矩阵求导浅析（二）

倚楼
早睡早起，多看paper，多吃多运动多睡觉

上次更新矩阵求导浅析（一）的时间还是10个月前。回想近一年的时间，科研上毫无进展，数学基础也没有提高，实在惭愧。在各种奇奇怪怪的论文模型中疲于奔命，不加思考地接受，缺乏见解（其实是懒）。在知识的海洋中，总是被海滩上的贝壳所吸引， 或是论坛、微博和朋友圈看到的东西， 分散了过多的注意力。不知道大家有没有这种感受：自己无论多么努力， 似乎都得不到社会的进一步认可； 相反， 如果按部就班地做事情， 好像也坏不到哪里去。

导言：本文主要介绍了标量对向量/矩阵求导的链式法则，相比于矩阵求导浅析（一）的全微分算子按部就班的分析方法，本文的链式法则方法针对常见形式的求导，能够一步给出求导的结果。

默认读者熟悉一元微积分求导的链式法则和多元微积分的求偏导的链式法则。先证明4条链式法则的应用。

符号说明：与矩阵求导浅析（一）一致。

1. $\nabla_{\mathbf{x}} f(\mathbf{A}\mathbf{x}+\mathbf{b})=\mathbf{A}^T\nabla f(\mathbf{A}\mathbf{x}+\mathbf{b})$

注意： $\nabla_{\mathbf{x}} f(\mathbf{A}\mathbf{x}+\mathbf{b})$ 表达的意思是 $\dfrac{\partial f(\mathbf{A}\mathbf{x}+\mathbf{b})}{\partial \mathbf{x}}$ ，而 $\nabla f(\mathbf{A}\mathbf{x}+\mathbf{b})$ 表示的是先求出梯度 $\nabla f(\cdot)$ ，然后将 $\mathbf{A}\mathbf{x}+\mathbf{b}$ 带入到 $\nabla f(\cdot)$ ，得到 $\nabla f(\mathbf{A}\mathbf{x}+\mathbf{b})$ 。

证明：令 $\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{b}$ ，则

$$
\nabla_{\mathbf{x}} f(\mathbf{A}\mathbf{x}+\mathbf{b})=\frac{\partial f(\mathbf{y})}{\partial \mathbf{x}}=\left[\frac{\partial f(\mathbf{y})}{\partial x_i}\right]
$$

$$
\begin{aligned}
\frac{\partial f(\mathbf{y})}{\partial x_i}&=\sum_{j=1}^{N}\frac{\partial f(\mathbf{y})}{\partial y_j}\frac{\partial y_j}{\partial x_i}(\text{by chain rule})=\left\langle\frac{\partial f(\mathbf{y})}{\partial \mathbf{y}},\frac{\partial \mathbf{y}}{\partial x_i}\right\rangle\\
&=\langle\nabla f(\mathbf{y}),\mathbf{a}_i\rangle=\mathbf{a}_i^T\nabla f(\mathbf{y})
\end{aligned}
$$

所以，

$$
\nabla_{\mathbf{x}} f(\mathbf{A}\mathbf{x}+\mathbf{b})=\begin{bmatrix}\mathbf{a}_1^T\nabla f(\mathbf{y})\\ \vdots\\ \mathbf{a}_M^T\nabla f(\mathbf{y})\end{bmatrix}=\begin{bmatrix}\mathbf{a}_1^T\\ \end{bmatrix}
$$

2. $\nabla_{\mathbf{X}} f(\mathbf{AX}+\mathbf{B}) = \mathbf{A}^T \nabla f(\mathbf{AX}+\mathbf{B})$

证明: $\nabla_{\mathbf{X}} f(\mathbf{AX}+\mathbf{B}) = \frac{\partial f(\mathbf{Y})}{\partial \mathbf{X}} = \left[\frac{\partial f(\mathbf{Y})}{\partial X_{ij}}\right]$

其中， $X_{ij} \to \mathbf{Y} \to f(\mathbf{Y})$,由链式法则

$$\frac{\partial f(\mathbf{Y})}{\partial X_{ij}} = \sum_{s,t} \frac{\partial f(\mathbf{Y})}{\partial Y_{s,t}} \frac{\partial Y_{s,t}}{\partial X_{ij}} = \left\langle \frac{\partial f(\mathbf{Y})}{\partial \mathbf{Y}}, \frac{\partial \mathbf{Y}}{\partial X_{ij}} \right\rangle$$

