符号求导原理

梯度 链式法则

问题原型

假设存在如下的运算：

$$Y_1,\cdots,Y_m=f(X_1,\cdots,X_n)$$

其中$Y_1,\cdots,Y_m,X_1,\cdots,X_n$均为张量,且$f$不一定可以解耦成函数列. 现在对于任意作用于$Y_1,\cdots,Y_m$的标量函数$g$：

$$z=g(Y_1,\cdots,Y_m)$$

若已知 $\frac{\partial z}{\partial Y_1},\cdots,\frac{\partial z}{\partial Y_m},Y_1,\cdots,Y_m,X_1,\cdots,X_n$ 问是否可以计算出 $\frac{\partial z}{\partial X_1},\cdots,\frac{\partial z}{\partial X_n}$？
其实问题是肯定的.

求导原理分析

设Y,X均为张量,那么$\frac{\partial Y}{\partial X}$依$Y$元素梯度的结果可以排列成一个新的张量,其shape为Y,X二者shape的叠加.则普通的链式法则可以表示成张量积的形式：

$$\frac{\partial z}{\partial X_i}=\sum_j\frac{\partial z}{\partial Y_j}\frac{\partial Y_j}{\partial X_i},\quad i=1,\cdots,n$$

由于$\frac{\partial Y_j}{\partial X_i}$表达式肯定是与$g$无关的,所以$(1)$可以被表达为：

$$\frac{\partial z}{\partial X_i}=\sum_j\frac{\partial z}{\partial Y_j}H_{ij}(X_1,\cdots,X_n)=\sum_j h_{ij}(\frac{\partial z}{\partial Y_j},X_1,\cdots,X_n),\quad i=1,\cdots,n$$

其中H_{ij},h_{ij}均为输出为张量的函数,H_{ij}的shape为Y_j和$X_i$shape的叠加,h_{ij}的shape为$X_i$shape的一致.从(2)可以看出H_{ij}的表达式与g 无关 ,进而h_{ij}的表达式也与g 无关 .

所以若已知 $\frac{\partial z}{\partial Y_1},\cdots,\frac{\partial z}{\partial Y_m},Y_1,\cdots,Y_m,X_1,\cdots,X_n$,是可以通过推导表达式h_{ij}计算出 $\frac{\partial z}{\partial X_1},\cdots,\frac{\partial z}{\partial X_n}$ 的.

变分原理与表达式h_{ij}的推导

我们参考PDE里的变分原理,来实现对表达式h_{ij}的推导.

多输入单输出时的梯度算法

首先考虑$m=1$的情况.不失一般性,记Y_1是一个含有三个上标的张量,构造一个下标与Y_1上标完全一致的任意张量$\Phi^{(1)}$.通过Einstein求和,这样就可以构造出不可数个g：

$$g(Y_1)=\Phi^{(1)}_{ijk}Y_1^{ijk}$$

我们这样做的目的是为了使得$\frac{\partial z}{\partial Y_1}$正好成为一个形状指定但是数值任意的张量：

$$\frac{\partial z}{\partial Y_1^{ijk}}=\Phi^{(1)}_{ijk}\quad\Rightarrow\quad\frac{\partial z}{\partial Y_1}=\Phi^{(1)}$$

通过将(2.2)带入1.2即可得到$h_{i1},i=1,\cdots,n$的表达式.

> 以下以f是矩阵乘法为例：

$$\begin{cases} \boldsymbol{Y}_1 = f(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{X}_1\boldsymbol{X}_2 \\ z = \sum\limits_{i,j} \Phi_{ij}^{(1)} \sum\limits_{k} \boldsymbol{X}_1^{ik}\boldsymbol{X}_2^{kj} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{Y}_1^{ij}} = \Phi_{ij}^{(1)} \\ \dfrac{\partial z}{\partial \boldsymbol{X}_1^{ik}} = \sum\limits_{j} \Phi_{ij}^{(1)} \boldsymbol{X}_2^{kj} \\ \dfrac{\partial z}{\partial \boldsymbol{X}_2^{kj}} = \sum\limits_{i} \Phi_{ij}^{(1)} \boldsymbol{X}_1^{ik} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{X}_1^{ik}} = \sum\limits_{j} \dfrac{\partial z}{\partial \boldsymbol{Y}_1^{ij}} \boldsymbol{X}_2^{kj} \\ \dfrac{\partial z}{\partial \boldsymbol{X}_2^{kj}} = \sum\limits_{i} \dfrac{\partial z}{\partial \boldsymbol{Y}_1^{ij}} \boldsymbol{X}_1^{ik} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{X}_1} = \dfrac{\partial z}{\partial \boldsymbol{Y}_1} \boldsymbol{X}_2^T \\ \dfrac{\partial z}{\partial \boldsymbol{X}_2} = \boldsymbol{X}_1^T \dfrac{\partial z}{\partial \boldsymbol{Y}_1} \end{cases}$$

多输出单输入时的梯度算法

考虑$\boldsymbol{n=1}$的情况.一般来说这种情况下$\boldsymbol{h_{m1}}$的表达式是很难推导出来的,但是如果假设$\boldsymbol{f}$可逆,则问题将变得简单，先根据多输入单输出时的梯度算法可以得到如下的表达式组:

$$\begin{cases} \dfrac{\partial z}{\partial \boldsymbol{Y}_1} &= \hat{\boldsymbol{h}}_{11}(\dfrac{\partial z}{\partial \boldsymbol{X}_1}, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_m) \\ &\vdots \\ \dfrac{\partial z}{\partial \boldsymbol{Y}_m} &= \hat{\boldsymbol{h}}_{m1}(\dfrac{\partial z}{\partial \boldsymbol{X}_1}, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_m) \end{cases}$$

通过对$(\boldsymbol{2.2.1})$式进行变形即可得到相应的$(\boldsymbol{1.2})$.

