5 Derivatives in General Vector Spaces

Matrix calculus requires us to generalize concepts of derivative and gradient further, to functions whose inputs and/or outputs are not simply scalars or column vectors. To achieve this, we extend the notion of the ordinary vector **dot product** and ordinary Euclidean vector "length" to general **inner products** and **norms** on **vector spaces**. Our first example will consider familiar matrices from this point of view.

Recall from linear algebra that we can call any set V a "vector space" if its elements can be added/subtracted $x \pm y$ and multiplied by scalars αx (subject to some basic arithmetic axioms, e.g. the distributive law). For example, the set of $m \times n$ matrices themselves form a vector space, or even the set of continuous functions $u(x)$ (mapping $\mathbb{R} \to \mathbb{R}$)—the key fact is that we can add/subtract/scale them and get elements of the same set. It turns out to be extraordinarily useful to extend differentiation to such spaces, e.g. for functions that map matrices to matrices or functions to numbers. Doing so crucially relies on our input/output vector spaces V having a **norm** and, ideally, an **inner product**.

5.1 A Simple Matrix Dot Product and Norm

Recall that for *scalar-valued* functions $f(x) \in \mathbb{R}$ with *vector inputs* $x \in \mathbb{R}^n$ (i.e. n-component "column vectors") we have that

$$df = f(x + dx) - f(x) = f'(x)[dx] \in \mathbb{R}.$$

Therefore, $f'(x)$ is a linear operator taking in the vector dx in and giving a scalar value out. Another way to view this is that $f'(x)$ is the row vector[3] $(\nabla f)^T$. Under this viewpoint, it follows that df is the dot product (or "inner product"):

$$df = \nabla f \cdot dx$$

We can generalize this to any vector space V with inner products! Given $x \in V$, and a scalar-valued function f, we obtain the linear operator $f'(x)[dx] \in \mathbb{R}$, called a "linear form." In order to define the gradient ∇f, we need an inner product for V, the vector-space generalization of the familiar dot product!

Given $x, y \in V$, the inner product $\langle x, y \rangle$ is a map $(\cdot)$ such that $\langle x, y \rangle \in \mathbb{R}$. This is also commonly denoted $x \cdot y$ or $\langle x \mid y \rangle$. More technically, an inner product is a map that is

1. **Symmetric**: i.e. $\langle x, y \rangle = \langle y, x \rangle$ (or conjugate-symmetric,[4] $\langle x, y \rangle = \overline{\langle y, x \rangle}$, if we were using complex numbers),
2. **Linear**: i.e. $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$, and
3. **Non-negative**: i.e. $\langle x, x \rangle := \|x\|^2 \geq 0$, and $= 0$ if and only if $x = 0$.

Note that the combination of the first two properties means that it must also be linear in the left vector (or conjugate-linear, if we were using complex numbers). Another useful consequence of these three properties, which is a bit trickier to derive, is the *Cauchy–Schwarz inequality* $|\langle x, y \rangle| \leq \|x\| \, \|y\|$.

[3] The concept of a "row vector" can be formalized as something called a "covector," a "dual vector," or an element of a "dual space," not to be confused with the *dual numbers* used in automatic differentiation (Sec. 8).

[4] Some authors distinguish the "dot product" from an "inner product" for complex vector spaces, saying that a dot product has no complex conjugation $x \cdot y = y \cdot x$ (in which case $x \cdot x$ need not be real and does not equal $\|x\|^2$), whereas the inner product must be conjugate-symmetric, via $\langle x, y \rangle = \bar{x} \cdot y$. Another source of confusion for complex vector spaces is that some fields of mathematics define $\langle x, y \rangle = x \cdot \bar{y}$, i.e. they conjugate the *right* argument instead of the left (so that it is linear in the left argument and conjugate-linear in the right argument). Aren't you glad we're sticking with real numbers?

Definition 31 (Hilbert Space)

A (complete) vector space with an inner product is called a *Hilbert space*. (The technical requirement of "completeness" essentially means that you can take limits in the space, and is important for rigorous proofs.[a])

[a] Completeness means that any Cauchy sequence of points in the vector space—any sequence of points that gets closer and closer together—has a limit lying within the vector space. This criterion usually holds in practice for vector spaces over real or complex scalars, but can get trickier when talking about vector spaces of functions, since e.g. the limit of a sequence of continuous functions can be a discontinuous function.

