矩阵微积分

耿修瑞

中国科学院空间信息创新研究院

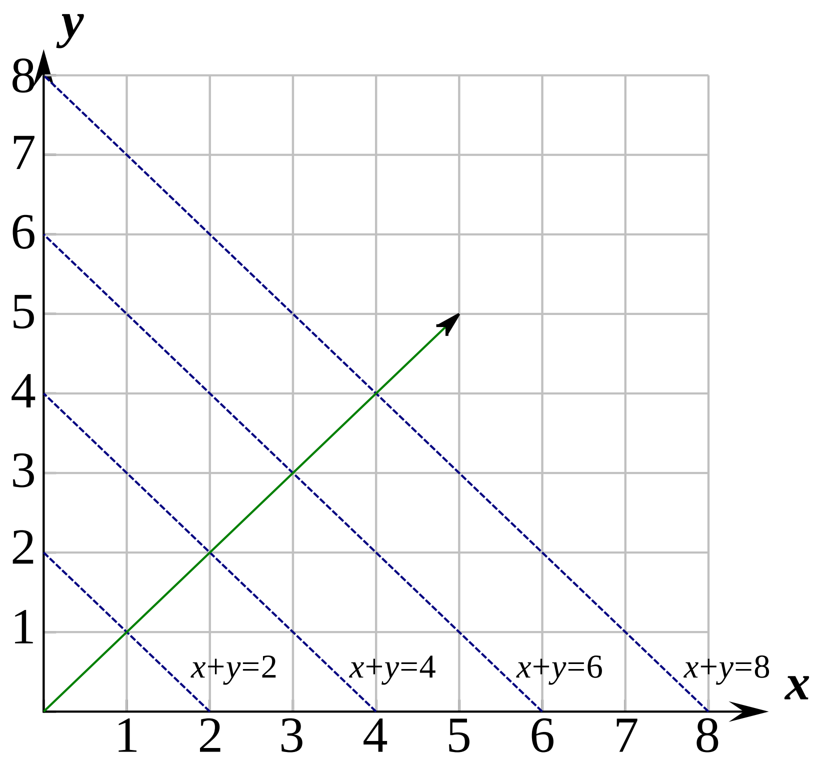
gengxr@sina.com.cn

2025.3

- 几个常用例子
- 实值（向量）函数相对于实向量的梯度
- 实值函数相对于矩阵的梯度
- 矩阵微分
- 迹函数的矩阵梯度
- 行列式的矩阵梯度
- Hessian矩阵和投影Hessian矩阵

常用例子-1

- 函数 $f(x,y)=5(x+y)$的梯度（线性）



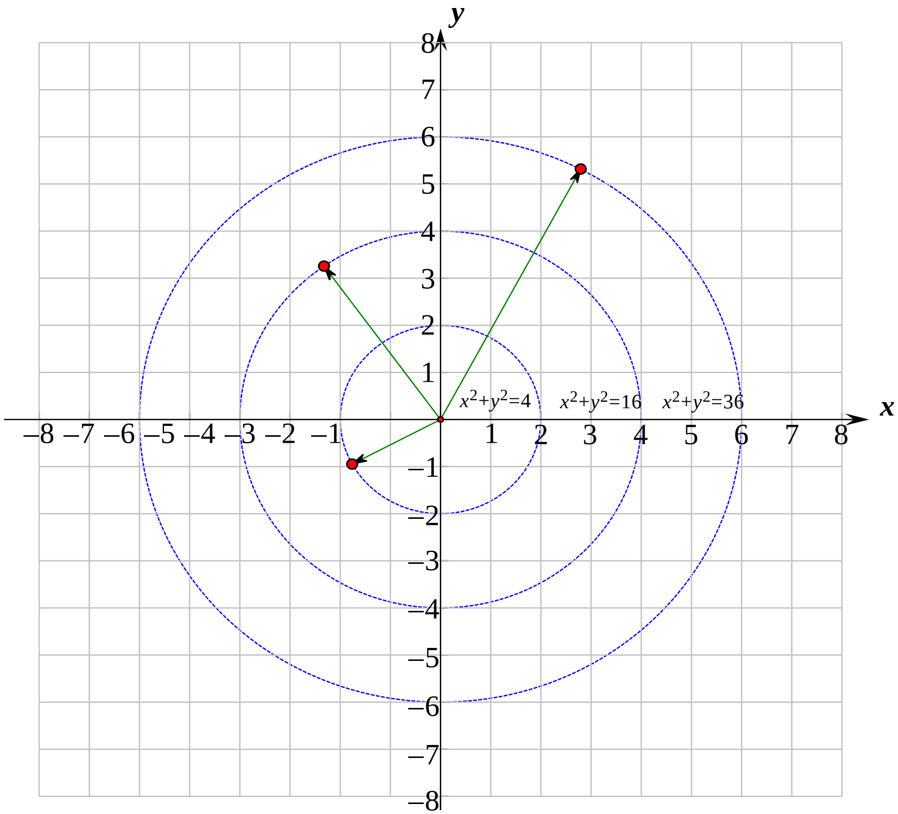
$$f(x,y)=5(x+y)=5\mathbf{x}^{\mathrm{T}}\mathbf{1}$$

$$\mathbf{x}=[x \quad y]^{\mathrm{T}},\mathbf{1}=[1 \quad 1]^{\mathrm{T}}$$

$$\nabla_{\mathbf{x}} f(\mathbf{x})=[5 \quad 5]^{\mathrm{T}}$$

常用例子-2

- 函数 $f(x,y)=\frac{1}{2}\left(x^2+y^2\right)$的梯度（二次型）



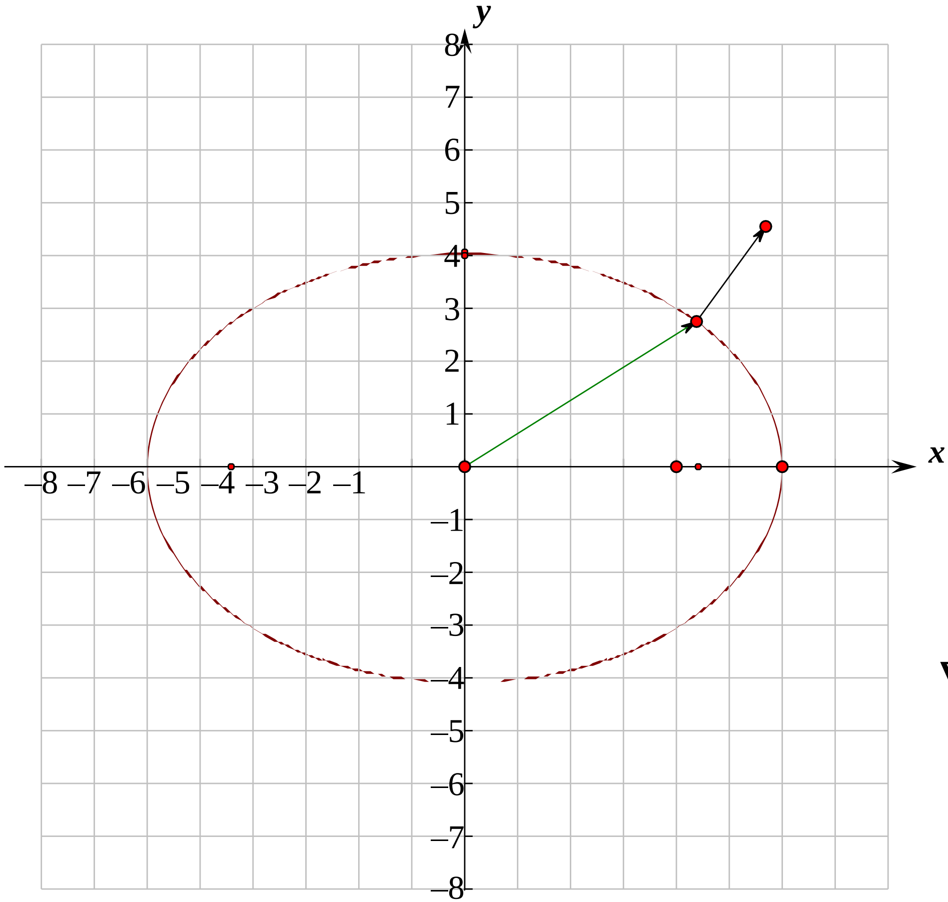
$$f(x,y)=\frac{1}{2}\left(x^2+y^2\right)=\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}$$

$$\mathbf{x}=\begin{bmatrix}x & y\end{bmatrix}^{\mathrm{T}}$$

$$\nabla_{\mathbf{x}} f(\mathbf{x})=\mathbf{x}$$

常用例子-3

- 函数 $f(x,y)=\dfrac{x^2}{a^2}+\dfrac{y^2}{b^2}$ 的梯度（二次型）



$$f(x,y)=\frac{x^2}{a^2}+\frac{y^2}{b^2}$$

$$=\begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} 1/a^2 & 0 \\ 0 & 1/b^2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

