数学笔记

深度学习相关数学

作者：左元

ElegantLaTeX Program

目录

第一章 标量导数法则

表 1.1: 标量导数法则

法则	$f(x)$	和 x 相关的标量导数记法	示例
常数	c	0	$\frac{d}{dx}99 = 0$
乘以常数	cf	$c\frac{df}{dx}$	$\frac{d}{dx}3x = 3$
指数法则	x^n	nx^{n-1}	$\frac{d}{dx}x^3 = 3x^2$
加法法则	$f + g$	$\frac{df}{dx} + \frac{dg}{dx}$	$\frac{d}{dx}(x^2 + 3x) = 2x + 3$
减法法则	$f - g$	$\frac{df}{dx} - \frac{dg}{dx}$	$\frac{d}{dx}(x^2 - 3x) = 2x - 3$
乘法法则	fg	$f\frac{dg}{dx} + \frac{df}{dx}g$	$\frac{d}{dx}x^2x = x^2 + x2x = 3x^2$
链式法则	$f(g(x))$	$\frac{df(u)}{du}\frac{du}{dx}$, let $u = g(x)$	$\frac{d}{dx}ln(x^2) = \frac{1}{x^2}2x = \frac{2}{x}$

举例:

$$\frac{d}{dx}9(x + x^2) = 9\frac{d}{dx}(x + x^2) = 9(\frac{d}{dx}x + \frac{d}{dx}x^2) = 9(1 + 2x) = 9 + 18x$$

第二章 矢量微积分与偏导数简介

神经网络层不是单一参数的单一函数，$f(x)$。因此，让我们继续讨论多参数函数，如 $f(x,y)$。例如，xy（即 x 和 y 的乘积）的导数是什么？换句话说，当我们调整变量时，乘积 xy 如何变化？这取决于我们是在改变 x 还是 y。我们一次计算一个变量（参数）的导数，从而为这个双参数函数得到两个不同的偏导数（一个关于 x，一个关于 y）。偏导数运算符不是使用运算符 $\frac{d}{dx}$，而是 $\frac{\partial}{\partial x}$（一个风格化的 d，而不是希腊字母 δ）。因此，$\frac{\partial}{\partial x}xy$ 和 $\frac{\partial}{\partial y}xy$ 是 xy 的偏导数；通常，这些简称为偏导数。对于单一参数的函数，运算符 $\frac{\partial}{\partial x}$ 等价于 $\frac{d}{dx}$（对于足够光滑的函数）。然而，最好使用 $\frac{d}{dx}$ 以明确你指的是标量导数。

关于 x 的偏导数就是通常的标量导数，只需将方程中的任何其他变量视为常数。考虑函数 $f(x,y) = 3x^2y$。关于 x 的偏导数写作 $3xy$。从 $\frac{\partial}{\partial x}$ 的角度来看，有三个常数：3、2 和 y。因此，$\frac{\partial}{\partial x}3yx^2 = 3y\frac{\partial}{\partial x}x^2 = 3y2x = 6yx$。关于 y 的偏导数将 x 视为常数：$\frac{\partial}{\partial y}3x^2y = 3x^2\frac{\partial}{\partial y}y = 3x^2\frac{\partial y}{\partial y} = 3x^2 \times 1 = 3x^2$。在继续之前，最好自己推导这些内容，否则文章的其余部分将无法理解。

为了明确我们正在进行的是向量微积分而不仅仅是多元微积分，让我们考虑如何处理偏导数 $\frac{\partial f(x,y)}{\partial x}$ 和 $\frac{\partial f(x,y)}{\partial y}$（另一种表示 $\frac{\partial}{\partial x}f(x,y)$ 和 $\frac{\partial}{\partial y}f(x,y)$ 的方式），我们为 $f(x,y) = 3x^2y$ 计算这些值。与其让它们随意漂浮且没有任何组织，不如将它们组织成一个水平向量。我们称这个向量为 $f(x,y)$ 的梯度，并将其写作：

$$\nabla f(x,y) = [\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}] = [6yx, 3x^2]$$

因此，$f(x,y)$ 的梯度就是其偏导数组成的向量。梯度是向量微积分世界的一部分，它处理的是将 n 个标量参数映射到单个标量的函数。现在，让我们大胆一点，同时考虑多个函数的导数。

