Multi-linear maps and Taylor's Theorem in higher dimensions

Steve Cheng

February 16, 2006

Contents

Introduction

It seems that nowadays the Taylor Theorem for multiple dimensions has been given the short shrift in the typical mathematics curriculum. At least, I did not see much of it, even though the single-variable case is, of course, familiar to everybody. It is not hard to understand why, because higher-order derivatives are much more complicated in multiple dimensions than in the single-variable case. Nevertheless, in this note, we show that, with the proper abstractions, Taylor's Theorem in higher dimensions has essentially the same form as the single-variable case.

Copyright matters

1 Multi-linear maps

Our first goal is to develop some formal devices that allow us to reduce the multi-dimensional case of Taylor's Theorem to the one-dimensional case fairly easily.

Definition 1.1. Let $V_1, \dots, V_k$ and W be vector spaces over a common field. A function $f\colon V_1 \times \dots \times V_k \to W$ is *multi-linear* if it is linear in each component separately: for all $1 \le i \le k$,

$$f(v_1, \dots, v_{i-1}, a_i v_i + u_i, v_{i+1}, \dots, v_k) = a_i f(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_k) \\ + f(v_1, \dots, v_{i-1}, u_i, v_{i+1}, \dots, v_k)$$

where a_i is a scalar, and $v_i, u_i \in V$.

In the special case that all the V_i are the same vector space, and W is the scalar field, a multi-linear map is also called a *tensor*.

Definition 1.2. $\mathcal{T}(V_1 \times \dots \times V_k \to W)$ denotes the set of all multi-linear maps $f\colon V_1 \times \dots \times V_k \to W$. This can be made into a vector space in the obvious way with point-wise addition and scalar multiplication. If all the V_i are the same, we also use the notation $\mathcal{T}(V^k \to W)$ for the set of multi-linear maps. If $k = 1$, we often revert to the linear map notation $\mathcal{L}(V \to W)$.

If $f \in \mathcal{T}(V_1 \times \dots \times V_k \to W)$, then for each fixed v_1, we can consider the multi-linear map of one lower order:

$$(v_2, \dots, v_k) \mapsto f(v_1, v_2, \dots, v_k)\,.$$

So there is a natural identification of $\mathcal{T}(V_1 \times \dots \times V_k \to W)$ with $\mathcal{L}(V_1 \to \mathcal{T}(V_2 \times \dots \times V_k \to W))$.

Furthermore, if $x \in V$ and $f \in \mathcal{T}(V^k \to W)$, let us use the short-hand:

$$f(x, \dots, x) = f(x^k)$$

Definition 1.3. Let V and W be normed (real or complex) vector spaces[1], and A be open in V. A function $f\colon A \to W$ is *(Fréchet-) differentiable* at $x \in V$ if there exists a continuous linear map $T \in \mathcal{L}(V \to W)$ such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - T(h)\|}{\|h\|} = 0\,.$$

The necessarily unique T is called the derivative of f at x, and will be denoted $\mathrm{D}f(x)$.

This is the usual definition of derivatives in $\mathbb{R}^n$, generalized to arbitrary normed vector spaces. It follows, as usual, that if f is differentiable at x, then it is continuous at x.

For clarity, when the linear map $\mathrm{D}f(x)$ is to be evaluated at a point v, I will denote that by juxtaposition instead of the usual functional notation: $\mathrm{D}f(x)v$ instead of $\bigl(\mathrm{D}f(x)\bigr)(v)$. This notation is intended to be suggestive of matrix multiplication, which is how one would compute things in a finite-dimensional setting.

Suppose $f\colon A \to W$, where A is open in V, is differentiable on A. Then the higher derivatives $\mathrm{DD}f(x), \mathrm{DDD}f(x), \dots, \mathrm{D}^k f(x)$ are multi-linear maps from V to W, provided we make the identification $\mathcal{L}\bigl(V \to \mathcal{T}(V^k \to W)\bigr) = \mathcal{T}(V^{k+1} \to W)$ as mentioned earlier.

It may help to think of $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, although we will prove as much as we can even for infinite-dimensional (!) vector spaces.

[1] Most definitions of the Fréchet derivative take the vector spaces to be complete normed vector spaces, a.k.a. Banach spaces. But we do not need completeness in our work here.

Definition 1.4. If $T \in \mathcal{T}(V_1 \times \cdots \times V_k \to W)$, when we say T is continuous, we mean that T is continuous as a function $V_1 \times \cdots \times V_k \to W$. Here, the vector space $V_1 \times \cdots \times V_k$ will always be given the product topology, induced by the maximum norm:

$$\|(v_1, \dots, v_k)\| = \max(\|v_1\|, \dots, \|v_k\|) \,.$$

It is readily shown, just as with linear maps, that T is continuous everywhere if and only if it is continuous at $(0, \dots, 0)$, and the latter happens if and only if

$$\begin{aligned} \|T\| &:= \sup_{\|(v_1,\dots,v_k)\| \neq 0} \frac{\|T(v_1, \dots, v_k)\|}{\|(v_1, \dots, v_k)\|^k} = \sup_{\|(v_1,\dots,v_k)\|=1} \|T(v_1, \dots, v_k)\| \\ &= \sup_{\substack{\|v_i\|=1 \\ 1 \leq i \leq k}} \|T(v_1, \dots, v_k)\| = \sup_{\substack{\|v_i\| \neq 0 \\ 1 \leq i \leq k}} \frac{\|T(v_1, \dots, v_k)\|}{\|v_1\| \cdots \|v_k\|} \end{aligned}$$

is $< \infty$. This always occurs when $V_1 \times \cdots \times V_k$ is finite-dimensional.

Note that $\|T\|$ as just defined is indeed a norm on the vector space $\mathcal{T}(V_1 \times \cdots \times V_k \to W)$.

Earlier on, we identified $\mathcal{L}\big(V \to \mathcal{T}(V^k \to W)\big) = \mathcal{T}(V^{k+1} \to W)$. When studying continuity and differentiability, it is crucial that, if T considered as an element of $\mathcal{L}\big(V \to \mathcal{T}(V^k \to W)\big)$ has a "small" norm there, then the norm of T is also "small" in $\mathcal{T}(V^{k+1} \to W)$. That this is true follows by applying the following fact repeatedly:

Lemma 1.1. *Let $U_1, \dots, U_k, V, W$ be normed vector spaces. Then the linear isomorphism $\Psi \colon \mathcal{T}(U_1 \times \cdots \times U_k \to \mathcal{L}(V, W)) \to \mathcal{T}(U_1 \times \cdots \times U_k \times V \to W)$, defined by $\Psi(L)(u_1, \dots, u_k, v) = L(u_1, \dots, u_k)(v)$, is an isometry. That is, $\|\Psi(L)\| = \|L\|$ for all L.*

Proof.

$$\begin{aligned} \|\Psi(L)\| &= \sup_{\substack{\|u_i\|=1,\, 1 \leq i \leq k \\ \|v\|=1}} \|\Psi(L)(u_1, \dots, u_k, v)\| = \sup_{\substack{\|u_i\|=1 \\ \|v\|=1}} \|L(u_1, \dots, u_k)(v)\| \\ &= \sup_{\|u_i\|=1} \sup_{\|v\|=1} \|L(u_1, \dots, u_k)(v)\| = \sup_{\|u_i\|=1} \|L(u_1, \dots, u_k)\| = \|L\| \,. \end{aligned}$$ ■

Algebraically, multi-linear maps generalize the concept of multiplication with the factors being from arbitrary vector spaces. A multi-linear map T is distributive in each argument, and scalar multiplication may be pulled outside of an argument. For example, if we write $u \cdot v$ for $T(u, v)$, and $u \in U$ and $v \in V$, then we have

$$\begin{aligned} (u_1 + u_2) \cdot v &= u_1 \cdot v + u_2 \cdot v && \text{left distributivity} \\ u \cdot (v_1 + v_2) &= u \cdot v_1 + u \cdot v_2 && \text{right distributivity} \\ (\lambda u) \cdot v &= c \cdot (\lambda v) = \lambda(u \cdot v) \,, \end{aligned}$$

which is just how ordinary multiplication behaves. In fact, in the next section we will show that continuous multi-linear maps obey an analogue of the product rule from calculus.

