm{d}\varphi_{ij}(\mathbf{X}_0) : \mathbb{R}^{m,n} \to \mathbb{R},\ H \mapsto H(i,j).$$

In this case, we have

$$\mathrm{d}f(\mathbf{X}_0) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial}{\partial x_{ij}} (\mathbf{X}_0) \mathrm{d}\varphi_{ij}.$$

After clearing the definition of differentiation, we begin to define the derivative of f with respect to a matrix. First note that

$$\begin{aligned} \mathrm{d}f(x_0) &= f'(x_0)\mathrm{d}x \\ &= \langle f'(x_0), \mathrm{d}x \rangle. \end{aligned}$$

And thus we apply the following the notation

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x_0) = f'(x_0).$$

to reflect this relationship.

Aslo, we define $\mathrm{d}\varphi = \mathrm{d}\mathbf{x}$ by

$$\mathrm{d}\mathbf{x} = \begin{bmatrix} \mathrm{d}\varphi_1 \\ \mathrm{d}\varphi_2 \\ \vdots \\ \mathrm{d}\varphi_n \end{bmatrix} .gv$$

Then the total differentiation of f can written as

$$\begin{aligned} \mathrm{d}f(\mathbf{x}_0) &= \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\mathbf{x}_0)\mathrm{d}\varphi_i \\ &= \langle \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}_0) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}_0) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}_0) \end{bmatrix}, \mathrm{d}\mathbf{x} \rangle \\ &= \mathrm{tr}(v'\mathrm{d}\mathbf{x}). \end{aligned}$$

Simlilary,

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}(\mathbf{x}_0) := \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}_0) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}_0) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}_0) \end{bmatrix}.$$

At last,

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}}(\mathbf{X}_0) := \begin{bmatrix} \frac{\partial}{\partial x_{11}} f(\mathbf{X}_0) & \frac{\partial}{\partial x_{12}} f(\mathbf{X}_0) & \cdots & \frac{\partial}{\partial x_{1m}} f(\mathbf{X}_0) \\ \frac{\partial}{\partial x_{21}} f(\mathbf{X}_0) & \frac{\partial}{\partial x_{22}} f(\mathbf{X}_0) & \cdots & \frac{\partial}{\partial x_{2m}} f(\mathbf{X}_0) \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial x_{n1}} f(\mathbf{X}_0) & \frac{\partial}{\partial x_{n2}} f(\mathbf{X}_0) & \cdots & \frac{\partial}{\partial x_{nm}} f(\mathbf{X}_0) \end{bmatrix}.$$

2 Differentiation of a matrix

Definition 2.1. Suppose $F = [f_{ij}]_{m\times n}$, where f is a map defined on $\mathbb{R}^p$. Then F can be seen as a map from $\mathbb{R}^p$ to $\mathbb{R}^{m\times n}$, and

$$F : \mathbb{R}^p \to \mathbb{R}^{m\times n},\ \mathbf{x} \mapsto [f_{ij}(\mathbf{x})]_{m\times n}.$$

Similarly, we define the differentiation of F at $\mathbf{x}_0$ by

$$\mathrm{d}F(\mathbf{x}_0) := [\mathrm{d}f_{ij}(\mathbf{x}_0)]_{m\times n}.$$

where

$$\mathrm{d}f_{ij}(\mathbf{x}_0) : \mathbb{R}^p \to \mathbb{R}, h \mapsto \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\mathbf{x}_0) h_k.$$

When $p = m \times n$ and $f_{ij} = \varphi_{ij}$. We have that

$$F(\mathbf{X}_0) = [\varphi_{ij}(\mathbf{X}_0)] = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}.$$

Example 2.1. Suppose $F : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}$, and

$$F : (s,t) \mapsto \begin{bmatrix} s+t & s^2+t^2 \\ e^s+t & \sin t + s \end{bmatrix}.$$