以下以$\boldsymbol{f}$是矩阵的最大主元LU分解为例：

$$\begin{cases} \boldsymbol{LU} &= \boldsymbol{P}f(\boldsymbol{X}) \\ z &= \sum\limits_{i,j} \Phi_{ij}\boldsymbol{X}_{ij} \\ &= \sum\limits_{i,j} \Phi_{ij} \sum\limits_{kl} \boldsymbol{P}_{ik}^T \boldsymbol{L}_{kl}\boldsymbol{U}_{lj} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{X}} = \Phi \\ \dfrac{\partial z}{\partial \boldsymbol{P}} = \boldsymbol{L}\boldsymbol{U}\Phi^T \\ \dfrac{\partial z}{\partial \boldsymbol{L}} = \boldsymbol{P}\Phi\boldsymbol{U}^T \\ \dfrac{\partial z}{\partial \boldsymbol{U}} = \boldsymbol{L}^T\boldsymbol{P}\Phi \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{X}} = (\dfrac{\partial z}{\partial \boldsymbol{P}})^T \boldsymbol{L}^{-T}\boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial \boldsymbol{X}} = \boldsymbol{P}^T \dfrac{\partial z}{\partial \boldsymbol{L}} \boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial \boldsymbol{X}} = \boldsymbol{P}^T\boldsymbol{L}^{-T} \dfrac{\partial z}{\partial \boldsymbol{U}} \end{cases}$$

本例说明了一个问题,也就是针对多输出的情况,$(\boldsymbol{1.2})$可以有多种不同的表达式且不一定需要所有的上一级梯度.

多输出多输入时的梯度算法

经过测试一般来说在这种情况下并没有什么非常的简单的方法可以导出$(\boldsymbol{1.2})$,若存在$\boldsymbol{f_X}$和可逆的$\boldsymbol{f_Y}$使得:

$$\boldsymbol{W} = f_Y(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_m) = f_X(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$$

则可以综合多输入单输出时的梯度算法和多输出单输入时的梯度算法仍然可以导出$(\boldsymbol{1.2})$.

以下以$\boldsymbol{f}$是对矩阵乘法结果的最大主元LU分解为例：

$$\begin{cases} \boldsymbol{LU} &= \boldsymbol{P}f(\boldsymbol{AB}) \\ z &= \sum\limits_{i,j} \Phi_{ij} \sum\limits_{k} \boldsymbol{A}_{ik}\boldsymbol{B}_{kj} \\ &= \sum\limits_{i,j} \Phi_{ij} \sum\limits_{kl} \boldsymbol{P}_{ik}^T \boldsymbol{L}_{kl}\boldsymbol{U}_{lj} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{A}} = \dfrac{\partial z}{\partial (\boldsymbol{AB})} \boldsymbol{B}^T \\ \dfrac{\partial z}{\partial \boldsymbol{B}} = \boldsymbol{A}^T \dfrac{\partial z}{\partial (\boldsymbol{AB})} \\ \dfrac{\partial z}{\partial (\boldsymbol{AB})} = (\dfrac{\partial z}{\partial \boldsymbol{P}})^T \boldsymbol{L}^{-T}\boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial (\boldsymbol{AB})} = \boldsymbol{P}^T \dfrac{\partial z}{\partial \boldsymbol{L}} \boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial (\boldsymbol{AB})} = \boldsymbol{P}^T\boldsymbol{L}^{-T} \dfrac{\partial z}{\partial \boldsymbol{U}} \end{cases} \Rightarrow \begin{cases} \dfrac{\partial z}{\partial \boldsymbol{A}} = (\dfrac{\partial z}{\partial \boldsymbol{P}})^T \boldsymbol{L}^{-T}\boldsymbol{U}^{-T}\boldsymbol{B}^T \\ \dfrac{\partial z}{\partial \boldsymbol{A}} = \boldsymbol{P}^T \dfrac{\partial z}{\partial \boldsymbol{L}} \boldsymbol{U}^{-T}\boldsymbol{B}^T \\ \dfrac{\partial z}{\partial \boldsymbol{A}} = \boldsymbol{P}^T\boldsymbol{L}^{-T} \dfrac{\partial z}{\partial \boldsymbol{U}} \boldsymbol{B}^T \\ \dfrac{\partial z}{\partial \boldsymbol{B}} = \boldsymbol{A}^T(\dfrac{\partial z}{\partial \boldsymbol{P}})^T \boldsymbol{L}^{-T}\boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial \boldsymbol{B}} = \boldsymbol{A}^T\boldsymbol{P}^T \dfrac{\partial z}{\partial \boldsymbol{L}} \boldsymbol{U}^{-T} \\ \dfrac{\partial z}{\partial \boldsymbol{B}} = \boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{L}^{-T} \dfrac{\partial z}{\partial \boldsymbol{U}} \end{cases}$$

离散变量求导法