L14: Chain rules

1. Chain rules

(1) Chain rules
For $z = f(y) \in R$, $y = g(x) \in R^n$ and $x \in R$, $\frac{dz}{dx} = \sum_{k=1}^{n} \frac{\partial z}{\partial y_k} \frac{dy_k}{dx}$ gives a chain rule.

(1) Three vectors
For $z(y) = \begin{pmatrix} z_1(y) \\ \vdots \\ z_m(y) \end{pmatrix} \in R^m$, $y(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{pmatrix} \in R^n$ and $x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in R^p$,
$\frac{\partial z}{\partial x^T} = \left(\frac{\partial z}{\partial y^T}\right)\left(\frac{\partial y}{\partial x^T}\right)$.

Pf: Recall: $AB = (a_{ij})_{m\times n}(b_{ij})_{n\times p} = (\sum_{k=1}^{n} a_{ik}b_{kj})_{m\times p}$.

$$\begin{aligned} \text{Now } \frac{\partial z}{\partial x^T} &= \left((z_i)'_{x_j}\right)_{m\times p} = \left(\sum_{k=1}^{n}(z_i)'_{y_k}\,(y_k)'_{x_j}\right)_{m\times p} = \left((z_i)'_{y_j}\right)_{m\times n}\left((y_i)'_{x_j}\right)_{n\times p} \\ &= \left(\frac{\partial z}{\partial y^T}\right)\left(\frac{\partial y}{\partial x^T}\right). \end{aligned}$$

Ex1: In $z = f(x) = \|b - Ax\|^2$, $b \in R^n$, $A \in R^{n\times p}$ and $x \in R^p$. Find the gradient vector $\nabla f(x) = \frac{\partial f(x)}{\partial x}$ and the equation equivalent to $\nabla f(x) = 0$.
Write $z = y^T y \in R$, $y = b - Ax \in R^n$ and $x \in R^n$. By chain rule

$$\frac{\partial z}{\partial x^T} = \frac{\partial z}{\partial y^T}\frac{\partial y}{\partial x^T} = (2y^T)(-A) = -2(b - Ax)^T A.$$

So $\nabla f(x) = \frac{\partial z}{\partial x} = \left(\frac{\partial z}{\partial x^T}\right)^T = -2A^T(b - Ax) = 2A^TAx - 2A^Tb$.
Clearly, $\nabla f(x) = 0 \iff A^TAx = A^Tb$.

(2) Two matrices
For $z(Y) \in R$, $Y(X) = (y_{st}(X))_{m\times n}$ and $X = (x_{ij})_{p\times q}$,

$$\frac{\partial z}{\partial X} = \sum_{s=1}^{m}\sum_{t=1}^{n}\frac{\partial z}{\partial y_{st}}\left(\frac{\partial y_{st}}{\partial X}\right) = \sum_{s=1}^{m}\sum_{t=1}^{n}\left(\frac{\partial z}{\partial Y}\right)_{st}\left(\frac{\partial y_{st}}{\partial X}\right).$$

Pf:
$$\begin{aligned} \frac{\partial z}{\partial X} &= \left(z'_{x_{ij}}\right)_{p\times q} = \left(\sum_{s=t}^{m}\sum_{t=1}^{n} z'_{y_{st}}(y_{st})'_{x_{ij}}\right)_{p\times q} = \sum_{s=1}^{m}\sum_{t=1}^{n} z'_{y_{st}}\left((y_{st})'_{x_{ij}}\right)_{p\times q}. \\ &= \sum_{s=1}^{m}\sum_{t=1}^{n}\frac{\partial z}{\partial y_{st}}\left(\frac{\partial y_{st}}{\partial X}\right)_{p\times q} = \sum_{s=1}\sum_{t=1}^{n}\left(\frac{\partial z}{\partial Y}\right)_{st}\left(\frac{\partial y_{st}}{\partial X}\right)_{p\times q}. \end{aligned}$$

Ex2: With $X \in R^{p\times q}$ and $AXB \in R^{n\times n}$ find $\frac{\partial |AXB|}{\partial X}$.
$z = |Y| \in R$, $Y = AXB \in R^{n\times n}$ and $X \in R^{p\times q}$.