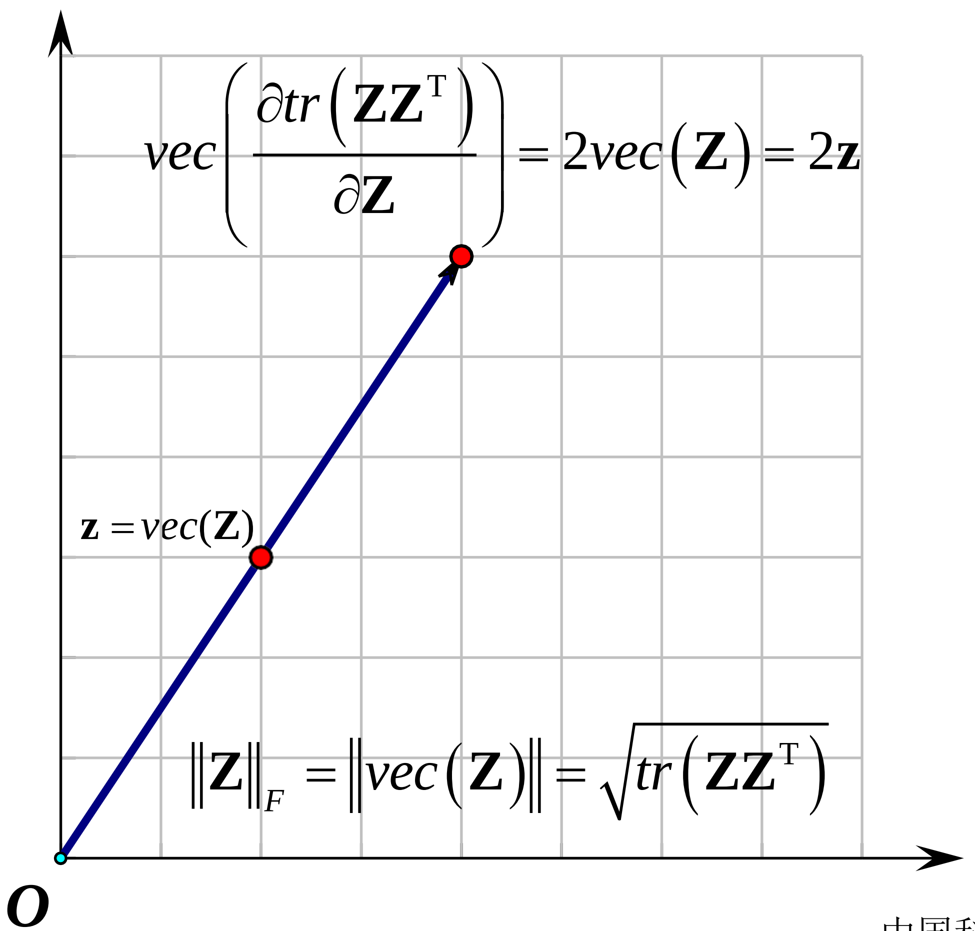
$$=\mathbf{x}^{\mathrm{T}}\begin{bmatrix} 1/a^2 & 0 \\ 0 & 1/b^2 \end{bmatrix}\mathbf{x}$$

$$\mathbf{x}=\begin{bmatrix} x & y \end{bmatrix}^{\mathrm{T}}$$

$$\nabla_{\mathbf{x}} f(\mathbf{x})=2\begin{bmatrix} 1/a^2 & 0 \\ 0 & 1/b^2 \end{bmatrix}\mathbf{x}=\begin{bmatrix} \dfrac{2x}{a^2} \\ \dfrac{2y}{b^2} \end{bmatrix}$$

常用例子-4

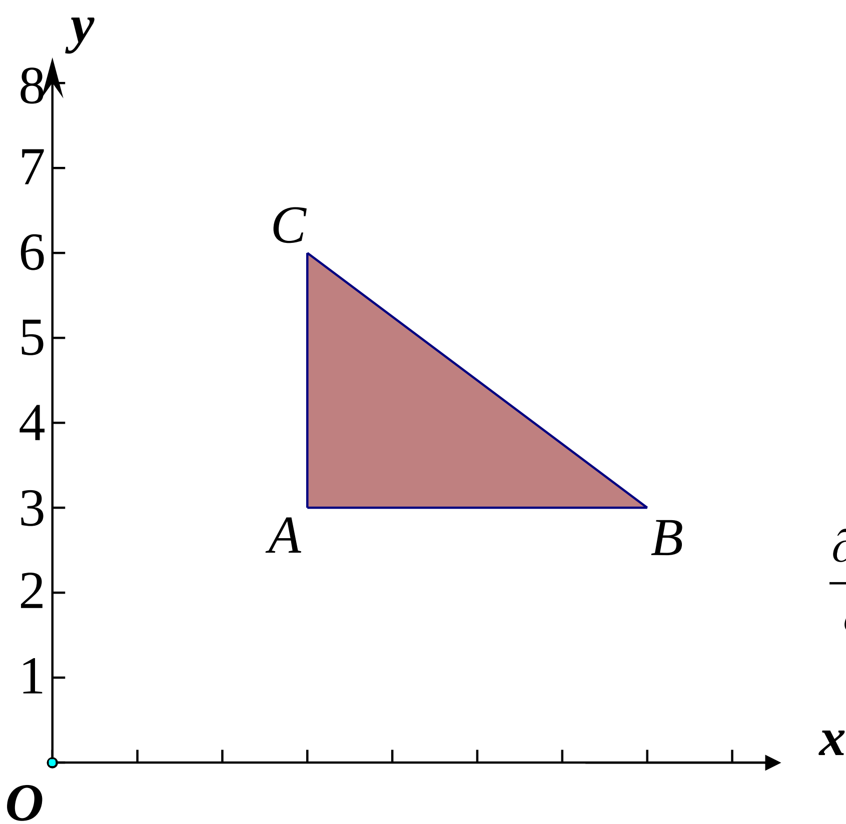
- 迹函数相对于矩阵的梯度



$$\frac{\partial tr\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)}{\partial \mathbf{Z}} = \frac{\partial tr\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)}{\partial \mathbf{Z}} = 2\mathbf{Z}$$

常用例子-5

- 行列式相对于矩阵的梯度 $\frac{\partial|\mathbf{Z}|}{\partial\mathbf{Z}}=|\mathbf{Z}|\mathbf{Z}^{-\mathrm{T}}$



$$\mathbf{Z}=\begin{bmatrix}1 & 1 & 1\\ x_A & x_B & x_C\\ y_A & y_B & y_C\end{bmatrix}=\begin{bmatrix}1 & 1 & 1\\ 3 & 7 & 3\\ 3 & 3 & 6\end{bmatrix}$$