2 Product rule for multi-linear maps

Theorem 2.1. *Let V_i and W be normed vector spaces, and $T \in \mathcal{T}(V_1 \times \cdots \times V_k \to W)$ be continuous. Then for all $a_i \in V_i$,*

$$\begin{aligned} \mathrm{D}T(a_1, \ldots, a_k) \cdot (v_1, \ldots, v_k) = T(v_1, a_2, \ldots, a_k) + T(a_1, v_2, a_3, \ldots, a_k) \\ + \cdots + T(a_1, \ldots, a_{k-1}, v_k)\,. \end{aligned}$$

In other words, differentiate one component at a time, just as with the product rule from calculus. In more abstract notation, the formula can be written as

$$\mathrm{D}T(a_1, \ldots, a_k) = \sum_{i=1}^{k} T(a_1, \ldots, \pi_i, \ldots, a_k)$$

where $\pi_i : V_1 \times \ldots \times V_k \to V_i$ denotes the linear projection function to the ith vector.

Proof. First note that the proposed derivative $\sum_{i=1}^{k} T(a_1, \ldots, \pi_i, \ldots, a_k)$ is indeed linear in $V_1 \times \cdots \times V_k$, and is continuous, so this theorem is not preposterous.

To confirm that $\sum_{i=1}^{k} T(a_1, \ldots, \pi_i, \ldots, a_k)$ is indeed the derivative of T, we have to show that the difference quotient occurring in the definition of the derivative,

$$\frac{\|T(a_1 + v_1, \ldots, a_k + v_k) - T(a_1, \ldots, a_k) - \sum_{i=1}^{k} T(a_1, \ldots, v_i, \ldots, a_k)\|}{\|(v_1, \ldots, v_k)\|} \tag{1}$$

goes to zero as $(v_1, \ldots, v_k) \to 0$.

Consider the numerator of (1). When we expand $T(a_1 + v_1, \ldots, a_k + v_k)$ by using multi-linearity, we get terms of the form $T(z_1, \ldots, z_k)$ where z_i is either a_i or v_i. If we subtract away the terms $T(a_1, \ldots, a_k)$ and $\sum_{i=1}^{k} T(a_1, \ldots, v_i, \ldots, a_k)$ from the above expansion, then we are left with terms of the form $T(z_1, \ldots, z_k)$ where *at least two* of the z_i are v_i — say $z_{i_1} = v_{i_1}$ and $z_{i_2} = v_{i_2}$.

The denominator (1) is $\|(v_1, \ldots, v_k)\| = \max(\|v_1\|, \ldots, \|v_k\|)$; abbreviate this scalar by λ. We move λ inside the i_1-th argument of the term $T(z_1, \ldots, z_{i_1}, \ldots, z_{i_2}, \ldots, z_k)$:

$$\begin{aligned} \frac{\|T(z_1, \ldots, z_{i_1}, \ldots, z_{i_2}, \ldots, z_k)\|}{\lambda} &= \|T(z_1, \ldots, \lambda^{-1} z_{i_1}, \ldots, z_{i_2}, \ldots, z_k)\| \\ &\leq \|T\|\, \|z_1\| \cdots \frac{\|z_{i_1}\|}{\lambda} \cdots \|z_{i_2}\| \cdots \|z_k\| \\ &\leq \|T\|\, \|z_1\| \cdots \frac{\|z_{i_1}\|}{\|z_{i_1}\|} \cdots \|z_{i_2}\| \cdots \|z_k\| \\ &= \|T\|\, \|z_1\| \cdots \quad 1 \quad \cdots \|z_{i_2}\| \cdots \|z_k\| \end{aligned}$$

Now as $(v_1, \ldots, v_k) \to 0$, $\|z_{i_2}\|$ goes to zero, while the other terms $\|z_i\|$ are bounded, so the term above also goes to zero.

Thus, by the triangle inequality, the entire quotient (1) goes to zero. ∎

Corollary 2.2 (Product rule). *Let $A \subseteq U$ open, and $f_i \colon A \to V_i$, $1 \leq i \leq k$, be differentiable at a. If $T \in \mathcal{T}(V_1 \times \cdots \times V_k \to W)$ is continuous, then*

$$\mathrm{D}T(f_1, \ldots, f_k)(a) = \sum_{i=1}^{k} T\big(f_1(a), \ldots, \mathrm{D}f_i(a), \ldots, f_k(a)\big)\,.$$

Proof. Let $F = (f_1, \dots, f_k)$. Then F is differentiable at a with $\mathrm{D}F(a) = (\mathrm{D}f_1(a), \dots, \mathrm{D}f_k(a))$. By Theorem 2.1 and the chain rule, we have:

$$\mathrm{D}(T \circ F)(a) = \mathrm{D}T\big(F(a)\big) \circ \mathrm{D}F(a) = \sum_{i=1}^{k} T\big(f_1(a), \dots, \mathrm{D}f_i(a), \dots, f_k(a)\big) . \qquad \blacksquare$$

We could have proven Theorem 2.1 by induction on k, using the two-argument product rule for the inductive step. This would mimic the usual proof of the product rule from calculus. But the proof would not be any easier.

Example 2.1. As a little consolation for the utter abstractness of our procedings so far, let us consider a concrete example for Corollary 2.2. Suppose $x_i = x_i(t)$, $1 \leq i \leq n$, are differentiable curves in $\mathbb{R}^n$. Consider the n-dimensional parallelopiped with end-points $(0, \dots, 0)$ and $(x_1, \dots, x_n)$. The rate of change of the volume of this parallelopiped is

$$\frac{d}{dt} \det(x_1, \dots, x_n) = \sum_{i=1}^{n} \det(x_1, \dots, \frac{dx_i}{dt}, \dots, x_n) ,$$

applying the corollary for the tensor det.

Example 2.2. The inner product and cross product are multi-linear, so they can be differentiated with the product rule. (In elementary courses, this fact is usually proven by an inelegant component-by-component analysis.)

In Corollary 2.2 we considered the multi-linear map T to be fixed. However, in the proof of Taylor's Theorem, we will need even T to vary — i.e. T is now a function on $A \subseteq U$. The derivative formula is just what we would expect, and in fact it is not hard to prove:

Corollary 2.3. *Let A, U, V_i, W and f_i as in the setting of Corollary 2.2. Suppose $T\colon U \to \mathcal{T}(V_1 \times \cdots \times V_k \to W)$ is differentiable at $a \in A$, and for each $u \in A$, $T(u)$ is a continuous multi-linear map $V_1 \times \cdots \times V_k \to W$. Then*

$$\mathrm{D}\big(T(\cdot)(f_1, \dots, f_k)\big)(a) = \mathrm{D}T(a)\big[f_1(a), \dots, f_k(a)\big] + \sum_{i=1}^{k} T(a)\big[f_1(a), \dots, \mathrm{D}f_i(a), \dots, f_k(a)\big] .$$

Proof. Define the evaluation multi-linear map $E\colon \mathcal{T}(V_1 \times \cdots \times V_k \to W) \times V_1 \times \cdots \times V_k \to W$, by

$$E[S, v_1, \dots, v_k] = S(v_1, \dots, v_k) ,$$

which is continuous because

$$\|E[S, v_1, \dots, v_k]\| = \|S(v_1, \dots, v_k)\| \leq \|S\| \, \|v_1\| \cdots \|v_k\| .$$

Next define $F(u) = \big(T(u), f_1(u), \dots, f_k(u)\big)$. Then

$$\begin{aligned} \mathrm{D}(E \circ F)(a) &= E\big[\mathrm{D}T(a), f_1(a), \dots, f_k(a)\big] + \sum_{i=1}^{k} E\big[T(a), f_1(a), \dots, \mathrm{D}f_i(a), \dots, f_k(a)\big] \\ &= \mathrm{D}T(a)\big[f_1(a), \dots, f_k(a)\big] + \sum_{i=1}^{k} T(a)\big[f_1(a), \dots, \mathrm{D}f_i(a), \dots, f_k(a)\big] . \qquad \blacksquare \end{aligned}$$

3 The general Taylor's Theorem

Theorem 3.1 (Taylor's Theorem). *Let A be open in a real normed vector space X, containing the line segment from a to x. Let $f\colon A \to \mathbb{R}$ be $n+1$ times differentiable ($n \geq 0$), and*

$$f(x) = f(a) + \mathrm{D}f(a)(x-a) + \frac{\mathrm{D}^2 f(a)}{2!}(x-a)^2 + \cdots + \frac{\mathrm{D}^n f(a)}{n!}(x-a)^n + R(x)\,.$$

Then there is some η and ξ on the line segment from a to x, such that

$$R(x) = \frac{D^{n+1}(\eta)}{n!}(x-\eta)^n(x-a)\,, \qquad \text{(Cauchy form of remainder)}$$
$$R(x) = \frac{D^{n+1}(\xi)}{(n+1)!}(x-a)^{n+1} \qquad \text{(Lagrange form of remainder).}$$

Moreover,

$$\begin{aligned} R(x) &= \int_0^1 \frac{\mathrm{D}^{n+1} f(a + t(x-a))}{n!}\big((1-t)(x-a)\big)^n \, dt\,, \qquad \text{(integral form of remainder)} \\ &= \int_0^1 \frac{\mathrm{D}^{n+1} f(ta + (1-t)x)}{n!}(t(x-a))^n \, dt\,, \end{aligned}$$

provided the integrals exist.

