Calculus on a Normed Linear Space

James S. Cook
Liberty University
Department of Mathematics

Fall 2017

introduction and scope

These notes are intended for someone who has completed the introductory calculus sequence and has some experience with matrices or linear algebra. This set of notes covers the first month or so of Math 332 at Liberty University. I intend this to serve as a supplement to our required text: *First Steps in Differential Geometry: Riemannian, Contact, Symplectic* by Andrew McInerney. Once we've covered these notes then we begin studying Chapter 3 of McInerney's text.

This course is primarily concerned with abstractions of calculus and geometry which are accessible to the undergraduate. This is not a course in real analysis and it does not have a prerequisite of real analysis so my typical students are not prepared for topology or measure theory. We defer the topology of manifolds and exposition of abstract integration in the measure theoretic sense to another course. Our focus for course as a whole is on what you might call the *linear algebra* of abstract calculus. Indeed, McInerney essentially declares the same philosophy so his text is the natural extension of what I share in these notes. So, what **do** we study here? In these notes:

How to generalize calculus to the context of a normed linear space

In particular, we study: basic linear algebra, spanning, linear independennce, basis, coordinates, norms, distance functions, inner products, metric topology, limits and their laws in a NLS, continuity of mappings on NLS's, linearization, Frechet derivatives, partial derivatives and continuous differentiability, linearization, properties of differentials, generalized chain and product rules, intuition and statement of inverse and implicit function theorems, implicit differentiation via the method of differentials, manifolds in $\mathbb{R}^n$ from an implicit or parametric viewpoint, tangent and normal spaces of a submanifold of Euclidean space, Lagrange multiplier technique, compact sets and the extreme value theorem, theory of quadratic forms, proof of real Spectral Theorem via method of Lagrange multipliers, higher derivatives and the multivariate Taylor theorem, multivariate power series, critical point analysis for multivariate functions, introduction to variational calculus.

In contrast to some previous versions of this course, I do not study contraction mappings, differentiating under the integral and other questions related to uniform convergence. I leave such topics to a future course which likely takes Math 431 (our real analysis course) as a prerequisite. Furthermore, these notes have little to say about further calculus (differential forms, vector fields, etc.). We read on those topics in McInerney once we exhaust these notes.

There are many excellent texts on calculus of many variables. Three which have had significant influence on my thinking and the creation of these notes are:

1. *Advanced Calculus of Several Variables* revised Dover Ed. by C.H. Edwards,
2. *Mathematical Analysis II*, Vladimir A. Zorich,
3. *Foundations of Modern Analysis*, by J. Dieudonné, Academic Press Inc. (1960)

These notes are a work in progress, do let me know about the errors. Thanks!

James S. Cook, August 14, 2017.

Contents

Chapter 1

on norms and limits

A normed linear space is a vector space which also has a concept of vector length. We use this length function to set-up limits for maps on normed linear spaces. The idea of the limit is the same as it was in first semester calculus; we say the map approaches a value when we can make values of the map arbitrary close to the value by taking inputs sufficiently close to the limit point. A map is continuous at a limit point in its domain if and only if its limiting value matches its actual value at the limit point. We derive the usual limit laws and work out results which are based on the component expansion with respect to a basis. We try to provide a fairly complete account of why common maps are continuous. For example, we argue why the determinant map is a continuous map from square matrices to real numbers.

We also introduce elementary concepts of topology. Open and closed sets are defined in terms of the metric topology induced from a given norm. We also discuss inner products and the more general concept of a distance function or metric. We explain why the set of invertible matrices is topologically open.

This Chapter concludes with a brief introduction into sequential methods. We define completeness of a normed linear space and hence introduce the concept of a Banach Space. Finally, the matrix exponential is shown to exist by an analytical appeal to the completeness of matrices.

Certain topics are not covered in depth in this work, I survey them here to attempt to provide context for the larger world of math I hope my students soon discover. In particular, while I introduce inner products, metric spaces and the rudiments of functional analysis, there is certainly far more to learn and indicate some future reading as we go. For future chapters we need to understand both linear algebra and limits carefully so my focus here is on normed linear spaces and limits. These suffice for us to begin our study of Frechet differentiation in the next chapter.

History is important and I must admit failure on this point. I do not know the history of these topics as deeply as I'd like. Similar comments apply to the next Chapter. I believe most of the linear algebra and analysis was discovered between about 1870 and 1910 by the likes of Frobenius, Frechet, Banach and other great analysts of that time, but, I have doubtless left out important work and names.

1.1 linear algebra

A real vector space is a set with operations of addition and scalar multiplication which satisfy a natural set of axioms. We call elements of the vector space **vectors**. We are primarily focused on **real** vector spaces which means the scalars are real numbers. Typical examples include:

(1.) $\mathbb{R}^n = \{(x_1, \ldots, x_n) \mid x_1, \ldots, x_n \in \mathbb{R}\}$ where for $x, y \in \mathbb{R}^n$ and $c \in \mathbb{R}$ we define $(x+y)_i = x_i + y_i$ and $(cx)_i = cx_i$ for each $i = 1, \ldots, n$. In words, these are real n-tuples formed as column vectors. The notation $(x_1, \ldots, x_n)$ is shorthand for $[x_1, \ldots, x_n]^T$ in order to ease the typesetting.

(2.) $\mathbb{R}^{m\times n}$ the set of $m \times n$ real matrices. If $A, B \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}$ then $(A+B)_{ij} = A_{ij} + B_{ij}$ and $(cA)_{ij} = cA_{ij}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$. Notice, an $m \times n$ matrix can be thought of as n-columns from $\mathbb{R}^m$ glued together, or as m-rows from $\mathbb{R}^{1\times n}$ glued together (sometimes I say the rows or columns are concatenated)

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} = [col_1(A)|col_2(A)|\cdots|col_n(A)] = \left[\begin{array}{c} row_1(A) \\ \hline row_2(A) \\ \hline \vdots \\ \hline row_m(A) \end{array}\right] \tag{1.1}$$

In particular, it is at times useful to note: $(col_j(A)_i = A_{ij}$ and $(row_i(A))_j = A_{ij}$. Furthermore, in addition to the vector space structure, we also have a **multiplication** of matrices; for $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ the matrix $AB \in \mathbb{R}^{m\times p}$ is defined by $(AB)_{ij} = row_i(A) \bullet col_j(B)$ which can be written in index notation as:

$$(AB)_{ij} = \sum_{k=1}^{p} A_{ik}B_{kj}. \tag{1.2}$$

(3.) $\mathbb{C}^n = \{(z_1, \ldots, z_n) \mid z_1, \ldots, z_n \in \mathbb{C}\}$. Once more define addition and scalar multiplication component-wise; for $z, w \in \mathbb{C}^n$ and $c \in \mathbb{C}$ define $(z+w)_i = z_i + w_i$ and $(cz)_i = cz_i$. Since $\mathbb{R} \subseteq \mathbb{C}$ the complex scalar multiplication in $\mathbb{C}^n$ also provides a real scalar multiplication. We can either view $\mathbb{C}^n$ as a real or complex vector space.

(4.) $\mathbb{C}^{m\times n}$ is the set of $m \times n$ complex matrices. If $Z, W \in \mathbb{C}^{m\times n}$ and $c \in \mathbb{C}$ then $(Z+W)_{ij} = Z_{ij} + W_{ij}$ and $(cZ)_{ij} = cZ_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Just as in the previous example, we can view $\mathbb{C}^{m\times n}$ either as a complex vector space or as a real vector space.

(5.) If V and W are real vector spaces then $Hom_{\mathbb{R}}(V, W)$ is the set of all linear transformations from V to W. This forms a vector space with respect to the usual pointwise addition of functions. If $V = W$ then we denote $Hom_{\mathbb{R}}(V, V) = End_{\mathbb{R}}(V)$ for **endomorphisms** of V. The set of endomorphisms forms an algebra with respect to composition of functions since the composite of linear maps is once more linear. The set of invertible endomorphisms of V forms $GL(V)$. In the particular case that $V = \mathbb{R}^n$ we denote $GL(\mathbb{R}^n) = GL(n, \mathbb{R})$. Notice $GL(V)$ is not a subspace since $Id_V \in GL(V)$ where $Id_V(x) = x$ for all $x \in V$ and $Id_V - Id_V = 0 \notin GL(V)$.

Definition 1.1.1.

If V is real vector space and $S \subseteq V$ then define the **span of** S by

$$span(S) = \{c_1 s_1 + \cdots + c_k s_k \mid s_1, \ldots, s_k \in S, c_1, \ldots, c_k \in \mathbb{R}, k \in \mathbb{N}\}.$$

In words, $span(S)$ is the set of all finite $\mathbb{R}$-linear combinations of vectors from S. Since the scalar multiple and linear combination of linear combinations is once more a linear combination we find that $span(S) \leq V$. That is, $span(S)$ forms a **subspace** of V. The set S is called a **spanning set** or **generating set** for $span(S)$.

Definition 1.1.2.

Let V be a real vector space and $S \subseteq V$. If $c_1, \ldots, c_k \in \mathbb{R}$ and $s_1, \ldots, s_k \in S$ with $c_1 s_1 + \cdots + c_k s_k = 0$ imply $c_1 = 0, \ldots, c_k = 0$ for each $k \in \mathbb{N}$ then S is **linearly independent** (LI). Otherwise, we say S is linearly dependent.

When generating sets are linearly independent they are minimal, if you remove any vector from a minimal spanning set then the resulting span is smaller. In contrast, if S is linearly dependent then there exists $S' \subset S$ for which $span(S') = span(S)$. Our convention is that $span(\emptyset) = \{0\}$.

Definition 1.1.3.

Let V be a real vector space. If β is a linearly independent spanning set for V then we say β is a **basis** for V. Furthermore, using $\#$ to denote cardnality, $\#(\beta)$ is the **dimension** of V. If $\#(\beta) = n \in \mathbb{N}$ then we say V is an n-dimensional vector space and write $dim(V) = n$.

Bases are very helpful for calculations. In particular, if $\beta = \{v_1, \ldots, v_n\}$ then

$$x_1 v_1 + \cdots + x_n v_n = y_1 v_1 + \cdots + y_n v_n \quad \Rightarrow \quad x_i = y_i \text{ for } i = 1, \ldots, n. \tag{1.3}$$

We call this calculation **equating coefficients** with respect to the basis β.

Definition 1.1.4.

Let V be a real finite dimensional vector space with basis $\beta = \{v_1, \ldots, v_n\}$ then for each $x \in V$ there exist $c_i \in \mathbb{R}$ for which $x = c_1 v_1 + \cdots + c_n v_n$. We write $[x]_\beta = (c_1, \ldots, c_n)$ and say $[x]_\beta$ is the **coordinate vector** of x with respect to the β basis. We also denote $\Phi_\beta(x) = [x]_\beta$ and say $\Phi_\beta : V \to \mathbb{R}^n$ is the **coordinate map** with respect to the basis β.

If $\beta = \{v_1, \ldots, v_n\}$ is a basis for the real vector space V and $\psi \in GL(V)$ then $\psi(\beta) = \{\psi(v_1), \ldots, \psi(v_n)\}$ forms a basis for V. Clearly the choice of basis is far from unique. That said, it is useful for us to settle on a **standard basis** for our usual real examples:

(1.) Let $(e_i)_j = \delta_{ij}$ hence $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$ and $e_n = (0, \ldots, 0, 1)$. If $\beta = \{e_1, \ldots, e_n\}$ then $x = (x_1, \ldots, x_n) = \sum_{i=1}^n x_i e_i$ and $[x]_\beta = x$. We say β is the **standard basis of column vectors** and note $\#(\beta) = n = dim(\mathbb{R}^n)$.

(2.) Let $(E_{ij})_{kl} = \delta_{ik}\delta_{jl}$ for $1 \leq i, k \leq m$ nad $1 \leq j, l \leq n$ define $E_{ij} \in \mathbb{R}^{m \times n}$. The matrix E_{ij} has a 1 in the ij-th entry and zeros elsewhere. For any $A \in \mathbb{R}^{m \times n}$ we have

$$A = \sum_{i=1}^m \sum_{j=1}^n A_{ij} E_{ij} \tag{1.4}$$

We order the **standard** $m \times n$ **matrix basis** $\beta = \{E_{ij} \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ by the usual lexographic ordering. For example, in the case of 2×3 matrices,

$$\beta = \{E_{11}, E_{12}, E_{13}, E_{21}, E_{22}, E_{23}\} \tag{1.5}$$

Following the notation from Equation 1.1,

$$\Phi_\beta(A) = (A_{11}, A_{12}, \ldots, A_{1n}, A_{21}, A_{21}, \ldots, A_{2n}, \ldots, A_{mn}) \tag{1.6}$$

The coordinate vector for $A \in \mathbb{R}^{m \times n}$ w.r.t. the standard basis is given by listing out the components of A row-by-row. Also, $\#(\beta) = mn = dim(\mathbb{R}^{m \times n})$.

Viewing $\mathbb{C}^n$ and $\mathbb{C}^{m\times n}$ as real vector spaces there are at least two natural choices for the basis,

(3.) For $\mathbb{C}^n$ notice $\beta = \{e_1, ie_1, \ldots, e_n, ie_n\}$ and $\gamma = \{e_1, \ldots, e_n, ie_1, \ldots, ie_n\}$ serve as natural bases. If $z = x + iy$ where $x, y \in \mathbb{R}^n$ then we **define** $Re(z) = x$ and $Im(z) = y$. Hence,[1]

$$\Phi_\gamma(z) = (x, y), \quad \& \quad \Phi_\beta(z) = (x_1, y_1, x_2, y_2, \ldots, x_n, y_n). \tag{1.7}$$

Note $dim_{\mathbb{R}}(\mathbb{C}^n) = 2n$.

(4.) For $\mathbb{C}^{m\times n}$ notice $\beta = \{E_{11}, iE_{11}, \ldots, E_{mn}, iE_{mm}\}$ and $\gamma = \{E_{11}, \ldots, E_{mn}, iE_{11}, \ldots, iE_{mn}\}$ serve as natural bases. If $Z = X + iY$ where $X, Y \in \mathbb{R}^{m\times n}$ then we **define** $Re(Z) = X$ and $Im(Z) = Y$. With this notation,

$$[Z]_\gamma = (X_{11}, \ldots, X_{mn}, Y_{11}, \ldots, Y_{mn}) \quad \& \quad [Z]_\beta = (X_{11}, Y_{11}, \ldots, X_{mn}, Y_{mn}). \tag{1.8}$$

For example,

$$A = \begin{bmatrix} 1+i & 2 \\ 3+4i & 5i \end{bmatrix} \quad \Rightarrow \quad [A]_\beta = (1, 1, 2, 0, 3, 4, 0, 5) \quad \& \quad [A]_\gamma = (1, 2, 3, 0, 1, 0, 4, 5).$$

Finally, note $dim_{\mathbb{R}}(\mathbb{C}^{m\times n}) = 2mn$

Naturally, $dim_{\mathbb{C}}(\mathbb{C}^n) = n$ and $dim_{\mathbb{C}}(\mathbb{C}^{m\times n}) = mn$, but, our primary interest is in the calculus of real vector spaces so we just need such formulas as a conceptual backdrop.

1.2 norms, metrics and inner products

The concept of norm, metric and inner product all strike a the same issue; how to describe distance abstractly. Of these the inner product is most special and the metric or **distance function** is most general. In particular, both norms and inner products require a background vector space. In constrast, distance functions can be given to all sorts of sets where there is no well-defined addition which closes on the set. The general study of distance functions belongs to real or functional analysis, however, I think it is important to mention them here for context.

1.2.1 normed linear spaces

This definition abstracts the concept of vector length:

Definition 1.2.1. *Normed Linear Space (NLS):*

Suppose V is a real vector space. If $||\cdot|| : V \times V \to \mathbb{R}$ is a function such that for all $x, y \in V$ and $c \in \mathbb{R}$:

(1.) $||cx|| = |c|\,||x||$

(2.) $||x + y|| \leq ||x|| + ||y||$ (triangle inequality)

(3.) $||x|| \geq 0$

(4.) $||x|| = 0$ if and only if $x = 0$

then we say $(V, ||\cdot||)$ is a normed vector space. When there is no danger of ambiguity we also say that V is a **normed vector space** or a **normed linear space (NLS)**.

Notice that we did not assume V was finite-dimensional in the definition above. Our current focus is on finite-dimensional cases.

[1] technically, this is an abuse of notation, I'm ignoring the distinction between a vector of vectors and a vector

(1.) the standard **euclidean** norm on $\mathbb{R}^n$ is defined by $||v|| = \sqrt{v \bullet v}$.

(2.) the **taxicab** norm on $\mathbb{R}^n$ is defined by $||v||_1 = \sum_{j=1}^n |v_j|$.

(3.) the **sup** norm on $\mathbb{R}^n$ is defined by $||v||_\infty = \max\{|v_j| \mid j = 1, \ldots, n\}$

(4.) the p-norm on $\mathbb{R}^n$ is defined by $||v||_p = \left(\sum_{j=1}^n |v_j|^p\right)^{1/p}$ for $p \in \mathbb{N}$. Notice $||v|| = ||v||_2$ and the taxicabl is the $p = 1$ case and finally the sup-norm appears as $p \to \infty$.

If we identify points with vectors based at the origin then it is natural to think about a **circle** of radius 1 as the set of vectors (points) which have norm (distance) one. Focusing on $n = 2$,

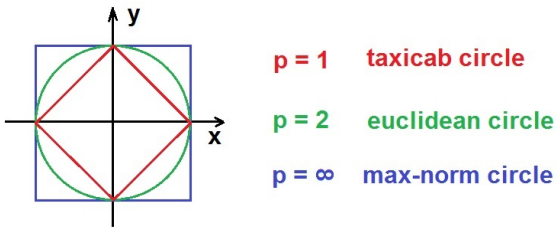
(1.) in the euclidean case a circle is the circle for $\mathbb{R}^2$ with $||v||_2 = 1$.

(2.) if we use the taxicab norm on $\mathbb{R}^2$ then the circle is a diamond.

(3.) for $\mathbb{R}^2$ with the p-norm the circle is something like a blown-up circle.

(4.) as $p \to \infty$ the circle expands to a square.

In other words, to **square the circle** we need only study p-norms in the plane.



We could play similar games for our other favorite examples, but primarily we just use the analog of the $p = 1, 2$ or ∞ norms in our application of norms. Let us agree by **convention** that $\|x\| = \sqrt{x \bullet x}$ for $x \in \mathbb{R}^n$, since the coordinate map yields real column vectors the r.h.s. makes use of this convention in each of the following examples:

(2.) the standard norm for $\mathbb{R}^{m\times n}$ is given by $||A|| = ||\Phi_\beta(A)||$ where Φ_β is the standard coordinate map for $\mathbb{R}^{m\times n}$ as defined in Equation 1.6.

(3.) the standard norm for $\mathbb{C}^n$ is given by $||z|| = ||\Phi_\beta(z)||$ where Φ_β is the standard coordinate map described in Equation 1.7.

(4.) the standard norm for $\mathbb{C}^{m\times n}$ is given by $||Z|| = ||\Phi_\beta(Z)||$ where Φ_β is the standard coordinate map for $\mathbb{C}^{m\times n}$ as defined in Equation 1.8.

In each case above there is some slick formula which hides the simple truth I described above; the length of matrices and complex vectors is simply the Euclidean length of the corresponding coordinate vectors.

$$||v||^2 = v^T \bar{v}, \qquad ||A||^2 = \text{trace}(A^T A), \qquad ||Z||^2 = \text{trace}(Z^\dagger Z)$$

where the complex vector $v = (v_1, \ldots, v_n)$ has conjugate vector $\bar{v} = (\bar{v}_1, \ldots, \bar{v}_n)$ and the complex matrix Z has conjugates $\bar{Z}$ defined by $(\bar{Z})_{ij} = \bar{Z}_{ij}$ and $Z^\dagger = \bar{Z}^T$ is the **Hermitian conjugate**. Again, to be clear, there is not just one choice of norm for $\mathbb{C}^n, \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$. The set paired with the norm is what gives us the structure of a normed space. We conclude this Section with norms which are a bit less obvious.

Example 1.2.2. *Let $C([a,b],\mathbb{R})$ denote the set of continuous real-valued functions with domain $[a,b]$. If $f \in C([a,b],\mathbb{R})$ then we define $\|f\| = max\{|f(x)| \mid x \in [a,b]\}$. It is not too difficult to check this defines a norm on the infinite dimensional vector space $C([a,b],\mathbb{R})$.*

Example 1.2.3. *Suppose V, W are normed linear spaces and $T : V \to W$ is a linear transformation. Then we may define the norm of $\|T\|$ as follows:*

$$\|T\| = sup\{\|T(x)\| \mid x \in V,\ \|x\| = 1\}$$

When V is infinite dimensional there is no reason that $\|T\|$ must be finite. In fact, the linear transformations with finite norm are special. I leave the completion of this thought to your functional analysis course. On the other hand, for finite dimensional V we can argue $\|T\|$ is finite.

Incidentally, given $T : V \to W$ with $\|T\| < \infty$ you can show $\|T(x)\| \leq \|T\|\|x\|$ for all $x \in V$. To see this claim, consider $x \neq 0$ has $\|x\| \neq 0$ hence:

$$\begin{aligned} \|T(x)\| &= \left\| T\left(\frac{\|x\|}{\|x\|}x\right)\right\| \\ &= \left\| \|x\| T\left(\frac{x}{\|x\|}\right)\right\| \\ &= \|x\| \left\| T\left(\frac{x}{\|x\|}\right)\right\| \\ &\leq \|x\|\|T\| \end{aligned} \tag{1.9}$$

as $\|x/\|x\|\| = 1$ so $\|T\|$ certainly provides the claimed bound.

I include the next example to give you a sense of what sort of calculation takes the place of coordinates in infinite dimensions. I'm mostly including these examples so we can appreciate the technical meaning of **continuously differentiable** in our later work.

Example 1.2.4. *Assume $a < b$. Define $T(f) = \int_a^b f(x)\,dx$ for each $f \in C([a,b], \mathbb{R})$. Observe T is a linear transformation. Also,*

$$|T(f)| = \left| \int_a^b f(x)dx \right| \leq \int_a^b |f(x)|dx.$$

Use the max-norm of Example 1.2.2. If $\|f\| = max\{|f(x)| \mid x \in [a,b]\} = 1$ then $|f(x)| \leq 1$ for $a \leq x \leq b$. Thus $|T(f)| \leq \int_a^b dx = b - a$. However, the constant function $f(x) = 1$ has $\|f\| = 1$ and $T(1) = \int_a^b dx = b - a$ thus $\|T\| = sup\{\|T(f)\| \mid f \in C([a,b], \mathbb{R}),\ \|f\| = 1\} = b - a$.

At this point I introduce some notation I found in Zorich. I think it's a useful addition to my standard notations. Pay attention to the semi-colon.

Definition 1.2.5. *multilinear maps*

Let $V_1, V_2, \ldots, V_k, W$ be real vector spaces then $T : V_1 \times V_2 \times \cdots \times V_k \to W$ is a **multilinear map** if T is linear in each of its k-variables while holding the other variables fixed. We write $T \in \mathcal{L}(V_1, V_2, \ldots, V_k; W)$ in this case.

In the case $k = 1$ and $V_1 = V_2 = V$ we say $T \in \mathcal{L}(V, V; W)$ is a W-valued bilinear map on V.

Example 1.2.6. *If $T \in \mathcal{L}(V_1, \ldots, V_k; W)$ where $V_1, \ldots, V_k, W$ are normed linear spaces then define*[2]

$$\|T\| = sup\{\|T(u_1, \ldots, u_k)\| \mid \|u_i\| = 1, i = 1, 2, \ldots, k\}. \tag{1.10}$$

[2]see page 52 of Zorich's *Mathematical Analysis II* for further discussion

Then we can argue, much as we did in Equation 1.9 that

$$\|T(x_1, x_2, \ldots, x_n)\| \leq \|T\|\|x_1\|\|x_2\| \cdots \|x_n\|. \tag{1.11}$$

Notice $det : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R} \in \mathcal{L}(\mathbb{R}^n, \ldots, \mathbb{R}^n; \mathbb{R})$. *Hence, as*

$$\|det\| = sup\{|det[u_1| \cdots |u_n]| \mid \|u_i\| = 1, i = 1, \ldots, n\} \tag{1.12}$$

and

$$|det(x_1, x_2, \ldots, x_n)| \leq \|det\|\|x_1\|\|x_2\| \cdots \|x_n\|. \tag{1.13}$$

But, $det(I) = 1$ *thus* $\|det\| = 1$.

I've probably done a bit more than we need here, I hope it is not too disturbing.

1.2.2 inner product space

There are generalized dot-products on many abstract vector spaces, we call them **inner-products**.

Definition 1.2.7. *Inner product space*

Suppose V is a real vector space. If $\langle\ ,\ \rangle : V \times V \to \mathbb{R}$ is a function such that for all $x, y, z \in V$ and $c \in \mathbb{R}$:

(1.) $\langle x, y\rangle = \langle y, x\rangle$ (symmetric)

(2.) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ (additive in the first slot)

(3.) $\langle cx, y\rangle = c\langle x, y\rangle$ (together with (2.) gives linearity of the first slot)

(4.) $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0$ if and only if $x = 0$.

then we say $(V, \langle\ ,\ \rangle)$ is an **inner-product space** with inner product $\langle\ ,\ \rangle$.

Given an inner-product space $(V, \langle\ ,\ \rangle)$ we can easily induce a norm for V by the formula $||x|| = \sqrt{\langle x, x\rangle}$ for all $x \in V$. Properties (1.), (3.) and (4.) in the definition of the norm are fairly obvious for the induced norm. Let's think throught the triangle inequality for the induced norm:

$$\begin{aligned}
||x + y||^2 &= \langle x + y, x + y\rangle && \text{def. of induced norm} \\
&= \langle x, x + y\rangle + \langle y, x + y\rangle && \text{additive prop. of inner prod.} \\
&= \langle x + y, x\rangle + \langle x + y, y\rangle && \text{symmetric prop. of inner prod.} \\
&= \langle x, x\rangle + \langle y, x\rangle + \langle x, y\rangle + \langle y, y\rangle && \text{additive prop. of inner prod.} \\
&= ||x||^2 + 2\langle x, y\rangle + ||y||^2
\end{aligned}$$

At this point we're stuck. A nontrivial identity[3] called the **Cauchy-Schwarz** identity helps us proceed; $\langle x, y\rangle \leq ||x||||y||$. It follows that $||x + y||^2 \leq ||x||^2 + 2||x||||y|| + ||y||^2 = (||x|| + ||y||)^2$. However, the induced norm is clearly positive so we find $||x + y|| \leq ||x|| + ||y||$.

Most linear algebra texts have a whole chapter on inner-products and their applications, you can look at my notes for a start if you're curious. That said, this is a bit of a digression for this course. Primarily we use the dot-product paired with $\mathbb{R}^n$ in certain applications. I should mention, $\mathbb{R}^n$ with the usual dot-product forms **Euclidean n-space**. We'll say more just before we use the theory of orthogonal complements to understand how to find extreme values on curves or surfaces.

[3] I prove this for the dot-product in my linear notes, however, the proof is written in such a way it equally well applies to a general inner-product

1.2.3 metric as a distance function

Given a set S a distance function describes the **distance** between points in S. This definition is a natural abstraction of our everyday idea of distance.

Definition 1.2.8.

A function $d : S \times S \to \mathbb{R}$ is a **metric** or **distance function** on S if d satisfies the following: for all $x, y, z \in S$,

(1.) $d(x, y) \geq 0$ (non-negativity)

(2.) $d(x, y) = d(y, x)$ (symmetric)

(3.) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

(4.) $d(x, y) = 0$ if and only if $x = y$

then we say (S, d) is a **metric space**.

There are many strange examples one may study, I leave those to your future courses. For our purposes, note any subset of a NLS forms a metric space via the distance function $d(x, y) = \|y - x\|$. Geometrically, the idea is that the distance from the point x to the point y is the length of the displacement vector $y - x$ which goes from x to y. Of course, we could just as well write $d(x, y) = \|x - y\|$ since $\|x - y\| = \|(-1)(y - x)\| = |-1|\|y - x\|$.

Remark 1.2.9. *There is another use of the term* **metric**. *In particular, $g : V \times V \to \mathbb{R}$ is a metric if it is symmetric, bilinear and nondegenerate. Then (V, g) forms a* **geometry**. *We say $T : V \to V$ is an* **isometry** *if $g(T(x), T(y)) = g(x, y)$ for all $x, y \in V$. For example, if $g(x, y) = -x^0y^0 + x^1y^1 + x^2y^2 + x^3y^3$ for $x, y \in \mathbb{R}^4$ then g is the* **Minkowski metric** *and isometries of this metric are called* **Lorentz transformations**. *To avoid confusion, I try to use the term* **scalar product** *rather than metric. An inner product is a scalar product which is positive definite. Riemannian geometry is based on an abstraction of inner products to curved space whereas semi-Riemannian geometry generalizes the Minkowski metric to curved space. The geometry of Eistein's General Relativity is semi-Riemannian geometry.*

1.3 topology and limits in normed linear spaces

The limit we describe here is the natural extension of the $\epsilon - \delta$-limit from elementary calculus. Recall, we say $f(x) \to L$ as $x \to a$ if for each $\epsilon > 0$ there exists $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - L| < \epsilon$. Essentially, this limit is constructed to zoom in on the values taken by f as they get close to a, yet, not $x = a$ itself. Avoidance of the limit point itself allows us to extend the algebra of limits past the confines of unqualified algebra. The same holds for NLS we simply need to replace absolute values with norms. This goes back to the metric space structure involved. For $\mathbb{R}$ we have distance function given by abolute value of the difference $d(x, y) = |x - y|$. For a NLS $(V, \| \ \|)$ we have distance function given by the norm of the difference $d(x, y) = \|x - y\|$. Keeping this analogy in mind it is not hard to see all the definitions in what follows as simple extensions of the analysis we learned in first semester calculus[4]

[4]at Liberty University we still cover elementary $\epsilon - \delta$ proofs in the beginning calculus course

Definition 1.3.1. *open and closed sets in an NLS*

Let $(V, \| \|)$ be a NLS. An **open ball centered at x_o with radius R** is:

$$B_R(x_o) = \{x \in V \mid \|x - x_o\| < R\}.$$

Likewise, a **closed ball centered at x_o with radius R**[5]

$$\overline{B_R(x_o)} = \{x \in V \mid \|x - x_o\| \leq R\}.$$

If $x_o \in U$ and there exists $R > 0$ for which $B_R(x_o) \subseteq U$ then we say x_o is an **interior point** of U. When each point in $U \subseteq V$ is an interior point we say U is an **open set**. If $S \subset V$ has $V - S$ open then we say S is a **closed set**.

In the case $V = \mathbb{R}^n$ with $n = 1, 2, 3$ we have other terms.

(1.) $n = 1$: an open ball is an open interval; $B_r(a) = (a - r, a + r)$,

(2.) $n = 2$: an open ball is an open disk,

(3.) $n = 3$: an open ball is an open ball,

Intuitively, an open set either has no edges, or, has only fuzzy edges whereas a closed set either has no edges, or, has solid edges. The larger problem of studying which sets are open and how that relates to the continuity of functions is known as **topology**. Briefly, a **topology** is a set paired with the set of all sets declared to be open. The topology we study here is **metric topology** as it is derived from a distance function. Moving on,

Definition 1.3.2. *limit points, isolated points and boundary points in an NLS.*

Let $(V, \| \|)$ be a NLS. We define a **deleted open ball centered at x_o with radius R** by:

$$B_R(x_o) - \{x_o\} = \{x \in V \mid 0 < \|x - x_o\| < R\}.$$

We say x_o is a **limit point** of a function f if and only if there exists a deleted open ball which is contained in the $dom(f)$. If $y_o \in dom(f)$ and there exists an open ball centered at y_o which contains no other points in $dom(f)$ then y_o is called an **isolated point** of $dom(f)$. A **boundary point** of $S \subseteq V$ is a point in $x_o \in V$ for which every open ball centered at x_o contains points outside S.

Notice a limit point of f need not be in the domain of f. Also, a boundary point of S need not be in S. Furthermore, if we consider $g : \mathbb{N} \to V$ then each point in $dom(g) = \mathbb{N}$ is isolated.

Definition 1.3.3. *limits and continuity in an NLS.*

If $f : dom(f) \subseteq V \to W$ is a function from normed space $(V, ||\cdot||_V)$ to normed vector space $(W, ||\cdot||_W)$ and x_o is either a limit point or an isolated point of $dom(f)$ and $L \in W$ then we say $\lim_{x \to x_o} f(x) = L$ if and only if for each $\epsilon > 0$ there exists $\delta > 0$ such that if $x \in V$ with $0 < ||x - x_o||_V < \delta$ then $||f(x) - f(x_o)||_W < \epsilon$. If $\lim_{x \to x_o} f(x) = f(x_o)$ then we say that f is a continuous function at x_o.

The definition above indicates functions are by default continuous at isolated points, my apologies if you find this bothersome. Let me give a few examples then we'll turn our attention to proving limit laws for an NLS.

Example 1.3.4. *Suppose V is an NLS and let $c \in \mathbb{R}$ with $c \neq 0$. Also fix $b_o \in V$. Let $F(x) = cx + b_o$ for each $x \in V$. We wish to calculate $\lim_{x\to a} F(x)$. Naturally, we expect the limit is simply $ca + b_o$ hence we work towards proving our intutiion is correct. If $\epsilon > 0$ then choose $\delta = \epsilon/|c|$ and note $0 < \|x - a\| < \delta = \epsilon/|c|$ provides $0 < |c|\|x - a\| < \epsilon$. With this estimate in mind we calculate:*

$$\|F(x) - F(a)\| = \|cx + b_o - (ax + b_o)\| = \|c(x - a)\| = |c|\|x - a\| < \epsilon.$$

Thus $F(x) \to F(a) = ca + b_o$ as $x \to a$. As $a \in V$ was arbitrary we've shown F is continuous on V. Specializing a bit, if we set $c = 1$ and $b_o = 0$ then $F = Id_V$ thus the **identity function** *on V is everywhere continuous.*

Example 1.3.5. *Let V and W be normed linear spaces. Fix $w_o \in W$ and define $F(x) = w_o$ for each $x \in V$. I leave it to the reader to prove $\lim_{x\to a}(F(x)) = w_o$ for any $a \in V$. In other words, a constant function is everywhere continuous in the context of a NLS.*

Example 1.3.6. *Let $F : \mathbb{R}^n - \{a\} \to \mathbb{R}^n$ be defined by $F(x) = \frac{1}{||x-a||}(x - a)$. In this case, certainly a is a limit point of F but geometrically it is clear that $\lim_{x\to a} F(x)$ does not exist. Notice for $n = 1$, the discontinuity of F at a can be understood by seeing that left and right limits exist, but are not equal. On the other hand, $G(x) = \frac{||x-a||}{||x-a||}(x - a)$ clearly has $\lim_{x\to a} G(x) = 0$ and we could classify the discontinuity of G at $x = a$ as removeable. Clearly $\tilde{G}(x) = x - a$ is a continuous extension of G to all of $\mathbb{R}^n$*

On occasion it is helpful to keep the following observation in mind:

Proposition 1.3.7. *norm is continuous with respect to itself.*

Suppose V has norm $||\cdot||$ then $f : V \to \mathbb{R}$ defined by $f(x) = ||x||$ is continuous.

Proof: Suppose $a \in V$ and let $\epsilon > 0$. Choose $\delta = \epsilon$ and consider $x \in V$ such that $0 < ||x - a|| < \delta$. Observe $||x|| = ||x - a + a|| \leq ||x - a|| + ||a|| = \delta + ||a||$ and hence

$$|f(x) - f(a)| = |||x|| - ||a||| < |\delta + ||a|| - ||a||| = |\delta| = \epsilon.$$

Thus $f(x) \to f(a)$ as $x \to a$ and as $a \in V$ was arbitrary the proposition follows $\square$.

It is generally quite challenging to prove limits directly from the definition. Fortunately, there are many useful properties which typically allow us to avoid direct attack.[6] One fun point to make here, if you missed the proof of the so-called *limit laws* in calculus then you can retroactively apply the arguments we soon offer here.

Proposition 1.3.8. *Linearity of the limit on a NLS.*

Let V, W be normed vector spaces. Let a be a limit point of mappings $F, G : U \subseteq V \to W$ and suppose $c \in \mathbb{R}$. If $\lim_{x\to a} F(x) = b_1 \in W$ and $\lim_{x\to a} G(x) = b_2 \in W$ then

(1.) $\lim_{x\to a}(F(x) + G(x)) = \lim_{x\to a} F(x) + \lim_{x\to a} G(x)$.

(2.) $\lim_{x\to a}(cF(x)) = c \lim_{x\to a} F(x)$.

Moreover, if F, G are continuous then $F + G$ and cF are continuous.

[6] of course, some annoying instructor probably will ask you to calculate a couple from the definition so you can learn the definition more deeply

Proof: Let $\epsilon > 0$ and suppose $\lim_{x\to a} f(x) = b_1 \in W$ and $\lim_{x\to a} g(x) = b_2 \in W$. Choose $\delta_1, \delta_2 > 0$ such that $0 < ||x-a|| < \delta_1$ implies $||f(x)-b_1|| < \epsilon/2$ and $0 < ||x-a|| < \delta_2$ implies $||g(x)-b_2|| < \epsilon/2$. Choose $\delta = min(\delta_1, \delta_2)$ and suppose $0 < ||x-a|| < \delta \leq \delta_1, \delta_2$ hence

$$||(f+g)(x) - (b_1+b_2)|| = ||f(x) - b_1 + g(x) - b_2|| \leq ||f(x) - b_1|| + ||g(x) - b_2|| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Item (2.) follows. To prove (2.) note that if $c = 0$ the result is clearly true so suppose $c \neq 0$. Suppose $\epsilon > 0$ and choose $\delta > 0$ such that $||f(x) - b_1|| < \epsilon/|c|$. Note that if $0 < ||x-a|| < \delta$ then

$$||(cf)(x) - cb_1|| = ||c(f(x) - b_1)|| = |c|||f(x) - b_1|| < |c|\epsilon/|c| = \epsilon.$$

The claims about continuity follow immediately from the limit properties. □

Induction easily extends the result above to linear combinations of three or more functions;

$$\lim_{x\to a} \sum_{i=1}^{n} c_i F_i(x) = \sum_{i=1}^{n} c_i \lim_{x\to a} F_i(x). \tag{1.14}$$

We now turn to analyzing limits of a map in terms of the limits of its component functions. First a Lemma which is a slight twist on what we already proved.

Lemma 1.3.9. *Constant vectors pull out of limit.*

Let V be a NLS and suppose $f : dom(f) \subseteq V \to \mathbb{R}$ is a function with $\lim_{x\to a} f(x) = L$. If W is a NLS with $w_o \in W$ then $\lim_{x\to a}(f(x)w_o) = Lw_o$.

Proof: if $w_o = 0$ then the Lemma is clearly true. Hence suppose $w_o \neq 0$ thus $\|w_o\| \neq 0$. Also, we assume $f(x) \to L$ as $x \to a$. Let $\epsilon > 0$ and note we are free to choose $\delta > 0$ such that $0 < \|x-a\| < \delta$ implies $|f(x) - L| < \epsilon/\|w_o\|$. Thus, for $x \in V$ with $0 < \|x-a\| < \delta$

$$\|f(x)w_o - Lw_o\| = \|(f(x) - L)w_o\| = |f(x) - L|\|w_o\| < \frac{\epsilon}{\|w_o\|}\|w_o\| = \epsilon. \tag{1.15}$$

Consequently $\lim_{x\to a}(f(x)w_o) = \lim_{x\to a}(f(x))w_o$. □

We soon need this Lemma to pull basis vectors out of a limit in the proof of Theorem 1.3.11.

Definition 1.3.10. *Component functions of map with values in an NLS.*

Suppose V, W are normed linear spaces and $dim(W) = m$. If $F : dom(F) \subseteq V \to W$ and $\gamma = \{w_1, w_2, \ldots, w_m\}$ is a basis for W and there exist $F_i : dom(F) \subseteq V \to \mathbb{R}$ for $i = 1, 2, \ldots, m$ such that $F = F_1 w_1 + \cdots + F_m w_m$ and we call $F_1, \ldots, F_m$ the **component functions** of F with respect to the γ basis.

Given the limits of each component function we may assemble the limit of the function. Notice, this is a comment about breaking up the limit in the range of the map. In contrast, there is no easy way to break a multivariate limit into one-dimensional limits in the domain, hopefully you saw examples in multivariable calculus which illustrate this subtle point. Only in one dimension to we have the luxury of reducing a full limit to a pair of path limits. See this question and answer, beware wolfram alpha not so good here, Maple wins and this master list of advice on how to calculate multivariate limits that arise in calculus III. There are many examples linked there if you need to see evidence of my claim here.

Theorem 1.3.11.

Suppose V, W are NLSs where W has basis $\gamma = \{w_1, \ldots, w_n\}$ and $F : dom(F) \subseteq V \to W$ has component functions $F_i : dom(F) \subseteq V \to \mathbb{R}$ for $i = 1, \ldots, m$. If $\lim_{x\to a} F_i(x) = L_i \in \mathbb{R}$ for $i = 1, \ldots, m$ then $\lim_{x\to a} F(x) = \sum_{i=1}^{m} L_i w_i$.

