梯度

定义 (梯度)

给定函数$f:\mathbb{R}^n \to \mathbb{R}$，且$f$在点$x$的一个邻域内有意义，若存在向量$g \in \mathbb{R}^n$ 满足

$$\lim_{p\to 0} \frac{f(x+p)-f(x)-g^{\mathrm{T}}p}{\|p\|} = 0,$$

其中$\|\cdot\|$是任意的向量范数，就称f 在点x 处**可微**（或Fréchet 可微）．此时g 称为f 在点x 处的**梯度**，记作$\nabla f(x)$．如果对区域D 上的每一个点x 都有$\nabla f(x)$ 存在，则称f 在D 上可微．

若f 在点x 处的梯度存在，在定义式中令$p=\varepsilon e_i$，e_i是第i个分量为1的单位向量，可知$\nabla f(x)$ 的第i 个分量为$\dfrac{\partial f(x)}{\partial x_i}$．因此，

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_n}\right]^{\mathrm{T}}.$$

海瑟矩阵

定义 (海瑟矩阵)

如果函数$f(x):\mathbb{R}^n \to \mathbb{R}$在点$x$处的二阶偏导数$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ $i,j=1,2,\cdots,n$都存在，则

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \frac{\partial^2 f(x)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

称为f 在点x 处的海瑟矩阵.

当$\nabla^2 f(x)$ 在区域D 上的每个点x 处都存在时，称f 在D 上二阶可微.
若$\nabla^2 f(x)$ 在D 上还连续，则称f 在D 上二阶连续可微，可以证明此时海瑟矩阵是一个对称矩阵.

矩阵变量函数的导数

多元函数梯度的定义可以推广到变量是矩阵的情形. 对于以$m \times n$ 矩阵X 为自变量的函数$f(X)$，若存在矩阵$G \in \mathbb{R}^{m\times n}$满足

$$\lim_{V\to 0} \frac{f(X+V) - f(X) - \langle G, V\rangle}{\|V\|} = 0,$$

其中$\|\cdot\|$是任意矩阵范数，就称矩阵变量函数f在X处Fréchet 可微，称G为f 在Fréchet 可微意义下的梯度. 类似于向量情形，矩阵变量函数$f(X)$的梯度可以用其偏导数表示为

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

其中$\frac{\partial f}{\partial x_{ij}}$ 表示f 关于x_{ij} 的偏导数.

矩阵变量函数的导数

在实际应用中，矩阵Fréchet 可微的定义和使用往往比较繁琐，为此我们需要介绍另一种定义——Gâteaux 可微.

定义 (Gâteaux 可微)

设$f(X)$为矩阵变量函数，如果对任意方向$V \in \mathbb{R}^{m\times n}$，存在矩阵$G \in \mathbb{R}^{m\times n}$ 满足

$$\lim_{t\to 0} \frac{f(X+tV) - f(X) - t\langle G, V\rangle}{t} = 0,$$

则称f 关于X 是Gâteaux 可微的. 满足上式的G称为f在X处在Gâteaux 可微意义下的梯度.

可以证明，当f是Fréchet 可微函数时，f也是Gâteaux 可微的，且这两种意义下的梯度相等.

矩阵变量函数的导数

- 线性函数：$f(X)=\mathrm{tr}(AX^{\mathrm{T}}B)$，其中$A\in\mathbb{R}^{p\times n},B\in\mathbb{R}^{m\times p},X\in\mathbb{R}^{m\times n}$
对任意方向$V\in\mathbb{R}^{m\times n}$以及$t\in\mathbb{R}$，有

$$\begin{aligned}\lim_{t\to 0}\frac{f(X+tV)-f(X)}{t}&=\lim_{t\to 0}\frac{\mathrm{tr}(A(X+tV)^{\mathrm{T}}B)-\mathrm{tr}(AX^{\mathrm{T}}B)}{t}\\&=\mathrm{tr}(AV^{\mathrm{T}}B)=\langle BA,V\rangle\,.\end{aligned}$$

因此，$\nabla f(X)=BA$.