Proof. The idea, of course, is to reduce the problem to the one-dimensional one by parameterizing the line segment from a to x. Let $\gamma(t) = a + t(x-a)$ for $t \in [0,1]$, and $g(t) = f\big(\gamma(t)\big)$.

We compute the derivatives of g:

$$g'(t) = \mathrm{D}f\big(\gamma(t)\big) \cdot \gamma'(t) = \mathrm{D}f\big(\gamma(t)\big) \cdot (x-a)$$

The chain rule is being used here. Also, strictly speaking, $\mathrm{D}g(t)$ should be a linear map $\mathbb{R} \to \mathbb{R}$, which sends any scalar θ to $\mathrm{D}f\big(\gamma(t)\big) \cdot (x-a)\,\theta$. In other words $\mathrm{D}g(t)$ is the linear map of "multiplication by $\mathrm{D}f\big(\gamma(t)\big) \cdot (x-a)$". But we will not make too much fuss about this distinction, and just think of $g'(t)$ as a real number, as in one-dimensional calculus.

From the second derivatives onwards, we will need tensor differentiation:

$$\begin{aligned} g''(t) &= \Big(\mathrm{D}^2 f\big(\gamma(t)\big) \cdot \gamma'(t)\Big) \cdot (x-a) + \mathrm{D}f\big(\gamma(t)\big) \cdot 0 \\ &= \mathrm{D}^2 f\big(\gamma(t)\big)(x-a)(x-a) + 0 \\ &= \mathrm{D}^2 f\big(\gamma(t)\big)(x-a)^2 \end{aligned}$$

(The two terms come from the product rule (Corollary 2.3). The 0 in the second term comes from the fact that the derivative of the constant $(x-a)$ is 0. And the extra $\gamma'(t) = (x-a)$ vector comes from the chain rule on $\mathrm{D}f\big(\gamma(t)\big)$.)

Continuing, we have, in general,

$$\begin{aligned} g^{(j)}(t) &= \mathrm{D}^j f\big(\gamma(t)\big)(x-a)\cdots(x-a) \\ &= \mathrm{D}^j f\big(\gamma(t)\big)(x-a)^j\,. \end{aligned}$$

By applying the one-dimensional Taylor's Theorem to g on $[0,1]$, we can write:

$$g(1) = g(0) + g'(0) + \frac{g''(0)}{2!} + \cdots + \frac{g^{(n)}(0)}{n!} + S\,.$$
$$f(x) = f(a) + \mathrm{D}f(a)(x-a) + \frac{\mathrm{D}^2 f(a)}{2!}(x-a)^2 + \cdots + \frac{\mathrm{D}^n f(a)}{n!}(x-a)^n + R(x)\,.$$

and

$$S = \frac{g^{(n)}(u)}{n!}(1-u)^n\,, \qquad \text{for some } u \in (0,1).$$
$$S = \frac{g^{(n+1)}(s)}{(n+1)!}\,, \qquad \text{for some } s \in (0,1).$$
$$S = \int_0^1 \frac{f^{(n+1)}(t)}{n!}(1-t)^n\,dt\,, \qquad \text{if this integral exists.}$$

But these are exactly the terms $R(x)$ in the statement of this theorem, where we set $\eta = a + u(x-a)$ and $\xi = a + s(x-a)$. For instance,

$$\begin{aligned} S &= \frac{(1-u)^n}{n!} g^{(n)}(u) \\ &= \frac{\mathrm{D}^n f\big(\gamma(u)\big)}{n!}(1-u)(x-a)\cdots(1-u)(x-a)(x-a) \\ &= \frac{\mathrm{D}^n f(\eta)}{n!}(x-\eta)\cdots(x-\eta)(x-a)\,. \end{aligned}$$

Note that $(1-u)$ is a scalar, so it may be moved into any of the arguments of a multi-linear map. The same arguments apply for the other two expressions for S. ■

In Taylor's Theorem, we restricted ourselves to the case that f takes values in $\mathbb{R}$ rather in an arbitrary normed vector space, because the theorem is not true in that case. If f has multiple "components" (think $\mathbb{R}^m$), then there is usually not going to be a single-point estimate that works for all components of f, as you can easily check.

Fortunately, the following theorem suffices for most purposes, when f is allowed to take values in an arbitrary normed vector space:

Theorem 3.2. *Let A be open in a real normed vector space X, containing the line segment from a to x. Let $f\colon A \to Y$ be $n+1$ times differentiable. Then there exists some ξ on the line segment from a to x, such that*

$$\left\| f(x) - \left[f(a) + \mathrm{D}f(a)(x-a) + \cdots + \frac{\mathrm{D}^n f(a)}{n!}(x-a)^n\right] \right\| \leq \left\| \frac{\mathrm{D}^{n+1} f(\xi)}{(n+1)!}(x-a)^{n+1} \right\|\,.$$

Proof. Let $\phi \in Y^*$ be a continuous *linear* functional on Y; then $\mathrm{D}^j(\phi \circ f) = \phi \circ \mathrm{D}^j f$. Applying Taylor's Theorem to $\phi \circ f\colon A \to \mathbb{R}$, we get:

$$\phi\big(f(x)\big) - \left[\phi\big(f(a)\big) + \phi\big(\mathrm{D}f(a)(x-a)\big) + \cdots + \phi\left(\frac{\mathrm{D}^n f(a)}{n!}(x-a)^n\right)\right] = \phi\left(\frac{\mathrm{D}^{n+1} f(\xi)}{(n+1)!}(x-a)^{n+1}\right)$$

Letting

$$v = f(x) - \left[f(a) + \cdots + \frac{\mathrm{D}^n f(a)}{n!}(x-a)^n\right],$$

the first equation can be rewritten:

$$\phi(v) = \phi\left(\frac{\mathrm{D}^{n+1} f(\xi)}{(n+1)!}(x-a)^{n+1}\right).$$

Assume $v \neq 0$; otherwise the theorem is trivial. By the Hahn-Banach Theorem, there exists a $\phi \in Y^*$ such that $\phi(v) = \|v\|$ and $\|\phi\| = 1$. Take this ϕ in particular when applying Taylor's theorem above. Then:

$$\|v\| = |\phi(v)| = \left|\phi\left(\frac{\mathrm{D}^{n+1} f(\xi)}{(n+1)!}(x-a)^{n+1}\right)\right| \leq \|\phi\| \left\|\frac{\mathrm{D}^{n+1} f(\xi)}{(n+1)!}(x-a)^{n+1}\right\|. \quad \blacksquare$$

The case $n = 0$ in Theorem 3.2 is a sort of "Generalized Mean Value Theorem" for vector-valued differentiable functions.

4 Characterization of Taylor polynomials

Theorem 4.1. *If $f\colon A \to Y$ is n times differentiable at a, then the Taylor polynomial of degree n for f at a,*

$$p_{n,f}(x) = f(a) + \mathrm{D}f(a)(x-a) + \cdots + \frac{\mathrm{D}^n f(a)}{n!}(x-a)^n$$

is equal to f up to order n at a. That is, $\|f(x) - p_{n,f}(x)\| = o(\|x-a\|^n)$ as $x \to a$.