Proof: assume F and its components are as described in the Proposition,

$$\begin{aligned}
\lim_{x\to a} F(x) &= \lim_{x\to a} \left(\sum_{i=1}^{m} F_i(x) w_i \right) && \text{: defn. of component functions} \\
&= \sum_{i=1}^{m} \lim_{x\to a} \left(F_i(x) w_i \right) && \text{: additivity of the limit} \\
&= \sum_{i=1}^{m} \left(\lim_{x\to a} F_i(x) \right) w_i && \text{: applying Lemma 1.3.9} \\
&= \sum_{i=1}^{m} L_i w_i.
\end{aligned} \tag{1.16}$$

Therefore, the limit of a map may be assembled from the limits of its component functions. □

It turns out the converse of this Theorem is also true, but, I need to prepare some preliminary ideas to give the proof in the desired generality. Basically, the trouble is that at one point in my proof I need the magnitude of a component to a vector $x = x_1 v_1 + \cdots + x_n v_n$ to be smaller than the norm of the whole vector; $|x_i| \leq \|x\|$. Certainly this is true for orthonormal bases, but, notice $\beta = \{(1, \varepsilon), (1, 0)\}$ is a basis for $\mathbb{R}^2$ which is not orthonormal in the euclidean sense for any $\varepsilon \neq 0$ and:

$$x = (1, \varepsilon) - (1, 0) = (0, \varepsilon) \tag{1.17}$$

hence $\|x\| = |\varepsilon|$ and $[x]_\beta = (1, -1)$ so both components of x in the β basis have magnitude 1. But, we can make $|\varepsilon|$ as small as we like. So, clearly, I cannot just assume for any basis of a NLS we have this property $|x_i| \leq \|x_1 v_1 + \cdots + x_n v_n\|$. It is a special property for certain nice bases. In fact, it is true for most examples we consider. You use it a great deal in study of complex analysis as it says $|Re(z)|, |Im(z)| \leq |z|$. But, we're trying to study abstract NLSs, so we must face the difficulty.

Lemma 1.3.12. *Coordinate change for component functions.*

Suppose $F : dom(F) \subseteq V \to W$ is a map on NLS where $dim(W) = m$ and W has bases $\overline{\gamma} = \{\overline{w}_1, \ldots, \overline{w}_m\}$ and $\gamma = \{w_1, \ldots, w_m\}$. Let $P_{ij} \in \mathbb{R}$ be such that $\overline{w}_i = \sum_{j=1}^{m} P_{ij} w_j$. If $F_1, \ldots, F_m$ are the component functions of F with respect to γ and $\overline{F_1}, \ldots, \overline{F}_m$ are the component functions of F with respect to $\overline{\gamma}$ then $F_j = \sum_{i=1}^{m} \overline{F}_i P_{ij}$ for $j = 1, \ldots, m$.

Proof: Since $\gamma, \overline{\gamma}$ are given bases of W we know there exist $P_{ij} \in \mathbb{R}$ such that $\overline{w}_i = \sum_{j=1}^{m} P_{ij} w_j$. Therefore, we can relate the component expansions in both bases as follows:

$$F = \sum_{i=1}^{m} \overline{F}_i \overline{w}_i = \sum_{j=1}^{m} F_j w_j \quad \Rightarrow \quad \sum_{i=1}^{m} \overline{F}_i \sum_{j=1}^{m} P_{ij} w_j = \sum_{j=1}^{m} F_j w_j \tag{1.18}$$

thus

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{m} \overline{F}_i P_{ij} \right) w_j = \sum_{j=1}^{m} F_j w_j \quad \Rightarrow \quad \sum_{i=1}^{m} \overline{F}_i P_{ij} = F_j \tag{1.19}$$

where we equated coefficients of w_j to obtain the result above. □

It always amuses me to see how the basis and components transform inversely. Continuing to use the notation of the previous Theorem and Lemma,

Proposition 1.3.13.

If $\lim_{x\to a} \overline{F}(x) = \overline{L}_i$ for $i = 1, \ldots, m$ then $\lim_{x\to a} F(x) = \sum_{i,j=1}^{m} P_{ij}\overline{L}_i$.

Proof: use Lemma 1.3.12 to see $F_i(x) = \sum_{j=1}^{m} P_{ij}\overline{F}_j(x)$. Then, by linearity of the limit,

$$\lim_{x\to a}(F_i(x)) = \sum_{j=1}^{m} P_{ij} \lim_{x\to a}\left(\overline{F}_j(x)\right) = \sum_{j=1}^{m} P_{ij}\overline{L}_j. \tag{1.20}$$

The Proposition follows by application of Theorem 1.3.11. □

The coordinate change results above are most interesting when paired with an additional freedom to analyze limits in finite dimensional vector spaces.

(1.) The metric topology for a finite dimensional normed linear space is independent of our choice of norm[7]. For example, in $\mathbb{R}^2$, if we find a point is interior with respect the euclidean norm then it's easy to see the point is also interior w.r.t. the taxicab or sup norm. I might assign a homework which helps you prove this claim.

(2.) Given normed linear spaces V, W and a function $F : dom(F) \subseteq V \to W$, we find F is continuous if and only if the inverse image under F of each open set in W is open in V.[8]

(3.) Since different choices of norm provide the same open sets it follows that the calculation of a limit in a finite dimensional NLS is in fact independent of the choice of norm.

Given any basis for finite dimensional real vector space we can construct an inner product by essentially mimicking the dot-product.

Lemma 1.3.14. *existence of inner product which makes given basis orthonormal.*

If $(V, \| \ \|)$ is a normed linear space with basis $\beta = \{v_1, \ldots, v_n\}$ then $\langle v_i, v_j\rangle = \delta_{ij}$ extended bilinearly serves to define an inner product for V where β is an orthonormal basis. Furthermore, if $\|x\|_2 = \sqrt{\langle x, x\rangle}$ and $x = x_1v_1 + \cdots + x_nv_n$ then

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

hence $|x_i| \leq \|x\|_2$ for any $x \in V$ and for each $i = 1, 2, \ldots, n$.

Proof: left to reader, essentially the claim is immediate once we show $\langle x, y\rangle = x_1y_1 + \cdots + x_ny_n$ where x_i, y_i are the coordinates of x, y with respect to β basis. □

[7] see this question and the answers for some interesting discussion of this point

[8] Notice, we insist that $\emptyset$ is open, my apologies if my earlier wording was insufficiently clear on this point.

Theorem 1.3.15.

Let V, W be normed vector spaces and suppose W has basis $\beta = \{w_j\}_{j=1}^m$. Let $a \in V$ then

$$\lim_{x\to a} F(x) = B = \sum_{j=1}^{m} B_j w_j \qquad \Leftrightarrow \qquad \lim_{x\to a} F_j(x) = B_j \ \text{ for all } j = 1, 2, \ldots m.$$

Proof: Suppose $\lim_{x\to a} F(x) = B \in W$. Construct the inner product $\langle , \rangle : W \times W \to \mathbb{R}$ which forces orthonormality of $\beta = \{w_1, \ldots, w_m\}$. That is, let $\langle w_i, w_j \rangle = \delta_{ij}$ and extend bilinearly. Let $\|y\| = \sqrt{\langle y, y \rangle}$ hence $y = y_1 w_1 + \cdots + y_m w_m$ has $\|y\| = \sqrt{y_1^2 + \cdots + y_n^2}$ thus $\|w_i\| = 1$ and $|y_i| \leq \|y\|$ for each $y \in W$ and $i = 1, \ldots, m$. Therefore,

$$|F_i(x) - B_i| = |F_i(x) - B_i|\|w_i\| = \|(F_i(x) - B_i)w_i\| \leq \|F(x) - B\|. \tag{1.21}$$

Hence, for $\epsilon > 0$ choose $\delta > 0$ such that $0 < \|x - a\| < \delta$ implies $\|F(x) - B\| < \epsilon$. Hence, by Inequality 1.21 we find $0 < \|x - a\| < \delta$ implies $|F_i(x) - B_i| < \epsilon$ for each $i = 1, 2, \ldots, m$. Thus $\lim_{x\to a} F_j(x) = B_j$ for each $j = 1, \ldots, m$ and this remains true when a different norm is given to W (here I use the result that the limit calculated in a finite dimensional NLS is independent of our choice of norm since all norms produce the same topology).

The converse direction follows from Theorem 1.3.11, but I include argument below since it's good to see. Conversely, suppose $\lim_{x\to a} F_j(x) = B_j$ as $x \to a$ for all $j \in \mathbb{N}_n$. Let $\epsilon > 0$ and choose $\delta_j > 0$ such that $0 < ||x - a|| < \delta_j$ implies $||F_j(x) - B_j|| < \frac{\epsilon}{||w_j||m}$. We are free to choose such δ_j by the given limits as clearly $\frac{\epsilon}{||w_j||m} > 0$ for each j. Choose $\delta = min(\delta_j \mid j \in \mathbb{N}_m\}$ and suppose $x \in V$ such that $0 < ||x - a|| < \delta$. Using properties $||x + y|| \leq ||x|| + ||y||$ and $||cx|| = |c|||x||$ multiple times yield:

$$||F(x) - B|| = ||\sum_{j=1}^{m}(F_j(x) - B_j)w_j|| \leq \sum_{j=1}^{m} |F_j(x) - B_j|||w_j|| < \sum_{j=1}^{m} \frac{\epsilon}{||w_j||m}||w_j|| = \sum_{j=1}^{m} \frac{\epsilon}{m} = \epsilon.$$

Therefore, $\lim_{x\to a} F(x) = B$ and this completes the proof $\square$.

Our next goal is to explain why polynomials in coordinates of an NLS are continuous. Many examples fall into this general category so it's worth the effort. The first result we need is the observation that we are free to pull limits out of continuous functions on an NLS:

Proposition 1.3.16. *Limit of composite functions.*

Suppose V_1, V_2, V_3 are normed vector spaces with norms $||\cdot||_1, ||\cdot||_2, ||\cdot||_3$ respective. Let $f : dom(f) \subseteq V_2 \to V_3$ and $g : dom(g) \subseteq V_1 \to V_2$ be mappings. Suppose that $\lim_{x\to x_o} g(x) = y_o$ and suppose that f is continuous at y_o then

$$\lim_{x\to x_o} (f \circ g)(x) = f\left(\lim_{x\to x_o} g(x)\right).$$

Proof: Let $\epsilon > 0$ and choose $\beta > 0$ such that $0 < ||y - b||_2 < \beta$ implies $||f(y) - f(y_o)||_3 < \epsilon$. We can choose such a β since Since f is continuous at y_o thus it follows that $\lim_{y\to y_o} f(y) = f(y_o)$. Next choose $\delta > 0$ such that $0 < ||x - x_o||_1 < \delta$ implies $||g(x) - y_o||_2 < \beta$. We can choose such a δ because we are given that $\lim_{x\to x_o} g(x) = y_o$. Suppose $0 < ||x - x_o||_1 < \delta$ and let $y = g(x)$

note $||g(x) - y_o||_2 < \beta$ yields $||y - y_o||_2 < \beta$ and consequently $||f(y) - f(y_o)||_3 < \epsilon$. Therefore, $0 < ||x - x_o||_1 < \delta$ implies $||f(g(x)) - f(y_o)||_3 < \epsilon$. It follows that $\lim_{x\to x_o}(f(g(x)) = f(\lim_{x\to x_o} g(x))$. □

The following functions are suprisingly useful as we seek to describe continuity of functions.

Definition 1.3.17.

The **sum** and **product** are functions from $\mathbb{R}^2$ to $\mathbb{R}$ defined by

$$s(x, y) = x + y \qquad p(x, y) = xy$$

Proposition 1.3.18.

The sum and product functions are continuous.

Proof: I leave to the reader. □

The proof that the product is continuous is not entirely trivial, but, once you have it, so many things follow:

Proposition 1.3.19.

Let V be an NLS. If $f : dom(f) \subseteq V \to \mathbb{R}$ and $g : dom(g) \subseteq V \to \mathbb{R}$ and $\lim_{x\to a} f(x), \lim_{x\to a} g(x) \in \mathbb{R}$ then $\lim_{x\to a}(f(x) \cdot g(x)) = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x)$.

Proof: Combining Propositions 1.3.19 and 1.3.16 we find

$$\begin{aligned} \lim_{x\to a}(f(x) \cdot g(x)) &= \lim_{x\to a}(p(f(x), g(x))) \\ &= p\left(\lim_{x\to a}(f(x), g(x))\right) \\ &= p\left(\lim_{x\to a} f(x), \lim_{x\to a} g(x)\right) \\ &= \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x). \quad \square \end{aligned} \tag{1.22}$$

Of course, we can continue to products of three or more factors by iterating the product:

$$fgh = p(fg, h) = p(p(f, g), h) \tag{1.23}$$

and by an argument much like that given in Equation 1.22 we can argue that the product of three continous real-valued functions on a subset of a NLS V is once more continuous. It should be clear we can extend by induction this result to any product of finitely many real-valued continuous functions.

Lemma 1.3.20.

Let V be a NLS with basis $\{v_1, \ldots, v_n\}$. Define coordinate function $x_i : V \to \mathbb{R}$ as follows: given $a = a_1v_1 + \cdots + a_nv_n$ set $x_i(a) = a_i$. Then $\Phi_\beta = (x_1, x_2, \ldots, x_n)$ and each coordinate function is continuous on V.

Proof: if $a = a_1v_1 + \cdots + a_nv_n$ then $\Phi_\beta(a) = (a_1, \ldots, a_n) = (x_1(a), \ldots, x_n(a))$ therefore $\Phi_\beta = (x_1, \ldots, x_n)$. I leave the proof that $x_i : V \to \mathbb{R}$ is continuous for each $i = 1, \ldots, m$ as a likely homework for the reader. □

Definition 1.3.21.

Let $x_1, \ldots, x_n$ be coordinate functions with respect to basis β for a NLS V then a function $f : V \to \mathbb{R}$ such that for constants $c_0, c_i, c_{ij}, \ldots, c_{i_1,\ldots,i_k} \in \mathbb{R}$,

$$f(x) = c_0 + \sum_{i=1}^{n} c_i x_i + \sum_{i,j=1}^{n} c_{ij} x_i x_j + \cdots + \sum_{i_1,\ldots,i_k} c_{i_1,\ldots,i_k} x_{i_1} \cdots x_{i_k}$$

is a k-th order multinomial in $x_1, \ldots, x_n$. We say $f(x) \in \mathbb{R}[x_1, \ldots, x_n]$.

The following Theorem is a clear consequence of the results we've thus far discussed in this Section:

Theorem 1.3.22.

Multinomials in the coordinates of a NLS V form continuous real-valued functions on V.

Example 1.3.23. *Define* $det : \mathbb{R}^{n\times n} \to \mathbb{R}$ *by*

$$det(A) = \sum_{i_1,\ldots,i_n=1}^{n} \epsilon_{i_1,\ldots,i_n} A_{1i_1} \cdots A_{ni_n}$$

hence $det(A) \in \mathbb{R}[A_{ij} \mid 1 \leq i, j \leq n]$ *is an* n*-th order multinomial in the coordinates* A_{ij} *with respect to the standard matrix basis for* $\mathbb{R}^{n\times n}$*. Thus the determinant is a continuous real-valued function of matrices.*

I'll let you explain why the complex-valued determinant function on $\mathbb{C}^{n\times n}$ is also continuous. Let's enjoy the application of these results:

Example 1.3.24. *The general linear group* $GL(n, \mathbb{R}) = \{A \in \mathbb{R}^{n\times n} \mid det(A) \neq 0\}$ *is an* **open** *subset of* $\mathbb{R}^{n\times n}$*. To see this notice that* $GL(n, \mathbb{R}) = det^{-1}((-\infty, 0) \cup (0, \infty))$*. But, the determinant is continuous and the inverse image of open sets is open. Clearly* $(-\infty, 0) \cup (0, \infty)$ *is open since each point is interior.*

To be picky, I have not shown the inverse image of open sets is open for a continuous map on an NLS, but, I will likely assign that as a homework, so, don't worry, you'll get a chance to ponder it.

Example 1.3.25. *Let* $T : \mathbb{R}^n \to \mathbb{R}^m$ *be a linear transformation. Then* $T(x)$ *has component functions which are formed from first order multinomials in* $x_1, \ldots, x_n$*. Thus* T *is continuous on* $\mathbb{R}^n$*. It is likely I'll ask you to prove* T *is continuous by direct application of the definition of the limit. It's a good problem to work through.*

The squeeze theorem relies heavily on the order properties of $\mathbb{R}$. Generally a normed vector space has no natural ordering. For example, is $1 > i$ or is $1 < i$ in $\mathbb{C}$? That said, we can state a squeeze theorem for real-valued functions whose domain reside in a normed vector space. This is a generalization of what we learned in calculus I. That said, the proof offered below is very similar to the typical proof which is not given in calculus I[9]

[9] this is lifted word for word from my calculus I notes, however here the meaning of open ball is considerably more general and the linearity of the limit which is referenced is the one proven earlier in this section

Theorem 1.3.26. *Squeeze Theorem.*

Suppose $f : dom(f) \subseteq V \to \mathbb{R}$, $g : dom(g) \subseteq V \to \mathbb{R}$, $h : dom(h) \subseteq V \to \mathbb{R}$ where V is a normed vector space with norm $||\cdot||$. Let $f(x) \leq g(x) \leq h(x)$ for all x on some $\delta > 0$ ball of $a \in V$ and suppose the limits of $f(x), g(x), h(x)$ all exist at limit point a then

$$\lim_{x\to a} f(x) \leq \lim_{x\to a} g(x) \leq \lim_{x\to a} h(x).$$

Furthermore, if the limits of $f(x)$ and $h(x)$ exist with $\lim_{x\to a} f(x) = \lim_{x\to a} h(x) = L \in \mathbb{R}$ then the limit of $g(x)$ likewise exists and $\lim_{x\to a} g(x) = L$.

Proof: Suppose $f(x) \leq g(x)$ for all[10] $x \in B_{\delta_1}(a)_o$ for some $\delta_1 > 0$ and also suppose $\lim_{x\to a} f(x) = L_f \in \mathbb{R}$ and $\lim_{x\to a} g(x) = L_g \in \mathbb{R}$. We wish to prove that $L_f \leq L_g$. Suppose otherwise towards a contradiction. That is, suppose $L_f > L_g$. Note that $\lim_{x\to a}[g(x) - f(x)] = L_g - L_f$ by the linearity of the limit. It follows that for $\epsilon = \frac{1}{2}(L_f - L_g) > 0$ there exists $\delta_2 > 0$ such that $x \in B_{\delta_2}(a)_o$ implies $|g(x) - f(x) - (L_g - L_f)| < \epsilon = \frac{1}{2}(L_f - L_g)$. Expanding this inequality we have

$$-\frac{1}{2}(L_f - L_g) < g(x) - f(x) - (L_g - L_f) < \frac{1}{2}(L_f - L_g)$$

adding $L_g - L_f$ yields,

$$-\frac{3}{2}(L_f - L_g) < g(x) - f(x) < -\frac{1}{2}(L_f - L_g) < 0.$$

Thus, $f(x) > g(x)$ for all $x \in B_{\delta_2}(a)_o$. But, $f(x) \leq g(x)$ for all $x \in B_{\delta_1}(a)_o$ so we find a contradiction for each $x \in B_\delta(a)$ where $\delta = min(\delta_1, \delta_2)$. Hence $L_f \leq L_g$. The same proof can be applied to g and h thus the first part of the theorem follows.

Next, we suppose that $\lim_{x\to a} f(x) = \lim_{x\to a} h(x) = L \in \mathbb{R}$ and $f(x) \leq g(x) \leq h(x)$ for all $x \in B_{\delta_1}(a)$ for some $\delta_1 > 0$. We seek to show that $\lim_{x\to a} f(x) = L$. Let $\epsilon > 0$ and choose $\delta_2 > 0$ such that $|f(x) - L| < \epsilon$ and $|h(x) - L| < \epsilon$ for all $x \in B_\delta(a)_o$. We are free to choose such a $\delta_2 > 0$ because the limits of f and h are given at $x = a$. Choose $\delta = min(\delta_1, \delta_2)$ and note that if $x \in B_\delta(a)_o$ then

$$f(x) \leq g(x) \leq h(x)$$

hence,

$$f(x) - L \leq g(x) - L \leq h(x) - L$$

but $|f(x) - L| < \epsilon$ and $|h(x) - L| < \epsilon$ imply $-\epsilon < f(x) - L$ and $h(x) - L < \epsilon$ thus

$$-\epsilon < f(x) - L \leq g(x) - L \leq h(x) - L < \epsilon.$$

Therefore, for each $\epsilon > 0$ there exists $\delta > 0$ such that $x \in B_\delta(a)_o$ implies $|g(x) - L| < \epsilon$ so $\lim_{x\to a} g(x) = L$. □

Our typical use of the theorem above applies to equations of norms from a normed vector space. The norm takes us from V to $\mathbb{R}$ so the theorem above is essential to analyze interesting limits. We shall make use of it in future analysis.

[10] I use the notation $B_{\delta_1}(a)_o$ to denote the deleted open ball of radius δ_1 centered at a; $B_{\delta_1}(a)_o = B_{\delta_1}(a) - \{a\}$.

1.4 sequential analysis

Let $(V, ||\cdot||_V)$ be a normed vector space, a function from $\mathbb{N}$ to V is a called a **sequence**. Limits of sequences play an important role in analysis in normed linear spaces. The real analysis course makes great use of sequences to tackle questions which are more difficult with only $\epsilon - \delta$ arguments. In fact, we can reformulate limits in terms of sequences and subsequences. Perhaps one interesting feature of abstract topological spaces is the appearance of spaces in which sequential convergence is insufficient to capture the concept of limits. In general, one needs *nets* and *filters*. I digress. More important to our context, the criteria of **completeness**. Let us settle a few definitions to make the words meaningful.

Definition 1.4.1.

Suppose $\{a_n\}$ is a sequence then we say $\lim_{n\to\infty} a_n = L \in V$ iff for each $\epsilon > 0$ there exists $M \in \mathbb{N}$ such that $||a_n - L||_V < \epsilon$ for all $n \in \mathbb{N}$ with $n > M$. If $\lim_{n\to\infty} a_n = L \in V$ then we say $\{a_n\}$ is a **convergent sequence**.

We spent some effort attempting to understand the definition above and its application to the problem of infinite summations in calculus II. It is less likely you have thought much about the following:

Definition 1.4.2.

We say $\{a_n\}$ is a **Cauchy sequence** iff for each $\epsilon > 0$ there exists $M \in \mathbb{N}$ such that $||a_m - a_n||_V < \epsilon$ for all $m, n \in \mathbb{N}$ with $m, n > M$.

In other words, a sequence is Cauchy if the terms in the sequence get arbitarily close as we go sufficiently far out in the list. Many concepts we cover in calculus II are made clear with proofs built around the concept of a Cauchy sequence. The interesting thing about Cauchy is that for some spaces of numbers we can have a sequence which converges but is not Cauchy. For example, if you think about the rational numbers $\mathbb{Q}$ we can construct a sequence of truncated decimal expansions of π:

$$\{a_n\} = \{3, 3.1, 3.14, 3.141, 3.1415 \dots\}$$

note that $a_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$ and yet the $a_n \to \pi \notin \mathbb{Q}$. When spaces are missing their limit points they are in some sense incomplete.

Definition 1.4.3.

If every Cauchy sequence in a metric space converges to a point within the space then we say the metric space is **complete**. If a normed vector space V is complete then we say V is a **Banach space**.

A metric space need not be a vector space. In fact, we can take any open set of a normed vector space and construct a metric space. Metric spaces require less structure.

Fortunately all the main examples of this course are built on the real numbers which are complete, this induces completeness for $\mathbb{C}, \mathbb{R}^n$ and $\mathbb{R}^{m\times n}$. The proof that $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ are Banach spaces follow from arguments similar to those given in the example below.

Example 1.4.4. Claim: $\mathbb{R}$ *complete implies* $\mathbb{R}^2$ *is complete.*

Proof: *suppose* (x_n, y_n) *is a Cauchy sequence in* $\mathbb{R}^2$*. Therefore, for each* $\epsilon > 0$ *there exists* $N \in \mathbb{N}$ *such that* $m, n \in \mathbb{N}$ *with* $N < m < n$ *implies* $||(x_m, y_m) - (x_n, y_n)|| < \epsilon$*. Consider that:*

$$||(x_m, y_m) - (x_n, y_n)|| = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}$$

Therefore, as $|x_m - x_n| = \sqrt{(x_m - x_n)^2}$*, it is clear that:*

$$|x_m - x_n| \leq ||(x_m, y_m) - (x_n, y_n)||$$

But, this proves that $\{x_n\}$ *is a Cauchy sequence of real numbers since for each* $\epsilon > 0$ *we can choose* $N > 0$ *such that* $N < m < n$ *implies* $|x_m - x_n| < \epsilon$*. The same holds true for the sequence* $\{y_n\}$*. By completeness of* $\mathbb{R}$ *we have* $x_n \to x$ *and* $y_n \to y$ *as* $n \to \infty$*. We propose that* $(x_n, y_n) \to (x, y)$*. Let* $\epsilon > 0$ *once more and choose* $N_x > 0$ *such that* $n > N_x$ *implies* $|x_n - x| < \epsilon/2$ *and* $N_y > 0$ *such that* $n > N_y$ *implies* $|y_n - y| < \epsilon/2$*. Let* $N = max(N_x, N_y)$ *and suppose* $n > N$*:*

$$||(x_n, y_n) - (x, y)|| = ||x_n - x, 0) + (0, y_n - y)|| \leq |x_n - x| + |y_n - y| < \epsilon/2 + \epsilon/2 = \epsilon.$$

The key point here is that components of a Cauchy sequence form Cauchy sequences in $\mathbb{R}$. That will also be true for sets of matrices and complex numbers.

Finally, I close with an appliction to the matrix exponential. We define:

$$e^A = I + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}A^k. \tag{1.24}$$

for such $A \in \mathbb{R}^{n \times n}$ as the series above converges. Convergence of a series of matrices is measured by the convergence of the sequence of partial sums. For e^A the n-th partial sum is simply:

$$S_n = \sum_{k=0}^{n-1} \frac{1}{k!}A^k = I + A + \cdots + \frac{1}{(n-1)!}A^{n-1} \tag{1.25}$$

Thus, assuming $m > n$,

$$S_m - S_n = \sum_{k=n}^{m-1} \frac{1}{k!}A^k = \frac{1}{n!}A^n + \cdots + \frac{1}{(m-1)!}A^{m-1} \tag{1.26}$$

The identity $\|AB\| \leq \|A\|\|B\|$ inductively extends to $\|A^k\| \leq \|A\|^k$ for $k \in \mathbb{N}$. With this identity and the triangle inequality we find:

$$\|S_m - S_n\| \leq \sum_{k=n}^{m-1} \frac{1}{k!}\|A\|^k = s_m - s_n \tag{1.27}$$

where $s_n = \sum_{k=0}^{n-1} \frac{1}{k!}\|a\|^k$ is the n-th partial sum of $\sum_{k=0}^{\infty} e^{\|A\|}$. Note s_n is convergence sequence in $\mathbb{R}$ hence it is Cauchy so as $m, n \to \infty$ we find $s_m - s_n \to 0$ and so by the squeeze theorem for sequences we deduce $\|S_m - S_n\| \to 0$ as $m, n \to \infty$. In other words, S_n forms a Cauchy sequence of matrices and thus by the completeness of $\mathbb{R}^{n \times n}$ we deduce the series defining the matrix exponential converges. Notice this argument holds for any matrix A.

I'm fond of the argument above, it was shown to me in some course I took with R.O Fulp, maybe a few courses. There is another argument from linear algebra which uses the real Jordan form. Since $A = P^{-1}JP$ for some $P \in GL(n, \mathbb{R})$ and e^J is easily calculated we obtain existence of e^A from the fact that $e^J = e^{PAP^{-1}} = Pe^AP^{-1}$. But, admittedly, it does take a little work to prove the existence of the real Jordan form for any $A \in \mathbb{R}^{n \times n}$. I bet there are many other arguments to show e^A is well-defined. The abstract concept of the exponential is much more useful than you might first expect. The past two summers I learned an exponential on the appropriate algebra solves any constant coefficient ODE, even when the coefficients are taken from algebras with all sorts of weird features.

Chapter 2

differentiation

Our goal in this chapter is to describe differentiation for functions to and from normed linear spaces. It turns out this is actually quite simple given the background of the preceding chapter. The differential at a point is a linear transformation which best approximates the change in a function at a particular point. We can quantify "best" by a limiting process which is naturally defined in view of the fact there is a norm on the spaces we consider.

The most important example is of course the case $f : \mathbb{R}^n \to \mathbb{R}^m$. In this context it is natural to write the differential as a matrix multiplication. The matrix of the differential is known as the Jacobian matrix. Partial derivatives are also defined in terms of directional derivatives. The directional derivative is sometimes defined where the differential fails to exist. We will discuss how the criteria of continuous differentiability allows us to build the differential from the directional derivatives. We study how the general concept of Frechet differentiation recovers all the derivatives you've seen previously in calculus and much more.

The general theory of differentiation is a bit of an adjustment from our previous experience differentiating. Dieudonne said it best: this is the introduction to his chapter on differentiation in *Modern Analysis* Chapter VIII.

> The subject matter of this Chapter is nothing else but the elementary theorems of Calculus, which however are presented in a way which will probably be new to most students. That presentation, which throughout adheres strictly to our general "geometric" outlook on Analysis, aims at keeping as close as possible to the fundamental idea of Calculus, namely the "local" approximation of functions by *linear* functions. **In the classical teaching of Calculus, the idea is immediately obscured by the accidental fact that, on a one-dimensional vector space, there is a one-to-one correspondence between *linear* forms and numbers, and therefore the derivative at a point is defined as a *number* instead of a *linear form*. This slavish subservience to the shibboleth[1] of numerical interpretation at any cost becomes much worse when dealing with functions of several variables...**

Dieudonne's then spends the next half page continuing this thought with explicit examples of how this custom of our calculus presentation injures the conceptual generalization. If you want to see

[1] from wikipedia: is a word, sound, or custom that a person unfamiliar with its significance may not pronounce or perform correctly relative to those who are familiar with it. It is used to identify foreigners or those who do not belong to a particular class or group of people. It also refers to features of language, and particularly to a word or phrase whose pronunciation identifies a speaker as belonging to a particular group.

differentiation written for mathematicians, that is the place to look. He proves many results for infinite dimensions because, well, why not?

In this chapter I define the Frechet differential and exhibit a number of abstract examples. Then we turn to proving the basic properties of the Frechet derivative including linearity and the chain rule. My proof of the chain rule has a bit of a gap, but, I hope the argument gives you some intuition as to why we should expect a chain rule. Next we explore partial derivatives in an NLS with respect to a given abstract basis. After that we focus on $\mathbb{R}^n$. Many many examples of Jacobians are given. We study a few perverse examples which fail to be continuously differentiable. We show continuous differentiability implies differentiability by a standard, but interesting, argument. I prove a quite general product rule, discuss the problem of higher derivatives in the abstract (I punt details to Zorich for now, sorry Fall 2017). Finally, I share some insights I've recently come to to understand about $\mathcal{A}$-Calculus. In particular, I discuss some of the rudiments of differentiating with respect to algebra variables.

2.1 the Frechet differential

The definition[2] below says that $\triangle F = F(a+h) - F(a) \cong dF_a(h)$ when h is close to zero.

Definition 2.1.1.

Let $(V, ||\cdot||_V)$ and $(W, ||\cdot||_W)$ be normed vector spaces. Suppose that U is open and $F : U \subseteq V \to W$ is a function the we say that F is **differentiable** at $a \in U$ iff there exists a linear mapping $L : V \to W$ such that

$$\lim_{h\to 0}\left[\frac{F(a+h) - F(a) - L(h)}{||h||_V} \right] = 0.$$

In such a case we call the linear mapping L the **differential at** a and we denote $L = dF_a$. In the case $V = \mathbb{R}^m$ and $W = \mathbb{R}^n$ the matrix of the differential is called the **derivative of** F **at** a or the **Jacobian matrix of** F **at** a and we denote $[dF_a] = F'(a) \in \mathbb{R}^{m\times n}$ which means that $dF_a(v) = F'(a)v$ for all $v \in \mathbb{R}^n$.

Notice this definition gives an equation which implicitly defines dF_a. For the moment the only way we have to calculate dF_a is educated guessing. We simply use brute-force calculation to suggest a guess for L which forces the Frechet quotient to vanish. In the next section we'll discover a systematic calculational method for functions on euclidean spaces. The purpose of this section is to understand the definition of the differential and to connect it to basic calculus. I'll begin with basic calculus as you probably are itching to understand where your beloved difference quotient has gone:

[2] Some authors might put a norm in the numerator of the quotient. That is an equivalent condition since a function $g : V \to W$ has $\lim_{h\to 0} g(h) = 0$ iff $\lim_{h\to 0} ||g(h)||_W = 0$

Example 2.1.2. *Suppose $f : dom(f) \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at x. It follows that there exists a linear function $df_x : \mathbb{R} \to \mathbb{R}$ such that*[3]

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

Note that

$$\lim_{h\to 0} \frac{f(x+h) - f(x) - df_x(h)}{|h|} = 0 \quad \Leftrightarrow \quad \lim_{h\to 0^{\pm}} \frac{f(x+h) - f(x) - df_x(h)}{|h|} = 0.$$

In the left limit $h \to 0^-$ we have $h < 0$ hence $|h| = -h$. On the other hand, in the right limit $h \to 0^+$ we have $h > 0$ hence $|h| = h$. Thus, differentiability suggests that $\lim_{h\to 0^{\pm}} \frac{f(x+h)-f(x)-df_x(h)}{\pm h} = 0$. But we can pull the minus out of the left limit to obtain $\lim_{h\to 0^-} \frac{f(x+h)-f(x)-df_x(h)}{h} = 0$. Therefore, after an algebra step, we find:

$$\lim_{h\to 0}\left[\frac{f(x+h) - f(x)}{h} - \frac{df_x(h)}{h}\right] = 0.$$

Linearity of $df_x : \mathbb{R} \to \mathbb{R}$ implies there exists $m \in \mathbb{R}^{1\times 1} = \mathbb{R}$ such that $df_x(h) = mh$. Observe that

$$\lim_{h\to 0} \frac{df_x(h)}{h} = \lim_{h\to 0} \frac{mh}{h} = m.$$

It is a simple exercise to show that if $\lim(A - B) = 0$ and $\lim(B)$ exists then $\lim(A)$ exists and $\lim(A) = \lim(B)$. Identify $A = \frac{f(x+h)-f(x)}{h}$ and $B = \frac{df_x(h)}{h}$. Therefore,

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

Consequently, we find the 1×1 matrix m of the differential is precisely $f'(x)$ as we defined it via a difference quotient in first semester calculus. In summary, we find $\boxed{df_x(h) = f'(x)h}$. In other words, if a function is differentiable in the sense we defined at the beginning of this chapter then it is differentiable in the terminology we used in calculus I. Moreover, the derivative at x is precisely the matrix of the differential.

Remark 2.1.3.

Incidentally, I should mention that df_x is the differential of f at the point x. The differential of f would be the mapping $x \mapsto df_x$. Technically, the differential df is a function from $\mathbb{R}$ to the set of linear transformations on $\mathbb{R}$. You can contrast this view with that of first semester calculus. There we say the mapping $x \mapsto f'(x)$ defines the derivative f' as a function from $\mathbb{R}$ to $\mathbb{R}$. This simplification in perspective is only possible because calculus in one-dimension is so special. More on this later. This distinction is especially important to understand if you begin to look at questions of higher derivatives.

Example 2.1.4. *Suppose $T : V \to W$ is a linear transformation of normed vector spaces V and W. I propose $L = T$. In other words, I think we can show the best linear approximation to the change in a linear function is simply the function itself. Clearly L is linear since T is linear. Consider the difference quotient:*

$$\frac{T(a+h) - T(a) - L(h)}{||h||_V} = \frac{T(a) + T(h) - T(a) - T(h)}{||h||_V} = \frac{0}{||h||_V}.$$

Note $h \neq 0$ implies $||h||_V \neq 0$ by the definition of the norm. Hence the limit of the difference quotient vanishes since it is identically zero for every nonzero value of h. We conclude that $dT_a = T$.

[3] unless we state otherwise, $\mathbb{R}^n$ is assumed to have the euclidean norm, in this case $||x||_{\mathbb{R}} = \sqrt{x^2} = |x|$.

Example 2.1.5. *Let $T : V \to W$ where V and W are normed vector spaces and define $T(v) = w_o$ for all $v \in V$. I claim the differential is the zero transformation. Linearity of $L(v) = 0$ is trivially verified. Consider the difference quotient:*

$$\frac{T(a+h) - T(a) - L(h)}{||h||_V} = \frac{w_o - w_o - 0}{||h||_V} = \frac{0}{||h||_V}.$$

Using the arguments to the preceding example, we find $dT_a = 0$.

Typically the difference quotient is not identically zero. The pair of examples above are very special cases. Our next example requires a bit more thought:

Example 2.1.6. *Suppose $F : \mathbb{R}^2 \to \mathbb{R}^3$ is defined by $F(x, y) = (xy,\ x^2,\ x + 3y)$ for all $(x, y) \in \mathbb{R}^2$. Consider the difference function $\triangle F$ at (x, y):*

$$\triangle F = F((x, y) + (h, k)) - F(x, y) = F(x + h, y + k) - F(x, y)$$

Calculate,

$$\triangle F = \big((x + h)(y + k),\ (x + h)^2,\ x + h + 3(y + k)\big) - \big(xy,\ x^2,\ x + 3y\big)$$

Simplify by cancelling terms which cancel with $F(x, y)$:

$$\triangle F = \big(xk + hy + hk,\ 2xh + h^2,\ h + 3k)\big)$$

Identify the linear part of $\triangle F$ as a good candidate for the differential. I claim that:

$$L(h, k) = \big(xk + hy,\ 2xh,\ h + 3k\big).$$

is the differential for f at (x,y). Observe first that we can write

$$L(h, k) = \begin{bmatrix} y & x \\ 2x & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix}.$$

therefore $L : \mathbb{R}^2 \to \mathbb{R}^3$ is manifestly linear. Use the algebra above to simplify the difference quotient below:

$$\lim_{(h,k)\to(0,0)} \left[\frac{\triangle F - L(h, k)}{||(h, k)||} \right] = \lim_{(h,k)\to(0,0)} \left[\frac{(hk, h^2, 0)}{||(h, k)||} \right]$$

Note $||(h, k)|| = \sqrt{h^2 + k^2}$ therefore we fact the task of showing that $\frac{1}{\sqrt{h^2+k^2}}(hk,\ h^2,\ 0) \to (0, 0, 0)$ as $(h, k) \to (0, 0)$. Notice that: $||(hk,\ h^2,\ 0)|| = |h|\sqrt{h^2 + k^2}$. Therefore, as $(h, k) \to 0$ we find

$$\left|\left| \frac{1}{\sqrt{h^2 + k^2}}(hk,\ h^2,\ 0) \right|\right| = |h| \to 0.$$

However, if $||v|| \to 0$ it follows $v \to 0$ so we derive the desired limit. Therefore,

$$df_{(x,y)}(h, k) = \begin{bmatrix} y & x \\ 2x & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix}.$$

Computation of less trivial multivariate limits is an art we'd like to avoid if possible. It turns out that we can actually avoid these calculations by computing partial derivatives. However, we still need a certain multivariate limit to exist for the partial derivative functions so in some sense it's unavoidable. The limits are there whether we like to calculate them or not.

Definition 2.1.7. *Linearization of a differentiable map.*

Let $(V, ||\cdot||_V)$ and $(W, ||\cdot||_W)$ be normed vector spaces and suppose $F : dom(F) \subseteq V \to W$ is differentiable at p then the **linearization of F at p** is given by $L_F^p(x) = F(p) + dF_p(x-p)$ for all $x \in V$. We also say $L_F^p : V \to W$ is the **affinization** of F at p.

Perhaps the term **linearization** is a holdover from the terminology *linear function* of the form $f(x) = mx + b$. Of course, this is an offense to the student of pure linear algebra. Unless $b = 0$ such a map is not technically **linear**. What is it? It's an affine function. So, I added the terminology **affinization of F** to the definition above. However, I must admit, I don't think that terminology is standard. Much can be said about affine maps of normed linear spaces, I probably fail to paint the big picture of affine maps in these notes. Maybe I should make it homework...

Example 2.1.8. *Suppose $F : \mathbb{R}^2 \to \mathbb{R}^3$ is defined by $F(x,y) = (xy,\ x^2,\ x+3y)$ for all $(x,y) \in \mathbb{R}^2$ then calculate the linearization of f at $(1,-2)$. Following Example 2.1.6 we find*

$$df_{(x,y)}(h,k) = \begin{bmatrix} y & x \\ 2x & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} \quad \Rightarrow \quad df_{(1,-2)}(h,k) = \begin{bmatrix} -2 & 1 \\ 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix}.$$

The linearization of f at $(1,-2)$ is constructed as follows:

$$\begin{aligned} L_f^{(1,-2)}(x,y) &= f(1,-2) + df_{(1,-2)}(x-1, y+2) \\ &= (-2,1,-5) + \begin{bmatrix} -2 & 1 \\ 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x-1 \\ y+2 \end{bmatrix} \\ &= (-2-2(x-1)+(y+2), 1+2(x-1), -5+(x-1)+3(y+2)) \\ &= (-2x+y+2, 2x-1, x+3y). \end{aligned} \tag{2.1}$$

Calculation of the differential simplifies considerably when the domain is one-dimensional. We already worked out the case of $f : \mathbb{R} \to \mathbb{R}$ in Example 2.1.2 and the following pair of examples work out the concrete case of $F : \mathbb{R} \to \mathbb{C}$ and then the general case $F : \mathbb{R} \to V$ for an arbitrary finite dimensional normed linear space V.

