

### 3 What is Differentiability?

There are a lot of ML engineers who brush off the mathematical details of what it means for a function to be differentiable. Algorithmic differentiation (sometimes misleadingly-called “automatic” differentiation) is only guaranteed to work when certain theoretical conditions about the *existence* of a gradient hold. This part of the class is dedicated to explaining that differentiability is not a freebie.

To start our discussion on differentiability, we will begin with a few preliminaries from analysis.

**Definition 3.1** Let  $A: \mathcal{H}_1 \rightarrow \mathcal{H}_2$ . Then  $A$  is **linear** if, for every  $\alpha \in \mathbb{R}$  and every  $x, y \in \mathcal{H}_1$ ,

$$A(\lambda x) = \lambda A(x) \quad \text{and} \quad A(x + y) = A(x) + A(y). \quad (29)$$

**Theorem 3.2 (Riesz-Fréchet representation)** Let  $A: \mathcal{H} \rightarrow \mathbb{R}$  be linear. Then there exists a unique vector  $u \in \mathcal{H}$  such that, for every  $x \in \mathcal{H}$ ,  $A(x) = \langle u | x \rangle$ .

Although at first-glance it looks unrelated, Theorem 3.2 is a central notion for defining the gradient. A necessary (albeit insufficient) condition for the existence of a gradient is the existence of a directional derivative, defined below.

**Definition 3.3** Let  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be proper. The **directional derivative** of  $f$  at  $x \in \text{dom } f$  in the direction  $y \in \mathcal{H}$  is

$$f'(x; y) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}. \quad (30)$$

From Definition 3.3, we point out a few things.

- (i) The limit in (30) might not exist.
- (ii) If  $f$  is convex, then  $f'(x; y) \in [-\infty, +\infty]$ .
- (iii) Even if a directional derivative exists, it might not exist in  $\mathbb{R}$  (since it could be  $+\infty$  or  $-\infty$ ).

**Definition 3.4** Let  $x \in \text{dom } f$ . If  $f'(x; \cdot)$  is linear, we say  $f$  is **differentiable at  $x$** . In this case, the unique vector provided by Theorem 3.2 is called the **gradient** of  $f$  at  $x$  and denoted  $\nabla f(x)$ .

$$f'(x; \cdot) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha \cdot) - f(x)}{\alpha} = \langle \nabla f(x) | \cdot \rangle \quad (31)$$

If  $f$  is differentiable at every  $x \in \text{dom } f$ , we say that  $f$  is **differentiable**.

**Exercise 3.5** Verify that  $\nabla(\frac{1}{2}\|\cdot\|^2)(x) = x$ .

All of the properties we know and love about differentiability (chain rule, product rule, etc.) have to be proven. Here is an example below.

**Proposition 3.6** Let  $A: \mathcal{H}_1 \rightarrow \mathcal{H}_2$  be a linear operator (with adjoint denoted  $A^*$ ), let  $b \in \mathcal{H}_2$ , and let  $f: \mathcal{H} \rightarrow \mathbb{R}$  be proper and differentiable. Set  $g = f(Ax + b)$ . Then  $g$  is differentiable and

$$\nabla g = A^*(\nabla f(A \cdot + b)). \quad (32)$$

*Proof.* Since  $\text{dom } f = \mathcal{H}_2$ ,  $\text{dom } g \neq \emptyset$  so we let  $x \in \text{dom } g$ . By definition,

$$g'(x; y) = \lim_{\alpha \searrow 0} \frac{g(x + \alpha y) - g(x)}{\alpha} \quad (33)$$

$$= \lim_{\alpha \searrow 0} \frac{f(A(x + \alpha y) + b) - f(Ax + b)}{\alpha} \quad (34)$$

$$= \lim_{\alpha \searrow 0} \frac{f(Ax + b + \alpha Ay) - f(Ax + b)}{\alpha} \quad (35)$$

$$= f'(Ax + b; Ay). \quad (36)$$

So the directional derivative of  $g$  exists. Now, since  $f$  is differentiable,

$$g'(x; y) = f'(Ax + b; Ay) = \langle \nabla f(Ax + b) \mid Ay \rangle = \langle A^*(\nabla f(Ax + b)) \mid y \rangle. \quad (37)$$