第三章 矩阵微积分

当我们从单一函数的导数转向多个函数的导数时，我们便从向量微积分的世界进入了矩阵微积分的领域。让我们计算两个函数的偏导数，这两个函数都接受两个参数。我们可以沿用上一节中的 $f(x,y) = 3x^2y$，但同时也引入 $g(x,y) = 2x + y^8$。g 的梯度有两个条目，每个参数对应一个偏导数：

$$\frac{\partial g(x,y)}{\partial y} = \frac{\partial 2x}{\partial y} + \frac{\partial y^8}{\partial y} = 0 + 8y^7 = 8y^7$$

得到梯度 $\nabla g(x,y) = [2, 8y^7]$ 。

梯度向量组织了一个特定标量函数的所有偏导数。如果我们有两个函数，我们也可以通过堆叠梯度将它们的梯度组织成一个矩阵。当我们这样做时，我们得到了雅可比矩阵（或简称为雅可比），其中梯度是行：

$$J = \begin{bmatrix} \nabla f(x,y) \\ \nabla g(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} & \frac{\partial f(x,y)}{\partial y} \\ \frac{\partial g(x,y)}{\partial x} & \frac{\partial g(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 6yx & 3x^2 \\ 2 & 8y^7 \end{bmatrix}$$

欢迎来到矩阵微积分！

请注意，有多种方法可以表示雅可比矩阵。我们使用的是所谓的分子布局，但许多论文和软件会使用分母布局。这只是分子布局雅可比矩阵的转置（沿其对角线翻转）：

$$\begin{bmatrix} 6yx & 2 \\ 3x^2 & 8y^7 \end{bmatrix}$$

3.1 雅可比矩阵的推广

到目前为止，我们已经看了一个雅可比矩阵的具体例子。为了更一般地定义雅可比矩阵，让我们将多个参数组合成一个单一的向量参数：$f(x,y,z) \Rightarrow f(\mathbf{x})$ 。（在文献中，你有时也会看到用 $\vec{x}$ 表示向量。）粗体小写字母如 $\mathbf{x}$ 表示向量，斜体小写字母如 x 表示标量。x_i 是向量 $\mathbf{x}$ 的第 i^{th} 个元素，并且是斜体，因为单个向量元素是标量。我们还必须为向量 $\mathbf{x}$ 定义一个方向。我们假设所有向量默认都是垂直的，大小为 $n \times 1$：

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

对于多个标量值函数，我们可以像处理参数一样将它们全部组合成一个向量。设 $\mathbf{y} = \mathbf{f}(\mathbf{x})$ 为一个包含 m 个标量值函数的向量，每个函数都接受一个长度为 $n = |\mathbf{x}|$ 的向量 $\mathbf{x}$ 作为参数，其中 $|\mathbf{x}|$ 表示 $\mathbf{x}$ 中元素的数量（基数）。$\mathbf{f}$ 中的每个函数 f_i 都像上一节中那样返回一个标量：

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) \\ y_2 &= f_2(\mathbf{x}) \\ &\vdots \\ y_m &= f_m(\mathbf{x}) \end{aligned}$$

例如，我们可以将 $f(x,y) = 3x^2y$ 和 $g(x,y) = 2x + y^8$ 表示为

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) = 3x_1^2 x_2 \qquad \text{(使用}x_1\text{替换掉}x, x_2\text{替换掉}y\text{)} \\ y_2 &= f_2(\mathbf{x}) = 2x_1 + x_2^8 \end{aligned}$$

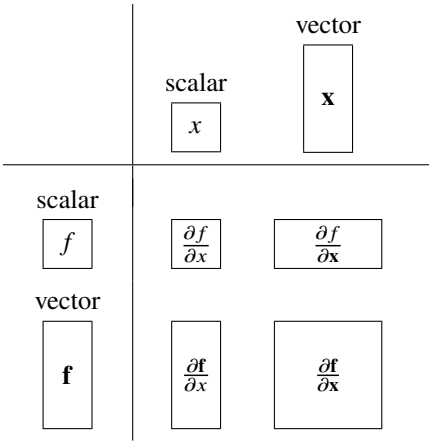
通常情况下，$m = n$ 是很常见的，因为我们会为 $\mathbf{x}$ 向量的每个元素得到一个标量函数结果。例如，考虑恒等函数 $\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{x}$：