Proof. The theorem is already true for $n = 1$, by the definition of the derivative at a.

Now suppose the theorem is true for Taylor polynomials of degree $n-1$; we aim to prove it for Taylor polynomials of degree n. Given f as in the hypotheses, let $R_{n,f} = f - p_{n,f}$ be the remainder term. Then differentiating with the product rule,

$$\mathrm{D}R_{n,f}(\xi) \cdot h = \mathrm{D}f(\xi) \cdot h - \sum_{k=1}^{n} \sum_{j=0}^{k-1} \frac{\mathrm{D}^k f(a)}{k!}(\xi - a)^j h (\xi - a)^{k-1-j}.$$

Since $\mathrm{D}^k f(a)$ were not assumed to be symmetric, the inner summation cannot be simplified. However, if $h = t(\xi - a)$ for some t, there is no problem:

$$\mathrm{D}R_{n,f}(\xi) \cdot h = \mathrm{D}f(\xi) \cdot t(\xi - a) - \sum_{k=1}^{n} \frac{\mathrm{D}^k f(a)}{(k-1)!} t(\xi - a)^k = R_{n-1,\mathrm{D}f}(\xi) \cdot h,$$

where $R_{n-1,\mathrm{D}f}$ is the remainder term for the Taylor polynomial of $\mathrm{D}f$ of degree $n-1$ at a.

Thus

$$\begin{aligned}
\|R_{n,f}(x)\| &= \|R_{n,f}(x) - R_{n,f}(a)\| && \text{since } R_{n,f}(a) = 0 \\
&\le \|\mathrm{D}R_{n,f}(\xi) \cdot (x-a)\| && \text{Generalized Mean Value Theorem} \\
&= \|R_{n-1,\mathrm{D}f}(\xi) \cdot (x-a)\| && \text{since } x - a = t(\xi - a) \text{ for some } t \\
&\le \|R_{n-1,\mathrm{D}f}(\xi)\| \cdot \|x-a\| \\
&= o(\|\xi - a\|^{n-1}) \cdot \|x - a\| && \text{induction hypothesis on } \mathrm{D}f \\
&= o(\|x-a\|^n) && \text{since } \|\xi - a\| \le \|x - a\|. \quad \blacksquare
\end{aligned}$$

The Taylor polynomial of degree n at a of a function $f\colon U \to Y$, is in fact the only polynomial g of degree n such that $g(x) - f(x) = o(\|x-a\|^n)$ as $x \to a$. To show this, the following suffices.

Theorem 4.2. *Let X, Y be normed vector spaces, and suppose that*

$$g(x) = \sum_{k=0}^{n} M_k (x-a)^k$$

for some continuous multi-linear mappings $M_k\colon X^k \to Y$. If $g(x) = o(\|x-a\|^n)$ as $x \to a$, then $g = 0$.

Proof. We prove by induction on k that $M_k(x-a)^k = 0$ for all x.

For the case $k = 0$, write

$$M_0 = g(x) - \sum_{j=1}^{n} M_j (x-a)^j \,,$$

and observe that the right side is $o(1)$. Therefore M_0 is also. But M_0 is a constant, so this means $M_0 = 0$.

Suppose that $M_k(x-a)^j = 0$ for $j = 0, \dots, k-1$. Then

$$\begin{aligned}
M_k(x-a)^k &= g(x) - \sum_{j=0}^{k-1} M_j (x-a)^k - \sum_{j=k+1}^{n} M_j (x-a)^j \\
&= g(x) - \sum_{j=k+1}^{n} M_j (x-a)^j \,,
\end{aligned}$$

and observe that the right side is $o(\|x-a\|^k)$. This means, for each $x \in X$,

$$\frac{\|M_k(x-a)^k\|}{\|x-a\|^k} = \frac{\|M_k(t(x-a))^k\|}{\|t(x-a)\|^k} \to 0 \,, \quad \text{as } t \to 0.$$

So $M_k(x-a)^k = 0$. ■

Without some additional assumptions on M_k, it is *not* true in Theorem 4.2 that $M_k = 0$, meaning $M_k(x_1, \dots, x_k) = 0$ for all $x_1, \dots, x_k \in X$. An easy counterexample: for $x, y \in \mathbb{R}$

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

and the bilinear forms represented by the two 2×2 matrices are clearly not the same, even though they agree whenever their two functional arguments are set equal.

However, if we require that the multi-linear maps M_k be symmetric, then the assertion is true.

Theorem 4.3. *Let X, Y be normed vector spaces, and $M\colon X^n \to Y$ be a continuous symmetric multi-linear map. Then $M(x^n) = 0$ for all $x \in X$ implies $M(h_1, \ldots, h_n) = 0$ for all $h_1, \ldots, h_n \in X$.*

Proof. Let $g(x) = M(x^n)$, and differentiate this function n times using the product rule:

$$\begin{aligned}
\mathrm{D}g(x)h_1 &= \sum_{k=0}^{n-1} M(x^k, h_1, x^{n-1-k}) = nM(x^{n-1}, h_1) \\
\mathrm{D}^2 g(x)(h_2, h_1) &= n \sum_{k=0}^{n-2} M(x^k, h_2, x^{n-2-k}, h_1) = n(n-1)M(x^{n-2}, h_2, h_1) \\
&\vdots \\
\mathrm{D}^n g(x)(h_n, \ldots h_1) &= n!M(h_n, \ldots, h_1)
\end{aligned}$$

If $g(x) = 0$, then obviously $\mathrm{D}^n g(x) = 0$, which implies $M = 0$. ■

5 The classical Taylor's formula

My presentation of the Taylor's Theorem is not the traditional one, which mentions partial derivatives explicitly, rather than expressing higher derivatives as tensors. The advantage of the current presentation is that tensor notation is more concise, and that our proof carries through without changes for infinite-dimensional real normed vector spaces (which have no obvious notion of "coordinates").

But now let us make the connection to the traditional presentation of Taylor's Theorem in the finite-dimensional case.

Definition 5.1. Let V, W be normed vector spaces, A be open in V, and $f\colon A \to W$ be a function. Suppose $a \in A$ and $v \in V$. Define the *directional derivative* of f at a, in the direction v, to be:

$$\mathrm{D}_v f(a) = \lim_{t \to 0} \frac{f(a + tv) - f(a)}{t} = \left.\frac{d}{dt}\right|_{t=0} f(a + tv)\,.$$

If f is differentiable at a, it follows just as in multi-variable calculus that $\mathrm{D}_v f(a) = \mathrm{D}f(a) \cdot v$.

Let us calculate $\mathrm{D}_v \mathrm{D}_u f(a)$ explicitly, given that f is twice differentiable.

$$\begin{aligned}
\mathrm{D}_v \mathrm{D}_u f(a) &= \mathrm{D}(\mathrm{D}_u f)(a) \cdot v && \text{since } \mathrm{D}_u f = \mathrm{D}f \cdot u \text{ is differentiable at } a \\
&= \mathrm{D}(\mathrm{D}f \cdot u)(a) \cdot v && \text{where } \mathrm{D}f \cdot u \text{ refers to the \textit{function} } x \mapsto \mathrm{D}f(x) \cdot u \\
&= \mathrm{D}^2 f(a) \cdot u \cdot v && \text{by product rule} \\
&= \mathrm{D}^2 f(a) \cdot (u, v) && \text{slight change of notation}
\end{aligned}$$

This formula may be extended (by induction) to kth order directional derivatives:

$$\mathrm{D}_{v_k} \cdots \mathrm{D}_{v_1} f(a) = \mathrm{D}^k f(a) \cdot (v_1, \ldots, v_k)$$

To proceed, we need some algebraic definitions.

Definition 5.2. Let V and W be vector spaces over a scalar field F, and $S \in \mathcal{T}(V^k \to F)$, $T \in \mathcal{T}(V^l \to F)$. The *tensor product* $S \otimes T \in \mathcal{T}(V^{k+l} \to F)$ is defined by

$$(S \otimes T)(v_1, \ldots, v_k, v_{k+1}, \ldots, v_{k+l}) = S(v_1, \ldots, v_k) \cdot T(v_{k+1}, \ldots, v_{k+l}) \,.$$

Let us fix a basis $B = \{e_1, \ldots, e_n\}$ for a n-dimensional vector space V. Let $\pi_1, \ldots, \pi_n$ be the projection functions $\pi_n \colon V \to F$ for this basis (i.e. they are the "dual functions").