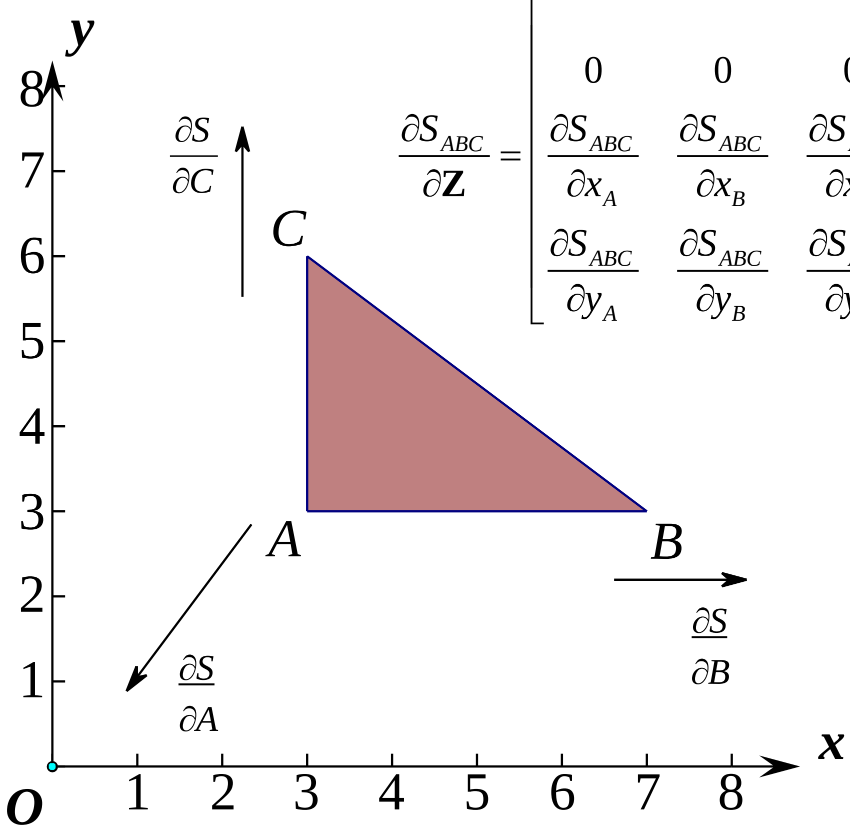
$$S_{ABC}=\frac{1}{2}|\mathbf{Z}|=\frac{1}{2}\det\left(\begin{bmatrix}1 & 1 & 1\\ 3 & 7 & 3\\ 3 & 3 & 6\end{bmatrix}\right)$$

$$\frac{\partial|\mathbf{Z}|}{\partial\mathbf{Z}}=\begin{vmatrix}1 & 1 & 1\\ 3 & 7 & 3\\ 3 & 3 & 6\end{vmatrix}\begin{bmatrix}1 & 1 & 1\\ 3 & 7 & 3\\ 3 & 3 & 6\end{bmatrix}^{-\mathrm{T}}=\begin{bmatrix}33 & -9 & -12\\ -3 & 3 & 0\\ -4 & 0 & 4\end{bmatrix}$$

常用例子-5

- 行列式相对于矩阵的梯度 $\dfrac{\partial|\mathbf{Z}|}{\partial\mathbf{Z}} = |\mathbf{Z}|\mathbf{Z}^{-\mathrm{T}}$

$$\frac{\partial S_{ABC}}{\partial \mathbf{Z}} = \begin{bmatrix} 0 & 0 & 0 \\ \dfrac{\partial S_{ABC}}{\partial x_A} & \dfrac{\partial S_{ABC}}{\partial x_B} & \dfrac{\partial S_{ABC}}{\partial x_C} \\ \dfrac{\partial S_{ABC}}{\partial y_A} & \dfrac{\partial S_{ABC}}{\partial y_B} & \dfrac{\partial S_{ABC}}{\partial y_C} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 \\ -3 & 3 & 0 \\ -4 & 0 & 4 \end{bmatrix}$$

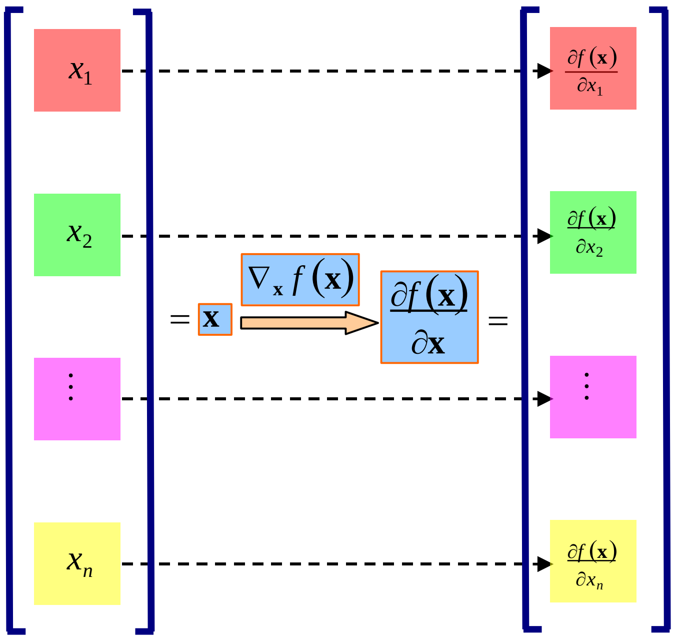


$$\begin{bmatrix} \dfrac{\partial S_{ABC}}{\partial A} & \dfrac{\partial S_{ABC}}{\partial B} & \dfrac{\partial S_{ABC}}{\partial C} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial S_{ABC}}{\partial x_A} & \dfrac{\partial S_{ABC}}{\partial x_B} & \dfrac{\partial S_{ABC}}{\partial x_C} \\ \dfrac{\partial S_{ABC}}{\partial y_A} & \dfrac{\partial S_{ABC}}{\partial y_B} & \dfrac{\partial S_{ABC}}{\partial y_C} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -3 & 3 & 0 \\ -4 & 0 & 4 \end{bmatrix}$$

实值函数相对于实向量的梯度

实值标量函数对于实向量的梯度

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} & \dfrac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}^{\mathrm{T}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$$



1. 以列（行）向量为自变量的标量函数，其对于自变量的梯度仍然为一阶数相同的列（行）向量
2. 梯度的每个分量代表着函数在该分量方向上的变化率。

实值函数相对于实向量的梯度

实值向量函数对于实向量的梯度

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_m(\mathbf{x}) \end{bmatrix}^{\mathrm{T}}$$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} \\ \dfrac{\partial f_2(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \dfrac{\partial f_m(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \dfrac{\partial \mathbf{f}(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \dfrac{\partial f_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_m(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_m(\mathbf{x})}{\partial x_2} & \cdots & \dfrac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} == \left(\frac{\partial \mathbf{f}^{\mathrm{T}}(\mathbf{x})}{\partial \mathbf{x}} \right)^{\mathrm{T}}$$

1. 向量函数对于向量的求导，相当于向量函数中的每一个分量函数对向量求导或者向量函数对向量的每一个分量求导。
2. 行向量函数对列向量自变量求导形成矩阵；列向量函数对行向量自变量求导也可以形成矩阵。

实值函数相对于实向量的梯度

- 例1 $f(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{a} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_1} & \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_2} & \cdots & \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_n} \end{bmatrix}^{\mathrm{T}}$$

$$= \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^{\mathrm{T}} = \mathbf{a}$$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} = \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_1} & \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_2} & \cdots & \frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{x}}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \mathbf{a}^{\mathrm{T}}$$

实值函数相对于实向量的梯度

- 例2 $\mathbf{f}(\mathbf{x}) = \mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^{\mathrm{T}}$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}^{\mathrm{T}}} = \mathbf{I}_{n\times n}$$

$$\frac{\partial (\mathbf{f}(\mathbf{x}))^{\mathrm{T}}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^{\mathrm{T}}}{\partial \mathbf{x}} = \mathbf{I}_{n\times n} = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} \right)^{\mathrm{T}}$$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}} = vec(\mathbf{I}_{n\times n})$$

$$\frac{\partial (\mathbf{f}(\mathbf{x}))^{\mathrm{T}}}{\partial \mathbf{x}^{\mathrm{T}}} = \frac{\partial \mathbf{x}^{\mathrm{T}}}{\partial \mathbf{x}^{\mathrm{T}}} = \left(vec(\mathbf{I}_{n\times n})\right)^{\mathrm{T}} = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right)^{\mathrm{T}}$$

实值函数相对于实向量的梯度

- 例3 $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{A}(1,:)\mathbf{x} \\ \mathbf{A}(2,:)\mathbf{x} \\ \vdots \\ \mathbf{A}(m,:)\mathbf{x} \end{bmatrix}$$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial \mathbf{A}(1,:)\mathbf{x}}{\partial \mathbf{x}^{\mathrm{T}}} \\ \frac{\partial \mathbf{A}(2,:)\mathbf{x}}{\partial \mathbf{x}^{\mathrm{T}}} \\ \vdots \\ \frac{\partial \mathbf{A}(m,:)\mathbf{x}}{\partial \mathbf{x}^{\mathrm{T}}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}(1,:) \\ \mathbf{A}(2,:) \\ \vdots \\ \mathbf{A}(m,:) \end{bmatrix} = \mathbf{A}$$

