(III.C) Linear Functionals I: Derivations and Jacobians

In this section we try to give some idea of how linear algebra is applied to multivariable calculus. In one-variable calculus, you look at the tangent line

\[ y = (y - f(a)) = f'(a) \cdot (x - a) = f'(a) \cdot x, \]

which is the linear transformation approximating the (nonlinear) function \( f : \mathbb{R} \to \mathbb{R} \) to first order at \( a \). We would like to say something similar for multivariate nonlinear functions \( f : \mathbb{R}^m \to \mathbb{R}^n \). Rather than just stating it, though, we will try to show how it arises from geometry. The actual formula for the linear transformation approximating \( f \) to first order will come at the end.

**Linear functionals.** Think back to §I.A, where we considered systems of equations of the form

\[ a_1 x_1 + \ldots + a_n x_n = b. \]

How should we think of the left-hand side? Is it \( \vec{a} \cdot \vec{x} \)? Or a \( 1 \times n \) matrix (i.e. row) operating on a (column) vector? Or perhaps, a linear transformation:

\[ \mathbf{a} : \mathbb{R}^n \to \mathbb{R}^1 \]

with the hyperplane (of vectors “perpendicular” to \( \vec{a} \))

\[ \mathcal{V} = \{ a_1 x_1 + \ldots + a_n x_n = 0 \} \subseteq \mathbb{R}^n \]

as its kernel (or null space). More generally, any linear transformation

\[ L : V \to \mathbb{R} \]
is called a linear functional (or form). If $V$ is $\mathbb{R}^n$ then all linear functionals are just “dot product” with a fixed vector $\vec{a} \in \mathbb{R}^n$, as above.

These examples seem pretty boring, huh? The famous mathematician G.H. Hardy called applied mathematics “dehydrated elephants.” Seems like the same could be said of abstractified linear algebra. This calls for a strong dose of mathematical culture!

Coordinate systems on manifolds. So let me ask the apparently unrelated question, “How would you differentiate a ’smooth’ function on a ’manifold’ (or curved space)”, for example on the $n$-sphere

$$S^n = \{ z_1^2 + \ldots + z_{n+1}^2 = 1 \} \subseteq \mathbb{R}^{n+1} ?$$

First you have to choose a direction against which to look at the rate of change of a function. So it seems plausible that we should denote derivatives by a vector. We will first describe them intrinsically, then explain how to use coordinates to “write them down” (exactly as we did with linear transformations).

Let $M$ be an $n$-dimensional manifold, $C_\infty(M)$ be the smooth functions on $M$ (the $\infty$ is for “infinitely differentiable”), and $p \in M$ any point. Then a derivation at $p \in M$ is any linear functional

$$\vec{w} : C_\infty(M) \to \mathbb{R}$$

that obeys the “Leibniz rule”

$$\vec{w}(f \cdot g) = f(p) \cdot \vec{w}(g) + g(p) \cdot \vec{w}(f).$$

From this rule you can derive the surprising property that $\vec{w}$ does not assign different values to $h_1$ and $h_2$ if they are equal in the smallest neighborhood of $p$. For let $g$ be a function that is 0 at $p$, and 1 outside that neighborhood: then $h_1 - h_2 = g \cdot (h_1 - h_2)$, and

$$\vec{w}(h_1 - h_2) = \vec{w}(g \cdot [h_1 - h_2]) =$$

$$(h_1 - h_2)(p) \cdot \vec{w}(g) + g(p) \cdot \vec{w}(h_1 - h_2) = 0;$$

since $\vec{w}$ is linear, $\vec{w}(h_1) = \vec{w}(h_2)$. Also we have

$$\vec{w}(1) = \vec{w}(1 \cdot 1) = 1 \cdot \vec{w}(1) + 1 \cdot \vec{w}(1) = 2 \cdot \vec{w}(1) \implies \vec{w}(1) = 0$$
and so $\bar{w}$ kills constants. We write $D_p(M)$ for the vector space of derivations at $p$.