Suppose a tensor $T \in \mathcal{T}(V^k \to F)$ is given. It is not hard to see, by multi-linearity, that T is completely determined by the values it takes on each $(e_{i_1}, \ldots, e_{i_k})$, for $1 \leq i_j \leq n$. In more formal terms, this means that the set $\{\pi_{i_1} \otimes \cdots \otimes \pi_{i_k}\}$ form a basis for $\mathcal{T}(V^k \to F)$.

In the context of higher-order derivatives, we have the following basis representation of $\mathrm{D}^k f(a)$:

$$\begin{aligned} \mathrm{D}^k f(a) &= \sum_{i_1, \ldots, i_k} \mathrm{D}^k f(a) \cdot (e_{i_1}, \ldots, e_{i_k}) \cdot \pi_{i_1} \otimes \cdots \otimes \pi_{i_k} \\ &= \sum_{i_1, \ldots, i_k} \mathrm{D}_{e_{i_k}} \cdots \mathrm{D}_{e_{i_1}} f(a) \cdot \pi_{i_1} \otimes \cdots \otimes \pi_{i_k} \end{aligned}$$

Let us use the multi-index I to represent $(i_1, \ldots, i_k)$, and write $\mathrm{D}_{e_{i_k}} \cdots \mathrm{D}_{e_{i_1}} = \mathrm{D}_I$, and $\pi_{i_1} \otimes \ldots \otimes \pi_{i_k} = \pi_I$. Then the formula reads:

$$\mathrm{D}^k f(a) = \sum_I \mathrm{D}_I f(a) \, \pi_I$$

If $\{e_1, \ldots, e_n\}$ is the standard basis of $\mathbb{R}^n$, then $\mathrm{D}_I f(x)$ is simply the mixed partial derivative usually denoted by:

$$\frac{\partial}{\partial x_{i_k}} \cdots \frac{\partial}{\partial x_{i_1}} f = \frac{\partial^k f}{\partial x_{i_k} \cdots \partial x_{i_1}}$$

Usually one sees a multi-index notation that is different from ours, that we will explain now. Let J denote a n-tuple $(j_1, \ldots, j_n)$. (Warning: this differs from I, which is a k-tuple.) For f possessing continuous derivatives, the order of how we take mixed partial derivatives is irrelevant, as is familiar from calculus:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

So we may write $\mathrm{D}^J f(a)$ to mean: take the partial derivative with respect to e_1 j_1 times, then take the partial derivative with respect to e_2 j_2 times, and so on up to e_n and j_n. This will not depend on the order in which the partial derivatives are taken.

Let $|J| = j_1 + \cdots + j_n = k$, and $J! = j_1! \cdots j_n!$. Then for each J, there are

$$\mathrm{Perm}(J) = \frac{|J|!}{J!} = \frac{k!}{j_1! \cdots j_n!}$$

different orders in which partial derivatives can be taken that will yield $D^J f(a)$. Each different order corresponds to exactly one k-tuple I, and the corresponding derivative is $D_I f(a)$, which are all equal by equality of mixed partials.

Finally, if we define π^J by $\pi^J(x) = x_1^{j_1} \cdots x_n^{j_n}$, where $x_\nu = \pi_\nu(x)$, then it is clear that $\pi_I(x^k) = \pi_I(x, \ldots, x) = \pi^J(x)$. (This is in analogy with D^J versus D_I.)

Therefore,

$$\begin{aligned}\frac{\mathrm{D}^k f(a)}{k!} \cdot (x-a)^k &= \frac{1}{k!} \sum_I \mathrm{D}_I f(a)\, \pi_I(x-a)^k \\ &= \frac{1}{k!} \sum_J \frac{k!}{J!} \mathrm{D}^J f(a)\, \pi^J(x-a) = \sum_J \frac{\mathrm{D}^J f(a)}{J!} \pi^J(x-a)\,,\end{aligned}$$

and we have recovered the classical formula for Taylor polynomials:

$$\begin{aligned}f(x) &= \sum_{|J| \leq n} \frac{\mathrm{D}^J f(a)}{J!} \pi^J(x-a) + R(x) \\ &= \sum_{|J| \leq n} \frac{1}{J!} \left.\frac{\partial^{|J|} f}{\partial x^J}\right|_{x=a} (x-a)^J + R(x)\,.\end{aligned}$$

6 The Gâteaux derivative

The definition of the Fréchet derivative we had given has a serious shortcoming for applications: you cannot find the derivative $\mathrm{D}f(a)$ from the definition. To get around this, we calculate the directional derivatives instead, just as one takes partial derivatives for the finite-dimensional case. If the Fréchet derivative of f exists at a point a, then the directional derivatives of f at a determine the linear map $\mathrm{D}f(a)$. What we aim to prove here, is that if the directional derivatives are *continuous*, then $\mathrm{D}f(a)$ *exists*.

Definition 6.1. Let V and W be normed vector spaces, and A be open in V. A function $f\colon A \to W$ is *Gâteaux-differentiable* at $a \in A$ if the function $\mathrm{D_G} f(a)\colon V \to W$ defined by

$$\mathrm{D_G} f(a) \cdot v = \mathrm{D}_v f(a) = \left.\frac{d}{dt}\right|_{t=0} f(a+tv) \qquad \text{(derivative in direction } v\text{)}$$

is a continuous linear map. In this case, $\mathrm{D_G} f(a)$ is called the *Gâteaux derivative* of f at a.

Theorem 6.1 (Mean Value Theorem). *Let $f\colon A \to \mathbb{R}$, where A is open in a real normed vector space X. For each $a, x \in A$, if A contains the line segment from a to x, then there exists some ξ between a and x such that*

$$f(x) - f(a) = \mathrm{D_G} f(\xi) \cdot (x-a)\,.$$

Proof. Define $g(t) = f\bigl(a + t(x-a)\bigr)$, and so $g'(t) = \mathrm{D_G} f\bigl(a + t(x-a)\bigr) \cdot (x-a)$. We apply the ordinary Mean Value Theorem to $g\colon [0,1] \to \mathbb{R}$. ■

Corollary 6.2 (Generalized Mean Value Theorem). *Let $f\colon A \to Y$, where Y is a real normed vector space. For a, x satisfying the same hypotheses as above, there is some ξ between a and x such that*

$$\|f(x) - f(a)\| \leq \|\mathrm{D_G} f(\xi) \cdot (x-a)\|\,.$$

Proof. Apply the same trick with $\phi \in Y^*$ that worked for Theorem 3.2. ∎

Theorem 6.3. *Let $f\colon A \to Y$, where A is open in a real normed vector space X. If f is Gâteaux-differentiable on A and the Gâteaux derivative is continuous at a (in the norm of $\mathcal{L}(X \to Y)$), then the Fréchet derivative of f exists at a and coincides with the Gâteaux derivative.*

Proof. Given h close enough to a, apply the Mean Value Theorem to $\phi \circ f$ on the line from a to h, where $\phi \in Y^*$ has $\phi(v) = 1$, $\|\phi\| = 1$, and $v = f(a+h) - f(a) - \mathrm{D_G} f(a)h$. Then there is some ξ_h with $\|\xi_h\| \leq \|h\|$, such that

$$\begin{aligned} \phi\bigl(f(a+h) - f(a)\bigr) &= \phi\bigl(\mathrm{D_G} f(\xi_h)h\bigr)\,, \\ \phi(v) &= \phi\bigl(\mathrm{D_G} f(\xi_h)h - \mathrm{D_G} f(a)h\bigr)\,, \\ \|f(a+h) - f(a) - \mathrm{D_G} f(a)h\| &\leq \|\mathrm{D_G} f(\xi_h) - \mathrm{D_G} f(a)\|\,\|h\|\,. \end{aligned}$$

Thus

$$\frac{\|f(a+h) - f(a) - \mathrm{D_G} f(a)h\|}{\|h\|} \leq \|\mathrm{D_G} f(\xi_h) - \mathrm{D_G} f(a)\| \to 0, \text{ as } h \to 0.$$