实值函数相对于实向量的梯度

- 例3 $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$

方法2：

$$\mathbf{f}\left(\mathbf{x}\right)=\mathbf{A}\mathbf{x}=x_1\mathbf{A}\left(:,1\right)+x_2\mathbf{A}\left(:,2\right)+\cdots+x_n\mathbf{A}\left(:,n\right)$$

$$\frac{\partial \mathbf{f}\left(\mathbf{x}\right)}{\partial \mathbf{x}^{\mathrm{T}}}=\begin{bmatrix}\frac{\partial \mathbf{f}\left(\mathbf{x}\right)}{\partial x_1} & \frac{\partial \mathbf{f}\left(\mathbf{x}\right)}{\partial x_2} & \cdots & \frac{\partial \mathbf{f}\left(\mathbf{x}\right)}{\partial x_n}\end{bmatrix}$$

$$=\begin{bmatrix}\mathbf{A}\left(:,1\right) & \mathbf{A}\left(:,2\right) & \cdots & \mathbf{A}\left(:,n\right)\end{bmatrix}$$

$$=\mathbf{A}$$

实值函数相对于实向量的梯度

- 例4　$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$

方法1：逐元素求导

$$f(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j$$

与x_k相关的项 $\displaystyle\sum_{j=1, j\neq k}^{n} a_{kj} x_k x_j + \sum_{i=1, i\neq k}^{n} a_{ik} x_i x_k + a_{kk} x_k^2$

$$\frac{\partial f(\mathbf{x})}{\partial x_k} = \sum_{j=1, j\neq k}^{n} a_{kj} x_j + \sum_{i=1, i\neq k}^{n} a_{ik} x_i + 2a_{kk} x_k = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$$

$$= \mathbf{A}(k,:)\mathbf{x} + \mathbf{A}(:,k)^{\mathrm{T}}\mathbf{x} = \left(\mathbf{A}(k,:) + \mathbf{A}(:,k)^{\mathrm{T}}\right)\mathbf{x}$$

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{A}^{\mathrm{T}}\mathbf{x} \overset{\text{当}\mathbf{A}\text{为对称矩阵时}}{=} 2\mathbf{A}\mathbf{x}$$

实值函数相对于实向量的梯度

常用梯度公式及求导法则

- 线性法则

$$\frac{\partial\left(c_1 f(\mathbf{x})+c_2 g(\mathbf{x})\right)}{\partial \mathbf{x}}=c_1\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}+c_2\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$$

- 乘积法则

$$\frac{\partial\left(f(\mathbf{x})g(\mathbf{x})\right)}{\partial \mathbf{x}}=g(\mathbf{x})\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}+f(\mathbf{x})\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$$

- 商法则

$$\frac{\partial\left(f(\mathbf{x})/g(\mathbf{x})\right)}{\partial \mathbf{x}}=\frac{1}{g^2(\mathbf{x})}\left(g(\mathbf{x})\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}-f(\mathbf{x})\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\right)$$

- 链式法则

$$\frac{\partial\left(f(\mathbf{g}(\mathbf{x}))\right)}{\partial \mathbf{x}}=\frac{\partial \mathbf{g}^{\mathrm{T}}(\mathbf{x})}{\partial \mathbf{x}}\frac{\partial f(\mathbf{g})}{\partial \mathbf{g}}$$

实值函数相对于实向量的梯度

- 常用梯度公式及求导法则

  - 函数 $f(\mathbf{x})=\mathbf{g}(\mathbf{x})^{\mathrm{T}}\mathbf{h}(\mathbf{x})$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}=\frac{\partial \mathbf{g}(\mathbf{x})^{\mathrm{T}}}{\partial \mathbf{x}}\mathbf{h}(\mathbf{x})+\frac{\partial \mathbf{h}(\mathbf{x})^{\mathrm{T}}}{\partial \mathbf{x}}\mathbf{g}(\mathbf{x})$$

  - 函数 $f(\mathbf{x})=\mathbf{x}^{\mathrm{T}}\mathbf{x}=\|\mathbf{x}\|^{2}$

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}}=2\mathbf{x}$$

实值函数相对于矩阵的梯度

- 实值函数$f(\mathbf{A})$相对于其自变量$m \times n$矩阵 $\mathbf{A}$的梯度定义为

$$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial f(\mathbf{A})}{\partial a_{11}} & \frac{\partial f(\mathbf{A})}{\partial a_{12}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial a_{1n}} \\ \frac{\partial f(\mathbf{A})}{\partial a_{21}} & \frac{\partial f(\mathbf{A})}{\partial a_{22}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{A})}{\partial a_{m1}} & \frac{\partial f(\mathbf{A})}{\partial a_{m2}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial a_{mn}} \end{bmatrix} = \nabla_{\mathbf{A}} f(\mathbf{A})$$

1.实值函数相对于矩阵的梯度仍然为一与矩阵同阶的矩阵

2.实值函数相对于矩阵的梯度矩阵的每一个分量对应于该函数在矩阵的每一个分量的变化率。

实值函数相对于矩阵的梯度

- 实值函数相对于矩阵的求导和相对于向量的求导本质上没有区别

$$\frac{\partial f\left(vec\left(\mathbf{A}\right)\right)}{\partial vec\left(\mathbf{A}\right)} = \begin{bmatrix} \dfrac{\partial f\left(\mathbf{A}\right)}{\partial a_{11}} \\ \vdots \\ \dfrac{\partial f\left(\mathbf{A}\right)}{\partial a_{m1}} \\ \vdots \\ \dfrac{\partial f\left(\mathbf{A}\right)}{\partial a_{1n}} \\ \vdots \\ \dfrac{\partial f\left(\mathbf{A}\right)}{\partial a_{mn}} \end{bmatrix} = \nabla_{vec\left(\mathbf{A}\right)} f\left(vec\left(\mathbf{A}\right)\right)$$

$$\nabla_{\mathbf{A}} f\left(\mathbf{A}\right) = unvec\left(\nabla_{vec\left(\mathbf{A}\right)} f\left(vec\left(\mathbf{A}\right)\right)\right)$$

实值函数相对于矩阵的梯度

- 例：求 $f(\mathbf{A}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$ 相对于矩阵 $\mathbf{A}$ 的梯度。

方法1：逐元素求导

$$f(\mathbf{A}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_i y_j$$

$$\frac{\partial f(\mathbf{A})}{\partial a_{ij}} = x_i y_j$$

$$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix} = \mathbf{x}\mathbf{y}^{\mathrm{T}}$$

实值函数相对于矩阵的梯度

- 例：求 $f(\mathbf{A}) = \mathbf{x}^T \mathbf{A}\mathbf{y}$ 相对于矩阵 $\mathbf{A}$ 的梯度。

  方法2：向量求导法

$$\boxed{f(\mathbf{A}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y} = vec(\mathbf{A})^{\mathrm{T}}\, kron(\mathbf{y},\mathbf{x}) = kron(\mathbf{y}^{\mathrm{T}},\mathbf{x}^{\mathrm{T}})\, vec(\mathbf{A})}$$

$$\frac{\partial f(vec(\mathbf{A}))}{\partial vec(\mathbf{A})} = kron(\mathbf{y},\mathbf{x})$$

$$\boxed{vec(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A})\, vec(\mathbf{B})}$$

$$\nabla_{\mathbf{A}} f(\mathbf{A}) = unvec\left(\nabla_{vec(\mathbf{A})} f(vec(\mathbf{A}))\right) = unvec(kron(\mathbf{y},\mathbf{x})) = \mathbf{x}\mathbf{y}^{\mathrm{T}}$$

$$kron(\mathbf{X},\mathbf{Y}) = \mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & \cdots & x_{1n}\mathbf{Y} \\ \vdots & \ddots & \vdots \\ x_{m1}\mathbf{Y} & \cdots & x_{mn}\mathbf{Y} \end{bmatrix}$$

实值函数相对于实向量的梯度

常用梯度公式及求导法则

- 线性法则

$$\frac{\partial\left(c_1 f(\mathbf{A})+c_2 g(\mathbf{A})\right)}{\partial \mathbf{A}}=c_1\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}+c_2\frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}$$

- 乘积法则

$$\frac{\partial f(\mathbf{A})g(\mathbf{A})}{\partial \mathbf{A}}=g(\mathbf{A})\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}+f(\mathbf{A})\frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}$$

- 商法则

$$\frac{\partial\left(f(\mathbf{A})/g(\mathbf{A})\right)}{\partial \mathbf{A}}=\frac{1}{g^2(\mathbf{A})}\left(g(\mathbf{A})\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}-f(\mathbf{A})\frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}\right)$$