By definition, we have

$$\mathrm{d}F(s_0,t_0) = \begin{bmatrix} \mathrm{d}s + \mathrm{d}t & 2s_0\mathrm{d}s + 2t_0\mathrm{d}t \\ e^{s_0}\mathrm{d}s + \mathrm{d}t & \mathrm{d}s + \cos t_0 \mathrm{d}t \end{bmatrix}$$

Theorem 2.1 (Properties of matrix diffrentiation)**.**
Suppose $F, G : \mathbb{R}^p \to \mathbb{R}^{m\times n}$. Then we have the followings,

- $\mathrm{d}(F+G) = \mathrm{d}F + \mathrm{d}G$;
- $\mathrm{d}(FG) = (\mathrm{d}F)G + F(\mathrm{d}G)$;
- $\mathrm{d}(AFB) = A(\mathrm{d}F)B$;
- $\mathrm{d}(F^{-1}) = -F^{-1}\mathrm{d}(F)F^{-1}$;

Proof. First note that,

$$\begin{aligned} \mathrm{d}(F+G)(\mathbf{x}_0) &= [\mathrm{d}(f_{ij}+g_{ij})(\mathbf{x}_0)]_{m\times n} \\ &= [\mathrm{d}(f_{ij})(\mathbf{x}_0)]_{m\times n} + [\mathrm{d}(g_{ij})(\mathbf{x}_0)]_{m\times n} \\ &= \mathrm{d}(F)(\mathbf{x}_0) + \mathrm{d}(G)(\mathbf{x}_0). \end{aligned}$$

□

Theorem 2.2 (Properties of matrix diffrentiation II)**.**
Suppose $F, G : \mathbb{R}^p \to \mathbb{R}^{m\times n}$. Then we have the followings,

- $\mathrm{d}(F') = \mathrm{d}(F)'$;
- $\mathrm{d}tr(F) = tr(\mathrm{d}F)$;

- $\mathrm{d}|F| = |F|tr(F^{-1}\mathrm{d}F)$

Before starting the proof, we need to ensure there notations make sense. For example, $tr(\mathrm{d}F)$ is defined by

$$tr(\mathrm{d}F)(\mathbf{x}_0) := tr(dF(\mathbf{x}_0))$$

and

$$\mathrm{d}|F|(\mathbf{x}_0) := |\mathrm{d}F(\mathbf{x}_0)|$$

Theorem 2.3 (Properties of matrix diffrentiation III)**.**
Suppose $F, G : \mathbb{R}^p \to \mathbb{R}^{m\times n}$. Then we have the followings,

- $\mathrm{d}(F \odot G) = \mathrm{d}(F) \odot G + F \odot \mathrm{d}(G)$;
- $\mathrm{d}(\sigma(F)) = \sigma'(F) \odot \mathrm{d}F$

Example 2.2. Suppose $f : \mathbf{X} \mapsto A\mathbf{X}B$.

By definition, we have

$$\begin{aligned} tr(\mathrm{d}f(\mathbf{X}_0)) &= tr(A\mathrm{d}\mathbf{X}B) \\ &= tr(A\mathrm{d}\mathbf{X}B) \\ &= tr(BA\mathrm{d}\mathbf{X}). \end{aligned}$$

Thus we have

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}}(\mathbf{X}_0) = A'B'.$$

3 Derivatives

Definition 3.1 (Scalar2Vector)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x}_0 \in \mathbb{R}^n$, then

$$\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_0) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}_0) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \end{bmatrix}.$$

In other words, $\frac{\partial f}{\partial \mathbf{x}}$ is defined to be a map from $\mathbb{R}^n \to \mathbb{R}^{n\times 1}$.