Example 2.1.9. *Suppose $F(t) = U(t) + iV(t)$ for all $t \in dom(f)$ and both U and V are differentiable functions on $dom(F)$. By the arguments given in Example 2.1.2 it suffices to find $L : \mathbb{R} \to \mathbb{C}$ such that*

$$\lim_{h \to 0} \left[\frac{F(t+h) - F(t) - L(h)}{h} \right] = 0.$$

I propose that on the basis of analogy to Example 2.1.2 we ought to have $dF_t(h) = (U'(t) + iV'(t))h$. Let $L(h) = (U'(t) + iV'(t))h$. Observe that, using properties of $\mathbb{C}$:

$$\begin{aligned} L(h_1 + ch_2) &= (U'(t) + iV'(t))(h_1 + ch_2) \\ &= (U'(t) + iV'(t))h_1 + c(U'(t) + iV'(t))h_2 \\ &= L(h_1) + cL(h_2). \end{aligned}$$

for all $h_1, h_2 \in \mathbb{R}$ and $c \in \mathbb{R}$. Hence $L : \mathbb{R} \to \mathbb{C}$ is linear. Moreover,

$$\begin{aligned} \tfrac{F(t+h)-F(t)-L(h)}{h} &= \tfrac{1}{h}\Big(U(t+h) + iV(t+h) - U(t) + iV(t) - (U'(t) + iV'(t))h \Big) \\ &= \tfrac{1}{h}\Big(U(t+h) - U(t) - U'(t)h \Big) + i\tfrac{1}{h}\Big(V(t+h) - V(t) - V'(t)h \Big) \end{aligned}$$

Consider the problem of calculating $\lim_{h\to 0}\frac{F(t+h)-F(t)-L(h)}{h}$. *We observe that a complex function converges to zero iff the real and imaginary parts of the function separately converge to zero (this is covered by Theorem 1.3.22). By differentiability of* U *and* V *we find again using Example 2.1.2*

$$\lim_{h\to 0}\frac{1}{h}\Big(U(t+h)-U(t)-U'(t)h\Big)=0 \qquad \lim_{h\to 0}\frac{1}{h}\Big(V(t+h)-V(t)-V'(t)h\Big)=0.$$

Therefore, $dF_t(h)=(U'(t)+iV'(t))h$. *Note that the quantity* $U'(t)+iV'(t)$ *is not a real matrix in this case. To write the derivative in terms of a real matrix multiplication we need to construct some further notation which makes use of the isomorphism between* $\mathbb{C}$ *and* $\mathbb{R}^2$. *Actually, it's pretty easy if you agree that* $a+ib=(a,b)$ *then* $dF_t(h)=(U'(t),V'(t))h$ *so the matrix of the differential is* $(U'(t),V'(t))\in\mathbb{R}^{1\times 2}$ *which makes since as* $F:\mathbb{C}\approx\mathbb{R}^2\to\mathbb{R}$.

Example 2.1.10. *Suppose* V *is a normed vector space with basis* $\beta=\{f_1,f_2,\ldots,f_n\}$. *Futhermore, let* $G:I\subseteq\mathbb{R}\to V$ *be defined by*

$$G(t)=\sum_{i=1}^{n}G_i(t)f_i$$

where $G_i:I\to\mathbb{R}$ *is differentiable on* I *for* $i=1,2,\ldots,n$. *Recall Theorem 1.3.22 revealed that* $T=\sum_{j=1}^{n}T_jf_j:\mathbb{R}\to V$ *then* $\lim_{t\to 0}T(t)=\sum_{j=1}^{n}l_jf_j$ *iff* $\lim_{t\to 0}T_j(t)=l_j$ *for all* $j=1,2,\ldots,n$. *In words, the limit of a vector-valued function can be parsed into a vector of limits. With this in mind consider (again we can trade* $|h|$ *for* h *as we explained in-depth in Example 2.1.2) the difference quotient* $\lim_{h\to 0}\left[\frac{G(t+h)-G(t)-h\sum_{i=1}^{n}\frac{dG_i}{dt}f_i}{h}\right]$, *factoring out the basis yields:*

$$\lim_{h\to 0}\left[\frac{\sum_{i=1}^{n}[G_i(t+h)-G_i(t)-h\frac{dG_i}{dt}]f_i}{h}\right]=\sum_{i=1}^{n}\left[\lim_{h\to 0}\frac{G_i(t+h)-G_i(t)-h\frac{dG_i}{dt}}{h}\right]f_i=0$$

where the zero above follows from the supposed differentiability of each component function. It follows that:

$$\boxed{dG_t(h)=h\sum_{i=1}^{n}\frac{dG_i}{dt}f_i}$$

The example above encompasses a number of cases at once:

(1.) $V=\mathbb{R}$, functions on $\mathbb{R}$, $f:\mathbb{R}\to\mathbb{R}$

(2.) $V=\mathbb{R}^n$, space curves in $\mathbb{R}$, $\vec{r}:\mathbb{R}\to\mathbb{R}^n$

(3.) $V=\mathbb{C}$, complex-valued functions of a real variable, $f=u+iv:\mathbb{R}\to\mathbb{C}$

(4.) $V=\mathbb{R}^{m\times n}$, matrix-valued functions of a real variable, $F:\mathbb{R}\to\mathbb{R}^{m\times n}$.

In short, when we differentiate a function which has a real domain then we can define the derivative of such a function by component-wise differentiation. It gets more interesting when the domain has several independent variables as Examples 2.1.6 and 2.1.11 illustrate.

Example 2.1.11. *Suppose $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is defined by $F(A) = A^2$. Notice*

$$\triangle F = F(A+H) - F(A) = (A+H)(A+H) - X^2 = AH + HA + H^2$$

I propose that F is differentiable at A and $L(H) = AH + HA$. Let's check linearity,

$$L(H_1 + cH_2) = A(H_1 + cH_2) + (H_1 + cH_2)A = AH_1 + H_1A + c(AH_2 + H_2A)$$

Hence $L : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is a linear transformation. By construction of L the linear terms in the numerator cancel leaving just the quadratic term,

$$\lim_{H\to 0} \frac{F(A+H) - F(A) - L(H)}{||H||} \;=\; \lim_{H\to 0} \frac{H^2}{||H||}.$$

It suffices to show that $\lim_{H\to 0} \frac{||H^2||}{||H||} = 0$ since $\lim(||g||) = 0$ iff $\lim(g) = 0$ in a normed vector space. Fortunately the normed vector space $\mathbb{R}^{n\times n}$ is actually a **Banach algebra**. *A vector space with a multiplication operation is called an algebra. In the current context the multiplication is simply matrix multiplication. A Banach algebra is a normed vector space with a multiplication that satisfies $||XY|| \leq ||X||\,||Y||$. Thanks to this inequality[4] we can calculate our limit via the squeeze theorem. Observe $0 \leq \frac{||H^2||}{||H||} \leq ||H||$. As $H \to 0$ it follows $||H|| \to 0$ hence $\lim_{H\to 0} \frac{||H^2||}{||H||} = 0$. We find $dF_A(H) = AH + HA$.*

Generally constructing the derivative matrix for a function $f : V \to W$ where $V, W \neq \mathbb{R}$ involves a fair number of relatively ad-hoc conventions because the constructions necessarily involving choosing coordinates. The situation is similar in linear algebra. Writing abstract linear transformations in terms of matrix multiplication takes a little thinking. If you look back you'll notice that I did not bother to try to write a matrix for the differential in Examples 2.1.4 or 2.1.5.

Example 2.1.12. *Find the linearization of $F(X) = X^2$ at $X = I$. In Example 2.1.11 we proved $dF_A(H) = AH + HA$. Hence, for $A = I$ we find $dF_I(H) = IH + HI = 2H$. Thus the linearization is fairly simple to assemble,*

$$\begin{aligned} L_F^I(X) &= F(I) + dF_I(X - I) \\ &= I + 2(X - I) \\ &= 2X - I. \end{aligned} \tag{2.2}$$

[4] it does take a bit of effort to prove this inequality holds for the matrix norm, I omit it since it would be distracting here

2.2 properties of the Frechet derivative

Linearity and the chain rule naturally generalize for Frechet derivatives on normed linear spaces. It is helpful for me to introduce some additional notation to analyze the convergence of the Frechet quotient: supposing that $F : dom(F) \subset V \to W$ is differentiable at a we set:

$$\eta_F(h) = F(a+h) - F(a) - dF_a(h) \tag{2.3}$$

hence the **Frechet quotient** can be written as:

$$\frac{\eta_F(h)}{\|h\|} = \frac{F(a+h) - F(a) - dF_a(h)}{\|h\|}. \tag{2.4}$$

Thus differentiability of F at a requires $\frac{\eta_F(h)}{\|h\|} \to 0$ as $h \to 0$. For $h \neq 0$ and $\|h\| < 1$ we have:

$$0 \leq \|\eta_F(h)\| < \frac{\|\eta_F(h)\|}{\|h\|}. \tag{2.5}$$

Thus $\|\eta_F(h)\| \to 0$ as $h \to 0$ by the squeeze theorem. Consequently,

$$\lim_{h\to 0} \eta_F(h) = 0. \tag{2.6}$$

Therefore, $\eta_F : V \to W$ is continuous at $h = 0$ since $\eta_F(0) = F(a) - F(a) - dF_a(0) = 0$ (I remind the reader that the linear transformation dF_a must map zero to zero). Continuity of η_F at $h = 0$ allows us to use theorems for continuous functions on η_F.

Theorem 2.2.1. *Linearity of the Frechet derivatives.*

Suppose V and W are normed linear spaces. If $F : dom(F) \subseteq V \to W$ and $G : dom(G) \subseteq V \to W$ are differentiable at a and $c \in \mathbb{R}$ then $cF + G$ is differentiable at a and

$$d(cF + G)_a = cdF_a + dG_a$$

Proof: Let $\eta_F(h) = F(a+h) - F(a) - dF_a(h)$ and $\eta_G(h) = G(a+h) - G(a) - dG_a(h)$ for all $h \in V$. Assume F and G differentiable at a hence $\lim_{h\to 0} \frac{\eta_F(h)}{\|h\|} = 0$ and $\lim_{h\to 0} \frac{\eta_G(h)}{\|h\|} = 0$. Moreover, $dF_a, dG_a : V \to W$ are linear hence $cdF_a + dG_a : V \to W$ is linear. Hence calculate,

$$\begin{aligned} \eta_{cF+G}(h) &= (cF+G)(a+h) - (cF+G)(a) - (cdF_a + dG_a)(h) \\ &= c\left(F(a+h) - F(a) - dF_a(h)\right) + G(a+h) - G(a) - dG_a(h) \\ &= c\eta_F(h) + \eta_G(h) \end{aligned} \tag{2.7}$$

Therefore, by Proposition 1.3.8, we complete the proof:

$$\lim_{h\to 0} \frac{\eta_{cF+G}(h)}{\|h\|} = \lim_{h\to 0} \frac{c\eta_F(h) + \eta_G(h)}{\|h\|} = c \lim_{h\to 0} \left(\frac{\eta_F(h)}{\|h\|}\right) + \lim_{h\to 0} \frac{\eta_G(h)}{\|h\|} = 0. \quad \square$$

Setting $c = 1$ or $G = 0$ we obtain important special cases:

$$d(F+G)_a = dF_a + dG_a \qquad \& \qquad d(cF)_a = cdF_a. \tag{2.8}$$

The chain rule is also a general rule of calculus on a NLS[5]. This single chain rule produces all the chain rules you saw in calculus I, II and III and much more. To appreciate this we need to understand partial differentiation for normed linear spaces.

[5] I state the rule with domains of the entire NLS, but, this can easily be stated for smaller domains like $F : U \subseteq V_1 \to V_2$ and $G : dom(G) \subseteq V_2 \to V_3$ where $F(U) \subset dom(G)$ so $F \circ G$ is well-defined, but, this has nothing to do with the theorem so I just made the domains uninteresting

Theorem 2.2.2. *chain rule for Frechet derivatives.*

Suppose $G : V_1 \to V_2$ is differentiable at a and $F : V_2 \to V_3$ is differentiable at $G(a)$ then $F \circ G$ is differentiable at a and $d(F \circ G)_a = dF_{G(a)} \circ dG_a$.

The proof I offer here is not quite complete. The main ideas are here, but, there is a pesky term at the end which I have not quite pinned down to my liking. I found these notes by J. C. M. Grajales on page 40 have a proof which appears complete.

Proof: since G is differentiable at a we have the existence of η_G continuous at $h = 0$ defined by:

$$\eta_G(h) = G(a+h) - G(a) - dG_a(h) \tag{2.9}$$

Also, by differentiability of F at $G(a)$ we have the existence of η_F continuous at $k = 0$ given by:

$$\eta_F(k) = F(G(a)+k) - F(G(a)) - dF_{G(a)}(k) \tag{2.10}$$

Furthermore, the differentials are linear transformations and thus their composite $dF_{G(a)} \circ dG_a$ is likewise linear. It remains to show $\eta_{F\circ G}$ formed with $dF_{G(a)} \circ dG_a$ has the needed limiting property. Thus consider,

$$\begin{aligned}
\eta_{F\circ G}(h) &= (F\circ G)(a+h) - (F\circ G)(a) - (dF_{G(a)} \circ dG_a)(h) \\
&= F(G(a+h)) - F(G(a)) - dF_{G(a)}(dG_a(h)) \\
&= F\left(G(a) + dG_a(h) + \eta_G(h)\right) - F(G(a)) - dF_{G(a)}(dG_a(h)) \\
&= F(G(a)) + dF_{G(a)}(dG_a(h) + \eta_G(h)) + \eta_F(dG_a(h) + \eta_G(h)) \\
&\quad - F(G(a)) - dF_{G(a)}(dG_a(h)) \\
&= dF_{G(a)}(\eta_G(h)) + \eta_F(dG_a(h) + \eta_G(h))
\end{aligned} \tag{2.11}$$

where I used Equation 2.10 to make the expansion marked in blue. I need a bit of notation to help guide the remainder of the proof:

$$\frac{\eta_{F\circ G}(h)}{\|h\|} = \underbrace{\frac{1}{\|h\|} dF_{G(a)}(\eta_G(h))}_{(I.)} + \underbrace{\frac{1}{\|h\|}\eta_F(dG_a(h) + \eta_G(h))}_{(II.)} \tag{2.12}$$

We can understand (I.) using linearity and continuity of the linear map $dF_{G(a)}$:

$$\lim_{h\to 0}\left(\frac{1}{\|h\|} dF_{G(a)}(\eta_G(h))\right) = \lim_{h\to 0} dF_{G(a)}\left(\frac{\eta_G(h)}{\|h\|}\right) = dF_{G(a)}\left(\lim_{h\to 0}\frac{\eta_G(h)}{\|h\|}\right) = dF_{G(a)}(0) = 0. \tag{2.13}$$

To understand (II.) a substitution is helpful. Notice $dG_a(h) + \eta_G(h) \to 0$ as $h \to 0$. Let $k = dG_a(h) + \eta_G(h)$ and note $\frac{\eta_F(k)}{\|k\|} \to 0$ as $k \to 0$. Unfortunately, (II.) is not quite $\frac{\eta_F(k)}{\|k\|}$ since it has a denominator $\|h\|$ not $\|k\|$. We need to find a relation which binds $\|h\|$ and $\|k\|$. In particular, if we can find $m > 0$ for which $\|k\| < m\|h\|$ then

$$0 < \frac{\|\eta_F(k)\|}{m\|h\|} < \frac{\|\eta_F(k)\|}{\|k\|} \tag{2.14}$$

and we could argue (II.) vanishes as $h \to 0$ by the squeeze theorem. I leave this gap as an exercise for the reader. □

Remark 2.2.3.

Other authors use the big and little O notation to help with the analysis of the proof above. It may be that if I adopted such notation is would help me will in the gap. For now I stick with my somewhat unusual η_F notation.

2.3 partial derivatives of differentiable maps

In the preceding sections we calculated the differential at a point via educated guessing. We should like to find better method to derive differentials. It turns out that we can systematically calculate the differential from partial derivatives of the component functions. However, certain topological conditions are required for us to properly paste together the partial derivatives of the component functions. We discuss the perils of constructing proving differentiability from partial derivatives in Section 2.4. The purpose of the current section is to define partial differentiation and to explain how partial derivatives relate to the differential of a given differentiable map. To understand partial derivatives we begin with a study of directional derivatives. Once more we generalize the usual calculus III.

Remark 2.3.1.

Certainly parts of what is done in this section naturally generalize to an infinite dimensional context. You can read more about the Gateaux derivative in your future studies. However, here I limit our attention in this section to finite dimensional normed linear spaces.

2.3.1 partial differentiation in a finite dimensional real vector space

The directional derivative of a mapping F at a point $a \in dom(F)$ along v is defined to be the derivative of the curve $\gamma(t) = F(a + tv)$. In other words, the directional derivative gives you the instantaneous vector-rate of change in the mapping F at the point a along v.

Definition 2.3.2.

Suppose V and W are real normed linear spaces. Let $F : dom(F) \subseteq V \to W$ and suppose the limit below exists for $a \in dom(F)$ and $v \in V$ then we define the **directional derivative of F at a along v** to be $D_vF(a) \in W$ where

$$D_vF(a) = \lim_{h\to 0} \frac{F(a + hv) - F(a)}{h}$$

One great contrast we should pause to note is that the definition of the directional derivative is explicit whereas the definition of the differential was implicit. Naturally, if we take $V = W = \mathbb{R}$ then we recover the first semester difference quotient definition of the derivative at a point. This also reproduces the directional derivatives you were shown in multivariate calculus, except, we do not insist v have $\|v\| = 1$. Don't be fooled by the proof of the next Theorem, it's easier than it looks. Summary: since differentiability at a point controls the change of the map in all directions at a point in terms of the differential we can control the change in the map in a particular direction at the given point via the differential.

Theorem 2.3.3. *Differentiability implies directional differentiability.*

Let V, W be real normed linear spaces. If $F : U \subseteq V \to W$ is differentiable at $a \in U$ then the directional derivative $D_vF(a)$ exists for each $v \in V$ and $D_vF(a) = dF_a(v)$.

Proof: Suppose $a \in U$ such that dF_a is well-defined then we are given that

$$\lim_{h\to 0} \frac{F(a + h) - F(a) - dF_a(h)}{||h||} = 0.$$

This is a limit in V, when it exists it follows that the limits that approach the origin along particular paths also exist and are zero. Consider the path $t \mapsto tv$ for $v \neq 0$ and $t > 0$, we find

$$\lim_{tv\to 0,\ t>0} \frac{F(a+tv) - F(a) - dF_a(tv)}{||tv||} = \frac{1}{||v||} \lim_{t\to 0^+} \frac{F(a+tv) - F(a) - tdF_a(v)}{|t|} = 0.$$

Hence, as $|t| = t$ for $t > 0$ we find

$$\lim_{t\to 0^+} \frac{F(a+tv) - F(a)}{t} = \lim_{t\to 0} \frac{tdF_a(v)}{t} = dF_a(v).$$

Likewise we can consider the path $t \mapsto tv$ for $v \neq 0$ and $t < 0$

$$\lim_{tv\to 0,\ t<0} \frac{F(a+tv) - F(a) - dF_a(tv)}{||tv||} = \frac{1}{||v||} \lim_{t\to 0^-} \frac{F(a+tv) - F(a) - tdF_a(v)}{|t|} = 0.$$

Note $|t| = -t$ thus the limit above yields

$$\lim_{t\to 0^-} \frac{F(a+tv) - F(a)}{-t} = \lim_{t\to 0^-} \frac{tdF_a(v)}{-t} \quad \Rightarrow \quad \lim_{t\to 0^-} \frac{F(a+tv) - F(a)}{t} = dF_a(v).$$

Therefore,

$$\lim_{t\to 0} \frac{F(a+tv) - F(a)}{t} = dF_a(v)$$

and we conclude that $D_vF(a) = dF_a(v)$ for all $v \in V$ since the $v = 0$ case follows trivially. □

Partial derivatives are just directional derivatives in standard directions. In particular, given a basis $\beta = \{v_1, \ldots, v_n\}$ with coordinate maps $x_1, \ldots, x_n$ there is a standard concept of partial differentiation on an NLS:

Definition 2.3.4. *partial derivative with respect to coordinate on an NLS.*

Let V be a NLS with basis $\beta = \{v_1, \ldots, v_n\}$ and coordinates $\Phi_\beta = (x_1, \ldots, x_n)$. Then if $F : dom(F) \subseteq V \to W$ we define, for such points $a \in dom(F)$ as the limit exists,

$$\frac{\partial F}{\partial x_i}(a) = D_{v_i}F(a) = \lim_{h\to 0} \frac{F(a+hv_i) - F(a)}{h}.$$

Alternatively, we can present the partial derivative in terms of an ordinary derivative:

$$\frac{\partial F}{\partial x_i}(a) = \frac{d}{dt}\left[F(a+tv_i)\right]\Big|_{t=0} \tag{2.15}$$

Let's revisit the map from Example 2.1.11 and see if we can recover the differential in terms of partial derivatives.

Example 2.3.5. *Let $F(X) = X^2$ for each $X \in \mathbb{R}^{n\times n}$. Let X_{ij} be the usual coordinates with respect to the standard matrix basis $\{E_{ij}\}$. Calculate the partial derivative of F with respect to X_{ij} at A: using Equation 2.15 with v_i replaced with E_{ij} and a with A,*

$$\begin{aligned}
\frac{\partial F}{\partial X_{ij}}(A) &= \frac{d}{dt}\left[(A+tE_{ij})^2\right]\Big|_{t=0} \\
&= \frac{d}{dt}\left[A^2 + tAE_{ij} + tE_{ij}A + t^2E_{ij}^2\right]\Big|_{t=0} \\
&= (AE_{ij} + E_{ij}A + 2tE_{ij}^2)\big|_{t=0} \\
&= AE_{ij} + E_{ij}A.
\end{aligned} \tag{2.16}$$

If we know a map of normed linear spaces is differentiable then we can express the differential in terms of partial derivatives.

Theorem 2.3.6. *differentials can be built from partial derivatives.*

Let V, W be real normed linear spaces where V has basis $\beta = \{v_1, \ldots, v_n\}$ with coordinates $x_1, \ldots, x_n$. If $F : dom(F) \subseteq V \to W$ is differentiable at a and $h = \sum_{i=1}^n h_i v_i$ then

$$dF_a(h) = \sum_{i=1}^n h_i \frac{\partial F}{\partial x_i}(a).$$

Proof: observe that

$$dF_a\left(\sum_{i=1}^n h_i v_i\right) = \sum_{i=1}^n h_i dF_a(v_i) = \sum_{i=1}^n h_i D_{v_i} F(a) = \sum_{i=1}^n h_i \frac{\partial F}{\partial x_i}(a). \tag{2.17}$$

follows immediately from linearity of differential paired with Theorem 2.3.3. □

Let's apply the above result to Example 2.3.5.

Example 2.3.7. *Consider $F(X) = X^2$ for $X \in \mathbb{R}^{n\times n}$. Construct the differential from the partial derivatives with respect to the standard basis matrix $\{E_{ij}\}$. Let $H = \sum_{i,j} H_{ij} E_{ij}$ and calculate using Equation 2.16*

$$dF_A(H) = \sum_{i,j} H_{ij}(AE_{ij} + E_{ij}A) = A\left(\sum_{i,j} H_{ij}E_{ij}\right) + \left(\sum_{i,j} H_{ij}E_{ij}\right)A = AH + HA.$$

I should emphasize, at this point in our development, we cannot conclude the differential exists merely from partial derivatives existing[6]. The example above is reasonable because we have already shown differentiability of the $F(A) = A^2$ map in Example 2.1.11.

Remark 2.3.8.

I have deliberately defined the derivative in slightly more generality than we need for this course. It's probably not much trouble to continue to develop the theory of differentiation for a normed vector space, however I will for the most part stop here modulo an example here or there. If you understand many of the theorems that follow from here on out for $\mathbb{R}^n$ then it is a simple matter to transfer arguments to the setting of a Banach space by using an appropriate isomorphism. Traditionally this type of course only covers continuous differentiability, inverse and implicit function theorems in the context of mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$.

For the reader interested in generalizing these results to the context of an abstract normed vector space feel free to discuss it with me sometime. Also, if you want to read a master on these topics you could look at the text by Shlomo Sternberg on Advanced Calculus. He develops many things for normed spaces. Or, take a look at Dieudonne's Modern Analysis which pays special attention to reaping infinite dimensional results from our finite-dimensional arguments. I also find Zorich's two volume set on Mathematical Analysis is quite helpful. I'm hoping to borrow some arguments from Zorich in this update to my notes. Any of these texts would be good to read to follow-up my course with something deeper.

[6] we study this in depth in Section 2.4.

2.3.2 partial differentiation for real

Consider $F : dom(F) \subseteq \mathbb{R}^n \to \mathbb{R}^n$, in this case the differential dF_a is a linear transformation from $\mathbb{R}^n \to \mathbb{R}^n$ and we can calculate the standard matrix for the differential using the preceding proposition. Recall that if $L : \mathbb{R}^n \to \mathbb{R}^m$ then the standard matrix was simply

$$[L] = [L(e_1)|L(e_2)|\cdots|L(e_n)]$$

and thus the action of L is expressed nicely as a matrix multiplication; $L(v) = [L]v$. Similarly, $dF_a : \mathbb{R}^n \to \mathbb{R}^m$ is linear transformation and thus $dF_a(v) = [dF_a]v$ where

$$[dF_a] = [dF_a(e_1)|dF_a(e_2)|\cdots|dF_a(e_n)].$$

Moreover, by the preceding proposition we can calculate $dF_a(e_j) = D_{e_j}F(a)$ for $j = 1, 2, \ldots, n$. Clearly the directional derivatives in the coordinate directions are of great importance. For this reason we make the following definition:

Definition 2.3.9. *Partial derivatives are directional derivatives in coordinate directions.*

Suppose that $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is a mapping the we say that F is **has partial derivative** $\frac{\partial F}{\partial x_i}(a)$ at $a \in U$ iff the directional derivative in the e_i direction exists at a. In this case we denote,

$$\frac{\partial F}{\partial x_i}(a) = D_{e_i}F(a).$$

Also, the notation $D_{e_i}F(a) = D_iF(a)$ or $\partial_i F = \frac{\partial F}{\partial x_i}$ is convenient. We construct the **partial derivative function** $\partial_i F : V \subseteq \mathbb{R}^n \to \mathbb{R}^m$ as the function defined pointwise for each $v \in V$ where $\partial_i F(v)$ exists.

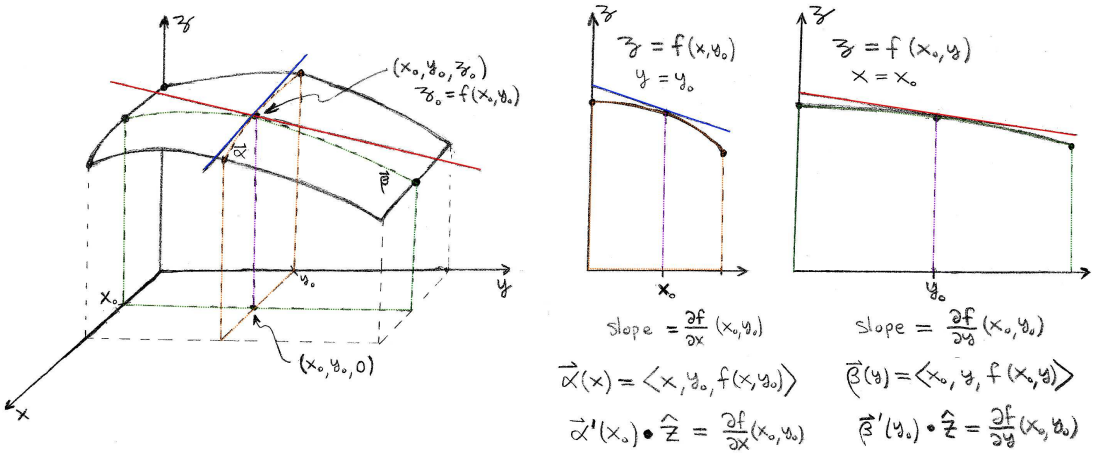
Let's expand this definition a bit. Note that if $F = (F_1, F_2, \ldots, F_m)$ then

$$D_{e_i}F(a) = \lim_{h\to 0} \frac{F(a + he_i) - F(a)}{h} \quad \Rightarrow \quad [D_{e_i}F(a)] \cdot e_j = \lim_{h\to 0} \frac{F_j(a + he_i) - F_j(a)}{h}$$

for each $j = 1, 2, \ldots m$. But then the limit of the component function F_j is precisely the directional derivative at a along e_i hence we find the result

$$\frac{\partial F}{\partial x_i} \cdot e_j = \frac{\partial F_j}{\partial x_i} \qquad \text{in other words,} \qquad \boxed{\partial_i F = (\partial_i F_1, \partial_i F_2, \ldots, \partial_i F_m).}$$

In the particular case $f : \mathbb{R}^2 \to \mathbb{R}$ the partial derivatives with respect to x and y at (x_o, y_o) are related to the graph $z = f(x, y)$ as illustrated below:



Similar pictures can be imagined for partial derivatives of more variables, even for vector-valued maps, but direct visualization is not possible (at least for me).

The proposition below shows how the differential of a m-vector-valued function of n-real variables is connected to a matrix of partial derivatives.

Proposition 2.3.10.

If $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $a \in U$ then the differential dF_a has derivative matrix $F'(a)$ and it has components which are expressed in terms of partial derivatives of the component functions:

$$[dF_a]_{ij} = \partial_j F_i$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Perhaps it is helpful to expand the derivative matrix explicitly for future reference:

$$F'(a) = \begin{bmatrix} \partial_1 F_1(a) & \partial_2 F_1(a) & \cdots & \partial_n F_1(a) \\ \partial_1 F_2(a) & \partial_2 F_2(a) & \cdots & \partial_n F_2(a) \\ \vdots & \vdots & \vdots & \vdots \\ \partial_1 F_m(a) & \partial_2 F_m(a) & \cdots & \partial_n F_m(a) \end{bmatrix}$$

Let's write the operation of the differential for a differentiable mapping at some point $a \in \mathbb{R}$ in terms of the explicit matrix multiplication by $F'(a)$. Let $v = (v_1, v_2, \ldots v_n) \in \mathbb{R}^n$,

$$dF_a(v) = F'(a)v = \begin{bmatrix} \partial_1 F_1(a) & \partial_2 F_1(a) & \cdots & \partial_n F_1(a) \\ \partial_1 F_2(a) & \partial_2 F_2(a) & \cdots & \partial_n F_2(a) \\ \vdots & \vdots & \vdots & \vdots \\ \partial_1 F_m(a) & \partial_2 F_m(a) & \cdots & \partial_n F_m(a) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

You may recall the notation from calculus III at this point, omitting the a-dependence,

$$\nabla F_j = grad(F_j) = \begin{bmatrix} \partial_1 F_j, & \partial_2 F_j, & \cdots, & \partial_n F_j \end{bmatrix}^T$$

So if the derivative exists we can write it in terms of a stack of gradient vectors of the component functions: (I used a transpose to write the stack side-ways),

$$F' = \begin{bmatrix} \nabla F_1 | \nabla F_2 | \cdots | \nabla F_m \end{bmatrix}^T$$

Finally, just to collect everything together,

$$F' = \begin{bmatrix} \partial_1 F_1 & \partial_2 F_1 & \cdots & \partial_n F_1 \\ \partial_1 F_2 & \partial_2 F_2 & \cdots & \partial_n F_2 \\ \vdots & \vdots & \vdots & \vdots \\ \partial_1 F_m & \partial_2 F_m & \cdots & \partial_n F_m \end{bmatrix} = \begin{bmatrix} \partial_1 F & | & \partial_2 F & | \cdots | & \partial_n F \end{bmatrix} = \begin{bmatrix} (\nabla F_1)^T \\ \hline (\nabla F_2)^T \\ \hline \vdots \\ \hline (\nabla F_m)^T \end{bmatrix}$$

Example 2.3.11. *Recall that in Example 2.1.6 we showed that $F : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $F(x, y) = (xy,\ x^2,\ x + 3y)$ for all $(x, y) \in \mathbb{R}^2$ was differentiable. In fact we calculated that*

$$dF_{(x,y)}(h, k) = \begin{bmatrix} y & x \\ 2x & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix}.$$

If you recall from calculus III the mechanics of partial differentiation it's simple to see that

$$\frac{\partial F}{\partial x} = \frac{\partial}{\partial x}(xy,\ x^2,\ x+3y) = (y,\ 2x, 1) = \begin{bmatrix} y \\ 2x \\ 1 \end{bmatrix}$$

$$\frac{\partial F}{\partial y} = \frac{\partial}{\partial y}(xy,\ x^2,\ x+3y) = (x,\ 0, 3) = \begin{bmatrix} x \\ 0 \\ 3 \end{bmatrix}$$

Thus $[dF] = [\partial_x F | \partial_y F]$ *(as we expect given the derivations in this section!)*

Directional derivatives and partial derivatives are of secondary importance in this course. They are merely the substructure of what is truly of interest: the differential. That said, it is useful to know how to construct directional derivatives via partial derivative formulas. In fact, in careless calculus texts it sometimes presented as the definition.

Proposition 2.3.12.

If $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $a \in U$ then the directional derivative $D_v F(a)$ can be expressed as a sum of partial derivative maps for each $v =< v_1, v_2, \ldots, v_n > \in \mathbb{R}^n$:

$$D_v F(a) = \sum_{j=1}^{n} v_j \partial_j F(a)$$

Proof: since F is differentiable at a the differential dF_a exists and $D_v F(a) = dF_a(v)$ for all $v \in \mathbb{R}^n$. Use linearity of the differential to calculate that

$$D_v F(a) = dF_a(v_1 e_1 + \cdots + v_n e_n) = v_1 dF_a(e_1) + \cdots + v_n dF_a(e_n).$$

Note $dF_a(e_j) = D_{e_j} F(a) = \partial_j F(a)$ and the prop. follows. □

Example 2.3.13. *Suppose* $f : \mathbb{R}^3 \to \mathbb{R}$ *then* $\nabla f = [\partial_x f, \partial_y f, \partial_z f]^T$ *and we can write the directional derivative in terms of*

$$D_v f = [\partial_x f, \partial_y f, \partial_z f]^T v = \nabla f \cdot v$$

if we insist that $||v|| = 1$ *then we recover the standard directional derivative we discuss in calculus III. Naturally the* $||\nabla f(a)||$ *yields the maximum value for the directional derivative at* a *if we limit the inputs to vectors of unit-length. If we did not limit the vectors to unit length then the directional derivative at* a *can become arbitrarily large as* $D_v f(a)$ *is proportional to the magnitude of* v*. Since our primary motivation in calculus III was describing rates of change along certain directions for some multivariate function it made sense to specialize the directional derivative to vectors of unit-length. The definition used in these notes better serves the theoretical discussion.*[7]

2.3.3 examples of Jacobian matrices

Our goal here is simply to exhibit the Jacobian matrix and partial derivatives for a few mappings. At the base of all these calculations is the observation that partial differentiation is just ordinary differentiation where we treat all the independent variable not being differentiated as constants. The criteria of independence is important. We'll study the case where variables are not independent in a later section (see implicit differentiation).

[7] If you read my calculus III notes you'll find a derivation of how the directional derivative in Stewart's calculus arises from the general definition of the derivative as a linear mapping. Look up page 305g.

Example 2.3.14. *Let $f(t) = (t, t^2, t^3)$ then $f'(t) = (1, 2t, 3t^2)$. In this case we have*

$$f'(t) = [df_t] = \begin{bmatrix} 1 \\ 2t \\ 3t^2 \end{bmatrix}$$

Example 2.3.15. *Let $f(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$ be a mapping from $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$. I'll denote the coordinates in the domain by $(x_1, x_2, x_3, y_1, y_2, y_3)$ thus $f(\vec{x}, \vec{y}) = x_1y_1 + x_2y_2 + x_3y_3$. Calculate,*

$$[df_{(\vec{x},\vec{y})}] = \nabla f(\vec{x}, \vec{y})^T = [y_1, y_2, y_3, x_1, x_2, x_3]$$

Example 2.3.16. *Let $f(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$ be a mapping from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. I'll denote the coordinates in the domain by $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ thus $f(\vec{x}, \vec{y}) = \sum_{i=1}^n x_iy_i$. Calculate,*

$$\frac{\partial}{x_j}\left[\sum_{i=1}^n x_iy_i\right] = \sum_{i=1}^n \frac{\partial x_i}{x_j} y_i = \sum_{i=1}^n \delta_{ij}y_i = y_j$$

Likewise,

$$\frac{\partial}{y_j}\left[\sum_{i=1}^n x_iy_i\right] = \sum_{i=1}^n x_i \frac{\partial y_i}{y_j} = \sum_{i=1}^n x_i\delta_{ij} = x_j$$

Therefore, noting that $\nabla f = (\partial_{x_1} f, \ldots, \partial_{x_n} f, \partial_{y_1} f, \ldots, \partial_{y_n} f)$,

$$[df_{(\vec{x},\vec{y})}]^T = (\nabla f)(\vec{x}, \vec{y}) = \vec{y} \times \vec{x} = (y_1, \ldots, y_n, x_1, \ldots, x_n)$$

Example 2.3.17. *Suppose $F(x, y, z) = (xyz, y, z)$ we calculate,*

$$\frac{\partial F}{\partial x} = (yz, 0, 0) \qquad \frac{\partial F}{\partial y} = (xz, 1, 0) \qquad \frac{\partial F}{\partial z} = (xy, 0, 1)$$

Remember these are actually column vectors in my sneaky notation; $(v_1, \ldots, v_n) = [v_1, \ldots, v_n]^T$. This means the **derivative** *or* **Jacobian matrix** *of F at (x, y, z) is*

$$F'(x, y, z) = [dF_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Example 2.3.18. *Suppose $F(x, y, z) = (x^2 + z^2, yz)$ we calculate,*

$$\frac{\partial F}{\partial x} = (2x, 0) \qquad \frac{\partial F}{\partial y} = (0, z) \qquad \frac{\partial F}{\partial z} = (2z, y)$$

The derivative is a 2×3 matrix in this example,

$$F'(x, y, z) = [dF_{(x,y,z)}] = \begin{bmatrix} 2x & 0 & 2z \\ 0 & z & y \end{bmatrix}$$

Example 2.3.19. *Suppose $F(x, y) = (x^2 + y^2, xy, x + y)$ we calculate,*

$$\frac{\partial F}{\partial x} = (2x, y, 1) \qquad \frac{\partial F}{\partial y} = (2y, x, 1)$$

The derivative is a 3×2 matrix in this example,

$$F'(x, y) = [dF_{(x,y)}] = \begin{bmatrix} 2x & 2y \\ y & x \\ 1 & 1 \end{bmatrix}$$

Example 2.3.20. *Suppose* $P(x,v,m) = (P_o, P_1) = (\frac{1}{2}mv^2 + \frac{1}{2}kx^2, mv)$ *for some constant* k*. Let's calculate the derivative via gradients this time,*

$$\nabla P_o = (\partial P_o/\partial x, \partial P_o/\partial v, \partial P_o/\partial m) = (kx, mv, \tfrac{1}{2}v^2)$$

$$\nabla P_1 = (\partial P_1/\partial x, \partial P_1/\partial v, \partial P_1/\partial m) = (0, m, v)$$

Therefore,

$$P'(x,v,m) = \begin{bmatrix} kx & mv & \frac{1}{2}v^2 \\ 0 & m & v \end{bmatrix}$$

Example 2.3.21. *Let* $F(r,\theta) = (r\cos\theta, r\sin\theta)$*. We calculate,*

$$\partial_r F = (\cos\theta, \sin\theta) \qquad \text{and} \qquad \partial_\theta F = (-r\sin\theta, r\cos\theta)$$

Hence,

$$F'(r,\theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

Example 2.3.22. *Let* $G(x,y) = (\sqrt{x^2+y^2}, \tan^{-1}(y/x))$*. We calculate,*

$$\partial_x G = \left(\tfrac{x}{\sqrt{x^2+y^2}}, \tfrac{-y}{x^2+y^2}\right) \qquad \text{and} \qquad \partial_y G = \left(\tfrac{y}{\sqrt{x^2+y^2}}, \tfrac{x}{x^2+y^2}\right)$$

Hence,

$$G'(x,y) = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} = \begin{bmatrix} \frac{x}{r} & \frac{y}{r} \\ \frac{-y}{r^2} & \frac{x}{r^2} \end{bmatrix} \qquad (\text{ using } r = \sqrt{x^2+y^2} \text{ })$$

Example 2.3.23. *Let* $F(x,y) = (x, y, \sqrt{R^2 - x^2 - y^2})$ *for a constant* R*. We calculate,*

$$\nabla\sqrt{R^2 - x^2 - y^2} = \left(\tfrac{-x}{\sqrt{R^2-x^2-y^2}}, \tfrac{-y}{\sqrt{R^2-x^2-y^2}} \right)$$

Also, $\nabla x = (1,0)$ *and* $\nabla y = (0,1)$ *thus*

$$F'(x,y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{\sqrt{R^2-x^2-y^2}} & \frac{-y}{\sqrt{R^2-x^2-y^2}} \end{bmatrix}$$

Example 2.3.24. *Let* $F(x,y,z) = (x, y, z, \sqrt{R^2 - x^2 - y^2 - z^2})$ *for a constant* R*. We calculate,*

$$\nabla\sqrt{R^2 - x^2 - y^2 - z^2} = \left(\tfrac{-x}{\sqrt{R^2-x^2-y^2-z^2}}, \tfrac{-y}{\sqrt{R^2-x^2-y^2-z^2}}, \tfrac{-z}{\sqrt{R^2-x^2-y^2-z^2}} \right)$$

Also, $\nabla x = (1,0,0)$*,* $\nabla y = (0,1,0)$ *and* $\nabla z = (0,0,1)$ *thus*

$$F'(x,y,z) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{-x}{\sqrt{R^2-x^2-y^2-z^2}} & \frac{-y}{\sqrt{R^2-x^2-y^2-z^2}} & \frac{-z}{\sqrt{R^2-x^2-y^2-z^2}} \end{bmatrix}$$

Example 2.3.25. *Let $f(x,y,z) = (x+y, y+z, x+z, xyz)$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ yz & xz & xy \end{bmatrix}$$

Example 2.3.26. *Let $f(x,y,z) = xyz$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \end{bmatrix}$$

Example 2.3.27. *Let $f(x,y,z) = (xyz, 1-x-y)$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \\ -1 & -1 & 0 \end{bmatrix}$$

Example 2.3.28. *Let $f : \mathbb{R}^3 \times \mathbb{R}^3$ be defined by $f(x) = x \times v$ for a fixed vector $v \neq 0$. We denote $x = (x_1, x_2, x_3)$ and calculate,*

$$\frac{\partial}{\partial x_a}(x \times v) = \frac{\partial}{\partial x_a}\Big(\sum_{i,j,k} \epsilon_{ijk} x_i v_j e_k \Big) = \sum_{i,j,k} \epsilon_{ijk} \frac{\partial x_i}{\partial x_a} v_j e_k = \sum_{i,j,k} \epsilon_{ijk} \delta_{ia} v_j e_k = \sum_{j,k} \epsilon_{ajk} v_j e_k$$

It follows,

$$\frac{\partial}{\partial x_1}(x \times v) = \sum_{j,k} \epsilon_{1jk} v_j e_k = v_2 e_3 - v_3 e_2 = (0, -v_3, v_2)$$

$$\frac{\partial}{\partial x_2}(x \times v) = \sum_{j,k} \epsilon_{2jk} v_j e_k = v_3 e_1 - v_1 e_3 = (v_3, 0, -v_1)$$

$$\frac{\partial}{\partial x_3}(x \times v) = \sum_{j,k} \epsilon_{3jk} v_j e_k = v_1 e_2 - v_2 e_1 = (-v_2, v_1, 0)$$

Thus the Jacobian is simply,

$$[df_{(x,y)}] = \begin{bmatrix} 0 & v_3 & -v_2 \\ -v_3 & 0 & -v_1 \\ v_2 & v_1 & 0 \end{bmatrix}$$

In fact, $df_p(h) = f(h) = h \times v$ for each $p \in \mathbb{R}^3$. The given mapping is linear so the differential of the mapping is precisely the mapping itself (we could short-cut much of this calculation and simply quote Example 2.1.4 where we proved $dT = T$ for linear T).