$$\begin{aligned} \frac{\partial |AXB|}{\partial X} &= \sum_{s=t}^{n}\sum_{t=1}^{n}\left(\frac{\partial z}{\partial Y}\right)_{st}\left(\frac{\partial y_{st}}{\partial X}\right) = \sum_{s=t}^{n}\sum_{t=1}^{n}\left(\frac{\partial |Y|}{\partial Y}\right)_{st}\left(\frac{\partial (AXB)_{st}}{\partial X}\right) \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n}|Y|\,\left[(Y^T)^{-1}\right]_{st}A^T\left[E_{n\times n}(s,t)\right]B^T \\ &= |Y|\,A^T(Y^{-1})^TB^T = |AXB|\,[B(AXB)^{-1}A]^T. \end{aligned}$$

Comment: $\frac{\partial \ln|AXB|}{\partial X} = \frac{d\ln|AXB|}{d|AXB|}\left(\frac{\partial |AXB|}{\partial X}\right)$.

2. Maximization and minimization

(1) Taylor expansion

For $y = f(x) \in R$ where $x \in R^p$, $\nabla f(x) = \frac{\partial f(x)}{\partial x} \in R^p$ is the gradient vector and $H_{f(x)} = \frac{\partial \nabla f(x)}{\partial x^T} \in R^{p \times p}$ is the Hessian matrix. By Taylor expansion of $f(x)$ at x_0

$$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2!}(x - x_0)^T H_{f(\xi)}(x - x_0)$$

where $\xi = \alpha x + (1 - \alpha)x_0$ with some $\alpha \in [0, 1]$.

Ex3: When $p = 1$, the Taylor expansion is

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(\xi).$$

(2) A sufficient condition for minimization

If $H_{f(x)} \geq 0$ for all x, then $f(x)$ is minimized at x_0 where x_0 is a solution to $\nabla f(x) = 0$ called the stationary points for $f(x)$.

Pf: For x in the domain of $f(\cdot)$, by Taylor expansion of $f(x)$ at x_0,

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)\nabla f(x_0) + \tfrac{1}{2!}(x - x_0)^T H_{f(\xi)}(x - x_0) \\ &= f(x_0) + \tfrac{1}{2!}(x - x_0)^T H_{f(\xi)}(x - x_0) \geq f(x_0). \end{aligned}$$

(3) A sufficient condition for maximization

If $H_{f(x)} \leq 0$ for all x, then $f(x)$ is maximized at x_0 where x_0 is a solution to $\nabla f(x) = 0$ called the stationary points for $f(x)$.

Pf: For x in the domain of $f(\cdot)$, by Taylor expansion of $f(x)$ at x_0,

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)\nabla f(x_0) + \tfrac{1}{2!}(x - x_0)^T H_{f(\xi)}(x - x_0) \\ &= f(x_0) + \tfrac{1}{2!}(x - x_0)^T H_{f(\xi)}(x - x_0) \leq f(x_0). \end{aligned}$$

3. An example: Minimization of $f(x) = \|b - Ax\|^2$ in Ex1.

(1) Calculus approach

From Ex1, $\nabla f(x) = 2A^T Ax - 2A^T b$. So

$$H_{f(x)} = \frac{\partial\ \nabla f(x)}{\partial\ x^T} = \frac{\partial\ (2A^T Ax - 2A^T b)}{\partial\ x^T} = 2A^T A \geq 0.$$

Also by Ex1 $\quad \nabla f(\widehat{x}) = 0 \Longleftrightarrow A^T A\widehat{x} = A^T b.$

Hence $\|b - Ax\|^2$ is minimized at $\widehat{x}$ where $A^T A\widehat{x} = A^T b$.

(2) Linear algebra approach

$$\|b - A\widehat{x}\|^2 \leq \|b - Ax\|^2 \text{ for all } x \Longleftrightarrow A\widehat{x} = \pi(b|\mathcal{R}(A)) = AA^+ b.$$

(3) Equivalence: $\quad A^T A\widehat{x} = A^T b \Longleftrightarrow A\widehat{x} = AA^+ b.$

⇒: $A^T A\widehat{x} = A^T b \Longrightarrow (A^+)^T A^T A\widehat{x} = (A^+)^T A^T b \Longrightarrow (AA^+)^T A\widehat{x} = (AA^+)^T b.$
So $A\widehat{x} = AA^+ b$.

⇐: $A\widehat{x} = AA^+ b \Longrightarrow A^T A\widehat{x} = A^T AA^+ b = A^T (AA^+)^T b = (AA^+ A)^T b = A^T b.$