- 二次函数：$f(X,Y)=\frac{1}{2}\|XY-A\|_F^2$，其中$(X,Y)\in\mathbb{R}^{m\times p}\times\mathbb{R}^{p\times n}$
对变量Y，取任意方向V以及充分小的$t\in\mathbb{R}$，有

$$\begin{aligned}f(X,Y+tV)-f(X,Y)&=\frac{1}{2}\|X(Y+tV)-A\|_F^2-\frac{1}{2}\|XY-A\|_F^2\\&=\langle tXV,XY-A\rangle+\frac{1}{2}t^2\|XV\|_F^2\\&=t\left\langle V,X^{\mathrm{T}}(XY-A)\right\rangle+\mathcal{O}(t^2).\end{aligned}$$

由定义可知$\frac{\partial f}{\partial Y}=X^{\mathrm{T}}(XY-A)$.
对变量X，同理可得$\frac{\partial f}{\partial X}=(XY-A)Y^{\mathrm{T}}$.

矩阵变量函数的导数

- *ln-det* 函数：$f(X)=\ln(\det(X))$, $X\in\mathcal{S}^n_{++}$，给定$X\succ 0$，对任意方向$V\in\mathcal{S}^n$ 以及$t\in\mathbb{R}$，我们有

$$\begin{aligned}
&f(X+tV)-f(X)\\
=&\ln(\det(X+tV))-\ln(\det(X))\\
=&\ln(\det(X^{1/2}(I+tX^{-1/2}VX^{-1/2})X^{1/2}))-\ln(\det(X))\\
=&\ln(\det(I+tX^{-1/2}VX^{-1/2})).
\end{aligned}$$

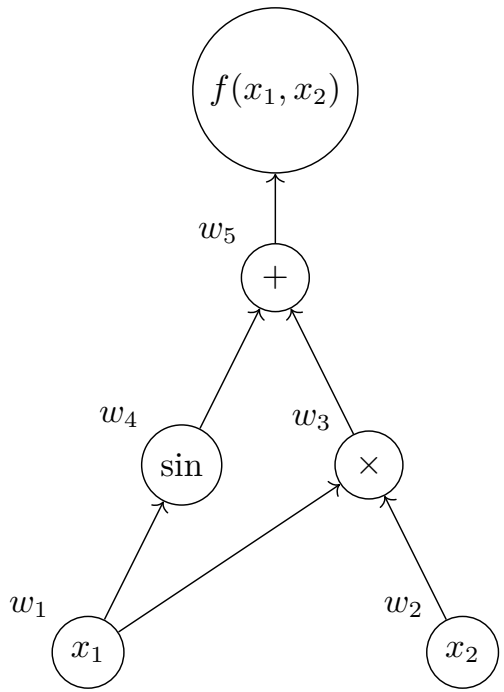
由于$X^{-1/2}VX^{-1/2}$是对称矩阵，所以它可以正交对角化，不妨设它的特征值为$\lambda_1,\lambda_2,\cdots,\lambda_n$，则

$$\begin{aligned}
\ln(\det(I+tX^{-1/2}VX^{-1/2}))=&\ln\prod_{i=1}^{n}(1+t\lambda_i)\\
=&\sum_{i=1}^{n}\ln(1+t\lambda_i)=\sum_{i=1}^{n}t\lambda_i+\mathcal{O}(t^2)\\
=&t\,\mathrm{tr}(X^{-1/2}VX^{-1/2})+\mathcal{O}(t^2)\\
=&t\left\langle(X^{-1})^{\mathrm{T}},V\right\rangle+\mathcal{O}(t^2).
\end{aligned}$$

因此，我们得到结论$\nabla f(X)=(X^{-1})^{\mathrm{T}}$.

自动微分

自动微分是使用计算机计算导数的算法. 在神经网络中，损失函数$f(x)$是由很多个简单函数复合而成的函数，根据复合函数的链式法则，可以通过每个简单函数的导数的乘积来计算对于各层变量的导数.
我们先考虑一个简单的例子$f(x_1, x_2) = x_1 x_2 + \sin x_1$. 计算该函数的过程可以用下图来表示.