Now let $V_1 \subseteq \mathbb{R}^n$, $U_1 \subseteq M$ be open sets containing the origin and $p$, respectively, and let

$$\rho_1 : V_1 \longrightarrow U_1$$

be a (smooth, bijective) function taking

$$(0, \ldots, 0) \longmapsto p.$$ You can think of this as a coordinate system (or parametrization map) with origin “at” $p$. There is a 1-to-1 correspondence between derivations

$$D_p(M) \longleftrightarrow D_0(\mathbb{R}^n)$$ sending

$$\bar{w} \longrightarrow (\rho_1^{-1})_* \bar{w}$$

$$(\rho_1)_* \bar{v} \longleftarrow \bar{v},$$ where $(\rho_1)_*$ and $(\rho_1^{-1})_*$ are defined by

$$(\rho_* \bar{v})(f) := \bar{v}(f \circ \rho) \quad \text{given } f \in C_\infty(M)$$

and $((\rho^{-1})_* \bar{w})(g) = \bar{w}(g \circ \rho^{-1})$. The following picture summarizes the situation:
In words: say I give you a derivation \( \mathbf{\bar{v}} \) on \([\text{functions on}] \ \mathbb{R}^n \) [at 0]. you can get a derivation “\( \rho_1 \mathbf{\bar{v}} \)” (at \( p \)) on functions \( f \in C_\infty(M) \) by (i) composing \( f \) with \( \rho \) to get a function on \( \mathbb{R}^n \), then (ii) applying the derivation \( \mathbf{\bar{v}} \) to the composite \( f \circ \rho \). More concisely, you should think of the lower star as “pushing” vectors in the same direction as the map.

Because of this 1 -to-1 correspondence, to study derivations \( \mathcal{D}_p(M) \) it suffices to understand the possibilities for \( \mathbf{\bar{v}} \in \mathcal{D}_0(\mathbb{R}^n) \). To this end, let’s consider an analytic function on \( M \), so that we may represent it locally by a Taylor series:\(^1\)

\[
g(x_1, \ldots, x_n) = a_0 + a_1 x_1 + \ldots + a_n x_n + \sum_{d \geq 2} \left( \sum_{d_1 + \ldots + d_n = d} a_{d_1 \ldots d_n} x_1^{d_1} \cdots x_n^{d_n} \right).
\]

On any monomial \( x_1^{d_1} \cdots x_n^{d_n} \) of degree \( \geq 2 \), \( \mathbf{\bar{v}} \) is zero by the Leibniz rule, since we are evaluating\(^2\) at \( 0 = (0, \ldots, 0) \). So by linearity (and because \( \mathbf{\bar{v}} \) also kills the constant term),

\[
\mathbf{\bar{v}}(g) = a_1 \mathbf{\bar{v}}(x_1) + \ldots + a_n \mathbf{\bar{v}}(x_n),
\]

which says that \( \mathbf{\bar{v}} \) is just the differential operator

\[
\mathbf{\bar{v}}(x_1) \frac{\partial}{\partial x_1} + \ldots + \mathbf{\bar{v}}(x_n) \frac{\partial}{\partial x_n} =: \sum a_i \frac{\partial}{\partial x_i}
\]

composed with evaluation at \( (0, \ldots, 0) \). So we have

\[
\dim(\mathcal{D}_p(M)) = \dim(\mathcal{D}_0(\mathbb{R}^n)) = n,
\]

and even a basis

\[
\left\{ (\rho_1)^* \frac{\partial}{\partial x_1}, \ldots, (\rho_1)^* \frac{\partial}{\partial x_n} \right\}
\]

\(^1\)not all smooth functions are analytic! See the exercise.
\(^2\)For example,

\[
\mathbf{\bar{v}}(x_1 x_2) = x_1(0) \cdot \mathbf{\bar{v}}(x_2) + x_2(0) \cdot \mathbf{\bar{v}}(x_1) = 0 \cdot \mathbf{\bar{v}}(x_2) + 0 \cdot \mathbf{\bar{v}}(x_1) = 0.
\]
for $\mathcal{D}_p(M)$ (we sometimes will omit the evaluation at $p$ [or 0] like this). If $\bar{\omega} \in \mathcal{D}_p(M)$ is any derivation, it can be written uniquely $\bar{\omega} = \sum \alpha_i \cdot (\rho_1)_* \frac{\partial}{\partial x_i}$, which means for $f \in C_\infty(M)$

$$\bar{\omega}(f) = \sum \alpha_i \frac{\partial (f \circ \rho_1)}{\partial x_i} \bigg|_{(0, \ldots, 0)}.$$ 