This means $\mathrm{D_G} f(a)$ is the Fréchet derivative of f at a ($= \mathrm{D} f(a)$). ∎

Theorem 6.4. *Let $f\colon A \to Y$, where A open in X, be Gâteaux-differentiable on A. Suppose for some fixed $v \in X$, and for each $u \in X$ and $x \in A$, $\mathrm{D}_v \mathrm{D}_u f(x)$ exists. Suppose furthermore that for any $\epsilon > 0$, there is a $\delta > 0$ such that $\|\mathrm{D}_v \mathrm{D}_u f(x) - \mathrm{D}_v \mathrm{D}_u(a)\| \leq \epsilon\|u\|$ for $\|x - a\| < \delta$. Then $\mathrm{D}_v \mathrm{D_G} f(a)$ exists, and $\bigl(\mathrm{D}_v \mathrm{D_G} f(a)\bigr) \cdot u = \mathrm{D}_v \mathrm{D}_u f(a)$.*

Proof. The proof is straightforward, though involving some unwrapping of the abstractions. For $t \neq 0$, let

$$L_t = \frac{\mathrm{D_G} f(a+tv) - \mathrm{D_G} f(a)}{t} \in \mathcal{L}(X \to Y)\,.$$

Let $u \in X$ be arbitrary. Then

$$\begin{aligned} &\|L_t u - \mathrm{D}_v \mathrm{D}_u f(a)\| \\ &= \left\| \frac{\mathrm{D_G} f(a+tv)u - \mathrm{D_G} f(a)u}{t} - \mathrm{D}_v \mathrm{D}_u f(a) \right\| && \text{(expand L_t and evaluate at u)} \\ &= \left\| \frac{\mathrm{D}_u f(a+tv) - \mathrm{D}_u f(a)}{t} - \mathrm{D}_v \mathrm{D}_u f(a) \right\| && \text{(by definition of $\mathrm{D_G}$)} \\ &\leq \bigl\|\mathrm{D}_v \mathrm{D}_u f(a+\tilde{t}v) - \mathrm{D}_v \mathrm{D}_u f(a)\bigr\| \end{aligned}$$

(Same old trick: apply the Mean Value Theorem 6.1 on $\phi \circ D_u f$ on the line segment from a to $a + tv$, where $\phi \in Y^*$ with $\|\phi\| = 1$, $\phi(w) = \|w\|$, and $w = \mathrm{D}_u f(a+tv) - \mathrm{D}_u f(a) - \mathrm{D}_v \mathrm{D}_u f(a)$.)

$$\leq \epsilon\,\|u\|$$

(For t small enough; more precisely, when $|t| \leq \delta/\|v\|$. Note that $\tilde{t}$ depends on u, but $0 < |\tilde{t}| < |t|$ always holds no matter what u is.)

This shows that, as $t \to 0$, L_t converges in the operator norm of $\mathcal{L}(X \to Y)$ to the linear map $u \mapsto D_v D_u f$, so $\mathrm{D_G} f$ is differentiable at a in the direction v.

Finally, fixing u,

$$\|\mathrm{D}_v \mathrm{D_G} f(a) \cdot u - L_t u\| \leq \|\mathrm{D}_v \mathrm{D_G} f(a) - L_t\| \, \|u\| \to 0$$

since $\|\mathrm{D}_v \mathrm{D_G} f(a) - L_t\| \to 0$ by the definition of the directional derivative $\mathrm{D}_v \mathrm{D_G} f(a)$. So $\lim_{t \to 0} L_t u = \mathrm{D}_v \mathrm{D}_u f(a)$. But $\lim_{t \to 0} L_t u = \mathrm{D}_v \mathrm{D}_u f(a)$ from the first calculation, so $\mathrm{D}_v \mathrm{D_G} f(a) = \mathrm{D}_v \mathrm{D}_u f(a)$. ■

Remark 6.5. The practical significance of this theorem is that, if we have some formula for $\mathrm{D}_v \mathrm{D}_u f(x)$, and we can show $\|\mathrm{D}_v \mathrm{D}_u f(x) - \mathrm{D}_v \mathrm{D}_u f(y)\| \leq \epsilon \|u\| \|v\|$ when x is close to y, then $\mathrm{D}_u f$ is Gâteaux-differentiable at a. Moreover, it will follow that $\|\mathrm{D}_v \mathrm{D} f(x) - \mathrm{D}_v \mathrm{D} f(y)\| \leq \epsilon \|v\|$, and hence $\mathrm{D} f$ is Fréchet-differentiable, i.e. f is twice Fréchet-differentiable.

Note that this condition is essentially that the multi-linear maps $(u, v) \mapsto \mathrm{D}_v \mathrm{D}_u f(x)$ should vary continuously as a function of x, *in the $\mathcal{T}(X \times X \to Y)$ norm.* This certainly occurs for the functional J appearing in the last section.

Remark 6.6. Even more generally, the theorem may be extended to the case when $\|\mathrm{D}_{v_{k+1}} \mathrm{D}_{v_k} \cdots \mathrm{D}_{v_1} f(x) - \mathrm{D}_{v_k} \mathrm{D}_{v_{k-1}} \cdots \mathrm{D}_{v_1} f(a)\| \leq \epsilon \|v_1\| \cdots \|v_k\|$, which would show that $\mathrm{D}^{\mathrm{k}}_{\mathrm{G}}$ is differentiable at a in the direction v.

The practical consequence is, of course, that if $(v_1, \ldots, v_k) \mapsto \mathrm{D}_{v_k} \cdots \mathrm{D}_{v_1} f(x)$ is continuous in x, in the $\mathcal{T}(X \times \cdots \times X \to Y)$ norm, then f is k times Fréchet-differentiable.

In the development of the classical Taylor's formula, we used the fact that the order of taking partial derivatives is insignificant. Actually, this holds even for the infinite-dimensional case. We prove it now: it involves a clever application of the Mean Value Theorem:

Theorem 6.7. *Let A be open in X, and $f \colon A \to Y$ be twice continuously differentiable at a. Then $\mathrm{D}^2 f(a)$ is a symmetric multi-linear map; that is, for all $u, v \in X$,*

$$\mathrm{D}^2 f(a) \cdot (u, v) = \mathrm{D}^2 f(a) \cdot (v, u) \,.$$

Proof. Given $u, v \in X$, define, for s, t real,

$$\begin{aligned} \Delta_t(s) &= f(a + su + tv) - f(a + su) \,. \\ \Delta(s, t) &= f(a + 0) - f(a + su) - f(a + tv) + f(a + su + tv) \\ &= \Delta_t(s) - \Delta_t(0) \,. \end{aligned}$$

Assume first that $Y = \mathbb{R}$. By the Mean Value Theorem in one dimension, for Δ_t on $[0, s]$, there exists s_1 such that

$$\begin{aligned} \Delta_t(s) - \Delta(0) &= s \, \Delta_t'(s_1) \\ &= s \big[\mathrm{D} f(a + s_1 u + tv) \cdot u - \mathrm{D} f(a + s_1 u) \cdot u \big] \end{aligned}$$

Now holding s_1 fixed and considering the function $t \mapsto \mathrm{D} f(a + s_1 u + tv)$, we apply the Mean Value Theorem again, for this function on $[0, t]$:

$$= s \big[t \big[\mathrm{D}^2 f(a + s_1 u + t_1 v) \cdot v \cdot u \big] \big]$$

To summarize, we have the equation:

$$\Delta(s,t) = st\big[\mathrm{D}^2 f(a + s_1 u + t_1 v) \cdot v \cdot u\big]$$

for some s_1 and t_1. But the roles of u and v are completely symmetric in our argument, so we may also write

$$\Delta(s,t) = st\big[\mathrm{D}^2 f(a + s_2 u + t_2 v) \cdot u \cdot v\big]$$

for some s_2 and t_2.