Hence the directional derivative of  $g$  is linear and  $g$  is differentiable. The specific form of the gradient is constructed in (37)  $\square$

Algorithmic differentiation tools use results like Proposition 3.6 to approximate a gradient of a function by reading its machine code. However, these subroutines do not check the theoretical conditions required for their theorems (e.g., *f must be differentiable*) – this must be done (and is oftentimes unjustly ignored) by the user.

**Definition 3.7** Let  $f$  be proper and differentiable.  $f$  is **smooth** (“ $L$ -smooth”) if there exists  $L > 0$  such that

$$(\forall x, y \in \mathcal{H}) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \quad (38)$$

**Exercise 3.8** Construct a function which is differentiable and nonsmooth.

**Proposition 3.9** Let  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be proper and convex. Then,

$$(\forall x \in \text{dom } f)(\forall y \in \mathcal{H}) \quad f'(x; y - x) + f(x) \leq f(y). \quad (39)$$

*Proof.* By Proposition 2.2, for every  $\alpha \in ]0, 1[$ ,

$$f(x + \alpha(y - x)) - f(x) = f((1 - \alpha)x + \alpha y) - f(x) \quad (40)$$

$$\leq (1 - \alpha)f(x) + \alpha f(y) - f(x) \quad (41)$$

$$= \alpha(f(y) - f(x)). \quad (42)$$

Therefore,

$$\frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \leq f(y) - f(x). \quad (43)$$

Taking the limit as  $\alpha \searrow 0$  implies  $f'(x; y) \leq f(y) - f(x)$ , which in turn yields (39).  $\square$

**Corollary 3.10** Let  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be proper and convex. If  $f$  is differentiable at an interior point  $x$  of its domain<sup>2</sup>, then

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid \nabla f(x) \rangle + f(x) \leq f(y). \quad (44)$$

When the lefthand side of (44) is viewed as a function of  $y$ , we see it is the first-order Taylor series approximation of  $f$ . Therefore, it follows from (39) that a convex differentiable function always remains above its first-order Taylor approximation! This is the motivating idea in defining a (convex) subgradient.

**Definition 3.11** Let  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$ . A vector  $g$  is a **subgradient** of  $f$  at  $x \in \mathcal{H}$  if

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid g \rangle + f(x) \leq f(y). \quad (45)$$

The **subdifferential** of  $f$  at  $x$  is the set of all subgradients, denoted  $\partial f(x)$ .

This leads to the following fundamental theorem for optimization.

**Theorem 3.12 (Fermat's Rule)** Let  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be proper. Then  $x$  is a minimizer of  $f$  if and only if  $0 \in \partial f(x)$ .

*Proof.* By definition,

$$0 \in \partial f(x) \Leftrightarrow (\forall y \in \mathcal{H}) \quad \langle 0 \mid y - x \rangle + f(x) \leq f(y) \quad (46)$$

$$\Leftrightarrow (\forall y \in \mathcal{H}) \quad f(x) \leq f(y). \quad (47)$$

□

Unlike differentiable functions, there are technical conditions we must check in order to get the “standard” rules one would hope for.

**Theorem 3.13 (Sum rule)** Let  $f, g \in \Gamma_0(\mathcal{H})$  and suppose that one of the following holds:

- (i) The interior of  $\text{dom } g$  intersects with  $\text{dom } f$
- (ii)  $\text{dom } g = \mathcal{H}$
- (iii) The relative interiors of  $\text{dom } f$  and  $\text{dom } g$  intersect.

Then  $\partial(f + g) = \partial f + \partial g$ .

Note: If  $f$  is convex and differentiable at  $x \in \mathcal{H}$ , then  $\partial f(x) = \{\nabla f(x)\}$ .

---

<sup>2</sup>the interior of the domain