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) &&= x_1 \\ y_2 &= f_2(\mathbf{x}) &&= x_2 \\ &\vdots \\ y_n &= f_n(\mathbf{x}) &&= x_n \end{aligned}$$

因此，在这种情况下，我们有 $m = n$ 个函数和参数。不过一般来说，雅可比矩阵是所有 $m \times n$ 个可能的偏导数的集合（m 行和 n 列），即关于 $\mathbf{x}$ 的 m 个梯度的堆叠：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_2(\mathbf{x}) \\ \dots \\ \nabla f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} f_2(\mathbf{x}) \\ \dots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}) \\ & & \dots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \frac{\partial}{\partial x_2} f_m(\mathbf{x}) & \dots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix}$$

每个 $\frac{\partial}{\partial \mathbf{x}} f_i(\mathbf{x})$ 是一个水平的 n-向量，因为偏导数是关于一个长度为 $n = |\mathbf{x}|$ 的向量 $\mathbf{x}$ 的。雅可比矩阵的宽度为 n，如果我们是关于 $\mathbf{x}$ 取偏导数，因为有 n 个参数可以调整，每个参数都有可能改变函数的值。因此，雅可比矩阵总是有 m 行，对应于 m 个方程。考虑可能的雅可比矩阵形状时，视觉化是很有帮助的：



恒等函数 $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ 的雅可比矩阵，其中 $f_i(\mathbf{x}) = x_i$，具有 n 个函数，每个函数有 n 个参数，这些参数保存在一个单一的向量 $\mathbf{x}$ 中。因此，雅可比矩阵是一个方阵，因为 $m = n$：

$$
\begin{aligned}
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} f_2(\mathbf{x}) \\ \cdots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} &= \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}) \\ & & \cdots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \frac{\partial}{\partial x_2} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} x_1 & \frac{\partial}{\partial x_2} x_1 & \cdots & \frac{\partial}{\partial x_n} x_1 \\ \frac{\partial}{\partial x_1} x_2 & \frac{\partial}{\partial x_2} x_2 & \cdots & \frac{\partial}{\partial x_n} x_2 \\ & & \cdots & \\ \frac{\partial}{\partial x_1} x_n & \frac{\partial}{\partial x_2} x_n & \cdots & \frac{\partial}{\partial x_n} x_n \end{bmatrix} \\
&\quad (\text{且因为对于} j \neq i, \frac{\partial}{\partial x_j} x_i = 0) \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} x_1 & 0 & \cdots & 0 \\ 0 & \frac{\partial}{\partial x_2} x_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \frac{\partial}{\partial x_n} x_n \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\
&= I \quad (I \text{ 是单位矩阵，也就是主对角线都是 1 的矩阵。})
\end{aligned}
$$

在继续之前，请确保你能够推导出上述每一步。如果遇到困难，可以单独考虑矩阵的每个元素，并应用通常的标量导数规则。这是一个通常有用的技巧：将向量表达式简化为一组标量表达式，然后对所有偏导数进行求解，最后适当地将结果组合成向量和矩阵。

还要注意跟踪矩阵是垂直的 $\mathbf{x}$ 还是水平的 $\mathbf{x}^T$，其中 $\mathbf{x}^T$ 表示 $\mathbf{x}$ 的转置。同时，请确保注意某个函数是标量值函数 $y = \ldots$ 还是函数向量（或向量值函数）$\mathbf{y} = \ldots$。

3.2 示例

假设

$$\mathbf{x} = [x_1, x_2, x_3]$$

以及

$$\mathbf{y} = [y_1, y_2]$$

和

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

其中

$$z_1 = g_1(\mathbf{y})$$
$$z_2 = g_2(\mathbf{y})$$
$$z_3 = g_3(\mathbf{y})$$

以及

$$y_1 = f_1(\mathbf{x})$$
$$y_2 = f_2(\mathbf{x})$$

使用向量的写法只是多元函数的简洁表示法。所以，

$$\begin{aligned} z_1 &= g_1(y_1, y_2) \\ &= g_1(f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3)) \end{aligned}$$