Example 2.3.29. *Let $f(x,y) = (x, y, 1-x-y)$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}$$

Example 2.3.30. *Let $X(u,v) = (x,y,z)$ where x, y, z denote functions of u, v and I prefer to omit the explicit depedendence to reduce clutter in the equations to follow.*

$$\frac{\partial X}{\partial u} = X_u = (x_u, y_u, z_u) \quad \textit{and} \quad \frac{\partial X}{\partial v} = X_v = (x_v, y_v, z_v)$$

Then the Jacobian is the 3×2 matrix

$$\big[dX_{(u,v)}\big] = \begin{bmatrix} x_u & x_v \\ y_u & y_v \\ z_u & z_v \end{bmatrix}$$

Remark 2.3.31.

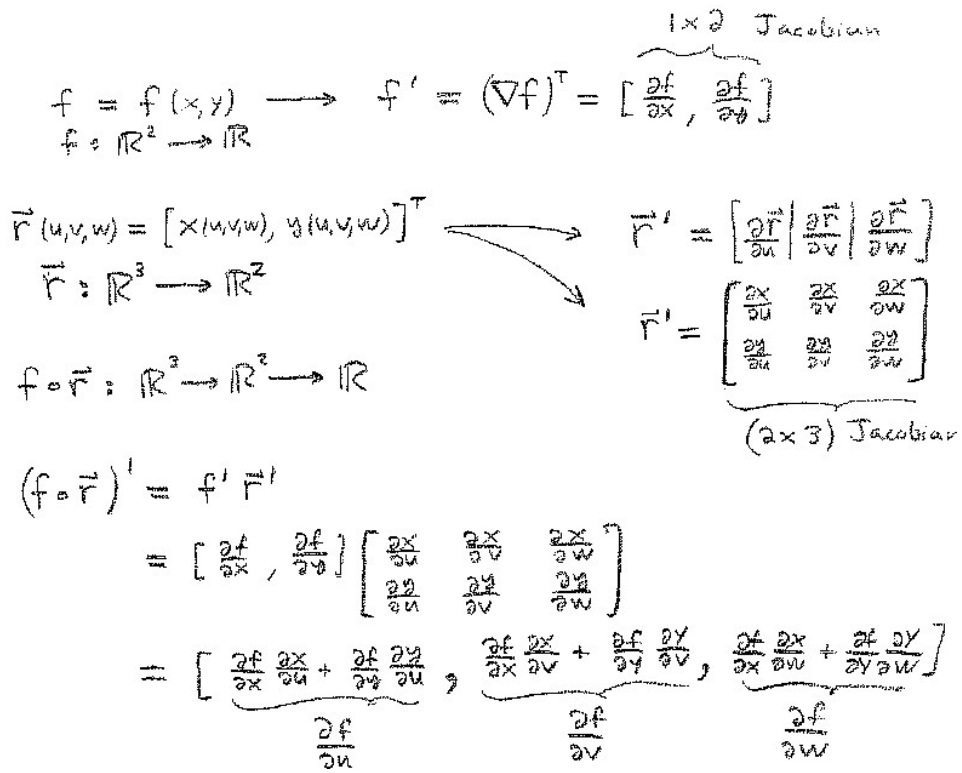
I return to these examples in the next chapter and we'll explore the geometric content of these formulas as they support the application of certain theorems. More on that later, for the remainder of this chapter we continue to focus on properties of differentiation.

2.3.4 on chain rule and Jacobian matrix multiplication

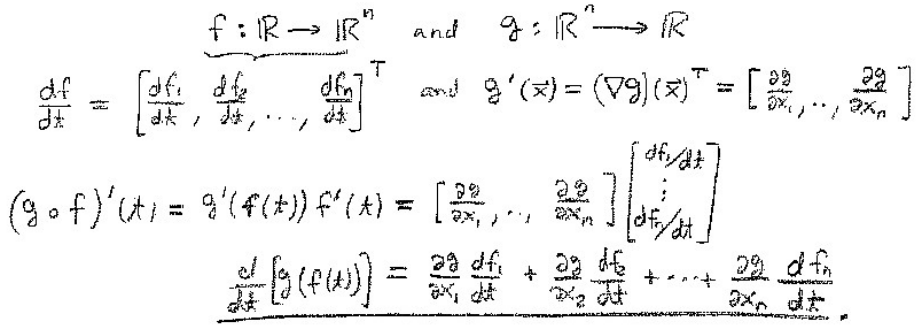
In calculus III you may have learned how to calculate partial derivatives in terms of tree-diagrams and intermediate variable etc... We now have a way of understanding those rules and all the other chain rules in terms of one over-arching calculation: matrix multiplication of the constituent Jacobians in the composite function. Of course once we have this rule for the composite of two functions we can generalize to n-functions by a simple induction argument. For example, for three suitably defined mappings F, G, H,

$$(F \circ G \circ H)'(a) = F'(G(H(a)))G'(H(a))H'(a)$$

Example 2.3.32. .



Example 2.3.33. .



Example 2.3.34. .

Let $f(x,y) = x^2y^2$ and $g(t) = (t, t^2)$

We have $f: \mathbb{R}^2 \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}^2$ note,

$$f'(x,y) = [2xy^2, 2x^2y] \quad \text{and} \quad g'(t) = \begin{bmatrix} 1 \\ 2t \end{bmatrix}$$

Note $f \circ g : \mathbb{R} \to \mathbb{R}^2 \to \mathbb{R}$ has

$$(f \circ g)'(t) = f'(g(t))g'(t) = f'(t,t^2)g'(t) = [2t^5, 2t^4]\begin{bmatrix} 1 \\ 2t \end{bmatrix} = 6t^5.$$

Note that $(f \circ g)(t) = f(t,t^2) = t^2t^4 = t^6$ so this result is not surprising!

Example 2.3.35. .

$$T(r,\theta) = (r\cos\theta, r\sin\theta) = (x,y)$$

$$T' = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

If $w = f(x,y)$ then $w = \underbrace{g(r,\theta) = f(T(r,\theta))}_{f \text{ rewritten in polars.}}$

$$\left[\frac{\partial g}{\partial r}, \frac{\partial g}{\partial \theta}\right] = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = \left[\cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y}, -r\sin\theta\frac{\partial f}{\partial x} + r\cos\theta\frac{\partial f}{\partial y}\right]$$

With the proper understanding we have derived,

$$\frac{\partial}{\partial r} = \cos\theta\frac{\partial}{\partial x} + \sin\theta\frac{\partial}{\partial y}$$

$$\frac{\partial}{\partial \theta} = -r\sin\theta\frac{\partial}{\partial x} + r\cos\theta\frac{\partial}{\partial y}$$

You can invert these, $r = \sqrt{x^2+y^2}$ & $\theta = \tan^{-1}(y/x)$

$$\frac{\partial}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta} = \frac{x}{r}\frac{\partial}{\partial r} - \frac{y}{r^2}\frac{\partial}{\partial \theta} = \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}$$

$$\frac{\partial}{\partial y} = \frac{\partial r}{\partial y}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial y}\frac{\partial}{\partial \theta} = \frac{y}{r}\frac{\partial}{\partial r} + \frac{x}{r^2}\frac{\partial}{\partial \theta} = \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta}$$

You can use these to change coordinates. For example

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \left(\cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}\right)\left(\cos\theta\frac{\partial f}{\partial r} - \frac{\sin\theta}{r}\frac{\partial f}{\partial \theta}\right) + \left(\sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta}\right)\left(\sin\theta\frac{\partial f}{\partial r} + \frac{\cos\theta}{r}\frac{\partial f}{\partial \theta}\right)$$

$$= \cos^2\theta\frac{\partial^2 f}{\partial r^2} - \cos\theta\sin\theta\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial f}{\partial \theta}\right] - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}\left[\cos\theta\frac{\partial f}{\partial r}\right] + \frac{\sin\theta}{r^2}\frac{\partial}{\partial \theta}\left[\sin\theta\frac{\partial f}{\partial \theta}\right]$$
$$+ \sin^2\theta\frac{\partial^2 f}{\partial r^2} + \sin\theta\cos\theta\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial f}{\partial \theta}\right] + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta}\left[\sin\theta\frac{\partial f}{\partial r}\right] + \frac{\cos\theta}{r^2}\frac{\partial}{\partial \theta}\left[\cos\theta\frac{\partial f}{\partial \theta}\right]$$

$$= \frac{\partial^2 f}{\partial r^2} + \frac{\sin^2\theta}{r}\frac{\partial f}{\partial r} + \frac{\cos^2\theta}{r}\frac{\partial f}{\partial r} - \frac{\sin\theta\cos\theta}{r}\frac{\partial^2 f}{\partial \theta \partial r} + \frac{\cos\theta\sin\theta}{r}\frac{\partial^2 f}{\partial \theta \partial r} + \cdots$$
$$\cdots + \frac{\sin\theta\cos\theta}{r^2}\frac{\partial f}{\partial \theta} - \frac{\cos\theta\sin\theta}{r^2}\frac{\partial f}{\partial \theta} + \frac{\sin^2\theta}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\cos^2\theta}{r^2}\frac{\partial^2 f}{\partial \theta^2}$$

$$\therefore \quad \boxed{\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2}}$$

2.4 continuous differentiability

We have noted that differentiablility on some set U implies all sorts of nice formulas in terms of the partial derivatives. Curiously the converse is not quite so simple. It is possible for the partial derivatives to exist on some set and yet the mapping may fail to be differentiable. We need an extra topological condition on the partial derivatives if we are to avoid certain pathological[8] examples.

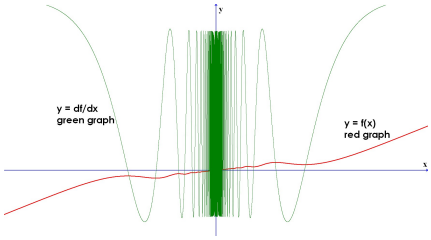
Example 2.4.1. *I found this example in Hubbard's advanced calculus text(see Ex. 1.9.4, pg. 123). It is a source of endless odd examples, notation and bizarre quotes. Let* $f(x) = 0$ *and*

$$f(x) = \frac{x}{2} + x^2 \sin\frac{1}{x}$$

for all $x \neq 0$*. I can be shown that the derivative* $f'(0) = 1/2$*. Moreover, we can show that* $f'(x)$ *exists for all* $x \neq 0$*, we can calculate:*

$$f'(x) = \frac{1}{2} + 2x \sin\frac{1}{x} - \cos\frac{1}{x}$$

Notice that $dom(f') = \mathbb{R}$*. Note then that the tangent line at* $(0,0)$ *is* $y = x/2$*.*



You might be tempted to say then that this function is increasing at a rate of 1/2 for x *near zero. But this claim would be false since you can see that* $f'(x)$ *oscillates wildly without end near zero. We have a tangent line at* $(0,0)$ *with positive slope for a function which is not increasing at* $(0,0)$ *(recall that increasing is a concept we must define in a open interval to be careful). This sort of thing cannot happen if the derivative is continuous near the point in question.*

The one-dimensional case is really quite special, even though we had discontinuity of the derivative we still had a well-defined tangent line to the point. However, many interesting theorems in calculus of one-variable require the function to be continuously differentiable near the point of interest. For example, to apply the 2nd-derivative test we need to find a point where the first derivative is zero and the second derivative exists. We cannot hope to compute $f''(x_o)$ unless f' is continuous at x_o. The next example is *sick*.

Example 2.4.2. *Let us define* $f(0,0) = 0$ *and*

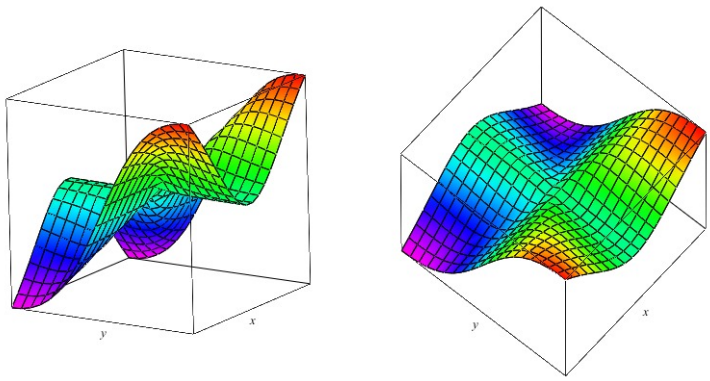
$$f(x,y) = \frac{x^2 y}{x^2 + y^2}$$

for all $(x,y) \neq (0,0)$ *in* $\mathbb{R}^2$*. It can be shown that* f *is continuous at* $(0,0)$*. Moreover, since* $f(x,0) = f(0,y) = 0$ *for all* x *and all* y *it follows that* f *vanishes identically along the coordinate axis. Thus the rate of change in the* e_1 *or* e_2 *directions is zero. We can calculate that*

$$\frac{\partial f}{\partial x} = \frac{2xy^3}{(x^2+y^2)^2} \qquad \text{and} \qquad \frac{\partial f}{\partial y} = \frac{x^4 - x^2y^2}{(x^2+y^2)^2}$$

If you examine the plot of $z = f(x,y)$ *you can see why the tangent plane does not exist at* $(0,0,0)$*.*

[8] "pathological" as in, "your clothes are so pathological, where'd you get them?"



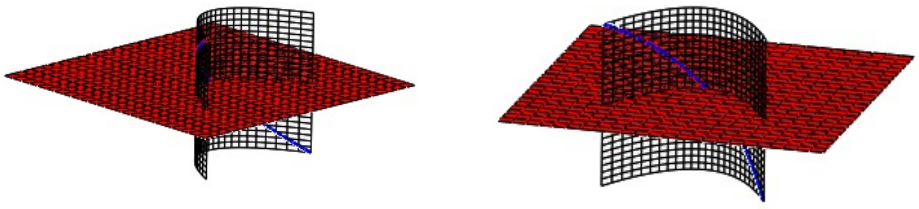
Notice the sides of the box in the picture are parallel to the x and y axes so the path considered below would fall on a diagonal slice of these boxes[9]*. Consider the path to the origin $t \mapsto (t,t)$ gives $f_x(t,t) = 2t^4/(t^2+t^2)^2 = 1/2$ hence $f_x(x,y) \to 1/2$ along the path $t \mapsto (t,t)$, but $f_x(0,0) = 0$ hence the partial derivative f_x is not continuous at $(0,0)$. In this example, the discontinuity of the partial derivatives makes the tangent plane fail to exist.*

One might be tempted to suppose that if a function is continuous at a given point and if all the possible directional derivatives exist then differentiability should follow. It turns out this is not sufficient since continuity of the function does not imply some continuity along the partial derivatives. For example:

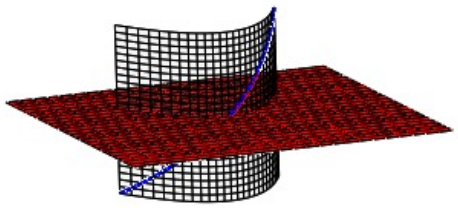
Example 2.4.3. *Let us define $f : \mathbb{R}^2 \to \mathbb{R}$ by $f(x,y) = 0$ for $y \neq x^2$ and $f(x,x^2) = x$. I invite the reader to verify that this function is continuous at the origin. Moreover, consider the directional derivatives at $(0,0)$. We calculate, if $v = \langle a,b \rangle$*

$$D_v f(0,0) = \lim_{h\to 0} \frac{f(0+hv) - f(0)}{h} = \lim_{h\to 0} \frac{f(ah,bh)}{h} = \lim_{h\to 0} \frac{0}{h} = 0.$$

To see why $f(ah,bh) = 0$, consider the intersection of $\vec{r}(h) = (ha,hb)$ and $y = x^2$ the intersection is found at $hb = (ha)^2$ hence, noting $h = 0$ is not of interest in the limit, $b = ha^2$. If $a = 0$ then clearly (ah,bh) only falls on $y = x^2$ at $(0,0)$. If $a \neq 0$ then the solution $h = b/a^2$ gives $f(ha,hb) = ha$ a nontrivial value. However, as $h \to 0$ we eventually reach values close enough to $(0,0)$ that $f(ah,bh) = 0$. Hence we find **all** *directional derivatives exist and are zero at $(0,0)$. Let's examine the graph of this example to see how this happened. The pictures below graph the xy-plane as red and the nontrivial values of f as a blue curve. The union of these forms the graph $z = f(x,y)$.*



[9] the argument to follow stands alone, you don't need to understand the picture to understand the math here, but it's nice if you do



Clearly, f is continuous at $(0,0)$ as I invited you to prove. Moreover, clearly $z = f(x,y)$ cannot be well-approximated by a tangent plane at $(0,0,0)$. If we capture the xy-plane then we lose the blue curve of the graph. On the other hand, if we use a tilted plane then we lose the xy-plane part of the graph.

The moral of the story in the last two examples is simply that derivatives at a point, or even all directional derivatives at a point do not necessarily tell you much about the function near the point. This much is clear: something else is required if the differential is to have meaning which extends beyond one point in a nice way. Therefore, we consider the following:

It would seem the trouble has something to do with discontinuity in the derivative. Continuity of the derivative requires the assignment $a \mapsto dF_a$ is continuous. Or,

$$\lim_{x \to a} dF_x = dF_a. \tag{2.18}$$

But, this is a limit of operators. Let us study this limit in view of the operator norm we discussed in the previous chapter. Let $\epsilon > 0$ then we must be able to find $\delta > 0$ such that $0 < \|x - a\| < \delta$ implies $\|dF_x - dF_a\| < \epsilon$. So, we need to control $\|dF_x - dF_a\|$ to be sure the derivative is continuous. Consider,

$$\begin{aligned} \|dF_x - dF_a\| &= \sup\{\|(dF_x - dF_a)(u)\| \ : \ \|u\| = 1\} \\ &= \sup\{\|dF_x(u) - dF_a(u)\| \ : \ \|u\| = 1\} \\ &= \sup\left\{ \left\| \sum_{i=1}^{n} u_i \frac{\partial F}{\partial x_i}(x) - \sum_{i=1}^{n} u_i \frac{\partial F}{\partial x_i}(a) \right\| \ : \ \|u\| = 1 \right\} \\ &\leq \sum_{i=1}^{n} \left\| \frac{\partial F}{\partial x_i}(x) - \frac{\partial F}{\partial x_i}(a) \right\| \end{aligned} \tag{2.19}$$

Therefore, the data $\lim_{x \to a} \frac{\partial F}{\partial x_i}(x) = \frac{\partial F}{\partial x_i}(a)$ for $i = 1, \ldots, n$ allows us to prove $\lim_{x \to a} dF_x = dF_a$. Naturally, when we teach multivariate calculus the preferred concept does not involve operator norms. Therefore, to be nice to the non-math majors we define:

Definition 2.4.4.

A mapping $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is **continuously differentiable** at $a \in U$ iff the partial derivative mappings $D_j F$ exist on an open set containing a and are continuous at a.

Equation 2.19 shows maps **continuously differentiable** at $x = a$ are those for which the mapping $x \to dF_x$ is a continuous mapping at $x = a$.

The import of the theorem below is that we can build the tangent plane from the Jacobian matrix provided the partial derivatives exist near the point of tangency and are continuous at the point of tangency. This is a very nice result because the concept of the linear mapping is quite abstract but partial differentiation of a given mapping is often easy. The proof that follows here is found in many texts, in particular see C.H. Edwards *Advanced Calculus of Several Variables* on pages 72-73.

Theorem 2.4.5.

If $F : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable at a then F is differentiable at a

Proof: Consider $a+h$ sufficiently close to a that all the partial derivatives of F exist. Furthermore, consider going from a to $a+h$ by traversing a hyper-parallel-piped travelling n-perpendicular paths:

$$\underbrace{a}_{p_o} \to \underbrace{a+h_1e_1}_{p_1} \to \underbrace{a+h_1e_1+h_2e_2}_{p_2} \to \cdots \underbrace{a+h_1e_1+\cdots+h_ne_n}_{p_n} = a+h.$$

Let us denote $p_j = a + b_j$ where clearly b_j ranges from $b_o = 0$ to $b_n = h$ and $b_j = \sum_{i=1}^{j} h_i e_i$. Notice that the difference between p_j and p_{j-1} is given by:

$$p_j - p_{j-1} = a + \sum_{i=1}^{j} h_i e_i - a - \sum_{i=1}^{j-1} h_i e_i = h_j e_j$$

Consider then the following identity,

$$F(a+h) - F(a) = F(p_n) - F(p_{n-1}) + F(p_{n-1}) - F(p_{n-2}) + \cdots + F(p_1) - F(p_o)$$

This is to say the change in F from $p_o = a$ to $p_n = a + h$ can be expressed as a sum of the changes along the n-steps. Furthermore, if we consider the difference $F(p_j) - F(p_{j-1})$ you can see that only the j-th component of the argument of F changes. Since the j-th partial derivative exists on the interval for h_j considered by construction we can apply the mean value theorem to locate c_j such that:

$$h_j \partial_j F(p_{j-1} + c_j e_j) = F(p_j) - F(p_{j-1})$$

Therefore, using the mean value theorem for each interval, we select $c_1, \ldots c_n$ with:

$$F(a+h) - F(a) = \sum_{j=1}^{n} h_j \partial_j F(p_{j-1} + c_j e_j)$$

It follows we should propose L to satisfy the definition of Frechet differentation as follows:

$$L(h) = \sum_{j=1}^{n} h_j \partial_j F(a)$$

It is clear that L is linear (in fact, perhaps you recognize this as $L(h) = (\nabla F)(a) \bullet h$). Let us prepare to study the Frechet quotient,

$$\begin{aligned} F(a+h) - F(a) - L(h) &= \sum_{j=1}^{n} h_j \partial_j F(p_{j-1} + c_j e_j) - \sum_{j=1}^{n} h_j \partial_j F(a) \\ &= \sum_{j=1}^{n} h_j \underbrace{\left[\partial_j F(p_{j-1} + c_j e_j) - \partial_j F(a)\right]}_{g_j(h)} \end{aligned}$$

Observe that $p_{j-1} + c_j e_j \to a$ as $h \to 0$. Thus, $g_j(h) \to 0$ by the continuity of the partial derivatives at $x = a$. Finally, consider the Frechet quotient:

$$\lim_{h \to 0} \frac{F(a+h) - F(a) - L(h)}{||h||} = \lim_{h \to 0} \frac{\sum_j h_j g_j(h)}{||h||} = \lim_{h \to 0} \sum_j \frac{h_j}{||h||} g_j(h)$$

Notice $|h_j| \leq ||h||$ hence $\left|\frac{h_j}{||h||}\right| \leq 1$ and

$$0 \leq \left|\frac{h_j}{||h||} g_j(h)\right| \leq |g_j(h)|$$

Apply the squeeze theorem to deduce each term in the sum $\star$ limits to zero. Consquently, $L(h)$ satisfies the Frechet quotient and we have shown that F is differentiable at $x = a$ and the differential is expressed in terms of partial derivatives as expected; $dF_x(h) = \sum_{j=1}^{n} h_j \partial_j F(a)$ $\square$.

Given the result above it is a simple matter to extend the proof to $F : \mathbb{R}^n \to \mathbb{R}^m$.

Theorem 2.4.6.

If $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable at a then F is differentiable at a

Proof: If F is continuously differentiable at a then clearly each component function $F^j : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable at a. Thus, by Theorem 2.4.5 we have F^j differentiable at a hence

$$\lim_{h\to 0} \frac{F^j(a+h) - F^j(a) - dF_a^j(h)}{||h||} = 0 \quad \text{for all } j \in \mathbb{N}_m \quad \Rightarrow \quad \lim_{h\to 0} \frac{F(a+h) - F(a) - dF_a(h)}{||h||} = 0$$

by Theorem 1.3.11. This proves F is differentiable at a $\square$.

2.5 the product rule

When I first wrote notes for advanced calculus I realized I was writing the same argument over and over. The result below is a result. This argument simultaneously covers derivatives of scalar multiplications, matrix multiplications, dot and cross products.

Theorem 2.5.1.

Let W_1, W_2, W_3, V be finite dimensional real normed linear spaces and suppose $U \subseteq V$ is open. Let $\beta = \{r_1, \ldots, r_n\}$ be a basis for V with coordinates $x_1, \ldots, x_n$. Let $\gamma_1 = \{w_1, \ldots, w_{m_1}\}$ be the basis for W_1. Let $\gamma_2 = \{v_1, \ldots, v_{m_2}\}$ be the basis for W_2. Let $\gamma_3 = \{\varepsilon_1, \ldots, \varepsilon_{m_3}\}$ be the basis for W_3. Assume there exists a product $\star : W_1 \times W_2 \to W_3$ such that

$$(cx + y) \star z = c(x \star z) + y \star z \qquad \& \qquad x \star (cz + w) = c(x \star z) + x \star w$$

for all $c \in \mathbb{R}$ and $x, y \in W_1$ and $z, w \in W_2$. Then, if $F : U \to W_1$ and $G : U \to W_2$ are continuously differentiable at $a \in U$ then $F \star G$ is continuously differentiable at $a \in U$ where $(F \star G)(a) = F(a) \star G(a)$. Moreover, denoting $\partial/\partial x_j$ by ∂_j we have

$$\partial_j(F \star G)(a) = (\partial_j F)(a) \star G(a) + F(a) \star (\partial_j G)(a).$$

Hence, for each $h \in V$,

$$d(F \star G)_a(h) = dF_a(h) \star G(a) + F(a) \star dG_a(h).$$

Proof: assume the notation given in the Theorem and define structure constants $c_{ijk} \in \mathbb{R}$ such that:

$$v_i \star w_j = \sum_{k=1}^{m_3} c_{ijk} \varepsilon_k. \tag{2.20}$$

These constants characterize the nature of the multiplication $\star$. Interestingly, they have little to do with the proof, essentially the play the role of bystanders. Assuming $F : U \to W_1$ and $G : U \to W_2$ are continuously differentiable at a means their component functions $F_1, \ldots, F_{m_1} : U \to \mathbb{R}$ with respect to γ_1 and $G_1, \ldots, G_{m_2} : U \to \mathbb{R}$ with respect to γ_2 are continuous at a. The component functions of $F \star G$ are naturally related to those of F and G as follows:

$$\begin{aligned} F \star G &= \left(\sum_{i=1}^{m_1} F_i v_i \right) \star \left(\sum_{j=1}^{m_2} G_j w_j \right) \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} F_i G_j \left(v_i \star w_j \right) \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} F_i G_j \left(\sum_{k=1}^{m_3} c_{ijk} \varepsilon_k \right) \\ &= \sum_{k=1}^{m_3} \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} F_i G_j c_{ijk} \right) \varepsilon_k \end{aligned} \tag{2.21}$$

Thus $F \star G$ has component function $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} F_i G_j c_{ijk}$. Observe this is the sum of products of continuously differentiable functions at a which is once again continuously differentiable a. Thus $F \star G$ is continuously differentiable at a as it has component functions whose partial derivative functions are continous at a. This becomes explicitly clear if we calculate the partial derivative of $F \star G$ with respect to x_l for points near a,

$$\begin{aligned} \partial_l (F \star G) &= \sum_{k=1}^{m_3} \partial_l \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} F_i G_j c_{ijk} \right) \varepsilon_k \qquad : \partial_l \text{ done componentwise} \\ &= \sum_{k=1}^{m_3} \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} c_{ijk} \partial_l (F_i G_j) \right) \varepsilon_k \qquad : \text{linearity of } \partial_l \\ &= \sum_{k=1}^{m_3} \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} c_{ijk} \left[(\partial_l F_i) G_j + F_i \partial_l G_j \right] \right) \varepsilon_k \qquad : \text{ordinary product rule} \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{m_3} c_{ijk} (\partial_l F_i) G_j \varepsilon_k + \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{m_3} c_{ijk} F_i (\partial_l G_j) \varepsilon_k \\ &= (\partial_l F) \star G + F \star (\partial_l G). \end{aligned} \tag{2.22}$$

where I used the calculation of Equation 2.21 in reverse in order to make the final step. The calculation makes it explicitly clear that the partial derivatives of $F \star G$ are sums and products of continuous functions hence $F \star G$ is continuously differentiable as claimed. Finally, we can construct

the differential from partial derivatives: for $h = \sum_{l=1}^{n} h_l r_l$ calculate:

$$\begin{aligned} d(F \star G)_a(h) &= \sum_{l=1}^{n} h_l \partial_l (F \star G)(a) \\ &= \sum_{l=1}^{n} h_l \left[(\partial_l F)(a) \star G(a) + F(a) \star (\partial_l G)(a) \right] \\ &= \left[\sum_{l=1}^{n} h_l (\partial_l F)(a) \right] \star G(a) + F(a) \star \left[\sum_{l=1}^{n} h_l (\partial_l G)(a) \right]. \\ &= dF_a(h) \star G(a) + F(a) \star dG_a(h). \end{aligned} \tag{2.23}$$

This completes the proof. □

Let's unwrap a few common cases of this general product rule. I'll continue to use the W_1, W_2, W_3 and V notation to connect directly to Theorem 2.5.1.

(1.) Set $W_1 = W_2 = W_3 = \mathbb{R}$ and $V = \mathbb{R}$ to produce the usual first semester calculus product rule:

$$\frac{d}{dt}(fg) = \frac{df}{dt} g + f \frac{dg}{dt}.$$

Of course, this was the heart of the proof.

(2.) Set $W_1 = W_2 = W_3 = \mathbb{R}$ and $V = \mathbb{R}^n$ to produce the usual product rule for real-valued functions of several variables:

$$\frac{\partial}{\partial x_i}(fg) = \frac{\partial f}{\partial x_i} g + f \frac{\partial g}{\partial x_i}.$$

(3.) Set $W_1 = \mathbb{R}$ and $W_2 = W_3$ and $V = \mathbb{R}^n$ to produce the usual product rule for a scalar function multiplied on a vector-valued function:

$$\frac{\partial}{\partial x_i}(f\vec{v}) = \frac{\partial f}{\partial x_i} \vec{v} + f \frac{\partial \vec{v}}{\partial x_i}.$$

(4.) Set $W_1 = W_2 = \mathbb{R}^n$ and $W_3 = \mathbb{R}$ and $V = \mathbb{R}$ to produce the product rule for dot-products of paths:

$$\frac{d}{dt}(\vec{v} \bullet \vec{w}) = \frac{d\vec{v}}{dt} \bullet \vec{w} + \vec{v} \bullet \frac{d\vec{w}}{dt}.$$

(5.) Set $W_1 = W_2 = \mathbb{R}^3$ and $W_3 = \mathbb{R}^3$ and $V = \mathbb{R}$ to produce the product rule for cross-products of paths:

$$\frac{d}{dt}(\vec{v} \times \vec{w}) = \frac{d\vec{v}}{dt} \times \vec{w} + \vec{v} \times \frac{d\vec{w}}{dt}.$$

(6.) Set $W_1 = W_2 = W_3 = \mathbb{R}^{n \times n}$ and $V = \mathbb{R}$ to produce the product rule for matrix-valued functions of a real variable: $t \mapsto A(t)$, $t \mapsto B(t)$,

$$\frac{d}{dt}(AB) = \frac{dA}{dt} B + A \frac{dB}{dt}.$$

(7.) Set $W_1 = W_2 = W_3 = \mathbb{C}$ and $V = \mathbb{C}$ with $z = x + iy$ we find for $f_1 = u_1 + iv_1$ and $f_2 = u_2 + iv_2$

$$\frac{\partial}{\partial x}(f_1 f_2) = \frac{\partial f_1}{\partial x} f_2 + f_1 \frac{\partial f_2}{\partial x} \qquad \& \qquad \frac{\partial}{\partial y}(f_1 f_2) = \frac{\partial f_1}{\partial y} f_2 + f_1 \frac{\partial f_2}{\partial y}.$$

Of course, there is much more. I simply wish to impress on you that these product rules are **all** simply the standard product rule married to the algebraic structure of the given product. So long as the product has the needed linearity properties, there will be a corresponding product rule for functions.

2.6 higher derivatives

Given normed linear spaces V, W and $U \subseteq V$ open and a differentiable map $F : U \to W$ we find a linear transformation $dF_a : V \to W$ for each $a \in U$. Therefore, we can define the map $f' : U \to \mathcal{L}(V; W)$ by the natural map $a \mapsto dF_a$. That is, $f'(a) = df_a$. Furthermore, since $\mathcal{L}(V; W)$ is itself a normed linear space we may study derivatives of f'. In particular, if $df'_a : V \to \mathcal{L}(V; \mathcal{L}(V; W))$ is linear for each $a \in U$ and satisfies the needed Frechet quotient then we may likewise define $f'' : U \to \mathcal{L}(V; \mathcal{L}(V; W))$ by $f''(a) = (f')'(a) = (df')_a \in \mathcal{L}(V; \mathcal{L}(V; W))$ for each $a \in U$. This all gets a bit meta, so, its helpful to make use of an isomorphism $\Psi : \mathcal{L}(V; \mathcal{L}(V; W)) \to \mathcal{L}(V, V; W)$ defined by:

$$\Psi(T)(x, y) = (T(x))(y) \tag{2.24}$$

for all $x, y \in V$ and $T \in \mathcal{L}(V, W)$. Typically the Ψ is not written. With this abuse of language, we have $f''(a) : V \times V \to W$ given by

$$f''(a)(h, k) = df'_a(h, k) = d(h \mapsto df_h)_a(k) \tag{2.25}$$

Thus, in stark contrast to first semester calculus, each added derivative brings out a new object. Using the isomorphism and its extension to higher derivatives, we find the n-th derivative of $f : V \to W$ is naturally understood as an n-linear map from V to W. What is beautiful is that we can capture this simply in terms of iterated partial derivatives provided a certain continuity is given. I'll attempt to explain this for the case of second derivatives this semester. For the sake of time, I'll let Zorich provide the many details I omit here. If I find time to prepare and Lecture, we may examine the proof that partial derivatives commute. Whether or not we have time for the proof, the fact that partial derivatives commute is a cornerstone of abstract calculus.

2.7 differentiation in an algebra variable

Here I share with you the rudiments of what I have come to call $\mathcal{A}$-calculus. We say $\mathcal{A}$ is an **algebra** if there is a multiplication $\star : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ which behaves like ordinary multiplication:

(1.) $(x + y) \star z = x \star z + y \star z$ and $x \star (y + z) = x \star y + x \star z$

(2.) $(cx) \star y = x \star (cy) = c(x \star y)$

(3.) $(x \star y) \star z = x \star (y \star z)$

(4.) there exists $1_{\mathcal{A}} \in \mathcal{A}$ for which $1_{\mathcal{A}} \star x = x = x \star 1_{\mathcal{A}}$ for each $x \in \mathcal{A}$

I usually think about **real algebras** which means there is essentially a copy of $\mathbb{R}$ in the center of the algebra. In item (2.) I assume $c \in \mathbb{R}$. However, the $1_{\mathcal{A}}$ may not appear manifestly as $1 \in \mathbb{R}$. Let me give a couple simple examples and forego the general theory.

Example 2.7.1. *$\mathcal{A} = \mathbb{C}$ is a nice example. If $a + ib, c + id \in \mathbb{C}$ then we define $(a + ib)(c + id) = ac - bd + i(ad + bc)$. Equivalently, we could just proclaim $i^2 = -1$ and otherwise, calculate like usual. Here $1_{\mathcal{A}} = 1$ as you might expect*[10]

[10] Using $\mathbb{C} = \mathbb{R}^2$ as a point set we note $1 = (1, 0)$ and $i = (0, 1)$ hence $c_{111} = 1$ and $c_{221} = -1$ and $c_{122} = c_{212} = 1$ whereas all other structure constants are zero. This has not much to do with anything, but, I thought it might be fun given the proof of the previous section

Example 2.7.2. *The* **direct product algebra** *of $\mathcal{A} = \mathbb{R} \times \mathbb{R}$ is defined by $(a,b)(x,y) = (ax,by)$. Here $(1,1)(x,y) = (x,y)$ for all $(x,y) \in \mathcal{A}$ and in fact $1_{\mathcal{A}} = (1,1)$.*

Example 2.7.3. *The* **hyperbolic numbers** *are of the form $a + bj$ where $j^2 = 1$. In particular, define $(a+bj)(c+jd) = ac + bd + j(ad+bc)$.*

Example 2.7.4. *The* **3-hyperbolic numbers** *are of the form $a + bj + cj^2$ where $j^3 = 1$. In particular, define*

$$(a+bj+cj^2)(x+jy+j^2z) = ax+by+cz+j(bx+ay+cz)+j^2(cx+by+az).$$

All the algebras I've listed thus far are **commutative**. There are also many noncommutative algebras like the quaternions or matrix algebras. Notice $\mathbb{R}^{n\times n}$ forms an algebra. Basically, I think of algebras as **generalized number systems**. So, given that, it is interesting to ask what it means to differentiate with respect to a variable which takes values in $\mathcal{A}$. In fact, we have a whole course devoted to studying what happens when you do calculus with respect to a complex variable. Many schools have such a course. What is less known, which is a shame since it's really pretty simple, is that you can differentiate with respect to an algebra variable in much the same way.

Definition 2.7.5. *Let $U \subseteq \mathcal{A}$ be an open set containing p. If $f : U \to \mathcal{A}$ is a function then we say f is* **$\mathcal{A}$-differentiable at** *p if there exists a linear function $d_pf \in \mathcal{R}_{\mathcal{A}}$ such that*

$$\lim_{h\to 0} \frac{f(p+h) - f(p) - d_pf(h)}{||h||} = 0. \tag{2.26}$$

When I say $d_pf \in \mathcal{R}_{\mathcal{A}}$ this simply means that $d_pf : \mathcal{A} \to \mathcal{A}$ is $\mathbb{R}$-linear mapping on $\mathcal{A}$ and $d_pf(v \star w) = d_pf(v) \star w$ for all $v, w \in \mathcal{A}$. In other words, $\mathcal{A}$-differentiability amounts to differentiability at p with an extra condition. Furthermore, we define the derivative at p as follows:

$$(d_pf)(h) = f'(p)h \tag{2.27}$$

But, since $(d_pf)(h) = d_pf(1 \star h) = d_pf(1) \star h = f'(p)h$ we have $f'(p) = d_pf(1)$. In contrast to the differential of an arbitrary real differentiable map on $\mathcal{A}$, the formula for d_pf is equivalent to the selection of a number in $\mathcal{A}$ for p. In other words, there is a natural manner to interpret the derivative of a function as a function once more. Furthermore, it can be shown for higher derivatives of an $\mathcal{A}$-differentiable function we have

$$d^nf(v_1, v_2, \ldots, v_n) = d^nf(1,1,\ldots,1) \star v_1 \star v_2 \star \cdots \star v_n \tag{2.28}$$

So the n-th derivative is also uniquely fixed by the value of $d^nf(1,1,\ldots,1)$. In fact, we can naturally identify the n-th derivative of a function as a function once more. In general, the n-th derivative is a symmetric n-linear functon. Finally, I must tell you a beautiful formula which makes $\mathcal{A}$-Calculus so very interesting: provided the basis for $\mathcal{A}$ has $1_{\mathcal{A}} = 1$ paired with coordinate x_1:

$$\frac{\partial^n f}{\partial x_{i_1}\partial x_{i_2}\cdots\partial x_{i_n}} = \frac{\partial^n f}{\partial x_1^n} \star v_{i_1} \star v_{i_2} \star \cdots \star v_{i_n} \tag{2.29}$$

If $\mathcal{A} = \mathbb{R}^n$ as a point set and $e_1 = 1$ then the formulas describing $\mathcal{A}$-calculus are quite nice.