Definition 3.2 (Scalar2Matrix)**.** Similarly, suppose $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ and $\mathbf{X}_0 \in \mathbb{R}^n$, then

$$\frac{\partial f}{\partial \mathbf{X}}(\mathbf{X}_0) := \begin{bmatrix} \frac{\partial f}{\partial x_{11}}(\mathbf{x}_0) & \frac{\partial f}{\partial x_{12}}(\mathbf{x}_0) & \cdots & \frac{\partial f}{\partial x_{1n}}(\mathbf{x}_0) \\ \frac{\partial f}{\partial x_{21}}(\mathbf{x}_0) & \frac{\partial f}{\partial x_{22}}(\mathbf{x}_0) & \cdots & \frac{\partial f}{\partial x_{2n}}(\mathbf{x}_0) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f}{\partial x_{m1}}(\mathbf{x}_0) & \frac{\partial f}{\partial x_{m2}}(\mathbf{x}_0) & \cdots & \frac{\partial f}{\partial x_{mn}}(\mathbf{x}_0) \end{bmatrix}$$

In the following part, we will give some examples of deritatives with respect to a function whose range is in $\mathbb{R}$. More precisely, suppose

$$f : \mathbb{R}^{m\times n} \to \mathbb{R},$$

the derivative of f with respect to $\mathbf{X}$ is given by

$$\mathrm{d}f(\mathbf{X}_0) = tr((\frac{\mathrm{d}f}{\mathrm{d}\mathbf{X}}(\mathbf{X}_0))^T \mathrm{d}\mathbf{X})$$

Definition 3.3 (Matrix2Matrix)**.** Suppose F is a map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{p\times q}$, or

$$F : A \mapsto [f_{ij}(A)]_{p\times q}.$$

then $\frac{\mathrm{d}F}{\mathrm{d}A}$ is defined by

$$\frac{\mathrm{d}F}{\mathrm{d}A}(A_0) = \begin{bmatrix} \frac{\mathrm{d}f_{11}}{\mathrm{d}A}(A_0) & \frac{\mathrm{d}f_{12}}{\mathrm{d}A}(A_0) & \cdots & \frac{\mathrm{d}f_{1q}}{\mathrm{d}A}(A_0) \\ \frac{\mathrm{d}f_{21}}{\mathrm{d}A}(A_0) & \frac{\mathrm{d}f_{22}}{\mathrm{d}A}(A_0) & \cdots & \frac{\mathrm{d}f_{2q}}{\mathrm{d}A}(A_0) \\ \vdots & \vdots & & \vdots \\ \frac{\mathrm{d}f_{p1}}{\mathrm{d}A}(A_0) & \frac{\mathrm{d}f_{p2}}{\mathrm{d}A}(A_0) & \cdots & \frac{\mathrm{d}f_{pq}}{\mathrm{d}A}(A_0) \end{bmatrix}$$

Example 3.1 (Gradient)**.**
Supppose $f : \mathbb{R}^n \to \mathbb{R}$, then

$$grad(f) = \bigtriangledown(f) = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}'} := \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n}. \end{bmatrix}$$

Definition 3.4 (Jacobian)**.**
Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$,and

$$\mathbf{f} : \mathbf{x} \mapsto (f_1(\mathbf{x}), \cdots, f_n(\mathbf{x}))',$$

then

$$Jacobian(\mathbf{f}) = \frac{\mathrm{d}\mathbf{f}}{\mathrm{d}\mathbf{x}'} := \begin{bmatrix} \frac{\mathrm{d}f_1}{\mathrm{d}\mathbf{x}} & \frac{\mathrm{d}f_2}{\mathrm{d}\mathbf{x}} & \cdots & \frac{\mathrm{d}f_m}{\mathrm{d}\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}f_1}{\mathrm{d}x_1} & \frac{\mathrm{d}f_2}{\mathrm{d}x_2} & \cdots & \frac{\mathrm{d}f_m}{\mathrm{d}\mathbf{x}} \end{bmatrix}$$