注意到 $Y_{s,t} = A_{s,:} X_{:,t} + B_{s,t}$ ，（ $A_{s,:}$ 表示矩阵 $\mathbf{A}$ 第 s 行， $X_{:,t}$ 表示矩阵 $\mathbf{X}$ 第 t 列），所以当 $t=j$ 时， $Y_{s,t}$ 中才会有 $X_{i,j}$ 项。所以， $\frac{\partial Y_{s,j}}{\partial X_{ij}} = A_{s,i}$. $\frac{\partial \mathbf{Y}}{\partial X_{ij}} = [0, \cdots, 0, A_{:,i}, 0 \cdots, 0]$ ，第 j 列的元素是矩阵 $\mathbf{A}$ 的第 i 列 $A_{:,i}$ ，其他列的元素为0。故，

$$\begin{aligned} \frac{\partial f(\mathbf{Y})}{\partial X_{ij}} &= \left\langle \frac{\partial f(\mathbf{Y})}{\partial \mathbf{Y}}, \frac{\partial \mathbf{Y}}{\partial X_{ij}} \right\rangle = \langle \nabla f(\mathbf{Y}), [0, \cdots, 0, A_{:,i}, 0 \cdots, 0] \rangle \\ &= \left\langle \nabla f(\mathbf{Y})_{:,j}, A_{:,i} \right\rangle = (A_{:,i})^T \nabla f(\mathbf{Y})_{:,j} \\ &= \left(A^T\right)_{i,:} \nabla f(\mathbf{Y})_{:,j} \end{aligned}$$

所以，

$$\frac{\partial f(\mathbf{Y})}{\partial \mathbf{X}} = \mathbf{A}^T \nabla f(\mathbf{Y})$$

3. $\nabla_{\mathbf{X}} f(\mathbf{XA}+\mathbf{B}) = \nabla f(\mathbf{XA}+\mathbf{B}) \mathbf{A}^T$ ，留给感兴趣的读者作为练习。

4. $\nabla_{\mathbf{X}} f(\mathbf{AXB}+\mathbf{C}) = \mathbf{A}^T \nabla f(\mathbf{AXB}+\mathbf{C}) \mathbf{B}^T$

证明: 计算 $\frac{\partial \mathbf{Y}}{\partial X_{ij}}$ ：提取出 $\mathbf{Y}$ 中跟 X_{ij} 有关的项。

$$\begin{aligned} \mathbf{Y} &= \mathbf{A}[\mathbf{X}_{:,1}, \cdots, \mathbf{X}_{:,j}, \cdots, \mathbf{X}_{:,N}]\mathbf{B} + \mathbf{C} \\ &= \mathbf{A}\mathbf{X}_{:,j}\mathbf{B}_{j,:} + \text{const} = \mathbf{A}_{:,i} X_{ij} \mathbf{B}_{j,:} + \text{const} \end{aligned}$$

所以，

$$\begin{aligned}\frac{\partial f(\mathbf{Y})}{\partial X_{ij}} &= \left\langle \frac{\partial f(\mathbf{Y})}{\partial \mathbf{Y}}, \frac{\partial \mathbf{Y}}{\partial X_{ij}} \right\rangle = \langle \nabla f(\mathbf{Y}), \mathbf{A}_{:,i}\mathbf{B}_{j,:} \rangle \\ &= Tr\left[\nabla f(\mathbf{Y}) (\mathbf{A}_{:,i}\mathbf{B}_{j,:})^T\right] \\ &= Tr\left[(\mathbf{A}_{:,i})^T \nabla f(\mathbf{Y}) (\mathbf{B}_{j,:})^T\right] \\ &= \left(\mathbf{A}^T\right)_{i,:} \nabla f(\mathbf{Y}) \left(\mathbf{B}^T\right)_{:,j}\end{aligned}$$

故 $\frac{\partial f(\mathbf{Y})}{\partial \mathbf{X}} = \mathbf{A}^T \nabla f(\mathbf{Y}) \mathbf{B}^T$