所以

$$\frac{\partial z_1}{\partial x_1} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_1}$$

同样的道理

$$\frac{\partial z_1}{\partial x_2} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_2}$$

还有

$$\frac{\partial z_1}{\partial x_3} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_3} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_3}$$

所以有，

$$\begin{aligned} \frac{\partial z_1}{\partial \mathbf{x}} &= [\frac{\partial z_1}{\partial x_1}, \frac{\partial z_1}{\partial x_2}, \frac{\partial z_1}{\partial x_3}] \\ &= [\frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_1}, \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_2}, \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_3} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_3}] \\ &= [\frac{\partial z_1}{\partial y_1}, \frac{\partial z_1}{\partial y_2}] \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} \end{aligned}$$

所以有

$$\begin{aligned} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} &= \begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} \\ \frac{\partial z_3}{\partial y_1} & \frac{\partial z_3}{\partial y_2} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} \\ &= \frac{\partial \mathbf{z}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}^{\mathbf{T}}}{\partial \mathbf{x}} \end{aligned}$$

所以一个 3×1 的向量 $\mathbf{z}$ 对一个 1×3 的向量 $\mathbf{x}$ 进行求导的结果是一个 3×3 的矩阵。
这里就是 $\mathbf{y}$ 作为输出时，应该是列向量。作为输入时，是行向量。所以这里有点疑问。

3.3 向量进行逐元素二元运算时的导数

对向量进行逐元素的二元操作，例如向量加法 $\mathbf{w}+\mathbf{x}$，是很重要的，因为我们可以将许多常见的向量操作（例如向量与标量的乘法）表示为逐元素的二元操作。我们所说的“逐元素二元操作”只是意味着将一个运算符应用于每个向量的第一个元素，以获得输出的第一个元素，然后对输入的第二个元素进行相同操作，以获得输出的第二个元素，依此类推。这就是 numpy 或 tensorflow 中所有基本数学运算符的默认应用方式。例如，在深度学习中经常出现的例子有 $max(\mathbf{w},\mathbf{x})$ 和 $\mathbf{w}>\mathbf{x}$（返回一个由 1 和 0 组成的向量）。

我们可以用符号 $\mathbf{y}=\mathbf{f}(\mathbf{w})\bigcirc\mathbf{g}(\mathbf{x})$ 来概括逐元素的二元操作，其中 $m=n=|y|=|w|=|x|$。（提示：$|x|$ 是 x 中的元素数量。）符号 $\bigcirc$ 表示任何逐元素运算符（例如 +），而不是 $\circ$ 函数复合运算符。当我们放大查看标量方程时，方程 $\mathbf{y}=\mathbf{f}(\mathbf{w})\bigcirc\mathbf{g}(\mathbf{x})$ 的样子如下：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{w})\bigcirc g_1(\mathbf{x}) \\ f_2(\mathbf{w})\bigcirc g_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{w})\bigcirc g_n(\mathbf{x}) \end{bmatrix}$$

在这里，我们将 n（而不是 m）个方程垂直书写，以强调逐元素运算符的结果给出了大小为 $m=n$ 的向量结果。

利用上一节的思想，我们可以看到关于 $\mathbf{w}$ 的雅可比矩阵的一般情况是一个方阵：

$$J_{\mathbf{w}}=\frac{\partial\mathbf{y}}{\partial\mathbf{w}}=\begin{bmatrix} \frac{\partial}{\partial w_1}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) & \frac{\partial}{\partial w_2}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) & \dots & \frac{\partial}{\partial w_n}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) \\ \frac{\partial}{\partial w_1}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) & \frac{\partial}{\partial w_2}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) & \dots & \frac{\partial}{\partial w_n}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) \\ & & \dots & \\ \frac{\partial}{\partial w_1}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) & \frac{\partial}{\partial w_2}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) & \dots & \frac{\partial}{\partial w_n}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) \end{bmatrix}$$