Example 3.2. Suppose $f : \mathbf{x} \mapsto \mathbf{x}'\Omega\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$. Then

$$\begin{aligned}
\mathrm{d}f(\mathbf{x}_0) &= tr(\mathrm{d}f(\mathbf{x}_0)) \\
&= tr((\mathrm{d}\mathbf{x})'\Omega\mathbf{x}_0) + tr(\mathbf{x}_0'\Omega\mathrm{d}\mathbf{x}) \\
&= tr((\Omega\mathbf{x}_0)'\mathrm{d}\mathbf{x}) + tr(\mathbf{x}_0'\Omega\mathrm{d}\mathbf{x}) \\
&= tr(\mathbf{x}_0'(\Omega' + \Omega)\mathrm{d}\mathbf{x}).
\end{aligned}$$

Thus we have

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}(\mathbf{x}_0) = (\Omega + \Omega')\mathbf{x}_0$$

Example 3.3 (OLS)**.**
Suppose $l : \mathbb{R}^{p+1} \to \mathbb{R},\ \beta \mapsto (\mathbf{X}\beta - \mathbf{y})'(\mathbf{X}\beta - \mathbf{y})$. Then we have

$$\begin{aligned}
\mathrm{d}(l)(\beta_0) &= (\mathbf{X}\mathrm{d}\beta)'(\mathbf{X}\beta_0 - \mathbf{y}) + (\mathbf{X}\beta_0 - \mathbf{y})'(\mathbf{X}\mathrm{d}\beta) \\
&= (\mathbf{X}\beta_0 - \mathbf{y})'(\mathbf{X}\mathrm{d}\beta) + (\mathbf{X}\beta_0 - \mathbf{y})'(\mathbf{X}\mathrm{d}\beta) \\
&= 2(\mathbf{X}\beta_0 - \mathbf{y})'(\mathbf{X}\mathrm{d}\beta),
\end{aligned}$$

and thus

$$\frac{\mathrm{d}l}{\mathrm{d}\beta}(\beta_0) = 2\mathbf{X}'(\mathbf{X}\beta_0 - \mathbf{y}) = 0$$

implies

$$\beta_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Example 3.4 (OLS with multiple outputs)**.**
Suppose $Y \in \mathbb{R}^{N\times K}$, $X \in \mathbb{R}^{N\times(p+1)}$ and $B \in \mathbb{R}^{(p+1)\times K}$. Then the RSS is a map from $\mathbb{R}^{(p+1)\times K} \to \mathbb{R}$. More precisely,

$$\begin{aligned}
RSS(B) &= \sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik} - f_k(\mathbf{x}_i)) \\
&= tr((Y - XB)'(Y - XB)).
\end{aligned}$$

Then we have

$$\mathrm{d}RSS(B_0) = tr((-X\mathrm{d}B)'(Y - XB) - (Y - XB)'\mathrm{d}XB)$$

Example 3.5. Suppose $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are random samples from $N(\mu, \Sigma)$. Then

$$l : \Sigma \mapsto \ln(|\Sigma|) + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})'\Sigma^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})$$

$$\begin{aligned}
\mathrm{d}l(\Sigma_0) &= tr(\Sigma_0^{-1}\mathrm{d}\Sigma) - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})'\Sigma_0^{-1}\mathrm{d}\Sigma\Sigma_0^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \\
&= tr(\Sigma_0^{-1}\mathrm{d}\Sigma) - tr(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})'\Sigma_0^{-1}\mathrm{d}\Sigma\Sigma_0^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})) \\
&= tr(\Sigma_0^{-1}\mathrm{d}\Sigma) - tr(\Sigma_0^{-1}S\Sigma_0^{-1}\mathrm{d}\Sigma) \\
&= tr(\Sigma_0^{-1}(I - S\Sigma_0^{-1})\mathrm{d}\Sigma).
\end{aligned}$$

Then

$$\frac{\mathrm{d}l}{\mathrm{d}\Sigma}(\Sigma_0) = (I - S\Sigma_0^{-1})'(\Sigma_0^{-1})' = 0$$

implies $\Sigma_0 = S$.

4 Applications

4.1 Regression Models

Example 4.1 (Ridge Regression)**.**
123

1223123