Example 2.7.6. *Consider $f = u + iv$ which is complex differentiable at $p \in \mathbb{C}$. Use $z = x + iy$ as the typical variable in $\mathbb{C}$. Notice, $d_pf(i) = d_pf(1)i$ implies that $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial x}i$. These are the famed* **Cauchy Riemann** *equations. To help the reader make the connection, note $f_y = u_y + iv_y$ and $f_x = u_x + iv_x$ hence $f_y = if_x$ amounts to $(u_y + iv_y) = i(u_x + iv_x)$ hence $u_y = -v_x$ and $v_y = u_x$. Jumping ahead a bit, with no intention of explaining why here, it is fun to note since $i^2 + 1 = 0$ it follows $f_{yy} + f_{xx} = 0$ hence the component functions of a complex differentiable function are solutions to Laplace's equation.*

Example 2.7.7. *Consider $f = u + jv$ which is hyperbolic differentiable at $p \in \mathcal{H} = \mathbb{R} \oplus j\mathbb{R}$ (this is just notation for the hyperbolic numbers). Use $z = x + jy$ as the typical variable in $\mathcal{H}$. Notice, $d_pf(j) = d_pf(1)j$ implies that $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial x}j$. These are the no so well-known hyperbolic Cauchy Riemann equations. To help the reader make the connection, note $f_y = u_y + jv_y$ and $f_x = u_x + jv_x$ hence $f_y = jf_x$ amounts to $(u_y + jv_y) = j(u_x + jv_x)$ hence $u_y = v_x$ and $v_y = u_x$. Jumping ahead a bit, with no intention of explaining why here, it is fun to note since $1 - j^2 = 0$ it follows $f_{xx} - f_{yy} = 0$ hence the component functions of a hyperbolic differentiable function are solutions to the one-dimensional wave equation.*

Basically, any identity which appears amongst the basis elements of an algebra will be mirrored in a PDE which is solve by each function differentiable over the algebra. Most familar case is with $\mathbb{C}$ where harmonic functions are a standard and beatiful topic. But, this is just one of many function theories. In ordinary real analysis essentially $\mathcal{A} = \mathbb{R}$ itself so this feature cannot be seen. However, once $\mathcal{A}$ is two or more dimensional, the differentiability with respect to $\mathcal{A}$ binds real variables together in such a way that the change in one real variable is necessarily coupled to the rest.

Ok, so, let's return to our uber product rule once more, assume f, g are $\mathcal{A}$-differentiable at p in a commutative algebra then note:

$$\begin{aligned} d_p(f \star g)(v \star w) &= (d_pf)(v \star w) \star g(p) + f(p) \star d_pg(v \star w) \\ &= d_pf(v) \star w \star g(p) + f(p) \star d_pg(v) \star w \\ &= (d_pf(v) \star g(p) + f(p) \star d_pg(v)) \star w \end{aligned} \tag{2.30}$$

We can argue $f \star g$ is real differentiable and $d_p(f \star g) \in \mathcal{R}_\mathcal{A}$ thus $f \star g$ is $\mathcal{A}$-differentiable at p. Moreover, as $(f \star g)'(p) = d_p(f \star g)(1)$ we derive from the result above that

$$(f \star g)'(p) = f'(p) \star g(p) + f(p) \star g'(p).$$

Many further results about the calculus over an algebra are known and many resemble closely the calculus you've already seen. However, I've also found a few suprises, mostly thanks to the students who've helped me study $\mathcal{A}$-calculus the past few years. If this section was a bit too terse, my apologies, I have much more to say in my primer on $\mathcal{A}$-calculus: Introduction to $\mathcal{A}$-Calculus and my $\mathcal{A}$-Calculus II paper with Daniel Freese and my differential equations on an algebra paper with Nathan BeDell. I will probably share some tidbits about these papers when the time seems right in this course. But, our main focus is elsewhere.

Chapter 3

inverse and implicit function theorems

It is tempting to give a complete and rigourous proof of these theorems at the outset, but I will resist the temptation in lecture. I'm actually more interested that the student understand what the theorem claims before I show the real proof. I will sketch the proof and show many applications. A nearly complete proof is found in Edwards where he uses an iterative approximation technique founded on the contraction mapping principle, we will go through that a bit later in the course. I probably will not have typed notes on that material this semester, but Edward's is fairly readable and I think we'll profit from working through those sections. That said, we develop an intuition for just what these theorems are all about to start. That is the point of this chapter: to grasp what the linear algebra of the Jacobian suggests about the local behaviour of functions and equations.

3.1 inverse function theorem

Consider the problem of finding a **local** inverse for $f : dom(f) \subseteq \mathbb{R} \to \mathbb{R}$. If we are given a point $p \in dom(f)$ such that there exists an open interval I containing p with $f|_I$ a one-one function then we can reasonably construct an inverse function by the simple rule $f^{-1}(y) = x$ iff $f(x) = y$ for $x \in I$ and $y \in f(I)$. A sufficient condition to insure the existence of a local inverse is that the derivative function is either strictly positive or strictly negative on some neighborhood of p. If we are give a continuously differentiable function at p then it has a derivative which is continuous on some neighborhood of p. For such a function if $f'(p) \neq 0$ then there exists some interval centered at p for which the derivative is strictly positive or negative. It follows that such a function is strictly monotonic and is hence one-one thus there is a local inverse at p. We should all learn in calculus I that the derivative informs us about the local invertibility of a function. Natural question to ask for us here: does this extend to higher dimensions? If so, how?

The arguments I just made are supported by theorems that are developed in calculus I. Let me shift gears a bit and give a direct calculational explaination based on the linearization approximation.

If $x \approx p$ then $f(x) \approx f(p) + f'(p)(x - p)$. To find the formula for the inverse we solve $y = f(x)$ for x:

$$y \approx f(p) + f'(p)(x - p) \quad \Rightarrow \quad x \approx p + \frac{1}{f'(p)}\big[y - f(p)\big]$$

Therefore, $\boxed{f^{-1}(y) \approx p + \frac{1}{f'(p)}\big[y - f(p)\big]}$ for y near $f(p)$.

Example 3.1.1. *Just to help you believe me, consider $f(x) = 3x - 2$ then $f'(x) = 3$ for all x. Suppose we want to find the inverse function near $p = 2$ then the discussion preceding this example suggests,*

$$f^{-1}(y) = 2 + \frac{1}{3}(y - 4).$$

I invite the reader to check that $f(f^{-1}(y)) = y$ and $f^{-1}(f(x)) = x$ for all $x, y \in \mathbb{R}$.

In the example above we found a global inverse exactly, but this is thanks to the linearity of the function in the example. Generally, inverting the linearization just gives the first approximation to the inverse.

Consider $F : dom(F) \subseteq \mathbb{R}^n \to \mathbb{R}^n$. If F is differentiable at $p \in \mathbb{R}^n$ then we can write $F(x) \approx F(p) + F'(p)(x - p)$ for $x \approx p$. Set $y = F(x)$ and solve for x via matrix algebra. This time we need to assume $F'(p)$ is an invertible matrix in order to isolate x,

$$y \approx F(p) + F'(p)(x - p) \quad \Rightarrow \quad x \approx p + (F'(p))^{-1}\big[y - f(p)\big]$$

Therefore,

$$\boxed{F^{-1}(y) \approx p + (F'(p))^{-1}\big[y - f(p)\big]}$$

for y near $F(p)$. Apparently the condition to find a local inverse for a mapping on $\mathbb{R}^n$ is that the derivative matrix is nonsingular[1] in some neighborhood of the point. Experience has taught us from the one-dimensional case that we must insist the derivative is continuous near the point in order to maintain the validity of the approximation.

Recall from calculus II that as we attempt to approximate a function with a power series it takes an infinite series of power functions to recapture the formula exactly. Well, something similar is true here. However, the method of approximation is through an iterative approximation procedure which is built off the idea of Newton's method. The product of this iteration is a nested sequence of composite functions. To prove the theorem below one must actually provide proof the recursively generated sequence of functions converges. See pages 160-187 of Edwards for an in-depth exposition of the iterative approximation procedure. Then see pages 404-411 of Edwards for some material on uniform convergence[2] The main analytical tool which is used to prove the convergence is called the **contraction mapping principle**. The proof of the principle is relatively easy to follow and interestingly the main non-trivial step is an application of the geometric series. For the student of analysis this is an important topic which you should spend considerable time really trying to absorb as deeply as possible. The contraction mapping is at the base of a number of interesting and nontrivial theorems. Read Rosenlicht's *Introduction to Analysis* for a broader and better organized exposition of this analysis. In contrast, Edwards' uses analysis as a tool to obtain results for advanced calculus but his central goal is not a broad or well-framed treatment of analysis. Consequently, if analysis is your interest then you really need to read something else in parallel to get a better ideas about sequences of functions and uniform convergence. I have some notes from a series of conversations with a student about Rosenlicht, I'll post those for the interested student. These notes focus on the part of the material I require for this course. This is Theorem 3.3 on page 185 of Edwards' text:

[1] nonsingular matrices are also called invertible matrices and a convenient test is that A is invertible iff $det(A) \neq 0$.

[2] actually that later chapter is part of why I chose Edwards' text, he makes a point of proving things in $\mathbb{R}^n$ in such a way that the proof naturally generalizes to function space. This is done by arguing with properties rather than formulas. The properties offen extend to infinite dimensions whereas the formulas usually do not.

Theorem 3.1.2. *(inverse function theorem)*

Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable in an open set W containing a and the derivative matrix $F'(a)$ is invertible. Then F is locally invertible at a. This means that there exists an open set $U \subseteq W$ containing a and V a open set containing $b = F(a)$ and a one-one, continuously differentiable mapping $G : V \to W$ such that $G(F(x)) = x$ for all $x \in U$ and $F(G(y)) = y$ for all $y \in V$. Moreover, the local inverse G can be obtained as the limit of the sequence of successive approximations defined by

$$G_o(y) = a \quad \text{and} \quad G_{n+1}(y) = G_n(y) - [F'(a)]^{-1}[F(G_n(y)) - y] \qquad \text{for all } y \in V.$$

The qualifier local is important to note. If we seek a global inverse then other ideas are needed. If the function is everywhere injective then logically $F(x) = y$ defines $F^{-1}(y) = x$ and F^{-1} so constructed in single-valued by virtue of the injectivity of F. However, for differentiable mappings, one might wonder how can the criteria of global injectivity be tested via the differential. Even in the one-dimensional case a vanishing derivative does not indicate a lack of injectivity; $f(x) = x^3$ has $f^{-1}(y) = \sqrt[3]{y}$ and yet $f'(0) = 0$ (therefore $f'(0)$ is not invertible). One the other hand, we'll see in the examples that follow that even if the derivative is invertible over a set it is possible for the values of the mapping to double-up and once that happens we cannot find a single-valued inverse function[3]

Remark 3.1.3. *James R. Munkres' Analysis on Manifolds good for a different proof.*

Another good place to read the inverse function theorem is in James R. Munkres *Analysis on Manifolds*. That text is careful and has rather complete arguments which are not entirely the same as the ones given in Edwards. Munkres' text does not use the contraction mapping principle, instead the arguments are more topological in nature.

To give some idea of what I mean by topological let be give an example of such an argument. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable and $F'(p)$ is invertible. Here's a sketch of the argument that $F'(x)$ is invertible for all x near p as follows:

1. the function $g : \mathbb{R}^n \to \mathbb{R}$ defined by $g(x) = det(F'(x))$ is formed by a multinomial in the component functions of $F'(x)$. This function is clearly continuous since we are given that the partial derivatives of the component functions of F are all continuous.

2. note we are given $F'(p)$ is invertible and hence $det(F'(p)) \neq 0$ thus the continuous function g is nonzero at p. It follows there is some open set U containing p for which $0 \notin g(U)$

3. we have $det(F'(x)) \neq 0$ for all $x \in U$ hence $F'(x)$ is invertible on U.

I would argue this is a topological argument because the key idea here is the continuity of g. Topology is the study of continuity in general.

Remark 3.1.4. *James J. Callahan's Advanced Calculus: a Geometric View, good reading.*

James J. Callahan's *Advanced Calculus: a Geometric View* has great merit in both visualization and well-thought use of linear algebraic techniques. In addition, many student will enjoy his staggered proofs where he first shows the proof for a simple low dimensional case and then proceeds to the general case.

[3]there are scientists and engineers who work with multiply-valued functions with great success, however, as a point of style if nothing else, we try to use functions in math.

Example 3.1.5. *Suppose* $F(x,y) = \big(\ \sin(y)+1,\ \sin(x)+2\ \big)$ *for* $(x,y) \in \mathbb{R}^2$. *Clearly* F *is continuously differentiable as all its component functions have continuous partial derivatives. Observe,*

$$F'(x,y) = [\ \partial_x F \mid \partial_y F\ ] = \begin{bmatrix} 0 & \cos(y) \\ \cos(x) & 0 \end{bmatrix}$$

Hence $F'(x,y)$ *is invertible at points* (x,y) *such that* $det(F'(x,y)) = -\cos(x)\cos(y) \neq 0$. *This means we* **may** *not be able to find local inverses at points* (x,y) *with* $x = \frac{1}{2}(2n+1)\pi$ *or* $y = \frac{1}{2}(2m+1)\pi$ *for some* $m,n \in \mathbb{Z}$. *Points where* $F'(x,y)$ *are singular are points where one or both of* $\sin(y)$ *and* $\sin(x)$ *reach extreme values thus the points where the Jacobian matrix are singular are in fact points where we cannot find a local inverse. Why? Because the function is clearly not 1-1 on any set which contains the points of singularity for* dF. *Continuing, recall from precalculus that sine has a standard inverse on* $[-\pi/2,\pi/2]$. *Suppose* $(x,y) \in [-\pi/2,\pi/2]^2$ *and seek to solve* $F(x,y) = (a,b)$ *for* (x,y)*:*

$$F(x,y) = \begin{bmatrix} \sin(y)+1 \\ \sin(x)+2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \ \Rightarrow\ \left\{ \begin{array}{l} \sin(y)+1 = a \\ \sin(x)+2 = b \end{array} \right\} \ \Rightarrow\ \left\{ \begin{array}{l} y = \sin^{-1}(a-1) \\ x = \sin^{-1}(b-2) \end{array} \right\}$$

It follows that $F^{-1}(a,b) = \big(\sin^{-1}(b-2),\ \sin^{-1}(a-1)\big)$ *for* $(a,b) \in [0,2]\times[1,3]$ *where you should note* $F([-\pi/2,\pi/2]^2) = [0,2]\times[1,3]$. *We've found a local inverse for* F *on the region* $[-\pi/2,\pi/2]^2$. *In other words, we just found a global inverse for the restriction of* F *to* $[-\pi/2,\pi/2]^2$. *Technically we ought not write* F^{-1}*, to be more precise we should write:*

$$(F|_{[-\pi/2,\pi/2]^2})^{-1}(a,b) = \big(\sin^{-1}(b-2),\ \sin^{-1}(a-1)\big).$$

It is customary to avoid such detail in many contexts. Inverse functions for sine, cosine, tangent etc... are good examples of this slight of langauge.

A **coordinate system** on $\mathbb{R}^n$ is an invertible mapping of $\mathbb{R}^n$ to $\mathbb{R}^n$. However, in practice the term coordinate system is used with less rigor. Often a coordinate system has various degeneracies. For example, in polar coordinates you could say $\theta = \pi/4$ or $\theta = 9\pi/4$ or generally $\theta = 2\pi k + \pi/4$ for any $k \in \mathbb{Z}$. Let's examine polar coordinates in view of the inverse function theorem.

Example 3.1.6. *Let* $T(r,\theta) = \big(\ r\cos(\theta),\ r\sin(\theta)\ \big)$ *for* $(r,\theta) \in [0,\infty)\times(-\pi/2,\pi/2)$. *Clearly* T *is continuously differentiable as all its component functions have continuous partial derivatives. To find the inverse we seek to solve* $T(r,\theta) = (x,y)$ *for* (r,θ). *Hence, consider* $x = r\cos(\theta)$ *and* $y = r\sin(\theta)$. *Note that*

$$x^2 + y^2 = r^2\cos^2(\theta) + r^2\sin^2(\theta) = r^2(\cos^2(\theta)+\sin^2(\theta)) = r^2$$

and

$$\frac{y}{x} = \frac{r\sin(\theta)}{r\cos(\theta)} = \tan(\theta).$$

It follows that $r = \sqrt{x^2+y^2}$ *and* $\theta = \tan^{-1}(y/x)$ *for* $(x,y) \in (0,\infty)\times\mathbb{R}$. *We find*

$$T^{-1}(x,y) = \left(\ \sqrt{x^2+y^2},\ \tan^{-1}(y/x)\ \right).$$

Let's see how the derivative fits with our results. Calcuate,

$$T'(r,\theta) = [\ \partial_r T \mid \partial_\theta T\ ] = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}$$

note that $det(T'(r,\theta)) = r$ hence we the inverse function theorem provides the existence of a local inverse around any point except the origin. Notice the derivative does not detect the defect in the angular coordinate. Challenge, find the inverse function for $T(r,\theta) = \big(\, r\cos(\theta),\ r\sin(\theta)\,\big)$ with $dom(T) = [0,\infty)\times(\pi/2, 3\pi/2)$. Or, find the inverse for polar coordinates in a neighborhood of $(0,-1)$.

Example 3.1.7. *Suppose $T:\mathbb{R}^3\to\mathbb{R}^3$ is defined by $T(x,y,z) = (ax, by, cz)$ for constants $a,b,c\in\mathbb{R}$ where $abc\neq 0$. Clearly T is continuously differentiable as all its component functions have continuous partial derivatives. We calculate $T'(x,y,z) = [\partial_x T|\partial_y T|\partial_z T] = [ae_1|be_2|ce_3]$. Thus $det(T'(x,y,z)) = abc \neq 0$ for all $(x,y,z)\in\mathbb{R}^3$ hence this function is locally invertible everywhere. Moreover, we calculate the inverse mapping by solving $T(x,y,z) = (u,v,w)$ for (x,y,z):*

$$(ax,by,cz) = (u,v,w) \quad\Rightarrow\quad (x,y,z) = (u/a, v/b, w/c) \quad\Rightarrow\quad \boxed{T^{-1}(u,v,w) = (u/a, v/b, w/c).}$$

Example 3.1.8. *Suppose $F:\mathbb{R}^n\to\mathbb{R}^n$ is defined by $F(x) = Ax+b$ for some matrix $A\in\mathbb{R}^{n\times n}$ and vector $b\in\mathbb{R}^n$.* **Under what conditions is such a function invertible ?**. *Since the formula for this function gives each component function as a polynomial in the n-variables we can conclude the function is continuously differentiable. You can calculate that $F'(x) = A$. It follows that a sufficient condition for local inversion is $det(A)\neq 0$. It turns out that this is also a necessary condition as $det(A) = 0$ implies the matrix A has nontrivial solutions for $Av = 0$. We say $v\in Null(A)$ iff $Av = 0$. Note if $v\in Null(A)$ then $F(v) = Av + b = b$. This is not a problem when $det(A)\neq 0$ for in that case the null space is contains just zero; $Null(A) = \{0\}$. However, when $det(A) = 0$ we learn in linear algebra that $Null(A)$ contains infinitely many vectors so F is far from injective. For example, suppose $Null(A) = span\{e_1\}$ then you can show that $F(a_1,a_2,\ldots,a_n) = F(x,a_2,\ldots,a_n)$ for all $x\in\mathbb{R}$. Hence any point will have other points nearby which output the same value under F. Suppose $det(A)\neq 0$, to calculate the inverse mapping formula we should solve $F(x) = y$ for x,*

$$y = Ax+b \quad\Rightarrow\quad x = A^{-1}(y-b) \quad\Rightarrow\quad \boxed{F^{-1}(y) = A^{-1}(y-b).}$$

Remark 3.1.9. *inverse function theorem holds for higher derivatives.*

In Munkres the inverse function theorem is given for r-times differentiable functions. In short, a C^r function with invertible differential at a point has a C^r inverse function local to the point. Edwards also has arguments for $r > 1$, see page 202 and arguments and surrounding arguments.

3.2 implicit function theorem

Consider the problem of solving $x^2+y^2 = 1$ for y as a function of x.

$$x^2+y^2 = 1 \quad\Rightarrow\quad y^2 = 1-x^2 \quad\Rightarrow\quad y = \pm\sqrt{1-x^2}.$$

A function cannot have two outputs for a single input, when we write $\pm$ in the expression above it simply indicates our ignorance as to which is chosen. Once further information is given then we may be able to choose a $+$ or a $-$. For example:

1. if $x^2+y^2 = 1$ and we want to solve for y near $(0,1)$ then $y = \sqrt{1-x^2}$ is the correct choice since $y > 0$ at the point of interest.

2. if $x^2+y^2=1$ and we want to solve for y near $(0,-1)$ then $y=-\sqrt{1-x^2}$ is the correct choice since $y<0$ at the point of interest.

3. if $x^2+y^2=1$ and we want to solve for y near $(1,0)$ then it's impossible to find a single function which reproduces $x^2+y^2=1$ on an **open disk** centered at $(1,0)$.

What is the defect of case (3.) ? The trouble is that no matter how close we zoom in to the point there are always two y-values for each given x-value. Geometrically, this suggests either we have a discontinuity, a kink, or a vertical tangent in the graph. The given problem has a vertical tangent and hopefully you can picture this with ease since its just the unit-circle. In calculus I we studied implicit differentiation, our starting point was to assume $y=y(x)$ and then we differentiated equations to work out implicit formulas for dy/dx. Take the unit-circle and differentiate both sides,

$$x^2+y^2=1 \quad \Rightarrow \quad 2x+2y\frac{dy}{dx}=0 \quad \Rightarrow \quad \frac{dy}{dx}=-\frac{x}{y}.$$

Note $\frac{dy}{dx}$ is not defined for $y=0$. It's no accident that those two points $(-1,0)$ and $(1,0)$ are precisely the points at which we cannot solve for y as a function of x. Apparently, the singularity in the derivative indicates where we may have trouble solving an equation for one variable as a function of the remaining variable.

We wish to study this problem in general. Given n-equations in $(m+n)$-unknowns when can we solve for the last n-variables as functions of the first m-variables ? Given a continuously differentiable mapping $G=(G_1,G_2,\ldots,G_n):\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}^n$ study the level set: (here $k_1,k_2,\ldots,k_n$ are constants)

$$\begin{aligned} G_1(x_1,\ldots,x_m,y_1,\ldots,y_n)&=k_1\\ G_2(x_1,\ldots,x_m,y_1,\ldots,y_n)&=k_2\\ &\vdots\\ G_n(x_1,\ldots,x_m,y_1,\ldots,y_n)&=k_n \end{aligned}$$

We wish to locally solve for $y_1,\ldots,y_n$ as functions of $x_1,\ldots x_m$. That is, find a mapping $h:\mathbb{R}^m\to\mathbb{R}^n$ such that $G(x,y)=k$ iff $y=h(x)$ near some point $(a,b)\in\mathbb{R}^m\times\mathbb{R}^n$ such that $G(a,b)=k$. In this section we use the notation $x=(x_1,x_2,\ldots x_m)$ and $y=(y_1,y_2,\ldots,y_n)$.

Before we turn to the general problem let's analyze the unit-circle problem in this notation. We are given $G(x,y)=x^2+y^2$ and we wish to find $f(x)$ such that $y=f(x)$ solves $G(x,y)=1$. Differentiate with respect to x and use the chain-rule:

$$\frac{\partial G}{\partial x}\frac{dx}{dx}+\frac{\partial G}{\partial y}\frac{dy}{dx}=0$$

We find that $\boxed{dy/dx=-G_x/G_y}=-x/y$. Given this analysis we should suspect that if we are given some level curve $G(x,y)=k$ then we may be able to solve for y as a function of x near p if $G(p)=k$ and $G_y(p)\neq 0$. This suspicion is valid and it is one of the many consequences of the implicit function theorem.

We again turn to the linearization approximation. Suppose $G(x,y)=k$ where $x\in\mathbb{R}^m$ and $y\in\mathbb{R}^n$ and suppose $G:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}^n$ is continuously differentiable. Suppose $(a,b)\in\mathbb{R}^m\times\mathbb{R}^n$ has $G(a,b)=k$. Replace G with its linearization based at (a,b):

$$G(x,y)\approx k+G'(a,b)(x-a,y-b)$$

here we have the matrix multiplication of the $n \times (m+n)$ matrix $G'(a,b)$ with the $(m+n) \times 1$ column vector $(x-a, y-b)$ to yield an n-component column vector. It is convenient to define partial derivatives with respect to a whole vector of variables,

$$\frac{\partial G}{\partial x} = \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \cdots & \frac{\partial G_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial G_n}{\partial x_1} & \cdots & \frac{\partial G_n}{\partial x_m} \end{bmatrix} \qquad \frac{\partial G}{\partial y} = \begin{bmatrix} \frac{\partial G_1}{\partial y_1} & \cdots & \frac{\partial G_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial G_n}{\partial y_1} & \cdots & \frac{\partial G_n}{\partial y_n} \end{bmatrix}$$

In this notation we can write the $n \times (m+n)$ matrix $G'(a,b)$ as the concatenation of the $n \times m$ matrix $\frac{\partial G}{\partial x}(a,b)$ and the $n \times n$ matrix $\frac{\partial G}{\partial y}(a,b)$

$$G'(a,b) = \left[\frac{\partial G}{\partial x}(a,b) \middle| \frac{\partial G}{\partial y}(a,b) \right]$$

Therefore, for points close to (a,b) we have:

$$G(x,y) \approx k + \frac{\partial G}{\partial x}(a,b)(x-a) + \frac{\partial G}{\partial y}(a,b)(y-b)$$

The nonlinear problem $G(x,y) = k$ has been (locally) replaced by the linear problem of solving what follows for y:

$$k \approx k + \frac{\partial G}{\partial x}(a,b)(x-a) + \frac{\partial G}{\partial y}(a,b)(y-b) \tag{3.1}$$

Suppose the square matrix $\frac{\partial G}{\partial y}(a,b)$ is invertible at (a,b) then we find the following approximation for the implicit solution of $G(x,y) = k$ for y as a function of x:

$$y = b - \left[\frac{\partial G}{\partial y}(a,b) \right]^{-1} \left[\frac{\partial G}{\partial x}(a,b)(x-a) \right].$$

Of course this is not a formal proof, but it does suggest that $det\left[\frac{\partial G}{\partial y}(a,b)\right] \neq 0$ is a necessary condition for solving for the y variables.

As before suppose $G : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$. Suppose we have a continuously differentiable function $h : \mathbb{R}^m \to \mathbb{R}^n$ such that $h(a) = b$ and $G(x, h(x)) = k$. We seek to find the derivative of h in terms of the derivative of G. This is a generalization of the implicit differentiation calculation we perform in calculus I. I'm including this to help you understand the notation a bit more before I state the implicit function theorem. Differentiate with respect to x_l for $l \in \mathbb{N}_m$:

$$\frac{\partial}{\partial x_l}\Big[G(x, h(x)) \Big] = \sum_{i=1}^{m} \frac{\partial G}{\partial x_i}\frac{\partial x_i}{\partial x_l} + \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_l} = \frac{\partial G}{\partial x_l} + \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_l} = 0$$

we made use of the identity $\frac{\partial x_i}{\partial x_k} = \delta_{ik}$ to squash the sum of i to the single nontrivial term and the zero on the r.h.s follows from the fact that $\frac{\partial}{\partial x_l}(k) = 0$. Concatenate these derivatives from $k = 1$ up to $k = m$:

$$\left[\frac{\partial G}{\partial x_1} + \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_1} \middle| \frac{\partial G}{\partial x_2} + \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_2} \middle| \cdots \middle| \frac{\partial G}{\partial x_m} + \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_m} \right] = [0|0|\cdots|0]$$

Properties of matrix addition allow us to parse the expression above as follows:

$$\left[\frac{\partial G}{\partial x_1} \middle| \frac{\partial G}{\partial x_2} \middle| \cdots \middle| \frac{\partial G}{\partial x_m} \right] + \left[\sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_1} \middle| \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_2} \middle| \cdots \middle| \sum_{j=1}^{n} \frac{\partial G}{\partial y_j}\frac{\partial h_j}{\partial x_m} \right] = [0|0|\cdots|0]$$

But, this reduces to

$$\frac{\partial G}{\partial x} + \left[\frac{\partial G}{\partial y}\frac{\partial h}{\partial x_1}\middle|\frac{\partial G}{\partial y}\frac{\partial h}{\partial x_2}\middle|\cdots\middle|\frac{\partial G}{\partial y}\frac{\partial h}{\partial x_m}\right] = 0 \in \mathbb{R}^{m\times n}$$

The concatenation property of matrix multiplication states $[Ab_1|Ab_2|\cdots|Ab_m] = A[b_1|b_2|\cdots|b_m]$ we use this to write the expression once more,

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\left[\frac{\partial h}{\partial x_1}\middle|\frac{\partial h}{\partial x_2}\middle|\cdots\middle|\frac{\partial h}{\partial x_m}\right] = 0 \quad \Rightarrow \quad \frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\frac{\partial h}{\partial x} = 0 \quad \Rightarrow \quad \boxed{\frac{\partial h}{\partial x} = -\frac{\partial G}{\partial y}^{-1}\frac{\partial G}{\partial x}}$$

where in the last implication we made use of the assumption that $\frac{\partial G}{\partial y}$ is invertible.

Theorem 3.2.1. *(Theorem 3.4 in Edwards's Text see pg 190)*

Let $G : dom(G) \subseteq \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in a open ball about the point (a, b) where $G(a, b) = k$ (a constant vector in $\mathbb{R}^n$). If the matrix $\frac{\partial G}{\partial y}(a, b)$ is invertible then there exists an open ball U containing a in $\mathbb{R}^m$ and an open ball W containing (a, b) in $\mathbb{R}^m \times \mathbb{R}^n$ and a continuously differentiable mapping $h : U \to \mathbb{R}^n$ such that $G(x, y) = k$ iff $y = h(x)$ for all $(x, y) \in W$. Moreover, the mapping h is the limit of the sequence of successive approximations defined inductively below

$$h_o(x) = b, \quad h_{n+1} = h_n(x) - [\tfrac{\partial G}{\partial y}(a, b)]^{-1}G(x, h_n(x)) \qquad \text{for all } x \in U.$$

We will not attempt a proof of the last sentence for the same reasons we did not pursue the details in the inverse function theorem. However, we have already derived the first step in the iteration in our study of the linearization solution.

Proof: Let $G : dom(G) \subseteq \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in a open ball B about the point (a, b) where $G(a, b) = k$ ($k \in \mathbb{R}^n$ a constant). Furthermore, assume the matrix $\frac{\partial G}{\partial y}(a, b)$ is invertible. We seek to use the inverse function theorem to prove the implicit function theorem. Towards that end consider $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ defined by $F(x, y) = (x, G(x, y))$. To begin, observe that F is continuously differentiable in the open ball B which is centered at (a, b) since G and x have continuous partials of their components in B. Next, calculate the derivative of $F = (x, G)$,

$$F'(x, y) = [\partial_x F|\partial_y F] = \left[\begin{array}{c|c} \partial_x x & \partial_y x \\ \hline \partial_x G & \partial_y G \end{array}\right] = \left[\begin{array}{c|c} I_m & 0_{m\times n} \\ \hline \partial_x G & \partial_y G \end{array}\right]$$

The determinant of the matrix above is the product of the deteminant of the blocks I_m and $\partial_y G$; $det(F'(x, y) = det(I_m)det(\partial_y G) = \partial_y G$. We are given that $\frac{\partial G}{\partial y}(a, b)$ is invertible and hence $det(\frac{\partial G}{\partial y}(a, b)) \neq 0$ thus $det(F'(x, y) \neq 0$ and we find $F'(a, b)$ is invertible. Consequently, the inverse function theorem applies to the function F at (a, b). Therefore, there exists $F^{-1} : V \subseteq \mathbb{R}^m \times \mathbb{R}^n \to U \subseteq \mathbb{R}^m \times \mathbb{R}^n$ such that F^{-1} is continuously differentiable. Note $(a, b) \in U$ and V contains the point $F(a, b) = (a, G(a, b)) = (a, k)$.

Our goal is to find the implicit solution of $G(x, y) = k$. We know that

$$F^{-1}(F(x, y)) = (x, y) \qquad \text{and} \qquad F(F^{-1}(u, v)) = (u, v)$$

for all $(x, y) \in U$ and $(u, v) \in V$. As usual to find the formula for the inverse we can solve $F(x, y) = (u, v)$ for (x, y) this means we wish to solve $(x, G(x, y)) = (u, v)$ hence $x = u$. The

formula for v is more elusive, but we know it exists by the inverse function theorem. Let's say $y = H(u, v)$ where $H : V \to \mathbb{R}^n$ and thus $F^{-1}(u, v) = (u, H(u, v))$. Consider then,

$$(u, v) = F(F^{-1}(u, v) = F(u, H(u, v)) = (u, G(u, H(u, v))$$

Let $v = k$ thus $(u, k) = (u, G(u, H(u, k))$ for all $(u, v) \in V$. Finally, define $h(u) = H(u, k)$ for all $(u, k) \in V$ and note that $k = G(u, h(u))$. In particular, $(a, k) \in V$ and at that point we find $h(a) = H(a, k) = b$ by construction. It follows that $y = h(x)$ provides a continuously differentiable solution of $G(x, y) = k$ near (a, b).

Uniqueness of the solution follows from the uniqueness for the limit of the sequence of functions described in Edwards' text on page 192. However, other arguments for uniqueness can be offered, independent of the iterative method, for instance: see page 75 of Munkres *Analysis on Manifolds*. □

Remark 3.2.2. *notation and the implementation of the implicit function theorem.*

We assumed the variables y were to be written as functions of x variables to make explicit a local solution to the equation $G(x, y) = k$. This ordering of the variables is convenient to argue the proof, however the real theorem is far more general. We can select any subset of n input variables to make up the "y" so long as $\frac{\partial G}{\partial y}$ is invertible. I will use this generalization of the formal theorem in the applications that follow. Moreover, the notations x and y are unlikely to maintain the same interpretation as in the previous pages. Finally, we will for convenience make use of the notation $y = y(x)$ to express the existence of a function f such that $y = f(x)$ when appropriate. Also, $z = z(x, y)$ means there is some function h for which $z = h(x, y)$. If this notation confuses then invent names for the functions in your problem.

Example 3.2.3. *Suppose $G(x, y, z) = x^2 + y^2 + z^2$. Suppose we are given a point (a, b, c) such that $G(a, b, c) = R^2$ for a constant R.* **Problem: For which variable can we solve? What, if any, influence does the given point have on our answer?** *Solution: to begin, we have one equation and three unknowns so we should expect to find one of the variables as functions of the remaining two variables. The implicit function theorem applies as G is continuously differentiable.*

1. *if we wish to solve $z = z(x, y)$ then we need $G_z(a, b, c) = 2c \neq 0$.*
2. *if we wish to solve $y = y(x, z)$ then we need $G_y(a, b, c) = 2b \neq 0$.*
3. *if we wish to solve $x = x(y, z)$ then we need $G_x(a, b, c) = 2a \neq 0$.*

The point has no local solution for z if it is a point on the intersection of the xy-plane and the sphere $G(x, y, z) = R^2$. Likewise, we cannot solve for $y = y(x, z)$ on the $y = 0$ slice of the sphere and we cannot solve for $x = x(y, z)$ on the $x = 0$ slice of the sphere.

Notice, algebra verifies the conclusions we reached via the implicit function theorem:

$$z = \pm\sqrt{R^2 - x^2 - y^2} \qquad y = \pm\sqrt{R^2 - x^2 - z^2} \qquad x = \pm\sqrt{R^2 - y^2 - z^2}$$

When we are at zero for one of the coordinates then we cannot choose $+$ or $-$ since we need both on an open ball intersected with the sphere centered at such a point[4]. Remember, when I talk about local solutions I mean solutions which exist over the intersection of the solution set and an open

[4]if you consider $G(x, y, z) = R^2$ as a space then the open sets on the space are taken to be the intersection with the space and open balls in $\mathbb{R}^3$. This is called the subspace topology in topology courses.

ball in the ambient space ($\mathbb{R}^3$ in this context). The preceding example is the natural extension of the unit-circle example to $\mathbb{R}^3$. A similar result is available for the n-sphere in $\mathbb{R}^n$. I hope you get the point of the example, if we have one equation then if we wish to solve for a particular variable in terms of the remaining variables then all we need is continuous differentiability of the level function and a nonzero partial derivative at the point where we wish to find the solution. Now, the implicit function theorem doesn't find the solution for us, but it does provide the existence. In the section on implicit differentiation, existence is really all we need since focus our attention on rates of change rather than actually solutions to the level set equation.

Example 3.2.4. *Consider the equation $e^{xy} + z^3 - xyz = 2$.* **Can we solve this equation for $z = z(x, y)$ near $(0, 0, 1)$?** *Let $G(x, y, z) = e^{xy} + z^3 - xyz$ and note $G(0, 0, 1) = e^0 + 1 + 0 = 2$ hence $(0, 0, 1)$ is a point on the solution set $G(x, y, z) = 2$. Note G is clearly continuously differentiable and*

$$G_z(x, y, z) = 3z^2 - xy \quad \Rightarrow \quad G_z(0, 0, 1) = 3 \neq 0$$

therefore, there exists a continuously differentiable function $h : dom(h) \subseteq \mathbb{R}^2 \to \mathbb{R}$ which solves $G(x, y, h(x, y)) = 2$ for (x, y) near $(0, 0)$ and $h(0, 0) = 1$.

I'll not attempt an explicit solution for the last example.

Example 3.2.5. *Let $(x, y, z) \in S$ iff $x + y + z = 2$ and $y + z = 1$.* **Problem: For which variable(s) can we solve?** *Solution: define $G(x, y, z) = (x + y + z, y + z)$ we wish to study $G(x, y, z) = (2, 1)$. Notice the solution set is not empty since $G(1, 0, 1) = (1 + 0 + 1, 0 + 1) = (2, 1)$ Moreover, G is continuously differentiable. In this case we have two equations and three unknowns so we expect two variables can be written in terms of the remaining free variable. Let's examine the derivative of G:*

$$G'(x, y, z) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Suppose we wish to solve $x = x(z)$ and $y = y(z)$ then we should check invertiblility of[5]

$$\frac{\partial G}{\partial (x, y)} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The matrix above is invertible hence the implicit function theorem applies and we can solve for x and y as functions of z. On the other hand, if we tried to solve for $y = y(x)$ and $z = z(x)$ then we'll get no help from the implicit function theorem as the matrix

$$\frac{\partial G}{\partial (y, z)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

is not invertible. Geometrically, we can understand these results from noting that $G(x, y, z) = (2, 1)$ is the intersection of the plane $x + y + z = 2$ and $y + z = 1$. Substituting $y + z = 1$ into $x + y + z = 2$ yields $x + 1 = 2$ hence $x = 1$ on the line of intersection. We can hardly use x as a free variable for the solution when the problem fixes x from the outset.

The method I just used to analyze the equations in the preceding example was a bit adhoc. In linear algebra we do much better for systems of linear equations. A procedure called Gaussian elimination naturally reduces a system of equations to a form in which it is manifestly obvious how

[5] this notation should not be confused with $\frac{\partial(x,y)}{\partial(u,v)}$ which is used to denote a particular determinant associated with coordinate change of integrals, or pull-back of a differential form as explained on page 100 of H.M Edward's Advanced Calculus: A differential Forms Approach, we should discuss it in a later chapter.

to eliminate redundant variables in terms of a minimal set of basic free variables. The "y" of the implicit function proof discussions plays the role of the so-called **pivotal variables** whereas the "x" plays the role of the remaining **free variables**. These variables are generally intermingled in the list of total variables so to reproduce the pattern assumed for the implicit function theorem we would need to relabel variables from the outset of a calculation. In the following example, I show how reordering the variables allows us to solve for various pairs. In short, put the dependent variable first and the independent variables second so the Gaussian elimination shows the solution with minimal effort. Here's how:

Example 3.2.6. *Consider $G(x,y,u,v) = (3x+2y-u,\ 2x+y-v) = (-1,3)$. We have two equations with four variables. Let's investigate which pairs of variables can be taken as independent or dependent variables. The most efficient method to dispatch these questions is probably Gaussian elimination. I leave it to the reader to verify that:*

$$rref\left[\begin{array}{cccc|c} 3 & 2 & -1 & 0 & -1 \\ 2 & 1 & 0 & -1 & 3 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 0 & 1 & -2 & 7 \\ 0 & 1 & -2 & 3 & -11 \end{array}\right]$$

We can immediately read from the result above that x, y can be taken to depend on u, v via the formulas:

$$x = -u + 2v + 7, \qquad y = 2u - 3v - 11$$

On the other hand, if we order the variables (u,v,x,y) then Gaussian elimination gives:

$$rref\left[\begin{array}{cccc|c} -1 & 0 & 3 & 2 & -1 \\ 0 & -1 & 2 & 1 & 3 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 0 & -3 & -2 & 1 \\ 0 & 1 & -2 & -1 & -3 \end{array}\right]$$

Therefore, we find $u(x,y)$ and $v(x,y)$ as follows:

$$u = 3x + 2y + 1, \qquad v = 2x + y - 3.$$

To solve for x, u as functions of y, v consider:

$$rref\left[\begin{array}{cccc|c} 3 & -1 & 2 & 0 & -1 \\ 2 & 0 & 1 & -1 & 3 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 0 & 1/2 & -1/2 & 3/2 \\ 0 & 1 & -1/2 & -3/2 & 11/2 \end{array}\right]$$

From which we can read,

$$x = -y/2 + v/2 + 3/2, \qquad u = y/2 + 3v/2 + 11/2.$$

I could solve the problem below in the efficient style above, but I will instead follow the method in which we discussed in the paragraphs surrounding Equation 3.1. In contrast to the general case, because the problem is linear the solution of Equation 3.1 is also a solution of the actual problem.