- 链式法则

$$\frac{\partial g\left(f(\mathbf{A})\right)}{\partial \mathbf{A}}=\frac{dg(y)}{dy}\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}$$

矩阵微分

- 对于一个以向量 $\mathbf{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$ 为变量的实值函数 $f(\mathbf{x})$，其微分公式定义如下

$$df(\mathbf{x})=\sum_{i=1}^{n}\frac{\partial f(\mathbf{x})}{\partial x_i}dx_i$$

- 对于一个以 $m\times n$ 阶矩阵 $\mathbf{X}$ 为变量的实值函数 $f(\mathbf{x})$，其微分公式定义如下

$$df(\mathbf{X})=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial f(\mathbf{X})}{\partial x_{ij}}dx_{ij}$$

矩阵微分

GIPAS

重要公式

$$
d\mathbf{X} = \begin{bmatrix} dx_{11} & dx_{12} & \cdots & dx_{1n} \\ dx_{21} & dx_{22} & \cdots & dx_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ dx_{m1} & dx_{m2} & \cdots & dx_{mn} \end{bmatrix} = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{e}_i \hat{\mathbf{e}}_j^{\mathrm{T}} dx_{ij}
$$

$$
\frac{\partial \mathbf{X}}{\partial x_{ij}} = \mathbf{e}_i \hat{\mathbf{e}}_j^{\mathrm{T}}
$$

非常重要的一个公式

$$
\boxed{df(\mathbf{X}) = tr\left(\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)^{\mathrm{T}} d\mathbf{X} \right)}
$$

矩阵微分

- 矩阵微分满足如下规则
  - 线性法则

$$d(\alpha\mathbf{X})=\alpha d\mathbf{X} \quad d(\mathbf{X}+\mathbf{Y})=d\mathbf{X}+d\mathbf{Y}$$

  - 乘积法则

$$\boxed{d(\mathbf{XY})=(d\mathbf{X})\mathbf{Y}+\mathbf{X}(d\mathbf{Y})}$$

证明：令 $\mathbf{Z}=\mathbf{XY}$，则有

$$z_{ij}=\sum_{k=1}^{n}x_{ik}y_{kj}$$

根据标量微分的乘积法则有

$$d(z_{ij})=d\sum_{k=1}^{n}x_{ik}y_{kj}=\sum_{k=1}^{n}d(x_{ik}y_{kj})=\sum_{k=1}^{n}((dx_{ik})y_{kj}+x_{ik}(dy_{kj}))$$

从而 $d(\mathbf{XY})=(d\mathbf{X})\mathbf{Y}+\mathbf{X}(d\mathbf{Y})$

,

矩阵微分

常用矩阵微分公式

- 矩阵的逆的微分

$$\boxed{d\left(\mathbf{X}^{-1}\right) = -\mathbf{X}^{-1}\left(d\mathbf{X}\right)\mathbf{X}^{-1}}$$

请尝试证明！

- 矩阵微分算子和迹算子的可交换性

$$\boxed{d\left(tr\left(\mathbf{X}\right)\right) = tr\left(d\left(\mathbf{X}\right)\right) = \sum_{i=1}^{n} dx_{ii}}$$

,

迹函数的矩阵梯度

- 任意 $n \times n$ 矩阵 $\mathbf{A}$ 的迹的定义为其对角线的和，即

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

因此，

$$\frac{\partial(tr(\mathbf{A}))}{\partial a_{ij}} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

这意味着

$$\frac{\partial(tr(\mathbf{A}))}{\partial \mathbf{A}} = \mathbf{I}_{n \times n}$$

迹函数的矩阵梯度

- 例：求 $tr(\mathbf{AXB})$ 相对于矩阵 $\mathbf{X}$ 的梯度。其中 $\mathbf{A},\mathbf{X},\mathbf{B}$ 分别是 $p\times m, m\times n, n\times p$ 矩阵

方法1：

$$d\left(tr\left(\mathbf{AXB}\right)\right)=tr\left(\mathbf{A}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{e}_i\hat{\mathbf{e}}_j^{\mathrm{T}}dx_{ij}\right)\mathbf{B}\right)=\sum_{i=1}^{m}\sum_{j=1}^{n}tr\left(\mathbf{Ae}_i\hat{\mathbf{e}}_j^{\mathrm{T}}\mathbf{B}\right)dx_{ij}$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\mathbf{e}}_j^{\mathrm{T}}\mathbf{BAe}_i dx_{ij}=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{BA}\right)_{ji}dx_{ij}=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\right)_{ij}dx_{ij}$$

因此，$\dfrac{\partial tr\left(\mathbf{AXB}\right)}{\partial x_{ij}}=\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\right)_{ij}$

即，$\dfrac{\partial tr\left(\mathbf{AXB}\right)}{\partial \mathbf{X}}=\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}$

$$\boxed{tr\left(\mathbf{AB}\right)=tr\left(\mathbf{BA}\right)}$$

迹函数的矩阵梯度

- 例：求 $tr(\mathbf{AXB})$ 相对于矩阵 $\mathbf{X}$ 的梯度。其中 $\mathbf{A},\mathbf{X},\mathbf{B}$ 分别是 $p\times m, m\times n, n\times p$ 矩阵

方法2：

$$d\left(tr(\mathbf{AXB})\right)=tr\left(\mathbf{A}(d\mathbf{X})\mathbf{B}\right)=tr\left(\mathbf{BA}(d\mathbf{X})\right)$$

因此，

$$\frac{\partial tr(\mathbf{AXB})}{\partial \mathbf{X}}=(\mathbf{BA})^{\mathrm{T}}=\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}$$

迹函数的矩阵梯度

- 例：求 $tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)$ 相对于矩阵 $\mathbf{X}$ 的梯度。其中 $\mathbf{A},\mathbf{X},\mathbf{B}$ 分别是 $p\times n, n\times n, n\times p$ 矩阵

方法1：$d\left(tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)\right)=tr\left(d\left(\mathbf{AX}^{-1}\mathbf{B}\right)\right)=tr\left(\mathbf{A}\left(d\mathbf{X}^{-1}\right)\mathbf{B}\right)$

$$=tr\left(-\mathbf{A}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{X}^{-1}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}\mathbf{X}^{-1}dx_{ij}\right)\mathbf{B}\right)=-\sum_{i=1}^{n}\sum_{j=1}^{n}tr\left(\mathbf{AX}^{-1}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}\mathbf{X}^{-1}\mathbf{B}\right)dx_{ij}$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{e}_j^{\mathrm{T}}\mathbf{X}^{-1}\mathbf{BAX}^{-1}\mathbf{e}_i dx_{ij}=-\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{X}^{-1}\mathbf{BAX}^{-1}\right)_{ji}dx_{ij}$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{X}^{-\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}\right)_{ij}dx_{ij}$$

因此，$$\frac{\partial tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)}{\partial x_{ij}}=-\left(\mathbf{X}^{-\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}\right)_{ij}$$

即，$$\frac{\partial tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)}{\partial \mathbf{X}}=-\mathbf{X}^{-\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}$$

迹函数的矩阵梯度

- 例：求 $tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)$ 相对于矩阵 $\mathbf{X}$ 的梯度。其中 $\mathbf{A},\mathbf{X},\mathbf{B}$ 分别是 $p\times n, n\times n, n\times p$ 矩阵

方法2：

$$d\left(tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)\right)=tr\left(\mathbf{A}\left(d\mathbf{X}^{-1}\right)\mathbf{B}\right)$$
$$=-tr\left(\mathbf{AX}^{-1}\left(d\mathbf{X}\right)\mathbf{X}^{-1}\mathbf{B}\right)=-tr\left(\mathbf{X}^{-1}\mathbf{BAX}^{-1}d\mathbf{X}\right)$$

因此，$\dfrac{\partial tr\left(\mathbf{AX}^{-1}\mathbf{B}\right)}{\partial \mathbf{X}}=-\left(\mathbf{X}^{-1}\mathbf{BAX}^{-1}\right)^{\mathrm{T}}=-\mathbf{X}^{-\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}$

迹函数的矩阵梯度

- 例：求 $tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$ 相对于矩阵 $\mathbf{X}$ 的梯度