There is absolutely nothing special about this basis. Let $\rho_2 : V_2 \sim U_2$ be another parametrization, with $\rho_2(0) = p$ for convenience. This gives rise to an alternate basis $\{(\rho_2)_* \frac{\partial}{\partial y_j}\}$, with respect to which we can write $\bar{\omega}$ to get (exactly as above)

$$\bar{\omega}(f) = \sum \beta_j \frac{\partial (f \circ \rho_2)}{\partial y_j} \bigg|_{(0, \ldots, 0)}.$$ 

So our goal should be a transition (change of basis) matrix relating the coefficient vectors $\bar{\alpha}$ and $\bar{\beta}$ . Restricting if necessary to open subsets of $V_1$ and $V_2$ (which map to $U_1 \cap U_2$), we define

$$\theta_{12} := \rho_2^{-1} \circ \rho_1,$$

and write

$$\theta_{12}(x_1, \ldots, x_n) =: (y_1(x_1, \ldots, x_n), \ldots, y_n(x_1, \ldots, x_n))$$

(note: $\theta_{12}(0, \ldots, 0) = (0, \ldots, 0)$ ). The picture has become:
Now let \( g \in C_\infty(V_2) \) be a smooth function, giving us the picture

\[ \begin{array}{c}
V_1 \xrightarrow{\theta_{12}} V_2 \xrightarrow{g \circ \theta_{12}} \mathbb{R} \\
M \xrightarrow{\rho_1} \mathbb{R}^n \xrightarrow{\rho_2} \mathbb{R}^n
\end{array} \]

Push \( \bar{w} \) (on \( M \)) to \( V_1 \) and \( V_2 \) by taking

\[ \bar{v}_1 = (\rho_1^{-1})_*\bar{w} = \sum \alpha_i \frac{\partial}{\partial x_i}, \quad \bar{v}_2 = (\rho_2^{-1})_*\bar{w} = \sum \beta_j \frac{\partial}{\partial y_j}, \]

and let \( \bar{v}_2 \) act on \( g \):

\[ \bar{v}_2(g) = \bar{w}(g \circ \rho_2^{-1}) = \bar{w}(g \circ \rho_2^{-1} \circ (\rho_1 \circ \rho_1^{-1})) = \]

\[ = \bar{w}((g \circ \theta_{12}) \circ \rho_1^{-1}) = \bar{v}_1(g \circ \theta_{12}). \]

That is,

\[ \bar{v}_2 = (\theta_{12})_*\bar{v}_1! \]
What this says concretely is:

\[ \sum \beta_j \frac{\partial g}{\partial y_j} \bigg|_0 = v_2(g) = v_1(g \circ \theta_{12}) = \sum \alpha_i \frac{\partial (g \circ \theta_{12})}{\partial x_i} \bigg|_0, \]

which by the Chain rule

\[ = \sum \alpha_i \sum_j \frac{\partial g}{\partial y_j} \frac{\partial y_j}{\partial x_i} \bigg|_0 = \sum_j \left( \sum_i \alpha_i \frac{\partial y_j}{\partial x_i} \bigg|_0 \right) \frac{\partial g}{\partial y_j} \bigg|_0. \]

So (switching \( i \) and \( j \))

\[ \beta_i = \sum_j \alpha_j \frac{\partial y_j}{\partial x_i} \]

or in matrix notation

\[
\begin{pmatrix}
\beta_1 \\
\vdots \\
\beta_n
\end{pmatrix} = 
\begin{pmatrix}
\frac{\partial y_1}{\partial x_1} \bigg|_0 & \cdots & \frac{\partial y_1}{\partial x_n} \bigg|_0 \\
\vdots & \ddots & \vdots \\
\frac{\partial y_n}{\partial x_1} \bigg|_0 & \cdots & \frac{\partial y_n}{\partial x_n} \bigg|_0
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\
\vdots \\
\alpha_n
\end{pmatrix},
\]