Example 3.2.7. *Solve the following system of equations near (1,2,3,4,5).*

$$G(x,y,z,a,b) = \begin{bmatrix} x+y+z+2a+2b \\ x+0+2z+2a+3b \\ 3x+2y+z+3a+4b \end{bmatrix} = \begin{bmatrix} 24 \\ 30 \\ 42 \end{bmatrix}$$

Differentiate to find the Jacobian:

$$G'(x,y,z,a,b) = \begin{bmatrix} 1 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 & 3 \\ 3 & 2 & 1 & 3 & 4 \end{bmatrix}$$

Let us solve $G(x,y,z,a,b) = (24,30,42)$ for $x(a,b), y(a,b), z(a,b)$ by the method of Equation 3.1. I'll omit the point-dependence of the Jacobian since it clearly has none.

$$G(x,y,z,a,b) = \begin{bmatrix} 24 \\ 30 \\ 42 \end{bmatrix} + \frac{\partial G}{\partial(x,y,z)} \begin{bmatrix} x-1 \\ y-2 \\ z-3 \end{bmatrix} + \frac{\partial G}{\partial(a,b)} \begin{bmatrix} a-4 \\ b-5 \end{bmatrix}$$

Let me make the notational chimera above explicit:

$$G(x,y,z,a,b) = \begin{bmatrix} 24 \\ 30 \\ 42 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \\ z-3 \end{bmatrix} + \begin{bmatrix} 2 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} a-4 \\ b-5 \end{bmatrix}$$

To solve $G(x,y,z,a,b) = (24,30,42)$ for (x,y,z) we may use the expression above. After a little calculation one finds:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix}^{-1} = \frac{1}{3} \begin{bmatrix} -4 & 1 & 2 \\ 5 & -2 & -1 \\ 2 & 1 & -1 \end{bmatrix}$$

The constant term cancels and we find:

$$\begin{bmatrix} x-1 \\ y-2 \\ z-3 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} -4 & 1 & 2 \\ 5 & -2 & -1 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} a-4 \\ b-5 \end{bmatrix}$$

Multiplying the matrices gives:

$$\begin{bmatrix} x-1 \\ y-2 \\ z-3 \end{bmatrix} = -\frac{1}{3} \begin{bmatrix} 0 & 3 \\ 3 & 0 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} a-4 \\ b-5 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} a-4 \\ b-5 \end{bmatrix} = \begin{bmatrix} 5-b \\ 4-a \\ 9-a-b \end{bmatrix}$$

Therefore,

$$\boxed{x = 6-b, \quad y = 6-a, \quad z = 12-a-b.}$$

Is it possible to solve for any triple of the variables x,y,z,a,b for the given system? In fact, no. Let me explain by linear algebra. We can calculate: the augmented coefficient matrix for $G(x,y,z,a,b) = (24,30,42)$ Gaussian eliminates as follows:

$$rref \left[\begin{array}{ccccc|c} 1 & 1 & 1 & 2 & 2 & 24 \\ 1 & 0 & 2 & 2 & 3 & 30 \\ 3 & 2 & 1 & 3 & 4 & 42 \end{array}\right] = \left[\begin{array}{ccccc|c} 1 & 0 & 0 & 0 & 1 & 6 \\ 0 & 1 & 0 & 1 & 0 & 6 \\ 0 & 0 & 1 & 1 & 1 & 12 \end{array}\right].$$

First, note this is consistent with the answer we derived above. Second, examine the columns of $rref[G']$. You can ignore the 6-th column in the interest of this thought extending to nonlinear systems. The question of the suitability of a triple amounts to the invertibility of the submatrix of G' which corresponds to the triple. Examine:

$$\frac{\partial G}{\partial(y,z,a)} = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ 2 & 1 & 3 \end{bmatrix}, \qquad \frac{\partial G}{\partial(x,z,b)} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 3 & 1 & 4 \end{bmatrix}$$

both of these are clearly singular since the third column is the sum of the first two columns. Alternatively, you can calculate the determinant of each of the matrices above is zero. In contrast,

$$\frac{\partial G}{\partial(z,a,b)} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 2 & 2 \\ 1 & 3 & 4 \end{bmatrix}$$

is non-singular. How to I know there is no linear dependence? Well, we could calculate the determinant is $1(8-6)-2(8-2)+2(6-2)=-2\neq 0$. *Or, we could examine the row reduction above. The column correspondance property*[6] *states that linear dependences amongst columns of a matrix are preserved under row reduction. This means we can easily deduce dependence (if there is any) from the reduced matrix. Observe that column* 4 *is clearly the sum of columns* 2 *and* 3. *Likewise, column* 5 *is the sum of columns* 1 *and* 3. *On the other hand, columns* $3,4,5$ *admit no linear dependence. In general, more calculation would be required to "see" the independence of the far right columns. One reorders the columns and performs a new reduction to ascertain dependence. No such calculation is needed here since the problem is not that complicated.*

I find calculating the determinant of sub-Jacobian matrices is the simplest way for most students to quickly understand. I'll showcase this method in a series of examples attached to a later section. I have made use of some matrix theory in this section. If you didn't learn it in linear (or haven't taken linear yet) it's worth learning. These are nice tools to keep for later problems in life.

Remark 3.2.8. *independent constraints*

Gaussian elimination on a system of linear equations may produce a row of zeros. For example, $x+y=0$ and $2x+2y=0$ gives $\text{rref}\begin{bmatrix}1&1&0\\2&2&0\end{bmatrix}=\begin{bmatrix}1&1&0\\0&0&0\end{bmatrix}$. The reason for this is quite obvious: the equations consider are not indpendent. In fact the second equation is a scalar multiple of the first. Generally, if there is some linear-dependence in a set of equations then we can expect this will happen. Although, if the equations are inhomogenous the last column might not be trivial because the system could be inconsistent (for example $x+y=1$ and $2x+2y=5$).

Consider $G:\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}^n$. As we linearize $G=k$ we arrive at a homogeneous system which can be written briefly as $G'\vec{r}=0$ (think about Equation 3.1 with the k's cancelled). We should **define** $G(\vec{r})=k$ is a **system of** n **independent equations at** $\vec{r}_o$ iff $G(\vec{r}_o)=k$ and $\text{rref}[G'(\vec{r}_o)]$ has zero row. In other terminology, we could say the system of (possibly nonlinear) equations $G(\vec{r})=k$ is built from n-independent equations near $\vec{r}_o$ iff the Jacobian matrix has **full-rank** at $\vec{r}_o$. If this full-rank condition is met then we can solve for n of the variables in terms of the remaining p variables. In general there will be many choices of how to do this, and some choices will be forbidden as we have seen in the examples already.

[6] I like to call it the CCP in my linear notes

3.3 implicit differentiation

Enough theory, let's calculate. In this section I apply previous theoretical constructions to specific problems. I also introduce standard notation for "constrained" partial differentiation which is also sometimes called "partial differentiation with a side condition". The typical problem is the following: given equations:

$$\begin{aligned} G_1(x_1,\dots,x_m,y_1,\dots,y_n) &= k_1 \\ G_2(x_1,\dots,x_m,y_1,\dots,y_n) &= k_2 \\ &\vdots \\ G_n(x_1,\dots,x_m,y_1,\dots,y_n) &= k_n \end{aligned}$$

calculate partial derivative of dependent variables with respect to independent variables. Continuing with the notation of the implicit function discussion we'll assume that y will be dependent on x. I want to recast some of our arguments via differentials[7]. Take the total differential of each equation above,

$$\begin{aligned} dG_1(x_1,\dots,x_m,y_1,\dots,y_n) &= 0 \\ dG_2(x_1,\dots,x_m,y_1,\dots,y_n) &= 0 \\ &\vdots \\ dG_n(x_1,\dots,x_m,y_1,\dots,y_n) &= 0 \end{aligned}$$

Hence,

$$\begin{aligned} \partial_{x_1}G_1dx_1+\cdots+\partial_{x_m}G_1dx_m+\partial_{y_1}G_1dy_1+\cdots+\partial_{y_n}G_1dy_n &= 0 \\ \partial_{x_1}G_2dx_1+\cdots+\partial_{x_m}G_2dx_m+\partial_{y_1}G_2dy_1+\cdots+\partial_{y_n}G_2dy_n &= 0 \\ &\vdots \\ \partial_{x_1}G_ndx_1+\cdots+\partial_{x_m}G_ndx_m+\partial_{y_1}G_ndy_1+\cdots+\partial_{y_n}G_ndy_n &= 0 \end{aligned}$$

Notice, this can be nicely written in column vector notation as:

$$\partial_{x_1}Gdx_1+\cdots+\partial_{x_m}Gdx_m+\partial_{y_1}Gdy_1+\cdots+\partial_{y_n}Gdy_n=0$$

Or, in matrix notation:

$$[\partial_{x_1}G|\cdots|\partial_{x_m}G]\begin{bmatrix} dx_1 \\ \vdots \\ dx_m \end{bmatrix}+[\partial_{y_1}G|\cdots|\partial_{y_n}G]\begin{bmatrix} dy_1 \\ \vdots \\ dy_n \end{bmatrix}=0$$

Finally, solve for dy, we assume $[\partial_{y_1}G|\cdots|\partial_{y_n}G]^{-1}$ exists,

$$\begin{bmatrix} dy_1 \\ \vdots \\ dy_n \end{bmatrix}=-[\partial_{y_1}G|\cdots|\partial_{y_n}G]^{-1}[\partial_{x_1}G|\cdots|\partial_{x_m}G]\begin{bmatrix} dx_1 \\ \vdots \\ dx_m \end{bmatrix}$$

Given all of this we can calculate $\frac{\partial y_i}{\partial x_j}$ by simply reading the coeffient dx_j in the i-th row. I will make this idea quite explicit in the examples that follow.

[7] in contrast, In the previous section we mostly used derivative notation

Example 3.3.1. *Let's return to a common calculus III problem. Suppose $F(x,y,z) = k$ for some constant k.* **Find partial derivatives of x, y or z with repsect to the remaining variables.** *Solution: I'll use the method of differentials once more:*

$$dF = F_x dx + F_y dy + F_z dz = 0$$

We can solve for dx, dy or dz provided F_x, F_y or F_z is nonzero respective and these differential expressions reveal various partial derivatives of interest:

$$dx = -\frac{F_y}{F_x}dy - \frac{F_z}{F_x}dz \qquad \Rightarrow \qquad \frac{\partial x}{\partial y} = -\frac{F_y}{F_x} \;\; \& \;\; \frac{\partial x}{\partial z} = -\frac{F_z}{F_x}$$

$$dy = -\frac{F_x}{F_y}dx - \frac{F_z}{F_y}dz \qquad \Rightarrow \qquad \frac{\partial y}{\partial x} = -\frac{F_x}{F_y} \;\; \& \;\; \frac{\partial y}{\partial z} = -\frac{F_z}{F_y}$$

$$dz = -\frac{F_x}{F_z}dx - \frac{F_y}{F_z}dy \qquad \Rightarrow \qquad \frac{\partial z}{\partial x} = -\frac{F_x}{F_z} \;\; \& \;\; \frac{\partial z}{\partial y} = -\frac{F_y}{F_z}$$

In each case above, the implicit function theorem allows us to solve for one variable in terms of the remaining two. If the partial derivative of F in the denominator are zero then the implicit function theorem does not apply and other thoughts are required. Often calculus text give the following as a homework problem:

$$\frac{\partial x}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial x} = -\frac{F_y}{F_x}\frac{F_z}{F_y}\frac{F_x}{F_z} = -1.$$

In the equation above we have x appear as a dependent variable on y, z and also as an independent variable for the dependent variable z. These mixed expressions are actually of interest to engineering and physics. The less mbiguous notation below helps better handle such expressions:

$$\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_y = -1.$$

In each part of the expression we have clearly denoted which variables are taken to depend on the others and in turn what sort of partial derivative we mean to indicate. Partial derivatives are not taken alone, they must be done in concert with an understanding of the totality of the indpendent variables for the problem. We hold all the remaining indpendent variables fixed as we take a partial derivative.

The explicit independent variable notation is more important for problems where we can choose more than one set of indpendent variables for a given dependent variables. In the example that follows we study $w = w(x,y)$ but we could just as well consider $w = w(x,z)$. Generally it will not be the case that $\left(\frac{\partial w}{\partial x}\right)_y$ is the same as $\left(\frac{\partial w}{\partial x}\right)_z$. In calculation of $\left(\frac{\partial w}{\partial x}\right)_y$ we hold y constant as we vary x whereas in $\left(\frac{\partial w}{\partial x}\right)_z$ we hold z constant as we vary x. There is no reason these ought to be the same[8].

Example 3.3.2. *Suppose $x+y+z+w = 3$ and $x^2-2xyz+w^3 = 5$.* **Calculate partial derivatives of z and w with respect to the independent variables x, y.** *Solution: we begin by calculation of the differentials of both equations:*

$$dx + dy + dz + dw = 0$$
$$(2x - 2yz)dx - 2xzdy - 2xydz + 3w^2 dw = 0$$

[8] a good exercise would be to do the example over but instead aim to calculate partial derivatives for y, w with respect to independent variables x, z

We can solve for (dz, dw). In this calculation we can treat the differentials as formal variables.

$$\begin{aligned} dz + dw &= -dx - dy \\ -2xydz + 3w^2dw &= -(2x - 2yz)dx + 2xzdy \end{aligned}$$

I find matrix notation is often helpful,

$$\begin{bmatrix} 1 & 1 \\ -2xy & 3w^2 \end{bmatrix} \begin{bmatrix} dz \\ dw \end{bmatrix} = \begin{bmatrix} -dx - dy \\ -(2x - 2yz)dx + 2xzdy \end{bmatrix}$$

Use Kramer's rule, multiplication by inverse, substitution, adding/subtracting equations etc... whatever technique of solving linear equations you prefer. Our goal is to solve for dz and dw in terms of dx and dy. I'll use Kramer's rule this time:

$$dz = \frac{det \left[\begin{array}{c|c} -dx - dy & 1 \\ -(2x - 2yz)dx + 2xzdy & 3w^2 \end{array}\right]}{det \begin{bmatrix} 1 & 1 \\ -2xy & 3w^2 \end{bmatrix}} = \frac{3w^2(-dx - dy) + (2x - 2yz)dx - 2xzdy}{3w^2 + 2xy}$$

Collecting terms,

$$dz = \left(\frac{-3w^2 + 2x - 2yz}{3w^2 + 2xy}\right)dx + \left(\frac{-3w^2 - 2xz}{3w^2 + 2xy}\right)dy$$

From the expression above we can read various implicit derivatives,

$$\boxed{\left(\frac{\partial z}{\partial x}\right)_y = \frac{-3w^2 + 2x - 2yz}{3w^2 + 2xy} \qquad \& \qquad \left(\frac{\partial z}{\partial y}\right)_x = \frac{-3w^2 - 2xz}{3w^2 + 2xy}}$$

The notation above indicates that z is understood to be a function of independent variables x, y. $\left(\frac{\partial z}{\partial x}\right)_y$ means we take the derivative of z with respect to x while holding y fixed. The appearance of the dependent variable w can be removed by using the equations $G(x, y, z, w) = (3, 5)$. Similar ambiguities exist for implicit differentiation in calculus I. Apply Kramer's rule once more to solve for dw:

$$dw = \frac{det \left[\begin{array}{c|c} 1 & -dx - dy \\ -2xy & -(2x - 2yz)dx + 2xzdy \end{array}\right]}{det \begin{bmatrix} 1 & 1 \\ -2xy & 3w^2 \end{bmatrix}} = \frac{-(2x - 2yz)dx + 2xzdy - 2xy(dx + dy)}{3w^2 + 2xy}$$

Collecting terms,

$$dw = \left(\frac{-2x + 2yz - 2xy}{3w^2 + 2xy}\right)dx + \left(\frac{2xzdy - 2xydy}{3w^2 + 2xy}\right)dy$$

We can read the following from the differential above:

$$\boxed{\left(\frac{\partial w}{\partial x}\right)_y = \frac{-2x + 2yz - 2xy}{3w^2 + 2xy} \qquad \& \qquad \left(\frac{\partial w}{\partial y}\right)_x = \frac{2xzdy - 2xydy}{3w^2 + 2xy}}$$

You should ask: where did we use the implicit function theorem in the preceding example? Notice our underlying hope is that we can solve for $z = z(x, y)$ and $w = w(x, y)$. The implicit function theorem states this is possible precisely when $\frac{\partial G}{\partial(z,w)} = \begin{bmatrix} 1 & 1 \\ -2xy & 3w^2 \end{bmatrix}$ is non singular. Interestingly this is the same matrix we must consider to isolate dz and dw. The calculations of the example are only meaningful if the $det \begin{bmatrix} 1 & 1 \\ -2xy & 3w^2 \end{bmatrix} \neq 0$. In such a case the implicit function theorem applies and it is reasonable to suppose z, w can be written as functions of x, y.

3.3.1 computational techniques for partial differentiation with side conditions

In this section I show you how I teach this to calculus III. In other words, we set-aside the explicit mention of the implicit function theorem and work out some typical calculations. If one desires rigor then the answer is found from the implicit function theorems careful application, that is how to justify what follows. These notes are taken from my calculus III notes, but I thought it wise to include them here since most calculus texts do not bother to show these calculations (which is sad since they actually matter to the application of multivariate analysis to many real world applications) To begin, we define[9] the total differential.

Definition 3.3.3.

If $f = f(x_1, x_2, \ldots, x_n)$ then $df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_n}dx_n$.

Example 3.3.4. *Suppose $E = pv + t^2$ then $dE = vdp + pdv + 2tdt$. In this example the dependent variable is E whereas the independent variables are p, v and t.*

Example 3.3.5. Problem: *what are $\partial F/\partial x$ and $\partial F/\partial y$ if we know that $F = F(x, y)$ and $dF = (x^2 + y)dx - \cos(xy)dy$.*
Solution: *if $F = F(x, y)$ then the total differential has the form $dF = F_x dx + F_y dy$. We simply compare the general form to the given $dF = (x^2 + y)dx - \cos(xy)dy$ to obtain:*

$$\frac{\partial F}{\partial x} = x^2 + y, \qquad \frac{\partial F}{\partial y} = -\cos(xy).$$

Example 3.3.6. *Suppose $w = xyz$ then $dw = yzdx + xzdy + xydz$. On the other hand, we can solve for $z = z(x, y, w)$*

$$z = \frac{w}{xy} \qquad \Rightarrow \qquad dz = -\frac{w}{x^2y}dx - \frac{w}{xy^2}dy + \frac{1}{xy}dw. \quad \star$$

If we solve $dw = yzdx + xzdy + xydz$ directly for dz we obtain:

$$dz = -\frac{z}{x}dx - \frac{z}{y}dy + \frac{1}{xy}dw \quad \star\star.$$

Are $\star$ and $\star\star$ consistent? Well, yes. Note $\frac{w}{x^2y} = \frac{xyz}{x^2y} = \frac{z}{x}$ and $\frac{w}{xy^2} = \frac{xyz}{xy^2} = \frac{z}{y}$.

Which variables are independent/dependent in the example above? It depends. In this initial portion of the example we treated x, y, z as independent whereas w was dependent. But, in the last half we treated x, y, w as independent and z was the dependent variable. Consider this, if I ask you what the value of $\frac{\partial z}{\partial x}$ is in the example above then this question is ambiguous!

$$\underbrace{\frac{\partial z}{\partial x} = 0}_{z \text{ indpendent of } x} \qquad \text{verses} \qquad \underbrace{\frac{\partial z}{\partial x} = \frac{-z}{x}}_{z \text{ depends on } x}$$

Obviously this sort of ambiguity is rather unpleasant. A natural solution to this trouble is simply to write a bit more when variables are used in multiple contexts. In particular,

$$\underbrace{\left.\frac{\partial z}{\partial x}\right|_{y,z} = 0}_{\text{means } x,y,z \text{ independent}} \qquad \text{is different than} \qquad \underbrace{\left.\frac{\partial z}{\partial x}\right|_{y,w} = \frac{-z}{x}}_{\text{means } x,y,w \text{ independent}} .$$

[9] I invite the reader to verify the notation "defined" in this section is in fact totally sympatico with our previous definitions

The key concept is that all the other independent variables are held fixed as an indpendent variable is partial differentiated. Holding y, z fixed as x varies means z does not change hence $\frac{\partial z}{\partial x}\big|_{y,z} = 0$. On the other hand, if we hold y, w fixed as x varies then the change in z need not be trivial; $\frac{\partial z}{\partial x}\big|_{y,w} = \frac{-z}{x}$. Let me expand on how this notation interfaces with the total differential.

Definition 3.3.7.

If w, x, y, z are variables then

$$dw = \left.\frac{\partial w}{\partial x}\right|_{y,z} dx + \left.\frac{\partial w}{\partial y}\right|_{x,z} dy + \left.\frac{\partial w}{\partial z}\right|_{x,y} dz.$$

Alternatively,

$$dx = \left.\frac{\partial x}{\partial w}\right|_{y,z} dw + \left.\frac{\partial x}{\partial y}\right|_{w,z} dy + \left.\frac{\partial x}{\partial z}\right|_{w,y} dz.$$

The larger idea here is that we can identify partial derivatives from the coefficients in equations of differentials. I'd say a differential equation but you might get the wrong idea... Incidentally, there is a whole theory of solving differential equations by clever use of differentials, I have books if you are interested.

Example 3.3.8. *Suppose $w = x + y + z$ and $x + y = wz$ then calculate $\frac{\partial w}{\partial x}\big|_y$ and $\frac{\partial w}{\partial x}\big|_z$. Notice we must choose dependent and independent variables to make sense of partial derivatives in question.*

1. *suppose w, z both depend on x, y. Calculate,*

$$\left.\frac{\partial w}{\partial x}\right|_y = \left.\frac{\partial}{\partial x}\right|_y (x + y + z) = \left.\frac{\partial x}{\partial x}\right|_y + \left.\frac{\partial y}{\partial x}\right|_y + \left.\frac{\partial z}{\partial x}\right|_y = 1 + 0 + \left.\frac{\partial z}{\partial x}\right|_y \quad \star$$

To calculate further we need to eliminate w by substituting $w = x + y + z$ into $x + y = wz$; thus $x + y = (x + y + z)z$ hence $dx + dy = (dx + dy + dz)z + (x + y + z)dz$

$$(2z + x + y)dz = (1 - z)dx + (1 - z)dy \quad \star\star$$

Therefore,

$$dz = \frac{1 - z}{2z + x + y}dx + \frac{1 - z}{2z + x + y}dy = \left.\frac{\partial z}{\partial x}\right|_y dx + \left.\frac{\partial z}{\partial y}\right|_x dy \quad \Rightarrow \quad \left.\frac{\partial z}{\partial x}\right|_y = \frac{1 - z}{2z + x + y}.$$

Returning to $\star$ we derive

$$\boxed{\left.\frac{\partial w}{\partial x}\right|_y = 1 + \frac{1 - z}{2z + x + y}.}$$

2. *suppose w, y both depend on x, z. Calculate,*

$$\left.\frac{\partial w}{\partial x}\right|_z = \left.\frac{\partial}{\partial x}\right|_z (x + y + z) = \left.\frac{\partial x}{\partial x}\right|_z + \left.\frac{\partial y}{\partial x}\right|_z + \left.\frac{\partial z}{\partial x}\right|_z = 1 + \left.\frac{\partial y}{\partial x}\right|_z + 0$$

To complete this calculation we need to eliminate w as before, using $\star\star$,

$$(1 - z)dy = (1 - z)dx - (2z + x + y)dz \quad \Rightarrow \quad \left.\frac{\partial y}{\partial x}\right|_z = 1.$$

Therefore,

$$\boxed{\left.\frac{\partial w}{\partial x}\right|_z = 2.}$$

I hope you can begin to see how the game is played. Basically the example above generalizes the idea of implicit differentiation to several equations of many variables. This is actually a pretty important type of calculation for engineering. The study of thermodynamics is full of variables which are intermittently used as either dependent or independent variables. The so-called equation of state can be given in terms of about a dozen distinct sets of state variables.

Example 3.3.9. *The ideal gas law states that for a fixed number of particles n the pressure P, volume V and temperature T are related by $PV = nRT$ where R is a constant. Calculate,*

$$\left.\frac{\partial P}{\partial V}\right|_T = \frac{\partial}{\partial V}\left[\frac{nRT}{V}\right]\Big|_T = -\frac{nRT}{V^2},$$

$$\left.\frac{\partial V}{\partial T}\right|_P = \frac{\partial}{\partial T}\left[\frac{nRT}{P}\right]\Big|_T = \frac{nR}{P},$$

$$\left.\frac{\partial T}{\partial P}\right|_V = \frac{\partial}{\partial P}\left[\frac{PV}{nR}\right]\Big|_T = \frac{V}{nR}.$$

You might expect that $\left.\frac{\partial P}{\partial V}\right|_T \left.\frac{\partial V}{\partial T}\right|_P \left.\frac{\partial T}{\partial P}\right|_V = 1$. Is it true?

$$\left.\frac{\partial P}{\partial V}\right|_T \left.\frac{\partial V}{\partial T}\right|_P \left.\frac{\partial T}{\partial P}\right|_V = -\frac{nRT}{V^2}\cdot\frac{nR}{P}\cdot\frac{V}{nR} = \frac{-nRT}{PV} = -1.$$

This is an example where naive cancellation of partials fails.

Example 3.3.10. *Suppose $F(x,y) = 0$ then $dF = F_x dx + F_y dy = 0$ and it follows that $dx = -\frac{F_y}{F_x}dy$ or $dy = -\frac{F_x}{F_y}dx$. Hence, $\frac{\partial x}{\partial y} = -\frac{F_y}{F_x}$ and $\frac{\partial y}{\partial x} = -\frac{F_x}{F_y}$. Therefore,*

$$\frac{\partial x}{\partial y}\frac{\partial y}{\partial x} = \frac{F_y}{F_x}\cdot\frac{F_x}{F_y} = 1$$

for (x,y) such that $F_x \neq 0$ and $F_y \neq 0$. The condition $F_x \neq 0$ suggests we can solve for $y = y(x)$ whereas the condition $F_y \neq 0$ suggests we can solve for $x = x(y)$.

3.4 the constant rank theorem

The implicit function theorem required we work with independent constraints. However, one does not always have that luxury. There is a theorem which deals with the slightly more general case. The base idea is that if the Jacobian has rank k then it locally injects a k-dimensional image into the codomain. If we are using a map as a parametrization then the rank k condition suggests the mapping does parametrize a k-fold, at least locally. On the other hand, if we are using the map to define a space as a level set then $F : \mathbb{R}^n \to \mathbb{R}^p$ has $F^{-1}(C)$ as a $(n-k)$-fold. Previously, we would have insisted $k = p$. I've run out of time for 2013 notes[10] so sadly I have no reference for this claim. In any event, theorems aside, I think the red comments are worth some discussion.

Remark 3.4.1.

I have put remarks about the rank of the derivative in red for the examples below.

[10] in 2017 it seems the situation has not changed, perhaps we'll find it together this semester

Example 3.4.2. *Let $f(t) = (t, t^2, t^3)$ then $f'(t) = (1, 2t, 3t^2)$. In this case we have*

$$f'(t) = [df_t] = \begin{bmatrix} 1 \\ 2t \\ 3t^2 \end{bmatrix}$$

The Jacobian here is a single column vector. It has rank 1 provided the vector is nonzero. We see that $f'(t) \neq (0,0,0)$ for all $t \in \mathbb{R}$. This corresponds to the fact that this space curve has a well-defined tangent line for each point on the path.

Example 3.4.3. *Let $f(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$ be a mapping from $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$. I'll denote the coordinates in the domain by $(x_1, x_2, x_3, y_1, y_2, y_3)$ thus $f(\vec{x}, \vec{y}) = x_1y_1 + x_2y_2 + x_3y_3$. Calculate,*

$$[df_{(\vec{x},\vec{y})}] = \nabla f(\vec{x}, \vec{y})^T = [y_1, y_2, y_3, x_1, x_2, x_3]$$

The Jacobian here is a single row vector. It has rank 6 provided all entries of the input vectors are nonzero.

Example 3.4.4. *Let $f(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$ be a mapping from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. I'll denote the coordinates in the domain by $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ thus $f(\vec{x}, \vec{y}) = \sum_{i=1}^n x_iy_i$. Calculate,*

$$\frac{\partial}{x_j}\left[\sum_{i=1}^n x_iy_i\right] = \sum_{i=1}^n \frac{\partial x_i}{x_j} y_i = \sum_{i=1}^n \delta_{ij} y_i = y_j$$

Likewise,

$$\frac{\partial}{y_j}\left[\sum_{i=1}^n x_iy_i\right] = \sum_{i=1}^n x_i \frac{\partial y_i}{y_j} = \sum_{i=1}^n x_i \delta_{ij} = x_j$$

Therefore, noting that $\nabla f = (\partial_{x_1} f, \ldots, \partial_{x_n} f, \partial_{y_1} f, \ldots, \partial_{y_n} f)$,

$$[df_{(\vec{x},\vec{y})}]^T = (\nabla f)(\vec{x}, \vec{y}) = \vec{y} \times \vec{x} = (y_1, \ldots, y_n, x_1, \ldots, x_n)$$

The Jacobian here is a single row vector. It has rank 2n provided all entries of the input vectors are nonzero.

Example 3.4.5. *Suppose $F(x, y, z) = (xyz, y, z)$ we calculate,*

$$\frac{\partial F}{\partial x} = (yz, 0, 0) \qquad \frac{\partial F}{\partial y} = (xz, 1, 0) \qquad \frac{\partial F}{\partial z} = (xy, 0, 1)$$

Remember these are actually column vectors in my sneaky notation; $(v_1, \ldots, v_n) = [v_1, \ldots, v_n]^T$. This means the **derivative** *or* **Jacobian matrix** *of F at (x, y, z) is*

$$F'(x, y, z) = [dF_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note, $rank(F'(x, y, z)) = 3$ for all $(x, y, z) \in \mathbb{R}^3$ such that $y, z \neq 0$. There are a variety of ways to see that claim, one way is to observe $det[F'(x, y, z)] = yz$ and this determinant is nonzero so long as neither y nor z is zero. In linear algebra we learn that a square matrix is invertible iff it has nonzero determinant iff it has linearly indpendent column vectors.

Example 3.4.6. *Suppose* $F(x,y,z) = (x^2+z^2, yz)$ *we calculate,*

$$\frac{\partial F}{\partial x} = (2x, 0) \qquad \frac{\partial F}{\partial y} = (0, z) \qquad \frac{\partial F}{\partial z} = (2z, y)$$

The derivative is a 2×3 *matrix in this example,*

$$F'(x,y,z) = [dF_{(x,y,z)}] = \begin{bmatrix} 2x & 0 & 2z \\ 0 & z & y \end{bmatrix}$$

The maximum rank for F' *is* 2 *at a particular point* (x,y,z) *because there are at most two linearly independent vectors in* $\mathbb{R}^2$*. You can consider the three square submatrices to analyze the rank for a given point. If any one of these is nonzero then the rank (dimension of the column space) is two.*

$$M_1 = \begin{bmatrix} 2x & 0 \\ 0 & z \end{bmatrix} \qquad M_2 = \begin{bmatrix} 2x & 2z \\ 0 & y \end{bmatrix} \qquad M_3 = \begin{bmatrix} 0 & 2z \\ z & y \end{bmatrix}$$

We'll need either $det(M_1) = 2xz \neq 0$ *or* $det(M_2) = 2xy \neq 0$ *or* $det(M_3) = -2z^2 \neq 0$*. I believe the only point where all three of these fail to be true simulataneously is when* $x = y = z = 0$*. This mapping has maximal rank at all points except the origin.*

Example 3.4.7. *Suppose* $F(x,y) = (x^2+y^2, xy, x+y)$ *we calculate,*

$$\frac{\partial F}{\partial x} = (2x, y, 1) \qquad \frac{\partial F}{\partial y} = (2y, x, 1)$$

The derivative is a 3×2 *matrix in this example,*

$$F'(x,y) = [dF_{(x,y)}] = \begin{bmatrix} 2x & 2y \\ y & x \\ 1 & 1 \end{bmatrix}$$

The maximum rank is again 2*, this time because we only have two columns. The rank will be two if the columns are not linearly dependent. We can analyze the question of rank a number of ways but I find determinants of submatrices a comforting tool in these sort of questions. If the columns are linearly dependent then all three sub-square-matrices of* F' *will be zero. Conversely, if even one of them is nonvanishing then it follows the columns must be linearly independent. The submatrices for this problem are:*

$$M_1 = \begin{bmatrix} 2x & 2y \\ y & x \end{bmatrix} \qquad M_2 = \begin{bmatrix} 2x & 2y \\ 1 & 1 \end{bmatrix} \qquad M_3 = \begin{bmatrix} y & x \\ 1 & 1 \end{bmatrix}$$

You can see $det(M_1) = 2(x^2-y^2)$*,* $det(M_2) = 2(x-y)$ *and* $det(M_3) = y-x$*. Apparently we have* $rank(F'(x,y,z)) = 2$ *for all* $(x,y) \in \mathbb{R}^2$ *with* $y \neq x$*. In retrospect this is not surprising.*

Example 3.4.8. *Let* $F(x,y) = (x, y, \sqrt{R^2-x^2-y^2})$ *for a constant* R*. We calculate,*

$$\nabla\sqrt{R^2-x^2-y^2} = \left(\frac{-x}{\sqrt{R^2-x^2-y^2}}, \frac{-y}{\sqrt{R^2-x^2-y^2}} \right)$$

Also, $\nabla x = (1,0)$ *and* $\nabla y = (0,1)$ *thus*

$$F'(x,y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{\sqrt{R^2-x^2-y^2}} & \frac{-y}{\sqrt{R^2-x^2-y^2}} \end{bmatrix}$$

This matrix clearly has rank 2 where is is well-defined. Note that we need $R^2 - x^2 - y^2 > 0$ for the derivative to exist. Moreover, we could define $G(y,z) = (\sqrt{R^2 - y^2 - z^2}, y, z)$ and calculate,

$$G'(y,z) = \begin{bmatrix} 1 & 0 \\ \frac{-y}{\sqrt{R^2-y^2-z^2}} & \frac{-z}{\sqrt{R^2-y^2-z^2}} \\ 0 & 1 \end{bmatrix}.$$

Observe that $G'(y,z)$ exists when $R^2 - y^2 - z^2 > 0$. Geometrically, F parametrizes the sphere above the equator at $z = 0$ whereas G parametrizes the right-half of the sphere with $x > 0$. These parametrizations overlap in the first octant where both x and z are positive. In particular, $dom(F') \cap dom(G') = \{(x,y) \in \mathbb{R}^2 \mid x, y > 0 \text{ and } x^2 + y^2 < R^2\}$

Example 3.4.9. *Let $F(x,y,z) = (x, y, z, \sqrt{R^2 - x^2 - y^2 - z^2})$ for a constant R. We calculate,*

$$\nabla\sqrt{R^2 - x^2 - y^2 - z^2} = \left(\frac{-x}{\sqrt{R^2-x^2-y^2-z^2}}, \frac{-y}{\sqrt{R^2-x^2-y^2-z^2}}, \frac{-z}{\sqrt{R^2-x^2-y^2-z^2}} \right)$$

Also, $\nabla x = (1,0,0)$, $\nabla y = (0,1,0)$ and $\nabla z = (0,0,1)$ thus

$$F'(x,y,z) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{-x}{\sqrt{R^2-x^2-y^2-z^2}} & \frac{-y}{\sqrt{R^2-x^2-y^2-z^2}} & \frac{-z}{\sqrt{R^2-x^2-y^2-z^2}} \end{bmatrix}$$

This matrix clearly has rank 3 where is is well-defined. Note that we need $R^2 - x^2 - y^2 - z^2 > 0$ for the derivative to exist. This mapping gives us a parametrization of the 3-sphere $x^2 + y^2 + z^2 + w^2 = R^2$ for $w > 0$. (drawing this is a little trickier)

Example 3.4.10. *Let $f(x,y,z) = (x + y, y + z, x + z, xyz)$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ yz & xz & xy \end{bmatrix}$$

This matrix clearly has rank 3 and is well-defined for all of $\mathbb{R}^3$.

Example 3.4.11. *Let $f(x,y,z) = xyz$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \end{bmatrix}$$

This matrix fails to have rank 3 if x, y or z are zero. In other words, $f'(x,y,z)$ has rank 3 in $\mathbb{R}^3$ provided we are at a point which is not on some coordinate plane. (the coordinate planes are $x = 0, y = 0$ and $z = 0$ for the yz, zx and xy coordinate planes respective)

Example 3.4.12. *Let $f(x,y,z) = (xyz, 1 - x - y)$. You can calculate,*

$$[df_{(x,y,z)}] = \begin{bmatrix} yz & xz & xy \\ -1 & -1 & 0 \end{bmatrix}$$

This matrix has rank 3 if either $xy \neq 0$ or $(x - y)z \neq 0$. In contrast to the preceding example, the derivative does have rank 3 on certain points of the coordinate planes. For example, $f'(1,1,0)$ and $f'(0,1,1)$ both give $rank(f') = 3$.

Example 3.4.13. *Let $X(u,v) = (x,y,z)$ where x,y,z denote functions of u,v and I prefer to omit the explicit depedendence to reduce clutter in the equations to follow.*

$$\frac{\partial X}{\partial u} = X_u = (x_u, y_u, z_u) \quad \textit{and} \quad \frac{\partial X}{\partial v} = X_v = (x_v, y_v, z_v)$$

Then the Jacobian is the 3×2 matrix

$$\left[dX_{(u,v)}\right] = \begin{bmatrix} x_u & x_v \\ y_u & y_v \\ z_u & z_v \end{bmatrix}$$

The matrix $\left[dX_{(u,v)}\right]$ has rank 2 if at least one of the determinants below is nonzero,

$$det \begin{bmatrix} x_u & x_v \\ y_u & y_v \end{bmatrix} \quad det \begin{bmatrix} x_u & x_v \\ z_u & z_v \end{bmatrix} \quad det \begin{bmatrix} y_u & y_v \\ z_u & z_v \end{bmatrix}$$

Chapter 4

two views of manifolds in $\mathbb{R}^n$

In this chapter we describe spaces inside $\mathbb{R}^n$ which are k-dimensional [1]. Technically, to make this precise we would need to study manifolds with boundary. Careful discussion of manifolds with boundary in euclidean space can be found in Munkres *Analysis on Manifolds*. In the interest of focusing on examples, I'll be a bit fuzzy about the defintion of a k-dimensional subspace S of euclidean space. This much we can say: there are two ways to envision the geometry of S:

(1.) **Parametrically**: provide a **patch** R such that $R : U \subseteq \mathbb{R}^k \to S \subseteq \mathbb{R}^n$. Here U is called the **parameter space** and R^{-1} is called a **coordinate chart**. The cannonical example:

$$R(x_1, \dots x_k) = (x_1, \dots x_k, 0, \dots, 0).$$

(2.) **Implicitly**: provide a **level function** $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ such that $S = G^{-1}\{c\} = S$. This viewpoint casts S as points in $x \in \mathbb{R}^k \times \mathbb{R}^p$ for which $G(x) = k$. The cannonical example:

$$G(x_1, \dots, x_{k+p}) = (x_{k+1}, \dots, x_{k+p}) = (0, \dots, 0).$$

The cannonical examples of (1.) and (2.) are both the $x_1 \dots x_k$-coordinate plane embedded in $\mathbb{R}^n$. Just to take it down a notch. If $n = 3$ then we could look at the xy-plane in either view as follows:

$$(1.)\ R(x, y) = (x, y, 0) \qquad (2.)\ G(x, y, z) = z = 0.$$

Which viewpoint should we adopt? What is the dimension of a given space S? How should we find tangent space to S? How should we find the normal space to S? These are the questions we set-out to answer in this chapter.

Orthogonal complements help us to understand how all of this fits together. This is possible since we deal with embedded manifolds for which the euclidean dot-product of $\mathbb{R}^n$ is available to sort out the geometry. Finally, we use this geometry and a few simple lemmas to justify the method of Lagrange multipliers. Lagrange's technique paired with the theory of multivariate Taylor polynomials form the basis for analyzing extrema for multivariate functions. In this chapter we deal with the question of extrema on the edges of a set. The second half of the story is found in the next chapter where we deal with the interior points via the theory of quadratic forms applied to the second-order approximation to a function of several variables.

[1] I'll try to stick with this notation for this chapter, $n \geq k$ and $n = p + k$

4.1 definition of level set

A level set is the solution set of some equation or system of equations. We confine our interest to level sets of $\mathbb{R}^n$. For example, the set of all (x, y) that satisfy

$$G(x, y) = c$$

is called a **level curve** in $\mathbb{R}^2$. Often we can use k to label the curve. You should also recall **level surfaces** in $\mathbb{R}^3$ are defined by an equation of the form

$$G(x, y, z) = c.$$

The set of all $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ which solve $G(x_1, x_2, x_3, x_4) = c$ is a **level volume** in $\mathbb{R}^4$. We can obtain lower dimensional objects by simultaneously imposing several equations at once. For example, suppose $G_1(x, y, z) = z = 1$ and $G_2(x, y, z) = x^2 + y^2 + z^2 = 5$, points (x, y, z) which solve both of these equations are on the intersection of the plane $z = 1$ and the sphere $x^2 + y^2 + z^2 = 5$. Let $G = (G_1, G_2)$, note that $G(x, y, z) = (1, 5)$ describes a circle in $\mathbb{R}^3$. More generally:

Definition 4.1.1.

Suppose $G : dom(G) \subseteq \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$. Let c be a vector of constants in $\mathbb{R}^p$ and suppose $S = \{x \in \mathbb{R}^k \times \mathbb{R}^p \mid G(x) = c\}$ is non-empty and G is continuously differentiable on an open set containing S. We say S is an **k-dimensional level** set iff $G'(x)$ has p linearly independent rows at each $x \in S$.

The condition of linear independence of the rows is give to eliminate possible redundancy in the system of equations. In the case that $p = 1$ the criteria reduces to $G'(x) \neq 0$ over the level set of dimension $n - 1$. Intuitively we think of each equation in $G(x) = c$ as removing one of the dimensions of the ambient space $\mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^p$. It is worthwhile to cite a useful result from linear algebra at this point:

Proposition 4.1.2.

Let $A \in \mathbb{R}^{m \times n}$. The number of linearly independent columns in A is the same as the number of linearly independent rows in A. This invariant of A is called the **rank** of A.