In the last two equations, set $s = t$; then we have

$$\mathrm{D}^2 f(a + s_2 u + t_2 v) \cdot u \cdot v = \frac{\Delta(t,t)}{t^2} = \mathrm{D}^2 f(a + s_1 u + t_1 v) \cdot v \cdot u$$

Now take $t \to 0$. Then $s_1, s_2, t_1, t_2 \to 0$. Since $\mathrm{D}^2 f$ is continuous at a, the left and right limits coincide:

$$\mathrm{D}^2 f(a) \cdot u \cdot v = \mathrm{D}^2 f(a) \cdot v \cdot u\,.$$

To prove the general case for $Y \neq \mathbb{R}$, just apply what we have just proved to $\phi \circ f$, for each $\phi \in Y^*$: for each $u, v \in X$,

$$\phi\left(\mathrm{D}^2 f(a) \cdot u \cdot v\right) = \phi\left(\mathrm{D}^2 f(a) \cdot v \cdot u\right)\,.$$

But this is true for all linear functionals $\phi \in Y^*$; hence $\mathrm{D}^2 f(a) \cdot u \cdot v$ must equal $\mathrm{D}^2 f(a) \cdot v \cdot u$. (According to the Hahn-Banach Theorem, the continuous linear functionals separate points in Y.) ■

Remark 6.8. Theorem 6.7 applies to the double directional derivative $\mathrm{D}^2_{\mathrm{G}} f$ in place of $\mathrm{D}f$ as well, since we only took directional derivatives in u and v in the proof.

Remark 6.9. It is clear, by induction, that every higher-order derivative (if continuous) is a symmetric multi-linear map.

7 An application: the hanging-chain problem

Before proceding further in the abstract theory, let us carry out a computation for this classic problem:

Find the shape taken by a rope or chain affixed to two poles of the same height, under the influence of constant gravity. (Assume the rope has uniform mass per unit length.)

Suppose $y(x)$ is the height of the rope at horizontal position x, $-a \leq x \leq +a$, where $y(-a) = c = y(a)$, the height of the poles. The solution y should minimize the potential energy of the rope:

$$\int_{-a}^{+a} gy\,\rho\,ds = g\rho \int_{-a}^{+a} y\sqrt{1 + \left(\frac{dy}{dx}\right)^2}\,dx\,,$$

where ρ is the density of the rope, g is the acceleration due to gravity, and ds denotes the infinitesimal element of arc length.

We abstract this problem: our goal is to minimize a functional $J\colon \mathcal{F} \to \mathbb{R}$:

$$J(f) = \int_a^b L(f(t), f'(t), t)\, dt$$

where L is some fixed function (assumed to be twice continuously differentiable), and $\mathcal{F} = \mathcal{C}^1[a,b]$ is the space of continuously differentiable functions on $[a,b]$, equipped with the norm

$$\|f\| = \sup_{[a,b]} |f| \; + \; \sup_{[a,b]} |f'| \, .$$

(The variable names in the original problem and the abstract problem do not quite match, unfortunately. To clarify, in the abstract problem we are to take $L(x,y,t) = x\sqrt{1+y^2}$ for all x, y, t.)

Actually we are not minimizing J over all of $\mathcal{F}$, but just over functions f with the end-points $f(a)$ and $f(b)$ held fixed. (In the original problem, this corresponds to the height of the poles being held fixed.)

To find the minimizing f, we take the derivative of J in a direction $\varphi \in \mathcal{F}$ and set it to zero. (The rationale is, in the expression for the directional derivative, $\mathrm{D}_\varphi J(f) = \frac{d}{dt}|_{t=0} J(f + t\varphi)$, the expression $J(f + t\varphi)$ is a real-valued function of the real variable t that is minimized at $t = 0$; hence from calculus we know its derivative there must be zero.) Since we are only minimizing only functions with fixed end-points, we must only take φ such that $\varphi(a) = \varphi(b) = 0$.

So we compute:

$$\begin{aligned}
\mathrm{D_G} J(f) \cdot \varphi &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} J(f + \epsilon\varphi) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int_a^b L(f + \epsilon\varphi, f' + \epsilon\varphi', t)\, dt \\
&= \int_a^b \left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0} L(f + \epsilon\varphi, f' + \epsilon\varphi', t)\, dt \\
&= \int_a^b \bigl[D_x L(f, f', t) \cdot \varphi + D_y L(f, f', t) \cdot \varphi'\bigr]\, dt
\end{aligned}$$

Here $D_x L$ and $D_y L$ denote the partial derivatives of L with respect to the first and second variables, respectively. The differentiation under the integral sign can be justified by the uniform continuity of functions on a compact set. Notice that $\mathrm{D_G} J(f) \cdot \varphi$ is linear in φ, as expected.

The last equation is often simply written

$$\frac{\partial J}{\partial \varphi} = \int_a^b \left[\frac{\partial L}{\partial x}\varphi + \frac{\partial L}{\partial y}\dot{\varphi}\right] dt \, ,$$

where it is implicitly understood that the partial derivatives of L are to be evaluated at $(x, y, t) = (f(t), f'(t), t)$.

Actually, we are supposed to be minimizing J subject to the condition that the end-points at a and b are fixed, so we must consider only functions φ with $\varphi(a) = \varphi(b) = 0$.

Exploiting this vanishing boundary condition, integration by parts on the last integral yields:

$$\frac{\partial J}{\partial \varphi} = \int_a^b \left[\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial y} \right] \varphi \, dt \,.$$

If this integral is to be minimized over all φ (with vanishing boundary), the factor that multiplies φ in the integrand must be equal to zero[2], and this leads to the well-known Euler-Lagrange differential equation,

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial y} = 0 \,,$$

that may be solved for the function f. If L is time-independent ($L(x, y, t) = L(x, y)$), then on integration by parts this equation is equivalent to:

$$L - \frac{dx}{dt}\frac{\partial L}{\partial y} = C \,,$$

for some constant C. (This is called the "Beltrami identity".)

It is not hard to prove, using uniform continuity, that $\mathrm{D_G}J(f)\varphi$ is "continuous in both f and φ". Precisely, given $\epsilon > 0$, there is a $\delta > 0$ such that if $\|f - f_0\| \leq \delta$, then $|\mathrm{D_G}J(f)\varphi - J(f_0)\varphi| \leq \epsilon\|\varphi\|$. It follows that $\|\mathrm{D_G}J(f) - J(f_0)\| \leq \epsilon$. Hence the Gâteaux derivative $\mathrm{D_G}J$ is continuous, and so by Theorem 6.3, this is in fact also the Frêchet derivative[3].

The reader is invited to work out the details, and also to compute the solution curve $y(x)$ to the original problem.

One question remains[4]: is the computed solution to $\mathrm{D}_{\mathrm{G}J}(f) = 0$ actually a (local) minimum point?

To answer this, we can try to analyze the second derivative for positive-definiteness, analogous to the procedure in multi-variable calculus. Computing directional derivatives like before, we obtain:

$$\begin{aligned} \mathrm{D}^2_{\mathrm{G}}J(f) \cdot \psi\varphi &= \mathrm{D}_\psi \mathrm{D}_\varphi J(f) \\ &= \int_a^b \left[\frac{\partial^2 L}{\partial x \partial x}\psi\varphi + \frac{\partial^2 L}{\partial x \partial y}\psi\dot{\varphi} + \frac{\partial^2 L}{\partial y \partial x}\dot{\psi}\varphi + \frac{\partial^2 L}{\partial y \partial y}\dot{\psi}\dot{\varphi} \right] dt \,. \end{aligned}$$

Here we differentiated with respect to ψ holding φ fixed. It is not yet clear whether the second Fréchet derivative exists, since the second derivative involves the differentiation of *linear maps*, i.e. we are not supposed to be holding φ fixed and then differentiating, but differentiating "for all φ at the same time". But it can be shown that if certain things are continuous, then it all works out — this motivates the results of the next section.

8 Conditions for minima and maxima

Before returning to the hanging-chain problem, we present a simple criterion for an equilibrium point a of f ($\mathrm{D}f(a) = 0$) to be a local minimum.

[2] This fact is sometimes referred to as the Fundamental Lemma of the calculus of variations. We have not given its formal proof here, it is not hard — it's a first-year calculus problem.

[3] Strictly speaking, the analysis does not really require the use of the Frêchet derivative, just only the directional derivative, but we got the Frêchet derivative for free anyway.

[4] Of course, if we were to wear a physicist's hat, then the problem need not be analyzed further.