and we define \( J_{\theta_{12}} \bigg|_0 \) to be the Jacobian matrix \( \left[ \frac{\partial y_j}{\partial x_i} \bigg|_0 \right] \). In short,

\[ \mathbf{\beta} = J_{\theta_{12}} \bigg|_0 \cdot \mathbf{\alpha}; \]

conversely, derivations \( \sum \alpha_i \frac{\partial}{\partial x_i} \) and \( \sum \beta_j \frac{\partial}{\partial y_j} \) on distinct coordinate patches which “agree” in the sense of this boxed formula, correspond to one and the same derivation on \( M \).

Now in reality, such equivalences classes of derivations on \( \mathbb{R}^n \) are (by definition!) the derivations \( \mathbb{w} \) on \( M \) at \( p \). How else are you going to differentiate a function at a point on a hyperbolic paraboloid \( z = xy \), other than something like “projecting” the function to the \( xy \)-plane, and differentiating it there?

Here is another shock: this is how you define vectors on \( M \) [at \( p \)] as well – as derivations! How about smoothly varying vector fields on all of \( M \)? Simply cover \( M \) with coordinate systems \( \rho_\ell : V_\ell \sim U_\ell \subseteq M \), and write down smoothly varying \( \sum \alpha_\ell^i (x_1, \ldots, x_n) \frac{\partial}{\partial x_i} \) on each \( V_\ell \), that “match up” in the sense that \( \mathbf{\alpha}^{\ell_2} = J_{\theta_{\ell_1 \ell_2}} \mathbf{\alpha}^{\ell_1} \). While you may wish to define vector fields say on the sphere \( S^2 \) as vectors in \( \mathbb{R}^3 \) tangent to \( S^2 \), for most purposes this is less convenient.
Remark 1. (i) The theory of manifolds takes this philosophy further, defining $M$ abstractly as a topological space with “coordinate chart” maps $\mathbb{R}^n \supset V_\ell \xrightarrow{\rho_\ell} U_\ell \subset M$ which are homeomorphisms (topological isomorphisms), whose images cover $M$ (i.e. $M = \bigcup U_\ell$), and such that the induced $\theta_j|_\ell$’s on overlaps are $C_\infty$-functions. We don’t “assume that $\rho_\ell$ is $C_\infty$” because that would be circular – the definition of a $C_\infty$ function on $M$ is one whose compositions with the $\rho_\ell$ are all $C_\infty$.

(ii) $D_p(M)$ is usually written $T_p(M)$, the space of “tangent vectors” = linear differential operators (on $M$) at $p$.

Computationally, what have we seen here that is relevant to linear algebra? That one can summarize the first-order (linear) behavior of a nonlinear transformation from $\mathbb{R}^n \to \mathbb{R}^n$ (namely, $\theta_{12}$) by its Jacobian matrix. This is the appropriate way to transform vectors, and it extends to maps $\theta : \mathbb{R}^n \to \mathbb{R}^m$ (except $J$ is then an $m \times n$ matrix). The induced map $\theta_*$ on vectors is linear at each point $p = (x_1, \ldots, x_n)$ of $\mathbb{R}^n$, and has matrix

$$J_\theta \bigg|_p = \left( \begin{array}{c c c}
\frac{\partial \theta_1}{\partial x_1} & \cdots & \frac{\partial \theta_1}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial \theta_m}{\partial x_1} & \cdots & \frac{\partial \theta_m}{\partial x_n}
\end{array} \right)$$

with respect to the “standard basis” (for vectors at each point). From this simpler perspective, the link with differential operators is that they transform (by the chain rule) as vectors, i.e. by the same matrix.

Exercises
(1) [optional] For a smooth function $G$ on $M$, check that any derivation kills $G - L_G$, where $L_G(\vec{x}) = G(0) + G_{x_1}(0)x_1 + \cdots + G_{x_n}(0)x_n$ is the linear approximation (defined only locally), thereby completing the argument above.

---

$^3$writing $\theta(x_1, \ldots, x_n) = (\theta_1(x_1, \ldots, x_n), \ldots, \theta_m(x_1, \ldots, x_n))$