Given the wisdom of linear algebra we see that we should require a k-dimensional level set $S = G^{-1}(c)$ to have a level function $G : \mathbb{R}^n \to \mathbb{R}^p$ whose derivative is of rank $n - k = p$ over all of S. We can either analyze linear independence of columns or rows.

Example 4.1.3. *Consider $G(x, y, z) = x^2 + y^2 - z^2$ and suppose $S = G^{-1}\{0\}$. Calculate,*

$$G'(x, y, z) = [2x, 2y, -2z]$$

Notice that $(0, 0, 0) \in S$ and $G'(0, 0, 0) = [0, 0, 0]$ hence G' is not rank one at the origin. At all other points in S we have $G'(x, y, z) \neq 0$ which means this is almost a $3 - 1 = 2$-dimensional level set. However, almost is not good enough in math. Under our definition the cone S is not a 2-dimensional level set since it fails to meet the full-rank criteria at the point of the cone.

A p-dimensional level set is an example of a p-dimensional manifold. The example above with the origin included is a manifold paired with a singular point, such spaces are known as **orbifolds**. The study of orbifolds has attracted considerable effort in recent years as the singularities of such orbifolds can be used to do physics in string theory. I digress. Let us examine another level set:

Example 4.1.4. *Let $G(x, y, z) = (x, y)$ and define $S = G^{-1}(a, b)$ for some fixed pair of constants $a, b \in \mathbb{R}$. We calculate that $G'(x, y, z) = I_2 \in \mathbb{R}^{2 \times 2}$. We clearly have rank two at all points in S hence S is a $3 - 2 = 1$-dimensional level set. Perhaps you realize S is the vertical line which passes through $(a, b, 0)$ in the xy-plane.*

4.2 tangents and normals to a level set

There are many ways to define a tangent space for some subset of $\mathbb{R}^n$. One natural definition is that the tangent space to $p \in S$ is simply the set of all tangent vectors to curves on S which pass through the point p. In this section we study the geometry of curves on a level-set. We'll see how the tangent space is naturally a vector space in the particular context of level-sets in $\mathbb{R}^n$.

Throughout this section we assume that S is a k-dimensional level set defined by $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ where $G^{-1}(c) = S$. This means that we can apply the implicit function theorem to S and for any given point $p = (p_x, p_y) \in S$ where $p_x \in \mathbb{R}^k$ and $p_y \in \mathbb{R}^p$. There exists a local continuously differentiable solution $h : U \subseteq \mathbb{R}^k \to \mathbb{R}^p$ such that $h(p_x) = p_y$ and for all $x \in U$ we have $G(x, h(x)) = c$. We can view $G(x, y) = c$ for x near p as the graph of $y = h(x)$ for $x \in U$. With the set-up above in mind, suppose that $\gamma : \mathbb{R} \to U \subseteq S$. If we write $\gamma = (\gamma_x, \gamma_y)$ then it follows $\gamma = (\gamma_x, h \circ \gamma_x)$ over the subset $U \times h(U)$ of S. More explicitly, for all $t \in \mathbb{R}$ such that $\gamma(t) \in U \times h(U)$ we have

$$\gamma(t) = (\gamma_x(t), h(\gamma_x(t))).$$

Therefore, if $\gamma(0) = p$ then $\gamma(0) = (p_x, h(p_x))$. Differentiate, use the chain-rule in the second factor to obtain:

$$\gamma'(t) = (\gamma_x'(t), h'(\gamma_x(t))\gamma_x'(t)).$$

We find that the tangent vector to $p \in S$ of γ has a rather special form which was forced on us by the implicit function theorem:

$$\gamma'(0) = (\gamma_x'(0), h'(p_x)\gamma_x'(0)).$$

Or to cut through the notation a bit, if $\gamma'(0) = v = (v_x, v_y)$ then $v = (v_x, h'(p_x)v_x)$. The second component of the vector is not free of the first, it essentially redundant. This makes us suspect that the tangent space to S at p is k-dimensional.

Theorem 4.2.1.

Let $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ be a level-mappping which defines a k-dimensional level set S by $G^{-1}(c) = S$. Suppose $\gamma_1, \gamma_2 : \mathbb{R} \to S$ are differentiable curves with $\gamma_1'(0) = v_1$ and $\gamma_2'(0) = v_2$ then there exists a differentiable curve $\gamma : \mathbb{R} \to S$ such that $\gamma'(0) = v_1 + v_2$ and $\gamma(0) = p$. Moreover, there exists a differentiable curve $\beta : \mathbb{R} \to S$ such that $\beta'(0) = cv_1$ and $\beta(0) = p$.

Proof: It is convenient to define a map which gives a **local parametrization** of S at p. Since we have a description of S locally as a graph $y = h(x)$ (near p) it is simple to construct the parameterization. Define $\Phi : U \subseteq \mathbb{R}^k \to S$ by $\Phi(x) = (x, h(x))$. Clearly $\Phi(U) = U \times h(U)$ and there is an inverse mapping $\Phi^{-1}(x, y) = x$ is well-defined since $y = h(x)$ for each $(x, y) \in U \times h(U)$. Let $w \in \mathbb{R}^k$ and observe that

$$\psi(t) = \Phi(\Phi^{-1}(p) + tw) = \Phi(p_x + tw) = (p_x + tw, h(p_x + tw))$$

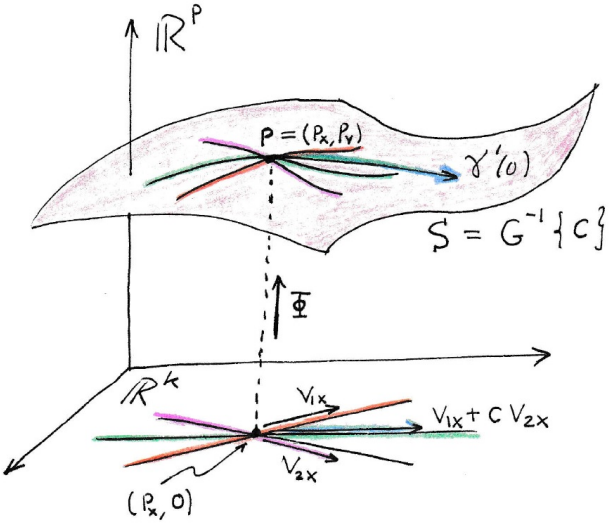
is a curve from $\mathbb{R}$ to $U \subseteq S$ such that $\psi(0) = (p_x, h(p_x)) = (p_x, p_y) = p$ and using the chain rule on the final form of $\psi(t)$:

$$\psi'(0) = (w, h'(p_x)w).$$

The construction above shows that any vector of the form $(v_x, h'(p_x)v_x)$ is the tangent vector of a particular differentiable curve in the level set (differentiability of ψ follows from the differentiability of h and the other maps which we used to construct ψ). In particular we can apply this to the case $w = v_{1x} + v_{2x}$ and we find $\gamma(t) = \Phi(\Phi^{-1}(p) + t(v_{1x} + v_{2x}))$ has $\gamma'(0) = v_1 + v_2$ and $\gamma(0) = p$.

Likewise, apply the construction to the case $w = cv_{1x}$ to write $\beta(t) = \Phi(\Phi^{-1}(p) + t(cv_{1x}))$ with $\beta'(0) = cv_1$ and $\beta(0) = p$. □

The idea of the proof is encapsulated in the picture below. This idea of mapping lines in a flat domain to obtain standard curves in a curved domain is an idea which plays over and over as you study manifold theory. The particular redundancy of the x and y sub-vectors is special to the discussion level-sets, however anytime we have a local parametrization we'll be able to construct curves with tangents of our choosing by essentially the same construction. In fact, there are infinitely many curves which produce a particular tangent vector in the tangent space of a manifold.



Theorem 4.2.1 shows that the definition given below is logical. In particular, it is not at all obvious that the sum of two tangent vectors ought to again be a tangent vector. However, that is just what the Theorem 4.2.1 told us for level-sets[2].

Definition 4.2.2.

Suppose S is a k-dimensional level-set defined by $S = G^{-1}\{c\}$ for $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$. We define the **tangent space at** $p \in S$ to be the set of pairs:

$$T_pS = \{(p, v) \mid \text{there exists differentiable } \gamma : \mathbb{R} \to S \text{ and } \gamma(0) = p \text{ where } v = \gamma'(0)\}$$

Moreover, we define (i.) **addition** and (ii.) **scalar multiplication** of vectors by the rules

$$(i.)\ \ (p, v_1) + (p, v_2) = (p, v_1 + v_2) \qquad (ii.)\ \ c(p, v_1) = (p, cv_1)$$

for all $(p, v_1), (p, v_2) \in T_pS$ and $c \in \mathbb{R}$.

When I picture T_pS in my mind I think of vectors pointing out from the base-point p. To make an explicit connection between the pairs of the above definition and the classical geometric form of the tangent space we simply take the image of T_pS under the mapping $\Psi(x, y) = x + y$ thus $\Psi(T_pS) = \{p + v \mid (p, v) \in T_pS\}$. I often picture T_pS as $\psi(T_pS)$[3]

[2] technically, there is another logical gap which I currently ignore. I wonder if you can find it.

[3] In truth, as you continue to study manifold theory you'll find at least three seemingly distinct objects which are all called "tangent vectors"; equivalence classes of curves, derivations, contravariant tensors.

We could set out to calculate tangent spaces in view of the definition above, but we are actually interested in more than just the tangent space for a level-set. In particular. we want a concrete description of all the vectors which are not in the tangent space.

Definition 4.2.3.

Suppose S is a k-dimensional level-set defined by $S = G^{-1}\{c\}$ for $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ and T_pS is the tangent space at p. Note that $T_pS \leq V_p$ where $V_p = \{p\} \times \mathbb{R}^k \times \mathbb{R}^p$ is given the natural vector space structure which we already exhibited on the subspace T_pS. We define the **inner product** on V_p as follows: for all $(p, v), (p, w) \in V_p$,

$$(p, v) \cdot (p, w) = v \cdot w.$$

The length of a vector (p, v) is naturally defined by $||(p, v)|| = ||v||$. Moreover, we say two vectors $(p, v), (p, w) \in V_p$ are **orthogonal** iff $v \cdot w = 0$. Given a set of vectors $R \subseteq V_p$ we define the **orthogonal complement** by

$$R^{\perp} = \{(p, v) \in V_p \mid (p, v) \cdot (p, r) \text{ for all } (p, r) \in R\}.$$

Suppose $W_1, W_2 \subseteq V_p$ then we say W_1 is **orthogonal** to W_2 iff $w_1 \cdot w_2 = 0$ for all $w_1 \in W_1$ and $w_2 \in W_2$. We denote orthogonality by writing $W_1 \perp W_2$. If every $v \in V_p$ can be written as $v = w_1 + w_2$ for a pair of $w_1 \in W_1$ and $w_2 \in W_2$ where $W_1 \perp W_2$ then we say that V_p is the **direct sum** of W_1 and W_2 which is denoted by $V_p = W_1 \oplus W_2$.

There is much more to say about orthogonality, however, our focus is not in that vein. We just need the langauge to properly define the normal space. The calculation below is probably the most important calculation to understand for a level-set. Suppose we have a curve $\gamma : \mathbb{R} \to S$ where $S = G^{-1}(c)$ is a k-dimensional level-set in $\mathbb{R}^k \times \mathbb{R}^p$. Observe that for all $t \in \mathbb{R}$,

$$G(\gamma(t)) = c \quad \Rightarrow \quad G'(\gamma(t))\gamma'(t) = 0.$$

In particular, suppose for $t = 0$ we have $\gamma(0) = p$ and $v = \gamma'(0)$ which makes $(p, v) \in T_pS$ with

$$G'(p)v = 0.$$

Recall $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ has an $p \times n$ derivative matrix where the j-th row is the gradient vector of the j-th component function. The equation $G'(p)v = 0$ gives us p-independent equations as we examine it componentwise. In particular, it reveals that (p, v) is orthogonal to $\nabla G_j(p)$ for $j = 1, 2, \ldots, p$. We have derived the following theorem:

Theorem 4.2.4.

Let $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ be a level-mappping which defines a k-dimensional level set S by $G^{-1}(c) = S$. The gradient vectors $\nabla G_j(p)$ are perpendicular to the tangent space at p; for each $j \in \mathbb{N}_p$

$$(p, \nabla(G_j(p))^T) \in (T_pS)^{\perp}.$$

It's time to do some counting. Observe that the mapping $\phi : \mathbb{R}^k \to T_pS$ defined by $\phi(v) = (p, v)$ is an isomorphism of vector spaces hence $dim(T_pS) = k$. But, by the same isomorphism we can see that $V_p = \phi(\mathbb{R}^k \times \mathbb{R}^p)$ hence $dim(V_p) = p + k$. In linear algebra we learn that if we have a k-dimensional subspace W of an n-dimensional vector space V then the orthogonal complement $W^{\perp}$ is a subspace of V with **codimension** k. The term **codimension** is used to indicate a loss

of dimension from the ambient space, in particular $dim(W^\perp) = n - k$. We should note that the direct sum of W and $W^\perp$ covers the whole space; $W \oplus W^\perp = V$. In the case of the tangent space, the codimension of $T_pS \leq V_p$ is found to be $p + k - k = p$. Thus $dim(T_pS)^\perp = p$. Any basis for this space must consist of p linearly independent vectors which are all orthogonal to the tangent space. Naturally, the subset of vectors $\{(p, (\nabla G_j(p))^T)_{j=1}^p$ forms just such a basis since it is given to be linearly independent by the $rank(G'(p)) = p$ condition. It follows that:

$$\boxed{(T_pS)^\perp \approx Row(G'(p))}$$

where equality can be obtained by the slightly tedious equation

$$\boxed{(T_pS)^\perp = \phi(Col(G'(p)^T))}.$$

That equation simply does the following:

1. transpose $G'(p)$ to swap rows to columns
2. construct column space by taking span of columns in $G'(p)^T$
3. adjoin p to make pairs of vectors which live in V_p

many wiser authors wouldn't bother. The comments above are primarily about notation. Certainly hiding these details would make this section prettier, however, would it make it better? Finally, I once more refer the reader to linear algebra where we learn that $(Row(A))^\perp = Null(A^T)$. Let me walk you through the proof: let $A \in \mathbb{R}^{m\times n}$. Observe $v \in Null(A^T)$ iff $A^T v = 0$ for $v \in \mathbb{R}^m$ iff $v^T A = 0$ iff $v^T col_j(A) = 0$ for $j = 1, 2, \ldots, n$ iff $v \cdot col_j(A) = 0$ for $j = 1, 2, \ldots, n$ iff $v \in Col(A)^\perp$. Another useful identity for the "perp" is that $(A^\perp)^\perp = A$. With those two gems in mind consider that:

$$(T_pS)^\perp \approx Row(G'(p)) \quad \Rightarrow \quad T_pS \approx Row(G'(p))^\perp = Null(G'(p)^T)$$

Let me once more replace $\approx$ by a more tedious, but explicit, procedure:

$$\boxed{T_pS = \phi(Null(G'(p)^T))}$$

Theorem 4.2.5.

Let $G : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$ be a level-mappping which defines a k-dimensional level set S by $G^{-1}(c) = S$. The **tangent space** T_pS and the **normal space** N_pS at $p \in S$ are given by

$$T_pS = \{p\} \times Null(G'(p)^T) \qquad \& \qquad N_pS = \{p\} \times Col(G'(p)^T).$$

Moreover, $V_p = T_pS \oplus N_pS$. Every vector can be uniquely written as the sum of a tangent vector and a normal vector.

The fact that there are only tangents and normals is the key to the method of Lagrange multipliers. It forces two seemingly distinct objects to be in the same direction as one another.

Example 4.2.6. *Let $g : \mathbb{R}^4 \to \mathbb{R}$ be defined by $g(x, y, z, t) = t+x^2+y^2-2z^2$ note that $g(x, y, z, t) = 0$ gives a three dimensional subset of $\mathbb{R}^4$, let's call it M. Notice $\nabla g =< 2x, 2y, -4z, 1 >$ is nonzero everywhere. Let's focus on the point $(2, 2, 1, 0)$ note that $g(2, 2, 1, 0) = 0$ thus the point is on M. The tangent plane at $(2, 2, 1, 0)$ is formed from the union of all tangent vectors to $g = 0$ at the point $(2, 2, 1, 0)$. To find the equation of the tangent plane we suppose $\gamma : \mathbb{R} \to M$ is a curve with $\gamma' \neq 0$ and $\gamma(0) = (2, 2, 1, 0)$. By assumption $g(\gamma(s)) = 0$ since $\gamma(s) \in M$ for all $s \in \mathbb{R}$. Define $\gamma'(0) =< a, b, c, d >$, we find a condition from the chain-rule applied to $g \circ \gamma = 0$ at $s = 0$,*

$$\begin{aligned} \frac{d}{ds}\big(\, g\circ\gamma(s)\,\big) = \big(\nabla g\big)(\gamma(s))\cdot\gamma'(s) = 0 \qquad &\Rightarrow \qquad \nabla g(2,2,1,0)\cdot < a,b,c,d >= 0 \\ &\Rightarrow \qquad < 4,4,-4,1 > \cdot < a,b,c,d >= 0 \\ &\Rightarrow \qquad 4a+4b-4c+d = 0 \end{aligned}$$

Thus the equation of the tangent plane is $4(x - 2) + 4(y - 2) - 4(z - 1) + t = 0$. In invite the reader to find a vector in the tangent plane and check it is orthogonal to $\nabla g(2, 2, 1, 0)$. However, this should not be surprising, the condition the chain rule just gave us is just the statement that $< a, b, c, d >\in Null(\nabla g(2, 2, 1, 0)^T)$ and that is precisely the set of vector orthogonal to $\nabla g(2, 2, 1, 0)$.

Example 4.2.7. *Let $G : \mathbb{R}^4 \to \mathbb{R}^2$ be defined by $G(x, y, z, t) = (z + x^2 + y^2 - 2, z + y^2 + t^2 - 2)$. In this case $G(x, y, z, t) = (0, 0)$ gives a two-dimensional manifold in $\mathbb{R}^4$ let's call it M. Notice that $G_1 = 0$ gives $z + x^2 + y^2 = 2$ and $G_2 = 0$ gives $z + y^2 + t^2 = 2$ thus $G = 0$ gives the intersection of both of these three dimensional manifolds in $\mathbb{R}^4$ (no I can't "see" it either). Note,*

$$\nabla G_1 =< 2x, 2y, 1, 0 > \qquad \nabla G_2 =< 0, 2y, 1, 2t >$$

It turns out that the inverse mapping theorem says $G = 0$ describes a manifold of dimension 2 if the gradient vectors above form a linearly independent set of vectors. For the example considered here the gradient vectors are linearly dependent at the origin since $\nabla G_1(0) = \nabla G_2(0) = (0, 0, 1, 0)$. In fact, these gradient vectors are colinear along along the plane $x = t = 0$ since $\nabla G_1(0, y, z, 0) = \nabla G_2(0, y, z, 0) =< 0, 2y, 1, 0 >$. We again seek to contrast the tangent plane and its normal at some particular point. Choose $(1, 1, 0, 1)$ which is in M since $G(1, 1, 0, 1) = (0 + 1 + 1 - 2, 0 + 1 + 1 - 2) = (0, 0)$. Suppose that $\gamma : \mathbb{R} \to M$ is a path in M which has $\gamma(0) = (1, 1, 0, 1)$ whereas $\gamma'(0) =< a, b, c, d >$. Note that $\nabla G_1(1, 1, 0, 1) =< 2, 2, 1, 0 >$ and $\nabla G_2(1, 1, 0, 1) =< 0, 2, 1, 1 >$. Applying the chain rule to both G_1 and G_2 yields:

$$\begin{aligned} (G_1\circ\gamma)'(0) = \nabla G_1(\gamma(0))\cdot < a,b,c,d >= 0 \qquad &\Rightarrow \qquad < 2,2,1,0 > \cdot < a,b,c,d >= 0 \\ (G_2\circ\gamma)'(0) = \nabla G_2(\gamma(0))\cdot < a,b,c,d >= 0 \qquad &\Rightarrow \qquad < 0,2,1,1 > \cdot < a,b,c,d >= 0 \end{aligned}$$

This is two equations and four unknowns, we can solve it and write the vector in terms of two free variables correspondant to the fact the tangent space is two-dimensional. Perhaps it's easier to use matrix techiques to organize the calculation:

$$\begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We calculate, $rref \begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -1/2 \\ 0 & 1 & 1/2 & 1/2 \end{bmatrix}$. It's natural to chose c, d as free variables then we can read that $a = d/2$ and $b = -c/2 - d/2$ hence

$$< a, b, c, d >=< d/2, -c/2 - d/2, c, d >= \tfrac{c}{2} < 0, -1, 2, 0 > + \tfrac{d}{2} < 1, -1, 0, 2 >$$

We can see a basis for the tangent space. In fact, I can give parametric equations for the tangent space as follows:

$$X(u,v) = (1,1,0,1) + u < 0,-1,2,0 > +v < 1,-1,0,2 >$$

Not surprisingly the basis vectors of the tangent space are perpendicular to the gradient vectors $\nabla G_1(1,1,0,1) =< 2,2,1,0 >$ *and* $\nabla G_2(1,1,0,1) =< 0,2,1,1 >$ *which span the* **normal plane** N_p *to the tangent plane* T_p *at* $p = (1,1,0,1)$*. We find that* T_p *is orthogonal to* N_p*. In summary* $T_p^{\perp} = N_p$ *and* $T_p \oplus N_p = \mathbb{R}^4$*. This is just a fancy way of saying that the normal and the tangent plane only intersect at zero and they together span the entire ambient space.*

4.3 tangent and normal space from patches

I use the term **parametrization** in courses more basic than this, however, perhaps the term **patch** would be better. It's certainly easier to say and in our current context has the same meaning. I suppose the term **parametrization** is used in a bit less technical sense, so it fits calculus III better. In any event, we should make a definition of patched k-dimensional surface for the sake of concrete discussion in this section.

Definition 4.3.1.

Suppose $R : dom(R) \subseteq \mathbb{R}^k \to S \subseteq \mathbb{R}^n$. We say S is an **k-dimensional patch** iff $R'(t)$ has rank k for each $t \in dom(R)$. We also call S a k-dimensional parametrized subspace of $\mathbb{R}^n$.

The condition $R'(t)$ is just a slick way to say that the k-tangent vectors to S obtained by partial differentiation with respect to $t_1, \ldots, t_k$ are linearly independent at $t = (t_1, \ldots, t_k)$. I spent considerable effort justifying the formulae for the level-set case. I believe what follows should be intuitively clear given our previous efforts. Or, if that leaves you unsatisfied then read on to the examples. It's really not that complicated. This theorem is dual to Theorem 4.2.5.

Theorem 4.3.2.

Suppose $R : dom(R) \subseteq \mathbb{R}^k \to S \subseteq \mathbb{R}^n$ defines a **k-dimensional patch** of S. The **tangent space** T_pS and the **normal space** at $p = R(t) \in S$ are given by

$$T_pS = \{p\} \times Col(R'(t)) \qquad \& \qquad N_pS = \{p\} \times Null(R'(t)^T).$$

Moreover, $V_p = T_pS \oplus N_pS$. Every vector can be uniquely written as the sum of a tangent vector and a normal vector.

Once again, the vector space structure of T_pS and N_pS is given by the addition of vectors based at p. Let us begin with a reasonably simple example.

Example 4.3.3. *Let* $R : \mathbb{R}^2 \to \mathbb{R}^3$ *with* $R(x,y) = (x,y,xy)$ *define* $S \subset \mathbb{R}^3$*. We calculate,*

$$R'(x,y) = [\partial_x R | \partial_y R] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ y & x \end{bmatrix}$$

If $p = (a,b,ab) \in S$ *then* $T_pS = \{(a,b,ab)\} \times span\{(1,0,b),(0,1,a)\}$*. The normal space is found from* $Null(R'(a,b)^T)$*. A short calculation shows that*

$$Null \begin{bmatrix} 1 & 0 & b \\ 0 & 1 & a \end{bmatrix} = span\{(-b,-a,1)\}$$

As a quick check, note $(1,0,b)\bullet(-b,-a,1) = 0$ *and* $(0,1,a)\bullet(-b,-a,1) = 0$. *We conclude, for* $p = (a,b,ab)$ *the normal space is simply:*

$$N_pS = \{(a,b,ab)\} \times span\{(-b,-a,1)\}.$$

In the previous example, we could rightly call T_pS the tangent plane at p and N_pS the normal line through p. Moreover, we could have used three-dimensional vector analysis to find the normal line from the cross-product. However, that will not be possible in what follows:

Example 4.3.4. *Let* $R : \mathbb{R}^2 \to \mathbb{R}^4$ *with* $R(s,t) = (s^2, t^2, t, s)$ *define* $S \subset \mathbb{R}^4$. *We calculate,*

$$R'(s,t) = [\partial_s R | \partial_t R] = \begin{bmatrix} 2s & 0 \\ 0 & 2t \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

If $p = (1,9,3,1) \in S$ *then* $T_pS = \{(1,9,3,1)\} \times span\{(2,0,0,1),(0,6,3,0)\}$. *The normal space is found from* $Null(R'(1,3)^T)$. *A short calculation shows that*

$$Null \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 6 & 3 & 0 \end{bmatrix} = span\{(-1,0,0,2),(0,-3,6,0)\}$$

We conclude, for $p = (1,9,3,1)$ *the normal space is simply:*

$$N_pS = \{(1,9,3,1)\} \times span\{(-1,0,0,2),(0,-3,6,0)\}.$$

4.4 summary of tangent and normal spaces

Let me briefly draw together what we did thus far in this chapter: the notation below given in I is also used in $II.$ and $III.$

(I.) a set S has dimension k if

(a) $\{\partial_1 R(t), \ldots, \partial_k R(t)\}$ is pointwise linearly independent at each $t \in U$ where $R : U \to S$ is a patch.

(b) $\text{rank}(F'(x)) = p$ for all $x \in \tilde{S}$ where $\tilde{S}$ is open and contains $S = F^{-1}\{c\}$ for continuously differentiable $F : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p$

(II.) the tangent space at x_o for the k-dimensional set S is found from:

(a) attaching the span of the vectors $\{\partial_1 R(t_o), \ldots, \partial_k R(t_o)\}$ to $x_o = R(t_o) \in S$.

(b) attaching the $Row(F'(x_o))^\perp$ to $x_o \in S$.

(III.) the normal space to a k-dimensional set S (embedded in $\mathbb{R}^n$) is found from:

(a) attaching $\{\partial_1 R(t_o), \ldots, \partial_k R(t_o)\}^\perp$ to $x_o = R(t_o)$.

(b) attaching $Row(F'(x_o))$ to $x_o \in S$.

4.5 method of Lagrange mulitpliers

Let us begin with a statement of the problem we wish to solve.

Problem: given an objective function $f : \mathbb{R}^n \to \mathbb{R}$ and continuously differentiable constraint function $G : \mathbb{R}^n \to \mathbb{R}^p$, find extreme values for the objective function f relative to the constraint $G(x) = c$.

Note that $G(x) = c$ is a vector notation for p-scalar equations. If we suppose $rank(G'(x)) = p$ then the constraint surface $G(x) = c$ will form an $(n-p)$-dimensional level set. Let us make that supposition throughout the remainder of this section.

In order to solve a problem it is sometimes helpful to find necessary conditions by assuming an answer exists. Let us do that here. Suppose x_o maps to the local extrema of $f(x_o)$ on $S = G^{-1}\{c\}$. This means there exists an open ball around x_o for which $f(x_o)$ is either an upper or lower bound of all the values of f over the ball intersected with S. One clear implication of this data is that if we take any continuously differentiable curve on S which passes through x_o, say $\gamma : \mathbb{R} \to \mathbb{R}^n$ with $\gamma(0) = x_o$ and $G(\gamma(t)) = c$ for all t, then the composite $f \circ \gamma$ is a function on $\mathbb{R}$ which takes an extreme value at $t = 0$. Fermat's theorem from calculus I applies and as $f \circ \gamma$ is differentiable near $t = 0$ we find $(f \circ \gamma)'(0) = 0$ is a necessary condition. But, this means we have two necessary conditions on γ:

1. $G(\gamma(t)) = c$
2. $(f \circ \gamma)'(0) = 0$

Let us expand a bit on both of these conditions:

1. $G'(x_o)\gamma'(0) = 0$
2. $f'(x_o)\gamma'(0) = 0$

The first of these conditions places $\gamma'(0) \in T_{x_o}S$ but then the second condition says that $f'(x_o) = (\nabla f)(x_o)^T$ is orthogonal to $\gamma'(0)$ hence $(\nabla f)(x_o)^T \in N_{x_o}$. Now, recall from the last section that the gradient vectors of the component functions to G span the normal space, this means any vector in N_{x_o} can be written as a linear combination of the gradient vectors. In particular, this means there exist constants $\lambda_1, \lambda_2, \ldots, \lambda_p$ such that

$$(\nabla f)(x_o)^T = \lambda_1(\nabla G_1)(x_o)^T + \lambda_2(\nabla G_2)(x_o)^T + \cdots + \lambda_p(\nabla G_p)(x_o)^T$$

We may summarize the method of Lagrange multipliers as follows:

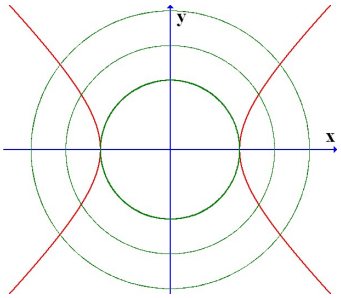
1. **choose n-variables which aptly describe your problem.**
2. **identify your objective function and write all constraints as level surfaces.**
3. **solve $\nabla f = \lambda_1 \nabla G_1 + \lambda_2 \nabla G_2 + \cdots + \lambda_p \nabla G_p$ subject to the constraint $G(x) = c$.**
4. **test the validity of your proposed extremal points.**

The obvious gap in the method is the supposition that an extrema exists for the restriction $f|_S$. Well examine a few examples before I reveal a sufficient condition. We'll also see how absence of that sufficient condition does allow the method to fail.

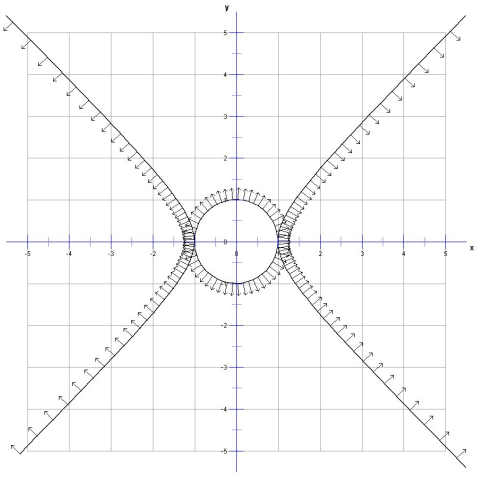
Example 4.5.1. *Suppose we wish to find maximum and minimum distance to the origin for points on the curve* $x^2 - y^2 = 1$. *In this case we can use the distance-squared function as our objective* $f(x, y) = x^2 + y^2$ *and the single constraint function is* $g(x, y) = x^2 - y^2$. *Observe that* $\nabla f =< 2x, 2y >$ *whereas* $\nabla g =< 2x, -2y >$. *We seek solutions of* $\nabla f = \lambda \nabla g$ *which gives us* $< 2x, 2y >= \lambda < 2x, -2y >$. *Hence* $2x = 2\lambda x$ *and* $2y = -2\lambda y$. *We must solve these equations subject to the condition* $x^2 - y^2 = 1$. *Observe that* $x = 0$ *is not a solution since* $0 - y^2 = 1$ *has no real solution. On the other hand,* $y = 0$ *does fit the constraint and* $x^2 - 0 = 1$ *has solutions* $x = \pm 1$. *Consider then*

$$2x = 2\lambda x \quad \textit{and} \quad 2y = -2\lambda y \qquad \Rightarrow \qquad x(1 - \lambda) = 0 \quad \textit{and} \quad y(1 + \lambda) = 0$$

Since $x \neq 0$ *on the constraint curve it follows that* $1 - \lambda = 0$ *hence* $\lambda = 1$ *and we learn that* $y(1 + 1) = 0$ *hence* $y = 0$. *Consequently,* $(1, 0$ *and* $(-1, 0)$ *are the two point where we expect to find extreme-values of* f. *In this case, the method of Lagrange multipliers served it's purpose, as you can see in the graph. Below the green curves are level curves of the objective function whereas the particular red curve is the given constraint curve.*



The picture below is a screen-shot of the Java applet created by David Lippman and Konrad Polthier to explore 2D and 3D graphs. Especially nice is the feature of adding vector fields to given objects, many other plotters require much more effort for similar visualization. See more at the website: http://dlippman.imathas.com/g1/GrapherLaunch.html.



Note how the gradient vectors to the objective function and constraint function line-up nicely at those points.

In the previous example, we actually got lucky. There are examples of this sort where we could get false maxima due to the nature of the constraint function.

Example 4.5.2. *Suppose we wish to find the points on the unit circle $g(x,y) = x^2 + y^2 = 1$ which give extreme values for the objective function $f(x,y) = x^2 - y^2$. Apply the method of Lagrange multipliers and seek solutions to $\nabla f = \lambda \nabla g$:*

$$< 2x, -2y >= \lambda < 2x, 2y >$$

We must solve $2x = 2x\lambda$ which is better cast as $(1-\lambda)x = 0$ and $-2y = 2\lambda y$ which is nicely written as $(1+\lambda)y = 0$. On the basis of these equations alone we have several options:

1. *if $\lambda = 1$ then $(1+1)y = 0$ hence $y = 0$*
2. *if $\lambda = -1$ then $(1-(1))x = 0$ hence $x = 0$*

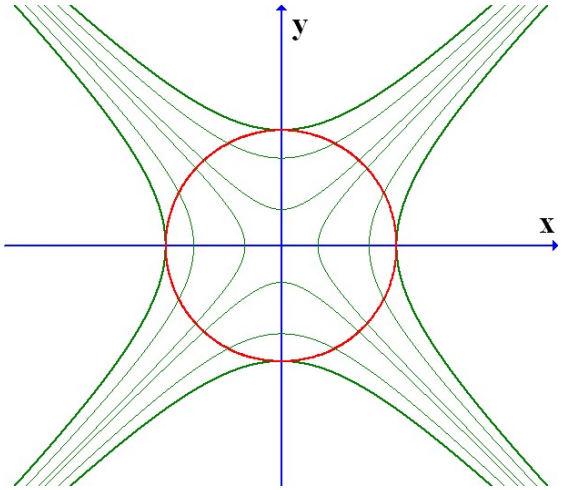
But, we also must fit the constraint $x^2 + y^2 = 1$ hence we find four solutions:

1. *if $\lambda = 1$ then $y = 0$ thus $x^2 = 1 \quad \Rightarrow \quad x = \pm 1 \quad \Rightarrow \quad (\pm 1, 0)$*
2. *if $\lambda = -1$ then $x = 0$ thus $y^2 = 1 \quad \Rightarrow \quad y = \pm 1 \quad \Rightarrow \quad (0, \pm 1)$*

We test the objective function at these points to ascertain which type of extrema we've located:

$$f(0, \pm 1) = 0^2 - (\pm 1)^2 = -1 \qquad \& \qquad f(\pm 1, 0) = (\pm 1)^2 - 0^2 = 1$$

When constrained to the unit circle we find the objective function attains a maximum value of 1 at the points $(1,0)$ and $(-1,0)$ and a minimum value of -1 at $(0,1)$ and $(0,-1)$. Let's illustrate the answers as well as a few non-answers to get perspective. Below the green curves are level curves of the objective function whereas the particular red curve is the given constraint curve.



The success of the last example was no accident. The fact that the constraint curve was a circle which is a closed and bounded subset of $\mathbb{R}^2$ means that is is a **compact** subset of $\mathbb{R}^2$. A well-known theorem of analysis states that any real-valued continuous function on a compact domain attains both maximum and minimum values. The objective function is continuous and the domain is compact hence the theorem applies and the method of Lagrange multipliers succeeds. In contrast, the constraint curve of the preceding example was a hyperbola which is not compact. We have no assurance of the existence of any extrema. Indeed, we only found minima but no maxima in Example 4.5.1.

The generality of the method of Lagrange multipliers is naturally limited to smooth constraint curves and smooth objective functions. We must insist the gradient vectors exist at all points of

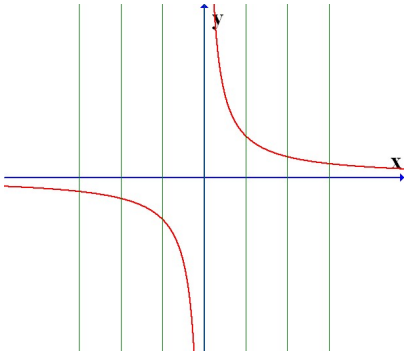
inquiry. Otherwise, the method breaks down. If we had a constraint curve which has sharp corners then the method of Lagrange breaks down at those corners. In addition, if there are points of discontinuity in the constraint then the method need not apply. This is not terribly surprising, even in calculus I the main attack to analyze extrema of function on $\mathbb{R}$ assumed continuity, differentiability and sometimes twice differentiability. Points of discontinuity require special attention in whatever context you meet them.

At this point it is doubtless the case that some of you are, to misquote an ex-student of mine, "not-impressed". Perhaps the following examples better illustrate the dangers of non-compact constraint curves.

Example 4.5.3. *Suppose we wish to find extrema of $f(x,y) = x$ when constrained to $xy = 1$. Identify $g(x,y) = xy = 1$ and apply the method of Lagrange multipliers and seek solutions to $\nabla f = \lambda \nabla g$:*

$$< 1, 0 >= \lambda < y, x > \;\Rightarrow\; 1 = \lambda y \;\; and \;\; 0 = \lambda x$$

If $\lambda = 0$ then $1 = \lambda y$ is impossible to solve hence $\lambda \neq 0$ and we find $x = 0$. But, if $x = 0$ then $xy = 1$ is not solvable. Therefore, we find no solutions. Well, I suppose we have succeeded here in a way. We just learned there is no extreme value of x on the hyperbola $xy = 1$. Below the green curves are level curves of the objective function whereas the particular red curve is the given constraint curve.



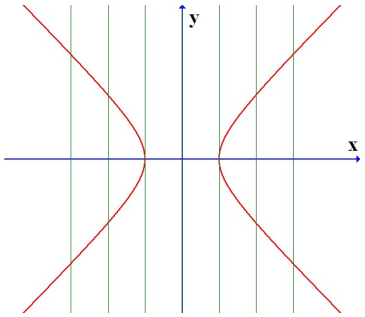
Example 4.5.4. *Suppose we wish to find extrema of $f(x,y) = x$ when constrained to $x^2 - y^2 = 1$. Identify $g(x,y) = x^2 - y^2 = 1$ and apply the method of Lagrange multipliers and seek solutions to $\nabla f = \lambda \nabla g$:*

$$< 1, 0 >= \lambda < 2x, -2y > \;\Rightarrow\; 1 = 2\lambda x \;\; and \;\; 0 = -2\lambda y$$

If $\lambda = 0$ then $1 = 2\lambda x$ is impossible to solve hence $\lambda \neq 0$ and we find $y = 0$. If $y = 0$ and $x^2 - y^2 = 1$ then we must solve $x^2 = 1$ whence $x = \pm 1$. We are tempted to conclude that:

1. *the objective function $f(x,y) = x$ attains a maximum on $x^2 - y^2 = 1$ at $(1,0)$ since $f(1,0) = 1$*
2. *the objective function $f(x,y) = x$ attains a minimum on $x^2 - y^2 = 1$ at $(-1,0)$ since $f(1,0) = -1$*

But, both conclusions are false. Note $\sqrt{2}^2 - 1^2 = 1$ hence $(\pm\sqrt{2}, 1)$ are points on the constraint curve and $f(\sqrt{2},1) = \sqrt{2}$ and $f(-\sqrt{2},1) = -\sqrt{2}$. The error of the method of Lagrange multipliers in this context is the supposition that there exists extrema to find, in this case there are no such points. It is possible for the gradient vectors to line-up at points where there are no extrema. Below the green curves are level curves of the objective function whereas the particular red curve is the given constraint curve.



Incidentally, if you want additional discussion of Lagrange multipliers for two-dimensional problems one very nice source I certainly profitted from was the YouTube video by Edward Frenkel of Berkley. See his website http://math.berkeley.edu/ frenkel/ for links.

Chapter 5

critical point analysis for several variables

In the typical calculus sequence you learn the first and second derivative tests in calculus I. Then in calculus II you learn about power series and Taylor's Theorem. Finally, in calculus III, in many popular texts, you learn an essentially ad-hoc procedure for judging the nature of critical points as minimum, maximum or saddle. These topics are easily seen as disconnected events. In this chapter, we connect them. We learn that the geometry of quadratic forms is ellegantly revealed by eigenvectors and more than that this geometry is precisely what elucidates the proper classifications of critical points of multivariate functions with real values.

5.1 multivariate power series

We set aside the issue of convergence for now. We will suppose the series discussed in this section exist on and converge on some domain, but we do not seek to treat that topic here. Our focus is computational. How should we calculate the Taylor series for $f(x,y)$ at (a,b)? Or, what about $f(x)$ at $x_o \in \mathbb{R}^n$?.

