Derivatives, gradients

Csaba Szepesvári

Friday 31st March, 2023

Gradients

Gradients guide us on our way,
In calculus and life, they say,
A path to take to reach the peak,
To optimize what we seek.

From slopes to vectors, we explore,
The changes that each direction bore,
In multivariate spaces, we find,
Gradients that lead to peace of mind.
– ChatGPT

1 Derivatives and gradients

The idea of a derivative is that it gives a linear approximation to a function:

Definition 1.1. Take a function $f : \mathbb{R}^d \to \mathbb{R}$. We say that f is differentiable at $x \in \mathbb{R}^d$ if there exists a vector $a \in \mathbb{R}^d$ such that

$$\lim_{h \to 0} \frac{f(x+h) - (f(x) + a^\top h)}{\|h\|} = 0\,, \tag{1}$$

where $h \in \mathbb{R}^d$. The vector a here is called the gradient of the function f at x and $a^\top$ is called the derivative of the function f at x. The gradient is denoted by $\nabla f(x)$, while the derivative is just denoted by $f'(x)$.[1]

Which norm $\|\cdot\|$ to use in the definition above? Any norm! This is because in Euclidean spaces (finite dimensional real normed vector spaces) all norms are equivalent in the sense that if $\|\cdot\|$ and $\|\cdot\|'$ are two norms, then there exist constants $0 < c \le C$ such that for any $x \in \mathbb{R}^d$,

$$c\|x\| \le \|x\|' \le C\|x\|\,.$$

Then, one can show that if f is differentiable at some point $x_0 \in \mathbb{R}^d$ if norm $\|\cdot\|$ is used and its gradient is $a \in \mathbb{R}^d$, then f is also differentiable at $x_0 \in \mathbb{R}^d$ when in the definition we switch to the norm $\|\cdot\|'$ and with this norm, we also get the same gradient $a \in \mathbb{R}^d$.

An equivalent way of writing Equation (1) uses the $o(\cdot)$ notation ("little-oh notation"):

$$f(x+h) - (f(x) + a^\top h) = o(\|h\|)\,. \tag{2}$$

Here, and in what follows, the meaning of the above is that the right-hand side, as a function of h, as h approaches zero converges to zero faster than $\|h\|$ converges to zero.[2]

It is wortwhile to also recall the definition of partial derivatives. With the help of the $o(\cdot)$ notation, the definition is as follows: The ith partial derivative of f, $\frac{\partial f}{\partial x_i}(x)$ is the unique value $a_1 \in \mathbb{R}$ such that

$$f(x + e_i h_1) - (f(x) + a_1 \cdot h_1) = o(|h_1|), \tag{3}$$

where $h_1 \in \mathbb{R}$ is a real number and e_i is the ith unit vector in the standard Euclidean basis (i.e., $e_{ij} = 0$ if $j \neq i$ while $e_{ii} = 1$). Also, compare Equation (3) with Equation (2).

Equation (2) has an intuitive geometric meaning as well: The (non-linear) function f is well approximated by the hyperplane $x+h \mapsto f(x)+a^\top h$ when h is small in that the error of this approximation decays to zero "strictly faster" than $\|h\|$ as h gets small.

The following theorem (that relates the gradient of a function to its partial derivatives) holds true:

Theorem 1.2. *Take a function $f : \mathbb{R}^d \to \mathbb{R}$ and let $x \in \mathbb{R}^d$. Assume that the partial derivatives of f all exist in a neighborhood of x and are also continuous at x. Then f is differentiable and its gradient satisfies*

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{pmatrix}.$$

Proof. Recall that a function $g : \mathbb{R}^d \to \mathbb{R}$ is continuous at a point x if $\lim_{h\to 0} g(x+h) - g(x) = 0$, or with our little-oh notation, we can write this concisely as $g(x+h) - g(x) = o(1)$. We will show that (2) holds, or, equivalently, $f(x+h) = f(x) + a^\top h + o(\|h\|)$, where it will be convenient to use the one-norm for $\|\cdot\|$ (i.e., $\|h\| = |h_1| + \cdots + |h_d|$). We start by writing $f(x+h)$ as a telescoping sum:

$$\begin{aligned} f(x+h) = f(x) &+ \big(f(x + e_1 h_1) - f(x)\big) \\ &+ \big(f(x + e_1 h_1 + e_2 h_2) - f(x + e_1 h_1)\big) \\ &\vdots \\ &+ \big(f(x + e_1 h_1 + \cdots + e_d h_d) - f(x + e_1 h_1 + \cdots + e_{d-1} h_{d-1})\big)\,, \end{aligned} \tag{4}$$