对于 $\mathbf{x}$ 的雅可比矩阵是：

$$J_{\mathbf{x}}=\frac{\partial\mathbf{y}}{\partial\mathbf{x}}=\begin{bmatrix} \frac{\partial}{\partial x_1}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) & \frac{\partial}{\partial x_2}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) & \dots & \frac{\partial}{\partial x_n}(f_1(\mathbf{w})\bigcirc g_1(\mathbf{x})) \\ \frac{\partial}{\partial x_1}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) & \frac{\partial}{\partial x_2}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) & \dots & \frac{\partial}{\partial x_n}(f_2(\mathbf{w})\bigcirc g_2(\mathbf{x})) \\ & & \dots & \\ \frac{\partial}{\partial x_1}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) & \frac{\partial}{\partial x_2}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) & \dots & \frac{\partial}{\partial x_n}(f_n(\mathbf{w})\bigcirc g_n(\mathbf{x})) \end{bmatrix}$$

这确实是个复杂的问题，但幸运的是，雅可比矩阵通常是一个对角矩阵，即除了对角线以外的地方都是零的矩阵。因为这大大简化了雅可比矩阵，我们来详细检查一下在什么情况下雅可比矩阵会简化为逐元素操作的对角矩阵。

在对角雅可比矩阵中，所有非对角元素都是零，即 $\frac{\partial}{\partial w_j}(f_i(\mathbf{w})\bigcirc g_i(\mathbf{x}))=0$，其中 $j\neq i$。（注意，我们是对 w_j 而不是 w_i 进行偏导数运算。）在什么条件下这些非对角元素为零？恰恰是在 f_i 和 g_i 对 w_j 是常数时，即 $\frac{\partial}{\partial w_j}f_i(\mathbf{w})=\frac{\partial}{\partial w_j}g_i(\mathbf{x})=0$。无论运算符是什么，如果这些偏导数为零，运算结果也将为零，$0\bigcirc 0=0$，无论如何，常数的偏导数为零。

当 f_i 和 g_i 不依赖于 w_j 时，这些偏导数为零。我们知道逐元素操作意味着 f_i 仅仅是 w_i 的函数，而 g_i 仅仅是 x_i 的函数。例如，$\mathbf{w}+\mathbf{x}$ 计算的是 w_i+x_i。因此，$f_i(\mathbf{w})\bigcirc g_i(\mathbf{x})$ 简化为 $f_i(w_i)\bigcirc g_i(x_i)$，目标变为 $\frac{\partial}{\partial w_j}f_i(w_i)=\frac{\partial}{\partial w_j}g_i(x_i)=0$。当 $j\neq i$ 时，$f_i(w_i)$ 和 $g_i(x_i)$ 对于关于 w_j 的偏导数运算符看起来像常数，因此非对角线上的偏导数为零。（符号 $f_i(w_i)$ 在技术上是对我们符号的滥用，因为 f_i 和 g_i 是向量的函数，而不是单个元素的函数。我们应该写成类似 $\hat{f}_i(w_i)=f_i(\mathbf{w})$ 的形式，但那样会使方程更加混乱，而程序员习惯于函数重载，所以我们还是继续使用这种符号。）

我们将在稍后利用这个简化，并将 $f_i(\mathbf{w})$ 和 $g_i(\mathbf{x})$ 至多访问 w_i 和 x_i 的约束称为*逐元素对角条件*。

在这种条件下，雅可比矩阵对角线上的元素为 $\frac{\partial}{\partial w_i}(f_i(w_i)\bigcirc g_i(x_i))$：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial}{\partial w_1}(f_1(w_1) \bigcirc g_1(x_1)) & & & \\ & \frac{\partial}{\partial w_2}(f_2(w_2) \bigcirc g_2(x_2)) & & \mathbf{0} \\ & & \dots & \\ \mathbf{0} & & & \frac{\partial}{\partial w_n}(f_n(w_n) \bigcirc g_n(x_n)) \end{bmatrix}$$

（大的 “0” 是一个简写，表示所有非对角线元素都是 0。）

更简洁地，我们可以写为：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = diag\left(\frac{\partial}{\partial w_1}(f_1(w_1) \bigcirc g_1(x_1)),\ \frac{\partial}{\partial w_2}(f_2(w_2) \bigcirc g_2(x_2)),\ \dots,\ \frac{\partial}{\partial w_n}(f_n(w_n) \bigcirc g_n(x_n))\right)$$

