Notation for Multivariable Derivatives

Pramook Khungurn

April 24, 2022

I have been reading many papers on deep learning and computer graphics, and its unavoidable to talk about multivariable derivatives: gradients, Jacobian matrices, and all that. It advantageous to develop a consistent system of notations when you talk about these things.

1 The Derivative

- We use the convention that inputs to functions are column vectors. So, if $\mathbf{x} \in \mathbb{R}^n$, we mean that

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

- To save space, we will also write $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. While the notation is horizontal, it denotes a column vector.
- If we want to denote a row vector horizontally, we can already write $\begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}$.
- In this section, we focus on a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. When we write $\mathbf{y} = \mathbf{f}(\mathbf{x})$, we mean that

$$(y_1, y_2, \ldots, y_m) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix}$$

  where $f_i(\mathbf{x})$ denotes the ith component of $\mathbf{f}(\mathbf{x})$. Each $f_i : \mathbb{R}^n \to \mathbb{R}$ is naturally call a **component function**.
- The derivative of $\mathbf{f}$ is a function that sends each point in the domain $\mathbb{R}^n$ to a matrix in $\mathbb{R}^{m \times n}$. If we denote this function by $F : \mathbb{R}^n \to \mathbb{R}^{m \times n}$, then $F(\mathbf{x})$ is a linear approximation of $\mathbf{f}(\mathbf{x})$ near $\mathbf{x}$. In other words, for any small vector $\boldsymbol{\varepsilon}$, we have that

$$\mathbf{f}(\mathbf{x} + \boldsymbol{\varepsilon}) \approx \mathbf{f}(\mathbf{x}) + F(\mathbf{x})\boldsymbol{\varepsilon}.$$

- The derivative can be denoted by a number of notations, including

$$D\mathbf{f}, \quad \nabla \mathbf{f}, \quad \frac{\partial \mathbf{f}}{\partial \mathbf{x}}.$$

- While $\partial \mathbf{f} / \partial \mathbf{x}$ can make the chain rule looks pretty, I believe it should not be used in complicated situations.

- If you want to evaluate the derivative at point $\mathbf{x}_0$, you would have to write

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_0) \quad \text{or} \quad \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0}$$

  Both of them are quite handful to write.

- If you want to discuss the derivative as a function of its input $\mathbf{x}$, the you may sometimes write

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}).$$

  However, you see that the two $\mathbf{x}$'s are not the same thing! The $\mathbf{x}$ after ∂ denotes the argument of $\mathbf{f}$. It is symbolic and should not be substituted with numbers. However, the $\mathbf{x}$ inside the parentheses is a free variable, which means that we can substitute numbers into it. This variable capture confused me countless times, and I wish to avoid it in the future.

- I, therefore, advocate the use of $\nabla \mathbf{f}$ because it makes $\nabla \mathbf{f}(\mathbf{x})$ totally unambiguous. It is also quite consistent with gradient of the scalar function f, which is written as ∇f.

- With the advocated notation, the chain rule can be written as follows.

  Theorem 1 (Chain rule). *Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^n$ be differentiable functions. Let $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$ be a function obtained by composing $\mathbf{f}$ with $\mathbf{g}$. In other words, for any $\mathbf{x} \in \mathbb{R}^p$, we have that*

$$\mathbf{h}(\mathbf{x}) = \mathbf{f}\big(\mathbf{g}(\mathbf{x})\big).$$

  Then, it follows that

$$\nabla \mathbf{h}(\mathbf{x}) = \nabla(\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = \nabla \mathbf{f}\big(\mathbf{g}(\mathbf{x})\big) \nabla \mathbf{g}(\mathbf{x}).$$

- Note that, in the above notation, it's totally clear where $\nabla \mathbf{f}(\cdot)$ and $\nabla \mathbf{g}(\cdot)$ are evaluated at.

- With the partial derivative notation, the chain rule may be written as

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}}.$$

  While it does look pretty and resembles the usual cancellation rule, it requires a lot of mental gymnastics to parse.

  - First, the $\mathbf{f}$ on the LHS is not the same one as that on the RHS. The $\mathbf{f}$ on the LHS is an alternative notation for $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$, which is a function with signature $\mathbb{R}^p \to \mathbb{R}^m$. On the other hand, the $\mathbf{f}$ on the RHS has signature $\mathbb{R}^n \to \mathbb{R}^m$.
  - Again, note that $\mathbf{g}$ is a function by definition. But the $\mathbf{g}$ in $\partial \mathbf{f}/\partial \mathbf{g}$ is not a function! It denotes the argument of $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. This is confusing as hell!
  - It is not clear at all where the derivative functions are evaluated at.