where $e_1 = (1, 0, \ldots, 0)^\top$, $e_2 = (0, 1, 0, \ldots, 0)^\top$, $\ldots$, $e_d = (0, \ldots, 0, 1)^\top$ are the unit vectors in the standard Euclidean basis of $\mathbb{R}^d$, and we use the standard convention that $h = (h_1, \ldots, h_d)^\top$ with $h_1, \ldots, h_d \in \mathbb{R}$. (Observe that this means that h can be decomposed as $h = e_1 h_1 + e_2 h_2 + \cdots + e_d h_d$, which is what we used while writing Equation (4).) Now, for all $2 \le i \le d$,

$$\begin{aligned}
&f(x + e_1h_1 + \cdots + e_ih_i) - f(x + e_1h_1 + \cdots + e_{i-1}h_{i-1}) \\
&\quad = \left(\frac{\partial f}{\partial x_i}(x + e_1h_1 + \cdots + e_{i-1}h_{i-1})\right) h_i + o(|h_i|) && \text{(by Equation (3))} \\
&\quad = \frac{\partial f}{\partial x_i}(x)h_i + \left(\frac{\partial f}{\partial x_i}(x + e_1h_1 + \cdots + e_{i-1}h_{i-1}) - \frac{\partial f}{\partial x_i}(x)\right) h_i + o(|h_i|) \\
&\quad = \frac{\partial f}{\partial x_i}(x)h_i + o_{h_1,\ldots,h_{i-1}}(1)h_i + o(|h_i|) && \text{(continuity of } \tfrac{\partial f}{\partial x_i} \text{ at } x) \\
&\quad = \frac{\partial f}{\partial x_i}(x)h_i + o_{h_1,\ldots,h_d}(|h_i|)\,.
\end{aligned}$$

Whereas for $i = 1$, using the continuity of $\frac{\partial f}{\partial x_1}$ at x, we would get

$$f(x + e_1h_1) - f(x) = \frac{\partial f}{\partial x_1}(x)h_1 + o_{h_1,\ldots,h_d}(|h_1|).$$

Plugging these results into Equation (4) gives us

$$f(x+h) - f(x) = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(x)h_i + \sum_{i=1}^{d} o_{h_1,\ldots,h_d}(|h_i|) = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(x)h_i + o(\|h\|)\,.$$

The above equation, along with Equation (2), immediately gives us the desired relation between the gradient of the function ∇f and its partial derivatives (why?). □

We cannot lift the requirement that the partial derivatives of f are continuous at x in the last result. Indeed, consider the following function:

$$f(x_1, x_2) = \begin{cases} x_1\,, & \text{if } x_2 = x_1^2\,; \\ 0\,, & \text{otherwise}\,. \end{cases}$$

We will consider the differentiability of f at $x = (0,0)^\top = 0$. What are the partial derivatives of f at 0? We have, for $h_1 \in \mathbb{R}$,

$$\begin{aligned}
f(h_1, 0) - f(0,0) &= 0 - 0 = 0 = 0 \cdot h_1 + o(|h_1|)\,, \\
f(0, h_1) - f(0,0) &= 0 - 0 = 0 = 0 \cdot h_1 + o(|h_1|)\,.
\end{aligned}$$

Hence both the partial derivatives exist and are zero. So if we did not need to worry about the existence and continuity of the partial derivatives in a small neighborhood of zero, we could think that the derivative of f at zero should be $(0,0)$. But can this be the derivative? If a is the gradient of f at 0, $f(h) - f(0) = a^\top h + o(\|h\|)$ must hold: no matter in what way we have $h \to 0$, we need to have that $|f(h) - f(0) - a^\top h|/\|h\| \to 0$. Let us now compute $f(h) - f(0)$ for the special choice of $h = (h_1, h_1^2)^\top$ (again using the one-norm $\|\cdot\|$). With this choice, we have

$$f(h_1, h_1^2) - f(0) = h_1 - 0 = h_1 \ne o(|h_1|) = o(|h_1| + |h_1^2|) = (0,0) \cdot \begin{pmatrix} h_1 \\ h_1^2 \end{pmatrix} + o(|h_1| + |h_1^2|)\,.$$

Indeed, if, for example, $h_1 \to 0^+$, $\lim_{h_1 \to 0^+} \frac{h_1}{|h_1|} = 1 \neq 0$. (This shows that for this choice of h, we have $f(h) - f(0) \neq a^\top h + o(\|h\|)$, i.e. $a = (0,0)^\top$ cannot be the gradient of f.)

Intuitively, the plane that best approximates the function f at 0 when we restrict the domain of f to $E = \{(x_1, x_1^2) \,:\, x_1 \in \mathbb{R}\}$ is $\mathcal{P} = \{(x_1, x_2, x_1) \,:\, x_1 \in \mathbb{R}, x_2 \in \mathbb{R}\}$. At the same time, the plane that best approximates the function f at 0 when we exclude from the domain of f the set E is $\mathcal{P}' = \{(x_1, x_2, 0) \,:\, x_1 \in \mathbb{R}, x_2 \in \mathbb{R}\}$. Since $\mathcal{P} \neq \mathcal{P}'$, f is not differentiable at 0.