和

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = diag\left(\frac{\partial}{\partial x_1}(f_1(w_1) \bigcirc g_1(x_1)),\ \frac{\partial}{\partial x_2}(f_2(w_2) \bigcirc g_2(x_2)),\ \dots,\ \frac{\partial}{\partial x_n}(f_n(w_n) \bigcirc g_n(x_n))\right)$$

其中 $diag(\mathbf{x})$ 构造一个矩阵，其对角元素来自向量 $\mathbf{x}$。

由于我们进行大量简单的向量运算，二元逐元素操作中的一般函数 $\mathbf{f}(\mathbf{w})$ 通常只是向量 $\mathbf{w}$。每当一般函数是一个向量时，我们知道 $f_i(\mathbf{w})$ 简化为 $f_i(w_i) = w_i$。例如，向量加法 $\mathbf{w} + \mathbf{x}$ 符合我们的逐元素对角条件，因为 $\mathbf{f}(\mathbf{w}) + \mathbf{g}(\mathbf{x})$ 有标量方程 $y_i = f_i(\mathbf{w}) + g_i(\mathbf{x})$，这简化为 $y_i = f_i(w_i) + g_i(x_i) = w_i + x_i$，其偏导数为：

$$\frac{\partial}{\partial w_i}(f_i(w_i) + g_i(x_i)) = \frac{\partial}{\partial w_i}(w_i + x_i) = 1 + 0 = 1$$

$$\frac{\partial}{\partial x_i}(f_i(w_i) + g_i(x_i)) = \frac{\partial}{\partial x_i}(w_i + x_i) = 0 + 1 = 1$$

这给我们带来了 $\frac{\partial(\mathbf{w}+\mathbf{x})}{\partial \mathbf{w}} = \frac{\partial(\mathbf{w}+\mathbf{x})}{\partial \mathbf{x}} = I$，即单位矩阵，因为对角线上的每个元素都是 1。$I$ 代表适当维度的平方单位矩阵，除了对角线以外的地方都是零，而对角线包含全 1。

考虑到这个特殊情况的简单性，即 $f_i(\mathbf{w})$ 简化为 $f_i(w_i)$，你应该能够推导出常见的逐元素二元运算在向量上的雅可比矩阵：

Op	**Partial with respect to w**
+	$\frac{\partial(\mathbf{w}+\mathbf{x})}{\partial \mathbf{w}} = diag(\dots \frac{\partial(w_i+x_i)}{\partial w_i} \dots) = diag(\vec{1}) = I$
−	$\frac{\partial(\mathbf{w}-\mathbf{x})}{\partial \mathbf{w}} = diag(\dots \frac{\partial(w_i-x_i)}{\partial w_i} \dots) = diag(\vec{1}) = I$
$\otimes$	$\frac{\partial(\mathbf{w}\otimes\mathbf{x})}{\partial \mathbf{w}} = diag(\dots \frac{\partial(w_i\times x_i)}{\partial w_i} \dots) = diag(\mathbf{x})$
$\oslash$	$\frac{\partial(\mathbf{w}\oslash\mathbf{x})}{\partial \mathbf{w}} = diag(\dots \frac{\partial(w_i/x_i)}{\partial w_i} \dots) = diag(\dots \frac{1}{x_i} \dots)$

Op	**Partial with respect to x**
+	$\frac{\partial(\mathbf{w}+\mathbf{x})}{\partial \mathbf{x}} = I$
−	$\frac{\partial(\mathbf{w}-\mathbf{x})}{\partial \mathbf{x}} = diag(\dots \frac{\partial(w_i-x_i)}{\partial x_i} \dots) = diag(-\vec{1}) = -I$
$\otimes$	$\frac{\partial(\mathbf{w}\otimes\mathbf{x})}{\partial \mathbf{x}} = diag(\mathbf{w})$
$\oslash$	$\frac{\partial(\mathbf{w}\oslash\mathbf{x})}{\partial \mathbf{x}} = diag(\dots \frac{-w_i}{x_i^2} \dots)$

$\otimes$ 和 $\oslash$ 运算符分别表示逐元素乘法和除法；$\otimes$ 有时被称为哈达玛积。对于逐元素乘法和除法没有标准的符号，因此我们使用与我们的一般二元运算符号一致的方法。