2 Partial Derivatives

- Since we have decided to do away with the partial derivate notation, we might as well do away with it at the scalar function level.

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a scalar function. The **partial derivative with respect to the ith argument**, denoted by $\nabla_i f$, is a function of signature $\mathbb{R}^n \to \mathbb{R}$ such that

$$\nabla_i f(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{f(x_1, \ldots, x_i + \varepsilon, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{\varepsilon}.$$

  In other words, if $\mathbf{e}_i$ be the one-hot vector whose ith component is 1, then

$$f(\mathbf{x} + \varepsilon \mathbf{e}_i) \approx f(\mathbf{x}) + \varepsilon \nabla_i f(\mathbf{x})$$

  for all small ε.

- Again, the advantage of using "$\nabla_i f$" over "$\partial f / \partial x_i$" is that we do not introduce a variable to denote the ith argument in "$\nabla_i f$."

- With the new notation for partial derivatives, the gradient $\nabla f(\mathbf{x})$ can be written as:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \nabla_1 f(\mathbf{x}) & \nabla_2 f(\mathbf{x}) & \cdots & \nabla_n f(\mathbf{x}) \end{bmatrix}$$

  It is clear that ∇f is a function of signagure $\mathbb{R}^n \to \mathbb{R}^{1 \times n}$.

- The derivative $\nabla \mathbf{f}(\mathbf{x})$ can now be written as

$$\nabla \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \nabla_1 f_1(\mathbf{x}) & \nabla_2 f_1(\mathbf{x}) & \cdots & \nabla_n f_1(\mathbf{x}) \\ \nabla_1 f_2(\mathbf{x}) & \nabla_2 f_2(\mathbf{x}) & \cdots & \nabla_n f_2(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla_1 f_m(\mathbf{x}) & \nabla_2 f_m(\mathbf{x}) & \cdots & \nabla_n f_m(\mathbf{x}) \end{bmatrix}.$$

- We can also talk about the partial derivative of $\mathbf{f}$ with respect to the ith argument:

$$\nabla_i \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \nabla_i f_1(\mathbf{x}) \\ \nabla_i f_2(\mathbf{x}) \\ \vdots \\ \nabla_i f_m(\mathbf{x}) \end{bmatrix}.$$

- **Theorem 2 (Law of total derivatives).** *Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^n$. Let $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$. Then,*

$$\nabla_j h_i(\mathbf{x}) = \sum_{k=1}^{n} \nabla_k f_i(\mathbf{g}(\mathbf{x})) \nabla_j g_k(\mathbf{x}).$$

  Proof. By the chain rule,

$$\nabla \mathbf{h}(\mathbf{x}) = \nabla \mathbf{f}(\mathbf{g}(\mathbf{x})) \nabla \mathbf{g}(\mathbf{x}).$$

  In other words

$$\begin{bmatrix} \nabla_1 h_1(\mathbf{x}) & \cdots & \nabla_p h_1(\mathbf{x}) \\ \nabla_1 h_2(\mathbf{x}) & \cdots & \nabla_p h_2(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \nabla_1 h_m(\mathbf{x}) & \cdots & \nabla_p h_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \nabla_1 f_1(\mathbf{g}(\mathbf{x})) & \cdots & \nabla_n f_1(\mathbf{g}(\mathbf{x})) \\ \nabla_1 f_2(\mathbf{g}(\mathbf{x})) & \cdots & \nabla_n f_2(\mathbf{g}(\mathbf{x})) \\ \vdots & \ddots & \vdots \\ \nabla_1 f_m(\mathbf{g}(\mathbf{x})) & \cdots & \nabla_n f_m(\mathbf{g}(\mathbf{x})) \end{bmatrix} \begin{bmatrix} \nabla_1 g_1(\mathbf{x}) & \cdots & \nabla_p g_1(\mathbf{x}) \\ \nabla_1 g_2(\mathbf{x}) & \cdots & \nabla_p g_2(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \nabla_1 g_n(\mathbf{x}) & \cdots & \nabla_p g_n(\mathbf{x}) \end{bmatrix}.$$

Hence,

$$\nabla_j h_i(\mathbf{x}) = \begin{bmatrix} \nabla_1 f_i(\mathbf{g}(\mathbf{x})) & \nabla_2 f_i(\mathbf{g}(\mathbf{x})) & \cdots & \nabla_n f_i(\mathbf{g}(\mathbf{x})) \end{bmatrix} \begin{bmatrix} \nabla_j g_1(\mathbf{x}) \\ \nabla_j g_2(\mathbf{x}) \\ \vdots \\ \nabla_j g_n(\mathbf{x}). \end{bmatrix}$$

$$= \sum_{k=1}^{n} \nabla_k f_i(\mathbf{g}(\mathbf{x})) \nabla_j g_k(\mathbf{x}).$$

as required. □

- What's nice about the above theorem is that everything is done in terms of partial derivatives. No mental gymnastics is required to distinguish between the total differential $\mathrm{d}x$ and the partial differential ∂x.