So what went wrong? As it turns out, the problem is not that the partial derivatives of f are discontinuous at 0: They do not even exist. For example, the partial derivative with respect to x_1 does not exist for any $x = (x_1, x_2) \in \mathbb{R}^2$ such that $x \neq (0,0)$ and $x_2 = x_1^2$. Indeed, for $x_1 \neq 0$, $s \mapsto f(x_1 + s, x_1^2)$ is not even continuous, let alone differentiable. Indeed, for $s \neq 0$,

$$f(x_1 + s, x_1^2) - f(x_1, x_1^2) = 0 - x_1 \neq o_s(1)\,.$$

Now, consider the example

$$f(x_1, x_2) = \begin{cases} x_1\,, & \text{if } x_2 \neq x_1^2\,; \\ 0\,, & \text{otherwise}\,. \end{cases}$$

As it is not hard to see, this function has the same problem as the previous one.

The canonical example of a function that is non-differentiable at (say) 0, and whose partial derivatives exist in a neighborhood of 0 but are discontinuous, is as follows:

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 x_2}{x_1^2 + x_2^2}\,, & \text{if} (x_1, x_2) \neq (0,0)\,; \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Some plots and the reasoning why this is the case can be found here. Note that any function that is non-differentiable at some point x must be such that either the partial derivatives of the function do not exist in some neighborhood of x or if they exist, some of them must be discontinuous at x.

Note however that the continuity of partial derivatives is not required for the differentiability of a function. The following function (from here) has the property that it is differentiable at zero but has discontinuous partial derivatives there:

$$f(x_1, x_2) = \begin{cases} (x_1^2 + x_2^2) \sin\left(\frac{1}{\sqrt{x_1^2 + x_2^2}}\right)\,, & \text{if } (x_1, x_2) \neq (0,0)\,; \\ 0, & \text{otherwise}\,. \end{cases}$$

Finally, note that the main difference between the derivative (or the gradient) and the partial derivatives is that the latter are tied to how a coordinate system is set up, while the former is independent of it.[3]

Notes

1. Why the distinction between the gradient and the derivative? Well, sometimes it is more convenient to work with column vectors (then we use the gradient), and sometimes it is more convenient to work with row vectors (then we use the derivative, also known as the total derivative).

2. There is a lot more to the $o(\cdot)$ notation, which we now review. This can be skipped at first reading, but it may be worthwhile to come back to later. First, let us look at the definition: Given a non-negative valued function u that is defined in a neighborhood of zero, $g(h) = o(u(h))$ means that $\lim_{h\to 0} g(h)/u(h)$ exists and is zero. We will use this definition with h being either a vector of some dimension, or even a scalar. If it becomes ambiguous which of the variables $h_1, \dots, h_d$ need to go to zero, we will list the variables that need to go to zero in the index of $o(\cdot)$. For instance, for $1 \le i_1 < \cdots < i_k \le d$, $o_{h_{i_1},\dots,h_{i_k}}(u(h))$ signifies that variables $h_{i_1}, \dots, h_{i_k}$ need to go to zero. In particular, $g(h) = o_{h_{i_1},\dots,h_{i_k}}(u(h))$ means that $\lim_{h_{i_1},\dots,h_{i_k}\to 0} g(h)/u(h) = 0$. Note that $g(h) = o_{h_{i_1},\dots,h_{i_k}}(u(h))$ implies that $g(h) = o(u(h))$ because $\lim_{h\to 0} g(h)/u(h) = \lim_{h_j\to 0, j\notin\{i_1,\dots,i_k\}} \left(\lim_{h_{i_1},\dots,h_{i_k}\to 0} g(h)/u(h)\right) = \lim_{h_j\to 0, j\notin\{i_1,\dots,i_k\}} 0 = 0$. In the special case when $u(h) = 1$, we write $g(h) = o(1)$, which, of course, just means then that $g(h) \to 0$ as $h \to 0$. We will also write expressions like $g(h) = f(h) + o(u(h))$, which should be interpreted as the statement that $g(h) - f(h) = o(u(h))$. We also allow to write $g(h) = f(h) + v(h)o(u(h))$ which just means $g(h) = f(h) + o(u(h)|v(h)|)$. (Similarly, $g(h) = f(h) + o(v(h)) + o(u(h))$ means $g(h) = f(h) + o(v(h) + u(h))$.)

3. For a fun and intuitive exposition on this point (with a lot of diagrams!), you can read Chapter 2, Feynman Lectures on Physics (Vol. 2). For a more in-depth introduction to vector differential calculus, you can refer Chapter 6, Analysis II by Terry Tao. However, for the purposes of our course, these notes are more than sufficient.