方法1：$d\left(tr\left(\mathbf{X}^{T}\mathbf{X}\right)\right) = tr\left(d\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\right) = tr\left(\left(d\mathbf{X}^{\mathrm{T}}\right)\mathbf{X} + \mathbf{X}^{\mathrm{T}} d\mathbf{X}\right)$

$$= tr\left(\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\mathbf{e}}_j \mathbf{e}_i^{\mathrm{T}} dx_{ij}\right)\mathbf{X} + \mathbf{X}^{\mathrm{T}}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{e}_i \hat{\mathbf{e}}_j^{\mathrm{T}} dx_{ij}\right)\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} tr\left(\hat{\mathbf{e}}_j \mathbf{e}_i^{\mathrm{T}}\mathbf{X}\right) dx_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{e}_i \hat{\mathbf{e}}_j^{\mathrm{T}}\right) dx_{ij}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{e}_i^{\mathrm{T}}\mathbf{X}\hat{\mathbf{e}}_j dx_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{\mathbf{e}}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{e}_i dx_{ij}$$

$$= 2\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{e}_i^{\mathrm{T}}\mathbf{X}\hat{\mathbf{e}}_j dx_{ij} = 2\sum_{i=1}^{m}\sum_{j=1}^{n} \left(\mathbf{X}\right)_{ij} dx_{ij}$$

因此，$\dfrac{\partial tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)}{\partial x_{ij}} = 2\left(\mathbf{X}\right)_{ij}$ 即 $\dfrac{\partial tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)}{\partial \mathbf{X}} = \dfrac{\partial tr\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)}{\partial \mathbf{X}} = 2\mathbf{X}$

迹函数的矩阵梯度

- 例：求$tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$相对于矩阵$\mathbf{X}$ 的梯度

方法2：

$$d\left(tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\right)=tr\left(d\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\right)=tr\left(\left(d\mathbf{X}^{\mathrm{T}}\right)\mathbf{X}+\mathbf{X}^{\mathrm{T}}d\mathbf{X}\right)=tr\left(\left(d\mathbf{X}^{\mathrm{T}}\right)\mathbf{X}\right)+tr\left(\mathbf{X}^{\mathrm{T}}d\mathbf{X}\right)$$

而

$$tr\left(\left(d\mathbf{X}^{\mathrm{T}}\right)\mathbf{X}\right)=tr\left(\left(d\mathbf{X}^{\mathrm{T}}\right)\mathbf{X}\right)^{\mathrm{T}}=tr\left(\mathbf{X}^{\mathrm{T}}d\mathbf{X}\right)$$

因此，$d\left(tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\right)=2tr\left(\mathbf{X}^{\mathrm{T}}d\mathbf{X}\right)$

所以，$\dfrac{\partial tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)}{\partial\mathbf{X}}=\dfrac{\partial tr\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)}{\partial\mathbf{X}}=2\left(\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}}=2\mathbf{X}$

常用公式 $\boxed{tr\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)=\left\|\mathbf{X}\right\|_F^2}$

迹函数的矩阵梯度

- 例：求$tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$相对于矩阵$\mathbf{X}$ 的梯度

方法3：

$$tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)=\left\|\mathbf{X}\right\|_F^2=vec\left(\mathbf{X}\right)^{\mathrm{T}}vec\left(\mathbf{X}\right)$$

$$\frac{\partial tr\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)}{\partial\mathbf{X}}=uvec\left(\frac{\partial\left(vec\left(\mathbf{X}\right)^{\mathrm{T}}vec\left(\mathbf{X}\right)\right)}{\partial vec\left(\mathbf{X}\right)}\right)$$

$$=2uvec\left(vec\left(\mathbf{X}\right)\right)=2\mathbf{X}$$

实值函数相对于矩阵的梯度

- 例：求 $f(\mathbf{A})=\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$ 相对于矩阵 $\mathbf{A}$ 的梯度。

方法3：

$$df(\mathbf{A})=d\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}\right)=d\left(tr\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}\right)\right)$$

$$=d\left(tr\left(\mathbf{y}\mathbf{x}^{\mathrm{T}}\mathbf{A}\right)\right)=tr\left(\mathbf{y}\mathbf{x}^{\mathrm{T}}d\mathbf{A}\right)$$

$$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}=\left(\mathbf{y}\mathbf{x}^{\mathrm{T}}\right)^{\mathrm{T}}=\mathbf{x}\mathbf{y}^{\mathrm{T}}$$

实值函数相对于实向量的梯度

- 例4　$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$

方法2：

$$d\left(f(\mathbf{x})\right) = d\left(tr\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right)\right) = tr\left(d\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right)\right)$$

$$= tr\left(\left(d\mathbf{x}^{\mathrm{T}}\right)\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}d\mathbf{x}\right) = tr\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}d\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}d\mathbf{x}\right)$$

$$= tr\left(\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\right)d\mathbf{x}\right)$$

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

迹函数的矩阵梯度

关于迹的梯度的常用公式

$$1.\ \frac{\partial tr(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{I}_{n\times n}$$

$$2.\ \frac{\partial tr(\mathbf{X}^{-1})}{\partial \mathbf{X}} = -(\mathbf{X}^{-2})^{\mathrm{T}}$$

$$3.\ \frac{\partial tr(\mathbf{XA})}{\partial \mathbf{X}} = \frac{\partial tr(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}^{\mathrm{T}}$$

$$4.\ \frac{\partial tr(\mathbf{X}^{\mathrm{T}}\mathbf{A})}{\partial \mathbf{X}} = \frac{\partial tr(\mathbf{AX}^{\mathrm{T}})}{\partial \mathbf{X}} = \mathbf{A}$$

$$5.\ \frac{\partial tr(\mathbf{X}^{2})}{\partial \mathbf{X}} = 2\mathbf{X}^{\mathrm{T}}$$

$$6.\ \frac{\partial tr(\mathbf{AX}^{-1})}{\partial \mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{AX}^{-1})^{\mathrm{T}}$$

$$7.\ \frac{\partial tr(\mathbf{X}^{\mathrm{T}}\mathbf{AX})}{\partial \mathbf{X}} = (\mathbf{A}+\mathbf{A}^{\mathrm{T}})\mathbf{X}$$

$$8.\ \frac{\partial tr(\mathbf{XAX}^{\mathrm{T}})}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{A}+\mathbf{A}^{\mathrm{T}})$$

$$9.\ \frac{\partial tr(\mathbf{AXX}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}})}{\partial \mathbf{X}} = 2\mathbf{A}^{\mathrm{T}}\mathbf{AX}$$

$$10.\ \frac{\partial tr(\mathbf{AXX}^{\mathrm{T}}\mathbf{B})}{\partial \mathbf{X}} = (\mathbf{BA}+\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}})\mathbf{X}$$

行列式的矩阵梯度

- 同矩阵的迹一样，矩阵的行列式也是以矩阵为自变量的标量函数，因此同样存在行列式相对于矩阵梯度问题。由于行列式明显的几何意义（体积），它在很多学科都有着广泛的应用。行列式相对于矩阵的梯度问题也几乎贯穿所有行列式优化问题。

行列式定义：$|\mathbf{X}| = \sum_{i=1}^{n} x_{ij} C_{ij}$ 或者 $|\mathbf{X}| = \sum_{j=1}^{n} x_{ij} C_{ij}$

代数余子式矩阵：$\mathbf{C} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix}$, $C_{ij} = (-1)^{i+j} M_{ij}$

伴随矩阵：$\mathbf{X}^* = \mathbf{C}^{\mathrm{T}}$

行列式的矩阵梯度

- 对于一个方阵 **X**，有

$$\frac{\partial|\mathbf{X}|}{\partial x_{ij}} = C_{ij}$$

因此，$$\boxed{\frac{\partial|\mathbf{X}|}{\partial \mathbf{X}} = \mathbf{C} = \left(\mathbf{X}^{*}\right)^{\mathrm{T}} = |\mathbf{X}|\mathbf{X}^{-\mathrm{T}}}$$

$$\mathbf{X}^{-1} = \frac{\mathbf{X}^{*}}{|\mathbf{X}|}$$

微分形式：$$d|\mathbf{X}| = tr\left(|\mathbf{X}|\mathbf{X}^{-1}d\mathbf{X}\right)$$

行列式的矩阵梯度

- 例. $|\mathbf{AXB}|$ 相对于矩阵 $\mathbf{X}$ 的梯度

$$d|\mathbf{AXB}| = tr\left(\left(|\mathbf{AXB}|(\mathbf{AXB})^{-1}\right)d(\mathbf{AXB})\right)$$
$$= tr\left(\left(|\mathbf{AXB}|(\mathbf{AXB})^{-1}\right)\mathbf{A}(d\mathbf{X})\mathbf{B}\right)$$
$$= tr\left(\left(|\mathbf{AXB}|\mathbf{B}(\mathbf{AXB})^{-1}\right)\mathbf{A}(d\mathbf{X})\right)$$