3 Block Notations

- Again, let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$.
- For $\mathbf{x} \in \mathbb{R}^n$, we can use the `numpy` notation $\mathbf{x}[i : j]$ and $\mathbf{x}_{i:j}$ to denote $(x_i, x_{i+1}, \ldots, x_j)$.
- Suppose we partition the range $1 : n$ into l blocks with the ranges

  $$1 : i_1, \quad (i_1 + 1) : i_2, \quad (i_2 + 1) : i_3, \quad \ldots, \quad (i_{l-1} + 1) : n.$$

  Then, the blocks of $\mathbf{x}$ are

  $$\mathbf{x}[1 : i_1], \quad \mathbf{x}[(i_1 + 1) : i_2], \quad \mathbf{x}[(i_2 + 1) : i_3], \quad \ldots, \quad \mathbf{x}[(i_{l-1} + 1) : n],$$

  or

  $$\mathbf{x}_{1:i_1}, \quad \mathbf{x}_{(i_1+1):i_2}, \quad \mathbf{x}_{(i_2+1):i_3}, \quad \ldots, \quad \mathbf{x}_{(i_{l-1}+1):n}.$$

- However, the range notations are long, and we also need the index variables $i_1, \ldots, i_{l-1}$. Instead, we shall denote the jth range by just $\S j$ (as in "Chapter j"). The blocks do become much shorter:

  $$\mathbf{x}[\S 1], \quad \mathbf{x}[\S 2], \quad \ldots, \quad \mathbf{x}[\S l],$$

  or

  $$\mathbf{x}_{\S 1}, \quad \mathbf{x}_{\S 2}, \quad \ldots, \quad \mathbf{x}_{\S l}.$$

  Now, we can succinctly write

  $$\mathbf{x} = (\mathbf{x}_{\S 1}, \mathbf{x}_{\S 2}, \ldots, \mathbf{x}_{\S l}).$$

- If we partition $\mathbb{R}^n$ into l blocks and $\mathbb{R}^m$ into k blocks, then we can view $\nabla \mathbf{f}(\mathbf{x})$ as a block matrix with $k \times l$ blocks. Naturally, the (i, j)-block is denoted by

  $$\nabla_{\S j} \mathbf{f}_{\S i}(\mathbf{x}).$$

4 Subscription as Differentiation

- For brevity, we sometimes see $\partial f/\partial x$ being abbreviated as just f_x. We shall create this type of abbreviated, but unambiguous notation in this section.
- It is tempting to use f_i to represent the partial derivative with respect to the ith argument. However, we already use f_i to denote the ith component of $\mathbf{f}$.
- To disambiguate, we shall use $f_{\nabla i}$ to denote the partial derivative in question.
- In this way, we can denote the partial derivative with respect to the jth argument of f_i with $(f_i)_{\nabla j}$ or just $f_{i\nabla j}$.
- In this way, the derivative $\nabla\mathbf{f}$ can also be written as $\mathbf{f}_\nabla$, and we have a very intuitive notation for the Jacobian matrix:

$$\mathbf{f}_\nabla(\mathbf{x}) = \begin{bmatrix} f_{1\nabla 1}(\mathbf{x}) & f_{1\nabla 2}(\mathbf{x}) & \cdots & f_{1\nabla n}(\mathbf{x}) \\ f_{2\nabla 1}(\mathbf{x}) & f_{2\nabla 2}(\mathbf{x}) & \cdots & f_{2\nabla n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{m\nabla 1}(\mathbf{x}) & f_{m\nabla 2}(\mathbf{x}) & \cdots & f_{m\nabla n}(\mathbf{x}). \end{bmatrix}$$

- The notation for second-order derivatives would be $\nabla_i\nabla_j f$ or $f_{\nabla j\nabla i}$. This is getting long, so we introduce the following shorthands:

$$\nabla_{i,j} f := \nabla_i\nabla_j f,$$
$$f_{\nabla j,i} := f_{\nabla j\nabla i}.$$

- Since the order of differentiation does not matter, we have that

$$f_{\nabla i,j} = f_{\nabla j,i}.$$

References