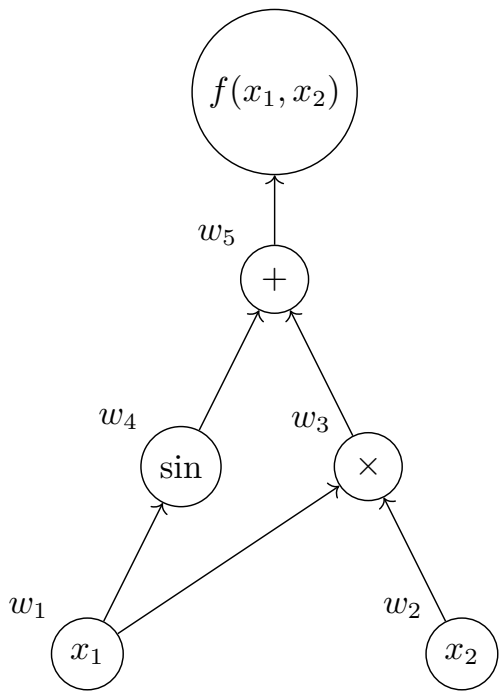


$$
\begin{aligned}
w_1 &= x_1 \\
w_2 &= x_2 \\
w_3 &= w_1 w_2 \\
w_4 &= \sin w_1 \\
w_5 &= w_3 + w_4
\end{aligned}
$$

Figure: 函数$f(x_1, x_2)$ 的计算过程

自动微分

在计算图中，w_1 和w_2 为自变量，w_3 和w_4 为中间变量，w_5代表最终的目标函数值．容易看出，函数f计算过程中涉及的所有变量和它们之间的依赖关系构成了一个有向图：每个变量w_i代表着图中的一个节点，变量的依赖关系为该图的边．如果有一条从节点w_i 指向w_j 的边，我们称w_i 为w_j 的父节点，w_j 为w_i 的子节点．一个节点的值由其所有的父节点的值确定．则称从父节点的值推子节点值的计算流为前向传播．



利用计算导数的链式法则，我们可以依次计算

$$\begin{aligned}
\frac{\partial f}{\partial w_5} &= 1, \\
\frac{\partial f}{\partial w_4} &= \frac{\partial f}{\partial w_5}\frac{\partial w_5}{\partial w_4} = 1, \\
\frac{\partial f}{\partial w_3} &= \frac{\partial f}{\partial w_5}\frac{\partial w_5}{\partial w_3} = 1, \\
\frac{\partial f}{\partial w_2} &= \frac{\partial f}{\partial w_3}\frac{\partial w_3}{\partial w_2} = w_1 = x_1, \\
\frac{\partial f}{\partial w_1} &= \frac{\partial f}{\partial w_3}\frac{\partial w_3}{\partial w_1} + \frac{\partial f}{\partial w_4}\frac{\partial w_4}{\partial w_1} = w_2 + \cos w_1 = \cos x_1 + x_2.
\end{aligned}$$

通过这种方式，就求得了导数

$$\frac{\partial f}{\partial x_1} = \cos x_1 + x_2, \quad \frac{\partial f}{\partial x_2} = x_1.$$

自动微分

自动微分有两种方式：前向模式和后向模式．在前向模式中，根据计算图，可以依次计算每个中间变量的取值及其对父变量的偏导数值．通过链式法则，可以复合得到每个中间变量对自变量的导数值．直至传播到最后一个子节点时，就得到了最终的目标函数值以及目标函数关于自变量的梯度值．

不同于前向模式，后向模式的节点求值和导数计算不是同时进行的．它是先利用前向模式计算各个节点的值，然后再根据计算图逆向计算对函数f关于各个中间变量的偏导数．

$$\frac{\partial f}{\partial w_i} = \sum_{w_j \text{ 是} w_i \text{的子节点}} \frac{\partial f}{\partial w_j} \frac{\partial w_j}{\partial w_i}.$$

对于前向模式而言，后向模式的梯度的计算复杂度更低．具体地，后向模式的梯度计算代价至多为函数值计算代价的5 倍，但是前向模式的计算代价可能多达函数值计算代价的n（n 为自变量维数）倍．因此对于神经网络中的优化问题，其自动微分采用的是后向模式．