因此，

$$\frac{\partial|\mathbf{AXB}|}{\partial\mathbf{X}} = \left(\left(|\mathbf{AXB}|\mathbf{B}\left(\mathbf{AXB}\right)^{-1}\right)\mathbf{A}\right)^{\mathrm{T}}$$
$$= |\mathbf{AXB}|\,\mathbf{A}^{\mathrm{T}}\left(\mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{B}^{\mathrm{T}}$$

行列式的矩阵梯度

常用的行列书关于矩阵求导的公式

1. $\dfrac{\partial\left|\mathbf{X}\right|}{\partial\mathbf{X}} = \left|\mathbf{X}\right|\mathbf{X}^{-\mathrm{T}}$

2. $\dfrac{\partial\left|\mathbf{X}^{-1}\right|}{\partial\mathbf{X}} = -\dfrac{\mathbf{X}^{-\mathrm{T}}}{\left|\mathbf{X}\right|}$

3. $\dfrac{\partial\log\left|\mathbf{X}\right|}{\partial\mathbf{X}} = \mathbf{X}^{-\mathrm{T}}$

4. $\dfrac{\partial\left|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right|}{\partial\mathbf{X}} = 2\left|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right|\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X} \qquad \mathrm{rank}\left(\mathbf{X}_{m\times n}\right) = m$

5. $\dfrac{\partial\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right|}{\partial\mathbf{X}} = 2\left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right|\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \qquad \mathrm{rank}\left(\mathbf{X}_{m\times n}\right) = n$

行列式的矩阵梯度

常用的行列书关于矩阵求导的公式

$$6.\ \frac{\partial\left|\mathbf{X}^{2}\right|}{\partial\mathbf{X}}=2\left|\mathbf{X}\right|^{2}\mathbf{X}^{-\mathrm{T}}\qquad \operatorname{rank}\left(\mathbf{X}_{m\times m}\right)=m$$

$$7.\ \frac{\partial\left|\mathbf{AXB}\right|}{\partial\mathbf{X}}=\left|\mathbf{AXB}\right|\mathbf{A}^{\mathrm{T}}\left(\mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{B}^{\mathrm{T}}$$

$$8.\ \frac{\partial\left|\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right|}{\partial\mathbf{X}}=\left|\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right|\left(\mathbf{AX}\left(\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right)^{-1}+\mathbf{A}^{T}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)\qquad \mathbf{A}\text{不是对称矩阵}$$

$$9.\ \frac{\partial\left|\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right|}{\partial\mathbf{X}}=2\left|\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right|\mathbf{AX}\left(\mathbf{X}^{\mathrm{T}}\mathbf{AX}\right)^{-1}\qquad \mathbf{A}\text{是对称矩阵}$$

Hessian矩阵

- 实值函数$f(\mathbf{x})$相对于实向量$\mathbf{x}$的二阶偏导数（称为Hessian矩阵），定义为：

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} = \frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}}\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)$$

或者写为梯度的梯度，

$$\nabla_{\mathbf{x}}^2 f(\mathbf{x}) = \nabla_{\mathbf{x}^{\mathrm{T}}}\left(\nabla_{\mathbf{x}} f(\mathbf{x})\right)$$

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

Hessian矩阵有许多重要应用，比如经典的SURF特征点检测算子就直接以其为核心：为了实现尺度不变性的特征点检测与匹配，SURF算法则先利用Hessian矩阵确定候选点，然后进行非极大抑制，大大降低了时间复杂度。

,

Hessian矩阵

- 实值函数$f\left(\mathbf{x}\right)$的泰勒级数展开

方向导数

$$f\left(\mathbf{x}+\Delta\mathbf{x}\right)=f\left(\mathbf{x}\right)+\boxed{\left(\nabla_{\mathbf{x}}f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\Delta\mathbf{x}}+\frac{1}{2}\boxed{\left(\Delta\mathbf{x}\right)^{\mathrm{T}}\nabla_{\mathbf{x}}^{2}f\left(\mathbf{x}\right)\Delta\mathbf{x}}+\cdots$$

$$\frac{\partial\left(\mathbf{g}^{\mathrm{T}}\left(\mathbf{x}\right)\mathbf{h}\left(\mathbf{x}\right)\right)}{\partial\mathbf{x}}=\frac{\partial\mathbf{g}\left(\mathbf{x}\right)^{\mathrm{T}}}{\partial\mathbf{x}}\mathbf{h}\left(\mathbf{x}\right)+\frac{\partial\mathbf{h}\left(\mathbf{x}\right)^{\mathrm{T}}}{\partial\mathbf{x}}\mathbf{g}\left(\mathbf{x}\right)$$

$$\frac{\partial\mathbf{g}^{\mathrm{T}}}{\partial\mathbf{x}}=\left(\frac{\partial\mathbf{g}}{\partial\mathbf{x}^{\mathrm{T}}}\right)^{\mathrm{T}}$$

- 讨论1：如何得到二阶方向导数的表达式?

- 讨论2：更高阶方向导数的表达式?

Hessian矩阵

实值函数 $f(\mathbf{x})$ 的局部极值条件

如果 $\mathbf{x}_*$ 是 $f(\mathbf{x})$ 的局部极小值，并且$\nabla_{\mathbf{x}}^2 f(\mathbf{x})$ 在 $\mathbf{x}_*$ 的开邻域 $\Delta\mathbf{x}$ 内连续，则

$$\nabla_{\mathbf{x}} f(\mathbf{x}_*) = \mathbf{0}, \quad \nabla_{\mathbf{x}}^2 f(\mathbf{x}_*) \geq \mathbf{0}$$

其中$\nabla_{\mathbf{x}}^2 f(\mathbf{x}_*) \geq \mathbf{0}$代表Hessian矩阵$\nabla_{\mathbf{x}}^2 f(\mathbf{x})$是半正定的。

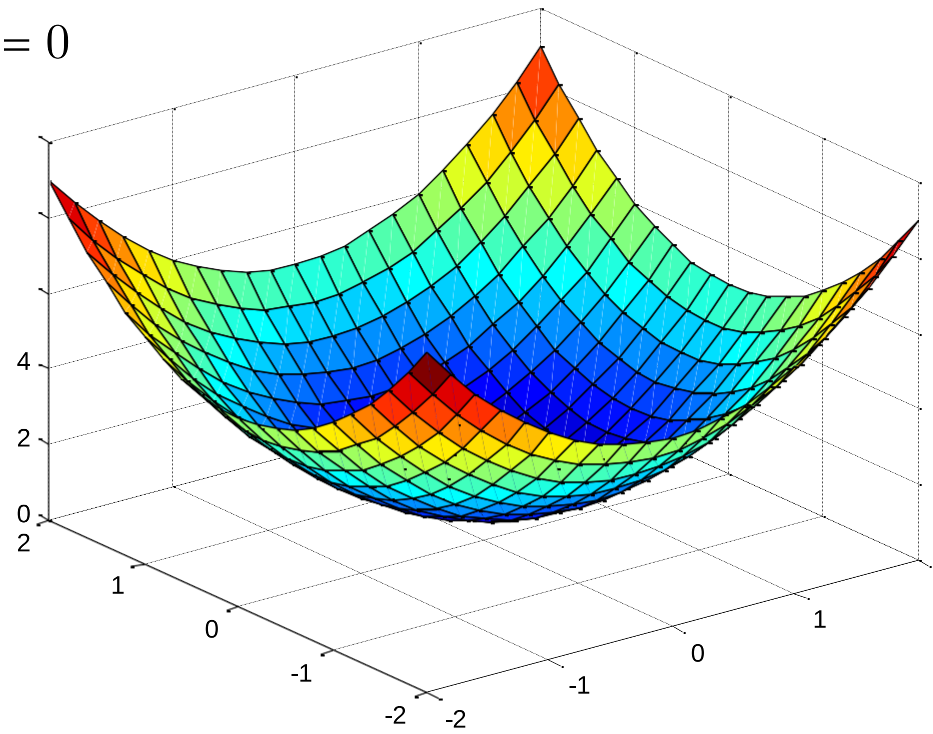
Hessian矩阵

- 例：实值函数$f\left(\mathbf{x}\right)$的极值

$$f\left(x,y\right)=x^2+y^2+1$$

$$\nabla_{\mathbf{x}} f\left(\mathbf{x}\right)=\begin{bmatrix}2x\\2y\end{bmatrix}\stackrel{令}{=}\mathbf{0}\Rightarrow x=0,y=0$$

$$\nabla_{\mathbf{x}}^2 f\left(\mathbf{x}\right)=\begin{bmatrix}2&0\\0&2\end{bmatrix}$$