Theorem 8.1 (Second Derivative Test). *Let A be open in X, $f\colon A \to \mathbb{R}$ twice differentiable, with equilibrium point a. If $\mathrm{D}^2 f(x)$ is positive semi-definite ($\mathrm{D}^2 f(x) \cdot h^2 \geq 0$ for all h) for x near a, then a is a local minimum.*

Proof. Expand with Taylor's formula: for $\|h\| < \delta$, then there exists a $\tilde{h}$, $\|\tilde{h}\| < \delta$, such that

$$f(a+h) - f(a) = f(a+h) - f(a) - \mathrm{D}f(a) \cdot h = \frac{1}{2}\mathrm{D}^2 f(\tilde{h}) \cdot h^2 \,,$$

and this quantity is ≥ 0 by hypothesis. ∎

Note that in the infinite-dimensional setting, it does not suffice in general that $\mathrm{D}^2 f$ be positive-definite only at a. This is because even if $\mathrm{D}^2 f(a)$, even if positive-definite, may get arbitrarily close to zero, and the $o(\|h\|^2)$ term may turn out to be bigger than this even when $\|h\|$ is small. (In finite-dimensional spaces, $\mathrm{D}^2 f(a)$ has a minimum on the unit sphere, because the sphere is compact, so what has been just described cannot happen.)

Remark 8.2. Needless to say, a local maximum is obtained if we replace "positive semi-definite" by "negative semi-definite". And a strict maximum or minimum is obtained if we replace "semi-definite" by "definite".

Remark 8.3. Sometimes h is to be restricted to some subset S of the whole space X; e.g. when we derived the Euler-Lagrange equation we restricted φ to be zero at the end-points a and b. In this case $\mathrm{D}f(a)$ may not be zero (i.e. $\mathrm{D}f(a) \cdot v$ is not always zero for vectors v), but the Second Derivative Test still works, provided, of course, that $\mathrm{D}f(a) \cdot h = 0$ for $h \in S$.

9 The hanging-chain problem, concluded

We now return to the problem of section 7.

It turns out that there are exactly two solutions to the Beltrami identity in the hanging-chain problem, with initial conditions $y(-a) = y(a) = c$: $y_0(x) = c$ and $y_1(x) = \lambda \cosh(x/\lambda)$. The former solution corresponds to a rope that is held tight at $x = -a$ and at $x = a$, making a horizontal line — the trivial situation. So we think we want the latter solution is

However, if we apply the Second Derivative Test, we will not get anywhere. I've spent hours trying to prove that y_1 is a minimizes J, until I realized that it couldn't be a minimum. Specific counterexamples can be constructed, but just consider this informal argument: if more rope were to be fed into the system at the left and right poles $x = -a$ and $x = a$, this will cause the rope to droop down even more. Now, this may not necessarily lower the potential energy of the rope, because the length of the rope has increased, so there is more mass. Nevertheless, we never said anything about the arc length of the rope in our equations — specifically, what we actually want is to minimize J over curves of a fixed length ℓ.

In this case it is not guaranteed that the Euler-Lagrange equation even holds, because as you will remember, to derive it we required that the integral $\mathrm{D}f(f) \cdot \varphi$ be equal to zero for all φ, but for most variations φ, the arc length is changed, and we don't want that.

The fix for our problem is to somehow introduce arc-length parameterizations into the integral we are minimizing.

For what follows, let

$$S(x) = \int_0^x \sqrt{1+f'(x)^2}\,dx$$

be the arc length function for the curve $y = f(x)$. Also let $X(s) = S^{-1}(s)$ be the inverse function, which exists since $S'(x) \geq 1$ for $0 \leq x \leq a$. We have $S(0) = 0$ and $S(a) = \ell$, the arc length that is to be held constant.

Consider the original integral (omitting the constant factor ρg)

$$\int_{-a}^{a} f(x)\sqrt{1+f'(x)^2}\,dx\,.$$

Let us restrict ourselves to solutions f that are horizontally symmetric, i.e. f is an even function. It is clear physically why the solution to the hanging-chain problem must be symmetric, but that can be proven rigorously too. (If the solution is not symmetric, then we can readily find another solution that is symmetric, and has an amount of energy not exceeding that of the original solution.)

So we change the limits of integration to $[0, a]$. Next, make the substitution $x = X(s)$. The integral transforms to:

$$\int_0^\ell f(X(s))\sqrt{1+f'(X(s))^2}X'(s)\,ds = \int_0^\ell f(X(s))\,ds\,.$$

(The two integrals are equal because $X'(s) = (1 + f'(X(s))^2)^{-\frac{1}{2}}$.)

We get rid of the f in this integral as follows:

$$\begin{aligned}
\frac{1}{X'(s)^2} &= 1 + f'(X(s))^2 \\
\sqrt{\frac{1}{X'(s)^2} - 1} &= f'(X(s)) \\
\sqrt{1 - X'(s)^2} &= f'(X(s)) \cdot X'(s) \\
\int_s^\ell \sqrt{1 - X'(\tilde{s})^2}\,d\tilde{s} &= \int_s^\ell f'(X(\tilde{s}))X'(\tilde{s})\,d\tilde{s} = f(X(\tilde{s}))\Big|_{\tilde{s}=\ell}^{\tilde{s}=s} \\
&= f(X(\ell)) - f(X(s)) \\
&= f(a) - f(X(s)) \\
&= c - f(X(s))
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_0^\ell f(X(s))\,ds &= \int_0^\ell \left[c - \int_s^\ell \sqrt{1 - X'(\tilde{s})^2}\,d\tilde{s}\right] ds \\
&= c\ell - \int_0^\ell \int_s^\ell \sqrt{1 - X'(\tilde{s})^2}\,d\tilde{s}\,ds \\
&= c\ell - \int_0^\ell \int_0^{\tilde{s}} \sqrt{1 - X'(\tilde{s})^2}\,ds\,d\tilde{s} \qquad \text{(switch order of integration)} \\
&= c\ell - \int_0^\ell \tilde{s}\sqrt{1 - X'(\tilde{s})^2}\,d\tilde{s}\,.
\end{aligned}$$

So our goal now should be to maximize

$$K(X) = \int_0^\ell s\sqrt{1 - X'(s)^2}\, ds$$

subject to the constraints $X(0) = 0$ and $X(\ell) = a$.

The Euler-Lagrange equation (for finding an equilibrium point of K) reads, after simplification:

$$0 = sX''(s) + X'(s) - X'(s)^3$$

Although this equation can be solved directly, it is easier to substitute what we think the solution is: $X = X_1$ where $S_1 = X_1^{-1}$ is the arc length function for the catenary $y_1(x) = \lambda \cosh(x/\lambda)$. X_1 and S_1 are found to be:

$$S_1(x) = \lambda \sinh(x/\lambda)\,, \quad X_1(s) = \lambda \sinh^{-1}(s/\lambda)\,, \quad X_1'(s) = \frac{1}{\sqrt{1 - (s/\lambda)^2}}\,.$$

(λ is supposed to satisfy $\ell = S(a) = \lambda \sinh(a/\lambda)$, but it cannot be determined in closed form.) You may verify easily that $X = X_1$ indeed satisfies the preceding differential equation.

Now we calculate second derivatives:

$$\begin{aligned} \mathrm{D}K(X) \cdot \varphi^2 &= \int_0^\ell \frac{\partial^2}{\partial u \partial u} s\sqrt{1 - u^2}\bigg|_{u = X'(s)} \varphi'(s)^2\, ds \\ &= -\int_0^\ell \frac{sX'(s)\, \varphi'(s)^2}{\big(1 - X'(s)^2\big)^{3/2}}\, ds\,. \end{aligned}$$

Aha! Every factor inside the last integral is non-negative, so with the negative sign, $\mathrm{D}K(X)$ is negative semi-definite. In fact, if $\varphi \neq 0$, the integral will be strictly positive (by continuity), so $\mathrm{D}K(X)$ is negative *definite*. What's even more interesting, $\mathrm{D}K(X)$ is negative definite independent of what X is, so we've just proved that $X = X_1$ is a *global* strict maximum for K (under the constraints). (In a more geometric language, what we have shown is that K is concave, so an equilibrium point is automatically a global maximum.)

So the curve y_1 is indeed our solution to the hanging-chain problem.