Once we have a Hilbert space, we can define the gradient for scalar-valued functions. Given $x \in V$ a Hilbert space, and $f(x)$ scalar, then we have the linear form $f'(x)[dx] \in \mathbb{R}$. Then, under these assumptions, there is a theorem known as the "Riesz representation theorem" stating that *any* linear form (including f') must be an inner product with *something*:

$$f'(x)[dx] = \Big\langle \underbrace{(\text{some vector})}_{\text{gradient } \nabla f\big|_x}, dx \Big\rangle = df.$$

That is, the gradient ∇f is *defined* as the thing you take the inner product of dx with to get df. Note that ∇f always has the "same shape" as x.

The first few examples we look at involve the usual Hilbert space $V = \mathbb{R}^n$ with different inner products.

Example 32

Given $V = \mathbb{R}^n$ with n-column vectors, we have the familiar Euclidean dot product $\langle x, y \rangle = x^T y$. This leads to the usual ∇f.

Example 33

We can have different inner products on $\mathbb{R}^n$. For instance,

$$\langle x, y \rangle_W = w_1 x_1 y_1 + w_2 x_2 y_2 + \ldots w_n x_n y_n = x^T \underbrace{\begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix}}_{W} y$$

for weights $w_1, \ldots, w_n > 0$.

More generally we can define a weighted dot product $\langle x, y \rangle_W = x^T W y$ for any symmetric-positive-definite matrix W ($W = W^T$ and W is positive definite, which is sufficient for this to be a valid inner product).

If we change the definition of the inner product, then we change the definition of the gradient! For example, with $f(x) = x^T A x$ we previously found that $df = x^T (A + A^T) dx$. With the ordinary Euclidean inner product, this gave a gradient $\nabla f = (A + A^T)x$. However, if we use the weighted inner product $x^T W y$, then we would obtain a different "gradient" $\nabla^{(W)} f = W^{-1}(A + A^T)x$ so that $df = \langle \nabla^{(W)} f, dx \rangle$.

In these notes, we will employ the Euclidean inner product for $x \in \mathbb{R}^n$, and hence the usual ∇f, unless noted otherwise. However, weighted inner products are useful in lots of cases, especially when the components of x have different scales/units.

We can also consider the space of $m \times n$ matrices $V = \mathbb{R}^{m \times n}$. There, is of course, a vector-space isomorphism from $V \ni A \to \text{vec}(A) \in \mathbb{R}^{mn}$. Thus, in this space we have the analogue of the familiar ("Frobenius") Euclidean inner product, which is convenient to rewrite in terms of matrix operations via the trace:

Definition 34 (Frobenius inner product)

The **Frobenius inner product** of two $m \times n$ matrices A and B is:

$$\langle A, B \rangle_F = \sum_{ij} A_{ij} B_{ij} = \text{vec}(A)^T \text{vec}(B) = \text{tr}(A^T B)\,.$$

Given this inner product, we also have the corresponding **Frobenius norm**:

$$\|A\|_F = \sqrt{\langle A, A \rangle_F} = \sqrt{\text{tr}(A^T A)} = \|\text{vec} A\| = \sqrt{\sum_{i,j} |A_{ij}|^2}\,.$$

Using this, we can now define the gradient of scalar functions with *matrix inputs*! This will be our default matrix inner product (hence defining our default matrix gradient) in these notes (sometimes dropping the F subscript).

Example 35

Consider the function

$$f(A) = \|A\|_F = \sqrt{\text{tr}(A^T A)}.$$

What is df?

Firstly, by the familiar scalar-differentiation chain and power rules we have that

$$df = \frac{1}{2\sqrt{\text{tr}(A^T A)}} d(\text{tr}\, A^T A).$$

Then, note that (by linearity of the trace)

$$d(\text{tr}\, B) = \text{tr}(B + dB) - \text{tr}(B) = \text{tr}(B) + \text{tr}(dB) - \text{tr}(B) = \text{tr}(dB).$$

Hence,

$$\begin{aligned} df &= \frac{1}{2\|A\|_F} \text{tr}(d(A^T A)) \\ &= \frac{1}{2\|A\|_F} \text{tr}(dA^T\, A + A^T\, dA) \\ &= \frac{1}{2\|A\|_F} (\text{tr}(dA^T\, A) + \text{tr}(A^T\, dA)) \\ &= \frac{1}{\|A\|_F} \text{tr}(A^T\, dA) = \left\langle \frac{A}{\|A\|_F}, dA \right\rangle. \end{aligned}$$

Here, we used the fact that $\text{tr}\, B = \text{tr}\, B^T$, and in the last step we connected df with a Frobenius inner product. In other words,

$$\nabla f = \nabla \|A\|_F = \frac{A}{\|A\|_F}.$$

Note that one obtains exactly the same result for column vectors x, i.e. $\nabla \|x\| = x/\|x\|$ (and in fact this is equivalent via $x = \text{vec}\, A$).