总结我们上述的准则（ J 是目标函数）：

$$\begin{aligned}\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} &\Rightarrow \frac{\partial J}{\partial \mathbf{x}} = \mathbf{A}^T \frac{\partial J}{\partial \mathbf{y}} \\ \mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{B} &\Rightarrow \frac{\partial J}{\partial \mathbf{X}} = \mathbf{A}^T \frac{\partial J}{\partial \mathbf{Y}} \\ \mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{B} &\Rightarrow \frac{\partial J}{\partial \mathbf{X}} = \frac{\partial J}{\partial \mathbf{Y}} \mathbf{A}^T \\ \mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{C} &\Rightarrow \frac{\partial J}{\partial \mathbf{X}} = \mathbf{A}^T \frac{\partial J}{\partial \mathbf{Y}} \mathbf{B}^T\end{aligned}$$

例1：已知 $J = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ ，求偏导数 $\frac{\partial J}{\partial \mathbf{A}}, \frac{\partial J}{\partial \mathbf{x}}, \frac{\partial J}{\partial \mathbf{b}}$

令 $\mathbf{y} = \mathbf{A}\mathbf{x} - \mathbf{b}$ ，显然 $\nabla_{\mathbf{y}} \|\mathbf{y}\|^2 = 2\mathbf{y}$ ，应用链式法则：

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{A}} &= 2\mathbf{y}\mathbf{x}^T = 2(\mathbf{A}\mathbf{x} - \mathbf{b})\mathbf{x}^T \\ \frac{\partial J}{\partial \mathbf{x}} &= 2\mathbf{A}^T\mathbf{y} = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ \frac{\partial J}{\partial \mathbf{b}} &= -2\mathbf{y} = -2(\mathbf{A}\mathbf{x} - \mathbf{b})\end{aligned}$$

例2:求 $\frac{\partial \left\|\mathbf{U}\mathbf{V}^T - \mathbf{Y}\right\|_F^2}{\partial \mathbf{U}}, \frac{\partial \left\|\mathbf{U}\mathbf{V}^T - \mathbf{Y}\right\|_F^2}{\partial \mathbf{V}}$

令 $\mathbf{W} = \mathbf{U}\mathbf{V}^T - \mathbf{Y}$ ，利用 $\nabla_{\mathbf{W}} \|\mathbf{W}\|_F^2 = 2\mathbf{W}$ 的运算，我们可以得到

$$\frac{\partial\left\|\mathbf{U}\mathbf{V}^T-\mathbf{Y}\right\|_F^2}{\partial\mathbf{U}}=2\mathbf{W}\mathbf{V}=2\left(\mathbf{U}\mathbf{V}^T-\mathbf{Y}\right)\mathbf{V}$$
$$\frac{\partial\left\|\mathbf{U}\mathbf{V}^T-\mathbf{Y}\right\|_F^2}{\partial\mathbf{V}^T}=2\mathbf{U}^T\mathbf{W}$$
$$\therefore\frac{\partial\left\|\mathbf{U}\mathbf{V}^T-\mathbf{Y}\right\|_F^2}{\partial\mathbf{V}}=2\mathbf{W}^T\mathbf{U}=2\left(\mathbf{U}\mathbf{V}^T-\mathbf{Y}\right)^T\mathbf{U}$$

由偏导数定义不难验证 $\frac{\partial\mathrm{Tr}\left(\mathbf{X}\right)}{\partial\mathbf{X}}=\mathbf{I}$ 。

例3： $\frac{\partial}{\partial\mathbf{X}}\mathrm{Tr}\left(\mathbf{A}\mathbf{X}\mathbf{B}\right)=\mathbf{A}^T\mathbf{B}^T$ （作为练习）

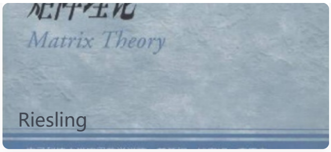
感兴趣的同学，可以进一步学习使用链式法则：反向传播（一层隐藏层，激活函数为Sigmoid函数。先写出前向传播的表达式，然后求出损失函数关于中间变量的偏导数，采用梯度下降方法更新）。

发布于 2018-06-27

矩阵运算 自然科学 矩阵

的专栏文章：矩阵求导术（上）矩阵求导术（下）机器学习中的矩阵/

倚楼



Riesling

Gan Pan

1 条评论 切换为时间排序

写下你的评论...

千雪湖 2018-08-19

想请教您一个矩阵的求导，能加下微信吗？这里输不了公式

赞

已赞同 14 ▼ 1 条评论 分享 收藏 ···