,

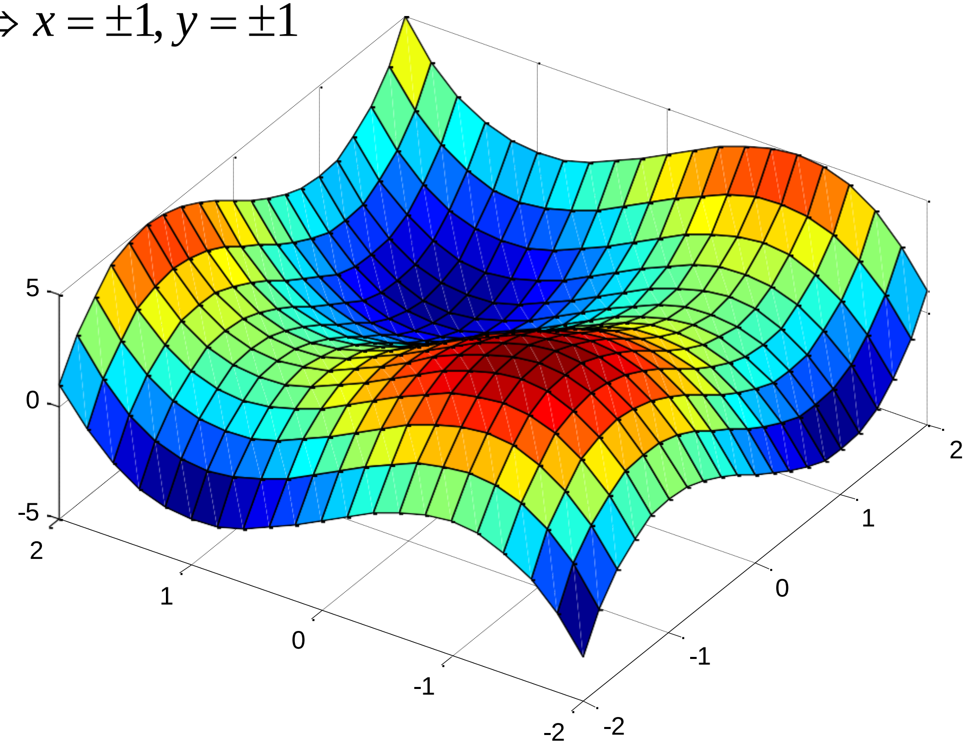
Hessian矩阵

例：实值函数$f\left(\mathbf{x}\right)$的极值

$$f\left(x,y\right)=x^3+y^3-3x-3y+1$$

$$\nabla_{\mathbf{x}} f\left(\mathbf{x}\right)=\begin{bmatrix}3x^2-3\\3y^2-3\end{bmatrix}\overset{令}{=}\mathbf{0}\Rightarrow x=\pm1, y=\pm1$$

$$\nabla_{\mathbf{x}}^2 f\left(\mathbf{x}\right)=\begin{bmatrix}6x & 0\\0 & 6y\end{bmatrix}$$



投影Hessian矩阵

一般的包含等式约束的优化模型形式如下：

$$\begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^{n\times 1}} \quad f(\mathbf{x}) \\ \text{s.t.} \quad \mathrm{g}(\mathbf{x}) = 0. \end{cases}$$

约束优化模型的拉格朗日函数定义如下：

$$\mathcal{L}(\mathbf{x},\lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

投影Hessian矩阵

定理 1 (二阶必要性条件) 假设对任意向量 $\mathbf{w} \in C$,

$$\mathbf{w}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathbf{x}^*, \lambda^*) \mathbf{w} \geq 0$$

成立，则 $\mathbf{x}^*$ 为优化问题的局部极值点。在此，集合 C 被定义为

$$C = \{\mathbf{w} \in \mathbb{R}^n | (\nabla g)^{\mathrm{T}} \mathbf{w} = 0\} = \mathrm{Null}[(\nabla \mathrm{g})^{\mathrm{T}}],$$

其中， $\mathrm{Null}(\mathbf{A})$ 表示 $\mathbf{A}$ 的零空间； $\nabla^2 \mathcal{L}(\mathbf{x}^*, \lambda^*)$为拉格朗日函数在$(\mathbf{x}^*, \lambda^*)$处的二阶Hessian矩阵； ∇g为约束函数的一阶梯度。

约束条件$g(\mathbf{x})$的一阶泰勒展开：

$$g(\mathbf{x} + \mathbf{w}) \approx g(\mathbf{x}) + (\nabla g)^{\mathrm{T}} \mathbf{w}$$

投影Hessian矩阵

由于 $\mathbf{w}$ 始终位于$(\nabla g)^{\mathrm{T}}$ 的零空间中，假设该零空间的维度为m $(m < n)$，矩阵$\mathbf{Z} \in \mathbb{R}^{n\times m}$ 的列向量构成零空间的一组基, 则 $\mathbf{w}$ 可以表示为

$$\mathbf{w} = \mathbf{Z}\mathbf{v}$$

其中， $\mathbf{v} \in \mathbb{R}^{m\times 1}$ 为相应的组合系数。

投影Hessian矩阵

$$\mathbf{w}^{\mathrm{T}}\nabla^2\mathcal{L}(\mathbf{x}^*,\lambda^*)\mathbf{w}\geq 0$$

$\mathbf{w}=\mathbf{Z}\mathbf{v}$

$$\Updownarrow$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\nabla^2\mathcal{L}(\mathbf{x}^*,\lambda^*)\mathbf{Z}\mathbf{v}\geq 0$$

$$\Updownarrow$$

$$\mathbf{Z}^{\mathrm{T}}\nabla^2\mathcal{L}(\mathbf{x}^*,\lambda^*)\mathbf{Z}\geq 0$$

投影Hessian矩阵

对于有约束优化模型，二阶必要性条件等价于判断投影Hessian矩阵的正定性

投影Hessian矩阵

获取矩阵 $\mathbf{Z}$ 的一种方法是对 ∇g执行QR 分解，其中有

$$\nabla g = \mathbf{Q}\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{bmatrix}\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix} = \mathbf{Q}_1\mathbf{R},$$

其中， $\mathbf{Q}$ 为 $n \times n$ 正交矩阵，进而可以令 $\mathbf{Z} = \mathbf{Q}_2$，因为

$$(\nabla g)^{\mathrm{T}}\mathbf{w} = \mathbf{R}^{\mathrm{T}}\mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_2\mathbf{v} = 0.$$

Reference:

J. Nocedal and S. J. Wright, Numerical Optimization. New York, NY, USA: Springer, 1999.

投影Hessian矩阵

无约束优化与约束优化模型二阶局部极值条件的对比

模型	二阶必要性条件	对比
$\min_{\mathbf{x}} \quad f(\mathbf{x})$	$\nabla^2 f(\mathbf{x}^*) \geq 0$	Hessian 矩阵
$\begin{cases} \min_{\mathbf{x}} & f(\mathbf{x}) \\ \text{s.t.} & g(\mathbf{x}) = 0. \end{cases}$	$\mathbf{Z}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathbf{x}^*, \lambda^*) \mathbf{Z} \geq 0$	投影Hessian 矩阵

Hessian矩阵

与Hessian矩阵相关的常用公式

$$1.\ \frac{\partial^2 \mathbf{a}^T\mathbf{x}}{\partial\mathbf{x}\partial\mathbf{x}^T} = \mathbf{O}_{n\times n}$$

$$2.\ \frac{\partial^2 \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{x}\partial\mathbf{x}^T} = \mathbf{A} + \mathbf{A}^T$$

$$3.\ \frac{\partial^2 \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{x}\partial\mathbf{x}^T} = 2\mathbf{A} \qquad \mathbf{A}\text{为对称矩阵时}$$

谢　谢

耿修瑞

中国科学院空间信息创新研究院

gengxr@sina.com.cn