Let's consider another simple example:

Example 36

Fix some constant $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, and consider the function $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ given by

$$f(A) = x^T A y.$$

What is ∇f?

We have that

$$\begin{aligned} df &= x^T \, dA \, y \\ &= \operatorname{tr}(x^T \, dA \, y) \\ &= \operatorname{tr}(yx^T \, dA) \\ &= \langle \underbrace{xy^T}_{\nabla f}, \, dA \rangle. \end{aligned}$$

More generally, for any scalar-valued function $f(A)$, from the definition of Frobenius inner product it follows that:

$$df = f(A + dA) - f(A) = \langle \nabla f, \, dA \rangle = \sum_{i,j} (\nabla f)_{i,j} \, dA_{i,j} \, ,$$

and hence the components of the gradient are exactly the elementwise derivatives

$$(\nabla f)_{i,j} = \frac{\partial f}{\partial A_{i,j}} \, ,$$

similar to the component-wise definition of the gradient vector from multivariable calculus! But for non-trivial matrix-input functions $f(A)$ it can be extremely awkward to take the derivative with respect to each entry of A individually. Using the "holistic" matrix inner-product definition, we will soon be able to compute even more complicated matrix-valued gradients, including $\nabla(\det A)$!

5.2 Derivatives, Norms, and Banach spaces

We have been using the term "norm" throughout this class, but what technically is a norm? Of course, there are familiar examples such as the Euclidean ("ℓ^2") norm $\|x\| = \sqrt{\sum_k x_k^2}$ for $x \in \mathbb{R}^n$, but it is useful to consider how this concept generalizes to other vector spaces. It turns out, in fact, that norms are crucial to the definition of a derivative!

Given a vector space V, a norm $\|\cdot\|$ on V is a map $\|\cdot\| : V \to \mathbb{R}$ satisfying the following three properties:

1. **Non-negative**: i.e. $\|v\| \geq 0$ and $\|v\| = 0 \iff v = 0$,
2. **Homogeneity**: $\|\alpha v\| = |\alpha| \|v\|$ for any $\alpha \in \mathbb{R}$, and
3. **Triangle inequality**: $\|u + v\| \leq \|u\| + \|v\|$.

A vector space that has a norm is called an *normed vector space*. Often, mathematicians technically want a slightly more precise type of normed vector space with a less obvious name: a *Banach* space.

Definition 37 (Banach Space)

A (complete) vector space with a norm is called a *Banach space*. (As with Hilbert spaces, "completeness" is a technical requirement for some types of rigorous analysis, essentially allowing you to take limits.)

For example, given any inner product $\langle u, v\rangle$, there is a corresponding norm $\|u\| = \sqrt{\langle u, u\rangle}$. (Thus, every Hilbert space is also a Banach space.[5])

To define derivatives, we technically need both the input *and* the output to be Banach spaces. To see this, recall our formalism

$$f(x + \delta x) - f(x) = \underbrace{f'(x)[\delta x]}_{\text{linear}} + \underbrace{o(\delta x)}_{\text{smaller}} .$$

To precisely define the sense in which the $o(\delta x)$ terms are "smaller" or "higher-order," we need norms. In particular, the "little-o" notation $o(\delta x)$ denotes any function such that

$$\lim_{\delta x \to 0} \frac{\|o(\delta x)\|}{\|\delta x\|} = 0 ,$$

i.e. which goes to zero faster than linearly in δx. This requires both the input δx and the output (the function) to have norms. This extension of differentiation to arbitrary normed/Banach spaces is sometimes called the **Fréchet derivative**.

[5] Proving the triangle inequality for an arbitrary inner product is not so obvious; one uses a result called the Cauchy–Schwarz inequality.

18.S096 Matrix Calculus for Machine Learning and Beyond
Independent Activities Period (IAP) 2023