5.1.1 taylor's polynomial for one-variable

If $f : U \subseteq \mathbb{R} \to \mathbb{R}$ is analytic at $x_o \in U$ then we can write

$$f(x) = f(x_o) + f'(x_o)(x-x_o) + \frac{1}{2}f''(x_o)(x-x_o)^2 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_o)}{n!}(x-x_o)^n$$

We could write this in terms of the operator $D = \frac{d}{dt}$ and the evaluation of $t = x_o$

$$f(x) = \left[\sum_{n=0}^{\infty} \frac{1}{n!}(x-t)^n D^n f(t)\right]_{t=x_o} =$$

I remind the reader that a function is called **entire** if it is analytic on all of $\mathbb{R}$, for example $e^x, \cos(x)$ and $\sin(x)$ are all entire. In particular, you should know that:

$$e^x = 1 + x + \frac{1}{2}x^2 + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}x^n$$

$$\cos(x) = 1 - \frac{1}{2}x^2 + \frac{1}{4!}x^4 \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}x^{2n}$$

$$\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}x^{2n+1}$$

Since $e^x = \cosh(x) + \sinh(x)$ it also follows that

$$\cosh(x) = 1 + \frac{1}{2}x^2 + \frac{1}{4!}x^4 \cdots = \sum_{n=0}^{\infty} \frac{1}{(2n)!}x^{2n}$$

$$\sinh(x) = x + \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \cdots = \sum_{n=0}^{\infty} \frac{1}{(2n+1)!}x^{2n+1}$$

The geometric series is often useful, for $a, r \in \mathbb{R}$ with $|r| < 1$ it is known

$$a + ar + ar^2 + \cdots = \sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$$

This generates a whole host of examples, for instance:

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots$$

$$\frac{1}{1-x^3} = 1 + x^3 + x^6 + x^9 + \cdots$$

$$\frac{x^3}{1-2x} = x^3(1 + 2x + (2x)^2 + \cdots) = x^3 + 2x^4 + 4x^5 + \cdots$$

Moreover, the term-by-term integration and differentiation theorems yield additional results in conjuction with the geometric series:

$$\tan^{-1}(x) = \int \frac{dx}{1+x^2} = \int \sum_{n=0}^{\infty} (-1)^n x^{2n} dx = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}x^{2n+1} = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 + \cdots$$

$$\ln(1-x) = \int \frac{d}{dx}\ln(1-x)dx = \int \frac{-1}{1-x}dx = -\int \sum_{n=0}^{\infty} x^n dx = \sum_{n=0}^{\infty} \frac{-1}{n+1}x^{n+1}$$

Of course, these are just the basic building blocks. We also can twist things and make the student use algebra,

$$e^{x+2} = e^x e^2 = e^2(1 + x + \frac{1}{2}x^2 + \cdots)$$

or trigonmetric identities,

$$\sin(x) = \sin(x - 2 + 2) = \sin(x-2)\cos(2) + \cos(x-2)\sin(2)$$

$$\Rightarrow \quad \sin(x) = \cos(2)\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(x-2)^{2n+1} + \sin(2)\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}(x-2)^{2n}.$$

Feel free to peruse my most recent calculus II materials to see a host of similarly sneaky calculations.

5.1.2 taylor's multinomial for two-variables

Suppose we wish to find the taylor polynomial centered at $(0,0)$ for $f(x,y) = e^x \sin(y)$. It is a simple as this:

$$f(x,y) = \left(1 + x + \frac{1}{2}x^2 + \cdots\right)\left(y - \frac{1}{6}y^3 + \cdots\right) = y + xy + \frac{1}{2}x^2y - \frac{1}{6}y^3 + \cdots$$

the resulting expression is called a multinomial since it is a polynomial in multiple variables. If all functions $f(x,y)$ could be written as $f(x,y) = F(x)G(y)$ then multiplication of series known from calculus II would often suffice. However, many functions do not possess this very special form. For example, how should we expand $f(x,y) = \cos(xy)$ about $(0,0)$?. We need to derive the two-dimensional Taylor's theorem.

We already know Taylor's theorem for functions on $\mathbb{R}$,

$$g(x) = g(a) + g'(a)(x-a) + \frac{1}{2}g''(a)(x-a)^2 + \cdots + \frac{1}{k!}g^{(k)}(a)(x-a)^k + R_k$$

and... If the remainder term vanishes as $k \to \infty$ then the function g is represented by the Taylor series given above and we write:

$$g(x) = \sum_{k=0}^{\infty} \frac{1}{k!}g^{(k)}(a)(x-a)^k.$$

Consider the function of two variables $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}$ which is smooth with smooth partial derivatives of all orders. Furthermore, let $(a,b) \in U$ and construct a line through (a,b) with direction vector (h_1,h_2) as usual:

$$\phi(t) = (a,b) + t(h_1,h_2) = (a + th_1, b + th_2)$$

for $t \in \mathbb{R}$. Note $\phi(0) = (a,b)$ and $\phi'(t) = (h_1,h_2) = \phi'(0)$. Construct $g = f \circ \phi : \mathbb{R} \to \mathbb{R}$ and choose $dom(g)$ such that $\phi(t) \in U$ for $t \in dom(g)$. This function g is a real-valued function of a real variable and we will be able to apply Taylor's theorem from calculus II on g. However, to differentiate g we'll need tools from calculus III to sort out the derivatives. In particular, as we differentiate g, note we use the chain rule for functions of several variables:

$$\begin{aligned} g'(t) = (f \circ \phi)'(t) &= f'(\phi(t))\phi'(t) \\ &= \nabla f(\phi(t)) \cdot (h_1,h_2) \\ &= h_1 f_x(a + th_1, b + th_2) + h_2 f_y(a + th_1, b + th_2) \end{aligned}$$

Note $g'(0) = h_1 f_x(a,b) + h_2 f_y(a,b)$. Differentiate again (I omit $(\phi(t))$ dependence in the last steps),

$$\begin{aligned} g''(t) &= h_1 f_x'(a + th_1, b + th_2) + h_2 f_y'(a + th_1, b + th_2) \\ &= h_1 \nabla f_x(\phi(t)) \cdot (h_1,h_2) + h_2 \nabla f_y(\phi(t)) \cdot (h_1,h_2) \\ &= h_1^2 f_{xx} + h_1 h_2 f_{yx} + h_2 h_1 f_{xy} + h_2^2 f_{yy} \\ &= h_1^2 f_{xx} + 2h_1 h_2 f_{xy} + h_2^2 f_{yy} \end{aligned}$$

Thus, making explicit the point dependence, $g''(0) = h_1^2 f_{xx}(a,b) + 2h_1 h_2 f_{xy}(a,b) + h_2^2 f_{yy}(a,b)$. We may construct the Taylor series for g up to quadratic terms:

$$\begin{aligned} g(0+t) &= g(0) + tg'(0) + \frac{1}{2}g''(0) + \cdots \\ &= f(a,b) + t[h_1 f_x(a,b) + h_2 f_y(a,b)] + \frac{t^2}{2}\left[h_1^2 f_{xx}(a,b) + 2h_1 h_2 f_{xy}(a,b) + h_2^2 f_{yy}(a,b)\right] + \cdots \end{aligned}$$

Note that $g(t) = f(a+th_1, b+th_2)$ hence $g(1) = f(a+h_1, b+h_2)$ and consequently,

$$f(a+h_1,b+h_2) = f(a,b) + h_1 f_x(a,b) + h_2 f_y(a,b) + \\ + \frac{1}{2}\Big[h_1^2 f_{xx}(a,b) + 2h_1h_2 f_{xy}(a,b) + h_2^2 f_{yy}(a,b)\Big] + \cdots$$

Omitting point dependence on the 2^{nd} derivatives,

$$\boxed{f(a+h_1,b+h_2) = f(a,b) + h_1 f_x(a,b) + h_2 f_y(a,b) + \tfrac{1}{2}\big[h_1^2 f_{xx} + 2h_1h_2 f_{xy} + h_2^2 f_{yy}\big] + \cdots}$$

Sometimes we'd rather have an expansion about (x, y). To obtain that formula simply substitute $x - a = h_1$ and $y - b = h_2$. Note that the point (a, b) is fixed in this discussion so the derivatives are not modified in this substitution,

$$f(x,y) = f(a,b) + (x-a) f_x(a,b) + (y-b) f_y(a,b) + \\ + \frac{1}{2}\Big[(x-a)^2 f_{xx}(a,b) + 2(x-a)(y-b) f_{xy}(a,b) + (y-b)^2 f_{yy}(a,b)\Big] + \cdots$$

At this point we ought to recognize the first three terms give the tangent plane to $z = f(z, y)$ at $(a, b, f(a, b))$. The higher order terms are nonlinear corrections to the linearization, these quadratic terms form a *quadratic form*. If we computed third, fourth or higher order terms we will find that, using $a = a_1$ and $b = a_2$ as well as $x = x_1$ and $y = x_2$,

$$\boxed{f(x,y) = \sum_{n=0}^{\infty} \sum_{i_1=0}^{2} \sum_{i_2=0}^{2} \cdots \sum_{i_n=0}^{2} \frac{1}{n!} \frac{\partial^{(n)} f(a_1,a_2)}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_n}} (x_{i_1} - a_{i_1})(x_{i_2} - a_{i_2}) \cdots (x_{i_n} - a_{i_n})}$$

Example 5.1.1. *Expand $f(x, y) = \cos(xy)$ about $(0, 0)$. We calculate derivatives,*

$$f_x = -y\sin(xy) \qquad f_y = -x\sin(xy)$$

$$f_{xx} = -y^2\cos(xy) \qquad f_{xy} = -\sin(xy) - xy\cos(xy) \qquad f_{yy} = -x^2\cos(xy)$$

$$f_{xxx} = y^3\sin(xy) \qquad f_{xxy} = -y\cos(xy) - y\cos(xy) + xy^2\sin(xy)$$

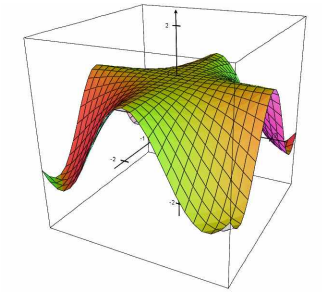
$$f_{xyy} = -x\cos(xy) - x\cos(xy) + x^2 y\sin(xy) \qquad f_{yyy} = x^3\sin(xy)$$

Next, evaluate at $x = 0$ and $y = 0$ to find $f(x, y) = 1 + \cdots$ to third order in x, y about $(0, 0)$. We can understand why these derivatives are all zero by approaching the expansion a different route: simply expand cosine directly in the variable (xy),

$$f(x,y) = 1 - \frac{1}{2}(xy)^2 + \frac{1}{4!}(xy)^4 + \cdots = 1 - \frac{1}{2}x^2y^2 + \frac{1}{4!}x^4y^4 + \cdots .$$

Apparently the given function only has nontrivial derivatives at $(0, 0)$ at orders $0, 4, 8, ...$ We can deduce that $f_{xxxxy}(0, 0) = 0$ without furthter calculation.

This is actually a very interesting function, I think it defies our analysis in the later portion of this chapter. The second order part of the expansion reveals nothing about the nature of the critical point $(0, 0)$. Of course, any student of trigonometry should recognize that $f(0, 0) = 1$ is likely a local maximum, it's certainly not a local minimum. The graph reveals that $f(0, 0)$ is a local maxium for f restricted to certain rays from the origin whereas it is constant on several special directions (the coordinate axes).



And, if you were wondering, yes, we could also derive this from subsitution of $u = xy$ into the standard expansion for $\cos(u) = 1 - \frac{1}{2}u^2 + \frac{1}{4!}u^4 + \cdots$. Often such subsitutions are the quickest way to generate interesting examples.

5.1.3 taylor's multinomial for many-variables

Suppose $f : dom(f) \subseteq \mathbb{R}^n \to \mathbb{R}$ is a function of n-variables and we seek to derive the Taylor series centered at $a = (a_1, a_2, \ldots, a_n)$. Once more consider the composition of f with a line in $dom(f)$. In particular, let $\phi : \mathbb{R} \to \mathbb{R}^n$ be defined by $\phi(t) = a + th$ where $h = (h_1, h_2, \ldots, h_n)$ gives the direction of the line and clearly $\phi'(t) = h$. Let $g : dom(g) \subseteq \mathbb{R} \to \mathbb{R}$ be defined by $g(t) = f(\phi(t))$ for all $t \in \mathbb{R}$ such that $\phi(t) \in dom(f)$. Differentiate, use the multivariate chain rule, recall here that $\nabla = e_1\frac{\partial}{\partial x_1} + e_2\frac{\partial}{\partial x_2} + \cdots + e_n\frac{\partial}{\partial x_n} = \sum_{i=1}^n e_i\partial_i$,

$$g'(t) = \nabla f(\phi(t)) \cdot \phi'(t) = \nabla f(\phi(t)) \cdot h = \sum_{i=1}^n h_i(\partial_i f)(\phi(t))$$

If we omit the explicit dependence on $\phi(t)$ then we find the simple formula $g'(t) = \sum_{i=1}^n h_i\partial_i f$. Differentiate a second time,

$$g''(t) = \frac{d}{dt}\left[\sum_{i=1}^n h_i\partial_i f(\phi(t))\right] = \sum_{i=1}^n h_i\frac{d}{dt}\Big[\big(\partial_i f\big)(\phi(t))\Big] = \sum_{i=1}^n h_i\big(\nabla\partial_i f\big)(\phi(t)) \cdot \phi'(t)$$

Omitting the $\phi(t)$ dependence and once more using $\phi'(t) = h$ we find

$$g''(t) = \sum_{i=1}^n h_i\nabla\partial_i f \cdot h$$

Recall that $\nabla = \sum_{j=1}^n e_j\partial_j$ and expand the expression above,

$$g''(t) = \sum_{i=1}^n h_i\left(\sum_{j=1}^n e_j\partial_j\partial_i f\right) \cdot h = \sum_{i=1}^n\sum_{j=1}^n h_ih_j\partial_j\partial_i f$$

where we should remember $\partial_j\partial_i f$ depends on $\phi(t)$. It should be clear that if we continue and take k-derivatives then we will obtain:

$$g^{(k)}(t) = \sum_{i_1=1}^n\sum_{i_2=1}^n\cdots\sum_{i_k=1}^n h_{i_1}h_{i_2}\cdots h_{i_k}\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}f$$

More explicitly,

$$g^{(k)}(t) = \sum_{i_1=1}^n\sum_{i_2=1}^n\cdots\sum_{i_k=1}^n h_{i_1}h_{i_2}\cdots h_{i_k}(\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}f)(\phi(t))$$

Hence, by Taylor's theorem, provided we are sufficiently close to $t = 0$ as to bound the remainder[1]

$$g(t) = \sum_{k=0}^{\infty} \frac{1}{k!} \Big(\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} h_{i_1} h_{i_2} \cdots h_{i_k} (\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f)(\phi(t)) \Big) t^k$$

Recall that $g(t) = f(\phi(t)) = f(a + th)$. Put[2] $t = 1$ and bring in the $\frac{1}{k!}$ to derive

$$\boxed{f(a+h) = \sum_{k=0}^{\infty} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \frac{1}{k!} \big(\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f\big)(a)\ h_{i_1} h_{i_2} \cdots h_{i_k}.}$$

Naturally, we sometimes prefer to write the series expansion about a as an expresssion in $x = a+h$. With this substitution we have $h = x - a$ and $h_{i_j} = (x-a)_{i_j} = x_{i_j} - a_{i_j}$ thus

$$\boxed{f(x) = \sum_{k=0}^{\infty} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \frac{1}{k!} \big(\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f\big)(a)\ (x_{i_1} - a_{i_1})(x_{i_2} - a_{i_2}) \cdots (x_{i_k} - a_{i_k}).}$$

Example 5.1.2. *Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ let's unravel the Taylor series centered at $(0,0,0)$ from the general formula boxed above. Utilize the notation $x = x_1, y = x_2$ and $z = x_3$ in this example.*

$$f(x) = \sum_{k=0}^{\infty} \sum_{i_1=1}^{3} \sum_{i_2=1}^{3} \cdots \sum_{i_k=1}^{3} \frac{1}{k!} \big(\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f\big)(0)\ x_{i_1} x_{i_2} \cdots x_{i_k}.$$

The terms to order 2 are as follows:

$$\begin{aligned} f(x) &= f(0) + f_x(0)x + f_y(0)y + f_z(0)z \\ &\quad + \tfrac{1}{2}\Big(f_{xx}(0)x^2 + f_{yy}(0)y^2 + f_{zz}(0)z^2 + \\ &\qquad + f_{xy}(0)xy + f_{xz}(0)xz + f_{yz}(0)yz + f_{yx}(0)yx + f_{zx}(0)zx + f_{zy}(0)zy \Big) + \cdots \end{aligned}$$

Partial derivatives commute for smooth functions hence,

$$\begin{aligned} f(x) &= f(0) + f_x(0)x + f_y(0)y + f_z(0)z \\ &\quad + \tfrac{1}{2}\Big(f_{xx}(0)x^2 + f_{yy}(0)y^2 + f_{zz}(0)z^2 + 2f_{xy}(0)xy + 2f_{xz}(0)xz + 2f_{yz}(0)yz \Big) \\ &\quad + \tfrac{1}{3!}\Big(f_{xxx}(0)x^3 + f_{yyy}(0)y^3 + f_{zzz}(0)z^3 + 3f_{xxy}(0)x^2y + 3f_{xxz}(0)x^2z \\ &\qquad + 3f_{yyz}(0)y^2z + 3f_{xyy}(0)xy^2 + 3f_{xzz}(0)xz^2 + 3f_{yzz}(0)yz^2 + 6f_{xyz}(0)xyz \Big) + \cdots \end{aligned}$$

[1] there exist smooth examples for which no neighborhood is small enough, the bump function in one-variable has higher-dimensional analogues, we focus our attention to functions for which it is possible for the series below to converge

[2] if $t = 1$ is not in the domain of g then we should rescale the vector h so that $t = 1$ places $\phi(1)$ in $dom(f)$, if f is smooth on some neighborhood of a then this is possible

Example 5.1.3. *Suppose* $f(x,y,z) = e^{xyz}$. *Find a quadratic approximation to* f *near* $(0,1,2)$. *Observe:*

$$f_x = yze^{xyz} \qquad f_y = xze^{xyz} \qquad f_z = xye^{xyz}$$
$$f_{xx} = (yz)^2e^{xyz} \qquad f_{yy} = (xz)^2e^{xyz} \qquad f_{zz} = (xy)^2e^{xyz}$$
$$f_{xy} = ze^{xyz} + xyz^2e^{xyz} \qquad f_{yz} = xe^{xyz} + x^2yze^{xyz} \qquad f_{xz} = ye^{xyz} + xy^2ze^{xyz}$$

Evaluating at $x = 0, y = 1$ *and* $z = 2$,

$$f_x(0,1,2) = 2 \qquad f_y(0,1,2) = 0 \qquad f_z(0,1,2) = 0$$
$$f_{xx}(0,1,2) = 4 \qquad f_{yy}(0,1,2) = 0 \qquad f_{zz}(0,1,2) = 0$$
$$f_{xy}(0,1,2) = 2 \qquad f_{yz}(0,1,2) = 0 \qquad f_{xz}(0,1,2) = 1$$

Hence, as $f(0,1,2) = e^0 = 1$ *we find*

$$f(x,y,z) = 1 + 2x + 2x^2 + 2x(y-1) + 2x(z-2) + \cdots$$

Another way to calculate this expansion is to make use of the adding zero trick,

$$f(x,y,z) = e^{x(y-1+1)(z-2+2)} = 1 + x(y-1+1)(z-2+2) + \frac{1}{2}\big[x(y-1+1)(z-2+2)\big]^2 + \cdots$$

Keeping only terms with two or less of x, $(y-1)$ *and* $(z-2)$ *variables,*

$$f(x,y,z) = 1 + 2x + x(y-1)(2) + x(1)(z-2) + \frac{1}{2}x^2(1)^2(2)^2 + \cdots$$

Which simplifies once more to $f(x,y,z) = 1 + 2x + 2x(y-1) + x(z-2) + 2x^2 + \cdots$.

5.2 a brief introduction to the theory of quadratic forms

Definition 5.2.1.

Generally, a **quadratic form** Q is a function $Q : \mathbb{R}^n \to \mathbb{R}$ whose formula can be written $Q(\vec{x}) = \vec{x}^T A\vec{x}$ for all $\vec{x} \in \mathbb{R}^n$ where $A \in \mathbb{R}^{n\times n}$ such that $A^T = A$. In particular, if $\vec{x} = (x,y)$ and $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ then

$$Q(\vec{x}) = \vec{x}^T A\vec{x} = ax^2 + bxy + byx + cy^2 = ax^2 + 2bxy + y^2.$$

The $n = 3$ case is similar,denote $A = [A_{ij}]$ and $\vec{x} = (x,y,z)$ so that

$$Q(\vec{x}) = \vec{x}^T A\vec{x} = A_{11}x^2 + 2A_{12}xy + 2A_{13}xz + A_{22}y^2 + 2A_{23}yz + A_{33}z^2.$$

Generally, if $[A_{ij}] \in \mathbb{R}^{n\times n}$ and $\vec{x} = [x_i]^T$ then the associated quadratic form is

$$Q(\vec{x}) = \vec{x}^T A\vec{x} = \sum_{i,j} A_{ij}x_ix_j = \sum_{i=1}^{n} A_{ii}x_i^2 + \sum_{i<j} 2A_{ij}x_ix_j.$$

In case you wondering, yes you could write a given quadratic form with a different matrix which is not symmetric, but we will find it convenient to insist that our matrix is symmetric since that

choice is always possible for a given quadratic form.

It is at times useful to use the dot-product to express a given quadratic form:

$$\vec{x}^T A\vec{x} = \vec{x} \cdot (A\vec{x}) = (A\vec{x}) \cdot \vec{x} = \vec{x}^T A^T \vec{x}$$

Some texts actually use the middle equality above to define a symmetric matrix.

Example 5.2.2.

$$2x^2 + 2xy + 2y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Example 5.2.3.

$$2x^2 + 2xy + 3xz - 2y^2 - z^2 = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 2 & 1 & 3/2 \\ 1 & -2 & 0 \\ 3/2 & 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Proposition 5.2.4.

The values of a quadratic form on $\mathbb{R}^n - \{0\}$ is completely determined by it's values on the $(n-1)$-sphere $S_{n-1} = \{\vec{x} \in \mathbb{R}^n \mid ||\vec{x}|| = 1\}$. In particular, $Q(\vec{x}) = ||\vec{x}||^2 Q(\hat{x})$ where $\hat{x} = \frac{1}{||\vec{x}||}\vec{x}$.

Proof: Let $Q(\vec{x}) = \vec{x}^T A\vec{x}$. Notice that we can write any nonzero vector as the product of its magnitude $||x||$ and its direction $\hat{x} = \frac{1}{||\vec{x}||}\vec{x}$,

$$Q(\vec{x}) = Q(||\vec{x}||\hat{x}) = (||\vec{x}||\hat{x})^T A||\vec{x}||\hat{x} = ||\vec{x}||^2 \hat{x}^T A\hat{x} = ||x||^2 Q(\hat{x}).$$

Therefore $Q(\vec{x})$ is simply proportional to $Q(\hat{x})$ with proportionality constant $||\vec{x}||^2$. □

The proposition above is very interesting. It says that if we know how Q works on unit-vectors then we can extrapolate its action on the remainder of $\mathbb{R}^n$. If $f : S \to \mathbb{R}$ then we could say $f(S) > 0$ iff $f(s) > 0$ for all $s \in S$. Likewise, $f(S) < 0$ iff $f(s) < 0$ for all $s \in S$. The proposition below follows from the proposition above since $||\vec{x}||^2$ ranges over all nonzero positive real numbers in the equations above.

Proposition 5.2.5.

If Q is a quadratic form on $\mathbb{R}^n$ and we denote $\mathbb{R}^n_* = \mathbb{R}^n - \{0\}$

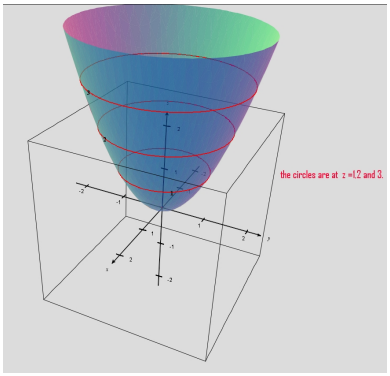
1.(negative definite) $Q(\mathbb{R}^n_*) < 0$ iff $Q(S_{n-1}) < 0$

2.(positive definite) $Q(\mathbb{R}^n_*) > 0$ iff $Q(S_{n-1}) > 0$

3.(non-definite) $Q(\mathbb{R}^n_*) = \mathbb{R} - \{0\}$ iff $Q(S_{n-1})$ has both positive and negative values.

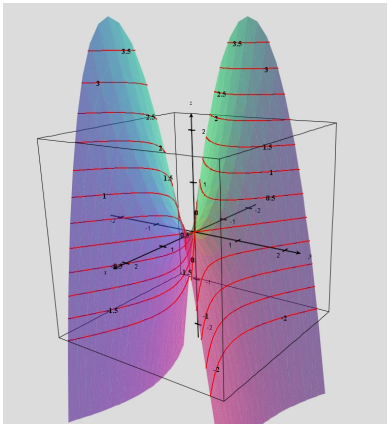
Before I get too carried away with the theory let's look at a couple examples.

Example 5.2.6. *Consider the quadric form $Q(x,y) = x^2 + y^2$. You can check for yourself that $z = Q(x,y)$ is a cone and Q has positive outputs for all inputs except $(0,0)$. Notice that $Q(v) = ||v||^2$ so it is clear that $Q(S_1) = 1$. We find agreement with the preceding proposition. Next, think about the application of $Q(x,y)$ to level curves; $x^2 + y^2 = k$ is simply a circle of radius $\sqrt{k}$ or just the origin. Here's a graph of $z = Q(x,y)$:*



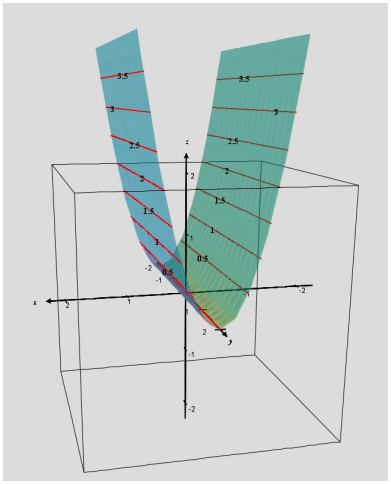
Notice that $Q(0,0) = 0$ is the absolute minimum for Q. Finally, let's take a moment to write $Q(x,y) = [x,y]\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$ in this case the matrix is diagonal and we note that the e-values are $\lambda_1 = \lambda_2 = 1$.

Example 5.2.7. *Consider the quadric form $Q(x,y) = x^2 - 2y^2$. You can check for yourself that $z = Q(x,y)$ is a hyperboloid and Q has non-definite outputs since sometimes the x^2 term dominates whereas other points have $-2y^2$ as the dominent term. Notice that $Q(1,0) = 1$ whereas $Q(0,1) = -2$ hence we find $Q(S_1)$ contains both positive and negative values and consequently we find agreement with the preceding proposition. Next, think about the application of $Q(x,y)$ to level curves; $x^2 - 2y^2 = k$ yields either hyperbolas which open vertically ($k > 0$) or horizontally ($k < 0$) or a pair of lines $y = \pm\frac{x}{2}$ in the $k = 0$ case. Here's a graph of $z = Q(x,y)$:*



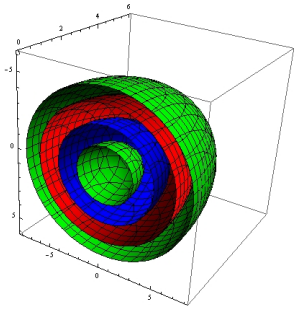
The origin is a **saddle point**. *Finally, let's take a moment to write $Q(x,y) = [x,y]\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$ in this case the matrix is diagonal and we note that the e-values are $\lambda_1 = 1$ and $\lambda_2 = -2$.*

Example 5.2.8. *Consider the quadric form $Q(x,y) = 3x^2$. You can check for yourself that $z = Q(x,y)$ is parabola-shaped trough along the y-axis. In this case Q has positive outputs for all inputs except $(0,y)$, we would call this form* **positive semi-definite**. *A short calculation reveals that $Q(S_1) = [0,3]$ thus we again find agreement with the preceding proposition (case 3). Next, think about the application of $Q(x,y)$ to level curves; $3x^2 = k$ is a pair of vertical lines: $x = \pm\sqrt{k/3}$ or just the y-axis. Here's a graph of $z = Q(x,y)$:*



Finally, let's take a moment to write $Q(x,y) = [x,y]\begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$ *in this case the matrix is diagonal and we note that the e-values are $\lambda_1 = 3$ and $\lambda_2 = 0$.*

Example 5.2.9. *Consider the quadric form $Q(x,y,z) = x^2+2y^2+3z^2$. Think about the application of $Q(x,y,z)$ to level surfaces; $x^2 + 2y^2 + 3z^2 = k$ is an ellipsoid. I can't graph a function of three variables, however, we can look at level surfaces of the function. I use Mathematica to plot several below:*



Finally, let's take a moment to write $Q(x,y,z) = [x,y,z]\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ *in this case the matrix is diagonal and we note that the e-values are $\lambda_1 = 1$ and $\lambda_2 = 2$ and $\lambda_3 = 3$.*

5.2.1 diagonalizing forms via eigenvectors

The examples given thus far are the simplest cases. We don't really need linear algebra to understand them. In contrast, e-vectors and e-values will prove a useful tool to unravel the later examples[3]

[3] this is the one place in this course where we need eigenvalues and eigenvector calculations, I include these to illustrate the structure of quadratic forms in general, however, as linear algebra is not a prerequisite you may find some things in this section mysterious. The homework and study guide will elaborate on what is required this semester

Definition 5.2.10.

Let $A \in \mathbb{R}^{n\times n}$. If $v \in \mathbb{R}^{n\times 1}$ is **nonzero** and $Av = \lambda v$ for some $\lambda \in \mathbb{C}$ then we say v is an **eigenvector** with **eigenvalue** λ of the matrix A.

Proposition 5.2.11.

Let $A \in \mathbb{R}^{n\times n}$ then λ is an eigenvalue of A iff $det(A-\lambda I) = 0$. We say $P(\lambda) = det(A-\lambda I)$ the **characteristic polynomial** and $det(A-\lambda I) = 0$ is the **characteristic equation**.

Proof: Suppose λ is an eigenvalue of A then there exists a nonzero vector v such that $Av = \lambda v$ which is equivalent to $Av - \lambda v = 0$ which is precisely $(A-\lambda I)v = 0$. Notice that $(A-\lambda I)0 = 0$ thus the matrix $(A-\lambda I)$ is singular as the equation $(A-\lambda I)x = 0$ has more than one solution. Consequently $det(A-\lambda I) = 0$.

Conversely, suppose $det(A-\lambda I) = 0$. It follows that $(A-\lambda I)$ is singular. Clearly the system $(A-\lambda I)x = 0$ is consistent as $x = 0$ is a solution hence we know there are infinitely many solutions. In particular there exists at least one vector $v \neq 0$ such that $(A-\lambda I)v = 0$ which means the vector v satisfies $Av = \lambda v$. Thus v is an eigenvector with eigenvalue λ for A. □

Remark 5.2.12.

I found a pretty derivation of the eigenvector condition from the method of Lagrange multipliers. I shared in the Lecture 10 part 1. It's likely I cover that argument again in Lecture this year, my apologies it has not made it to these notes at this time.

Example 5.2.13. *Let* $A = \begin{bmatrix} 3 & 1 \\ 3 & 1 \end{bmatrix}$ *find the e-values and e-vectors of* A.

$$det(A-\lambda I) = det\begin{bmatrix} 3-\lambda & 1 \\ 3 & 1-\lambda \end{bmatrix} = (3-\lambda)(1-\lambda) - 3 = \lambda^2 - 4\lambda = \lambda(\lambda-4) = 0$$

We find $\lambda_1 = 0$ *and* $\lambda_2 = 4$. *Now find the e-vector with e-value* $\lambda_1 = 0$, *let* $u_1 = [u,v]^T$ *denote the e-vector we wish to find. Calculate,*

$$(A-0I)u_1 = \begin{bmatrix} 3 & 1 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 3u+v \\ 3u+v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Obviously the equations above are redundant and we have infinitely many solutions of the form $3u+v = 0$ *which means* $v = -3u$ *so we can write,* $u_1 = \begin{bmatrix} u \\ -3u \end{bmatrix} = u\begin{bmatrix} 1 \\ -3 \end{bmatrix}$. *In applications we often make a choice to select a particular e-vector. Most modern graphing calculators can calculate e-vectors. It is customary for the e-vectors to be chosen to have length one. That is a useful choice for certain applications as we will later discuss. If you use a calculator it would likely give* $u_1 = \frac{1}{\sqrt{10}}\begin{bmatrix} 1 \\ -3 \end{bmatrix}$ *although the* $\sqrt{10}$ *would likely be approximated unless your calculator is smart.*

Continuing we wish to find eigenvectors $u_2 = [u,v]^T$ *such that* $(A-4I)u_2 = 0$. *Notice that* u, v *are disposable variables in this context, I do not mean to connect the formulas from the* $\lambda = 0$ *case with the case considered now.*

$$(A-4I)u_1 = \begin{bmatrix} -1 & 1 \\ 3 & -3 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -u+v \\ 3u-3v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Again the equations are redundant and we have infinitely many solutions of the form $v = u$. Hence, $u_2 = \begin{bmatrix} u \\ u \end{bmatrix} = u \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ *is an eigenvector for any $u \in \mathbb{R}$ such that $u \neq 0$.*

Theorem 5.2.14.

A matrix $A \in \mathbb{R}^{n \times n}$ is symmetric iff there exists an orthonormal eigenbasis for A.

There is a geometric proof of this theorem in Edwards[4] (see Theorem 8.6 pgs 146-147) . I prove half of this theorem in my linear algebra notes by a non-geometric argument (full proof is in Appendix C of Insel,Spence and Friedberg). It might be very interesting to understand the connection between the geometric verse algebraic arguments. We'll content ourselves with an example here:

Example 5.2.15. *Let* $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix}$. *Observe that* $det(A - \lambda I) = -\lambda(\lambda + 1)(\lambda - 3)$ *thus* $\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = 3$. *We can calculate orthonormal e-vectors of* $v_1 = [1, 0, 0]^T$, $v_2 = \frac{1}{\sqrt{2}}[0, 1, -1]^T$ *and* $v_3 = \frac{1}{\sqrt{2}}[0, 1, 1]^T$. *I invite the reader to check the validity of the following equation:*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Its really neat that to find the inverse of a matrix of orthonormal e-vectors we need only take the transpose; note

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Proposition 5.2.16.

If Q is a quadratic form on $\mathbb{R}^n$ with matrix A and e-values $\lambda_1, \lambda_2, \ldots, \lambda_n$ with orthonormal e-vectors $v_1, v_2, \ldots, v_n$ then

$$Q(v_i) = {\lambda_i}^2$$

for $i = 1, 2, \ldots, n$. Moreover, if $P = [v_1|v_2|\cdots|v_n]$ then

$$Q(\vec{x}) = (P^T\vec{x})^T P^T A P P^T \vec{x} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

where we defined $\vec{y} = P^T\vec{x}$.

Let me restate the proposition above in simple terms: we can transform a given quadratic form to a diagonal form by finding orthonormalized e-vectors and performing the appropriate coordinate transformation. Since P is formed from orthonormal e-vectors we know that P will be either a rotation or reflection. This proposition says we can remove "cross-terms" by transforming the quadratic forms with an appropriate rotation.

[4] think about it, there is a 1-1 correspondance between symmetric matrices and quadratic forms

Example 5.2.17. *Consider the quadric form $Q(x,y) = 2x^2 + 2xy + 2y^2$. It's not immediately obvious (to me) what the level curves $Q(x,y) = k$ look like. We'll make use of the preceding proposition to understand those graphs. Notice $Q(x,y) = [x,y]\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$. Denote the matrix of the form by A and calculate the e-values/vectors:*

$$det(A - \lambda I) = det\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = (\lambda - 2)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) = 0$$

Therefore, the e-values are $\lambda_1 = 1$ and $\lambda_2 = 3$.

$$(A - I)\vec{u}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

I just solved $u + v = 0$ to give $v = -u$ choose $u = 1$ then normalize to get the vector above. Next,

$$(A - 3I)\vec{u}_2 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

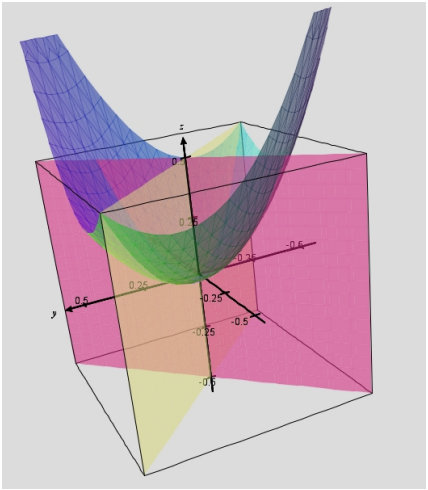
I just solved $u - v = 0$ to give $v = u$ choose $u = 1$ then normalize to get the vector above. Let $P = [\vec{u}_1|\vec{u}_2]$ and introduce new coordinates $\vec{y} = [\bar{x}, \bar{y}]^T$ defined by $\vec{y} = P^T\vec{x}$. Note these can be inverted by multiplication by P to give $\vec{x} = P\vec{y}$. Observe that

$$P = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \Rightarrow \quad \begin{array}{rl} x &= \frac{1}{2}(\bar{x} + \bar{y}) \\ y &= \frac{1}{2}(-\bar{x} + \bar{y}) \end{array} \quad or \quad \begin{array}{rl} \bar{x} &= \frac{1}{2}(x - y) \\ \bar{y} &= \frac{1}{2}(x + y) \end{array}$$

The proposition preceding this example shows that substitution of the formulas above into Q yield[5]*:*

$$\tilde{Q}(\bar{x}, \bar{y}) = \bar{x}^2 + 3\bar{y}^2$$

It is clear that in the barred coordinate system the level curve $Q(x,y) = k$ is an ellipse. If we draw the barred coordinate system superposed over the xy-coordinate system then you'll see that the graph of $Q(x,y) = 2x^2 + 2xy + 2y^2 = k$ is an ellipse rotated by 45 degrees. Or, if you like, we can plot $z = Q(x,y)$:



[5] technically $\tilde{Q}(\bar{x}, \bar{y})$ is $Q(x(\bar{x}, \bar{y}), y(\bar{x}, \bar{y}))$

Example 5.2.18. *Consider the quadric form $Q(x,y) = x^2+2xy+y^2$. It's not immediately obvious (to me) what the level curves $Q(x,y) = k$ look like. We'll make use of the preceding proposition to understand those graphs. Notice $Q(x,y) = [x,y]\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$. Denote the matrix of the form by A and calculate the e-values/vectors:*

$$det(A-\lambda I) = det\begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix} = (\lambda-1)^2 - 1 = \lambda^2 - 2\lambda = \lambda(\lambda-2) = 0$$

Therefore, the e-values are $\lambda_1 = 0$ and $\lambda_2 = 2$.

$$(A-0)\vec{u}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

I just solved $u+v=0$ to give $v=-u$ choose $u=1$ then normalize to get the vector above. Next,

$$(A-2I)\vec{u}_2 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

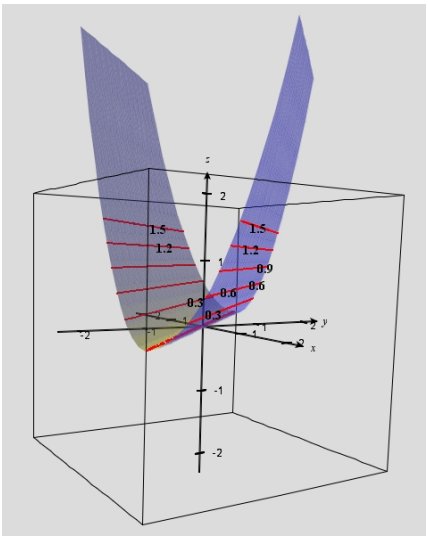
I just solved $u-v=0$ to give $v=u$ choose $u=1$ then normalize to get the vector above. Let $P = [\vec{u}_1|\vec{u}_2]$ and introduce new coordinates $\vec{y} = [\bar{x},\bar{y}]^T$ defined by $\vec{y} = P^T\vec{x}$. Note these can be inverted by multiplication by P to give $\vec{x} = P\vec{y}$. Observe that

$$P = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \Rightarrow \quad \begin{array}{rl} x &= \frac{1}{2}(\bar{x}+\bar{y}) \\ y &= \frac{1}{2}(-\bar{x}+\bar{y}) \end{array} \quad or \quad \begin{array}{rl} \bar{x} &= \frac{1}{2}(x-y) \\ \bar{y} &= \frac{1}{2}(x+y) \end{array}$$

The proposition preceding this example shows that substitution of the formulas above into Q yield:

$$\tilde{Q}(\bar{x},\bar{y}) = 2\bar{y}^2$$

It is clear that in the barred coordinate system the level curve $Q(x,y) = k$ is a pair of paralell lines. If we draw the barred coordinate system superposed over the xy-coordinate system then you'll see that the graph of $Q(x,y) = x^2+2xy+y^2 = k$ is a line with slope -1. Indeed, with a little algebraic insight we could have anticipated this result since $Q(x,y) = (x+y)^2$ so $Q(x,y) = k$ implies $x+y = \sqrt{k}$ thus $y = \sqrt{k}-x$. Here's a plot which again verifies what we've already found:



Example 5.2.19. *Consider the quadric form $Q(x,y) = 4xy$. It's not immediately obvious (to me) what the level curves $Q(x,y) = k$ look like. We'll make use of the preceding proposition to understand those graphs. Notice $Q(x,y) = [x,y]\begin{bmatrix} 0 & 2 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$. Denote the matrix of the form by A and calculate the e-values/vectors:*

$$det(A - \lambda I) = det\begin{bmatrix} -\lambda & 2 \\ 2 & -\lambda \end{bmatrix} = \lambda^2 - 4 = (\lambda + 2)(\lambda - 2) = 0$$

Therefore, the e-values are $\lambda_1 = -2$ and $\lambda_2 = 2$.

$$(A + 2I)\vec{u}_1 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

I just solved $u + v = 0$ to give $v = -u$ choose $u = 1$ then normalize to get the vector above. Next,

$$(A - 2I)\vec{u}_2 = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \vec{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

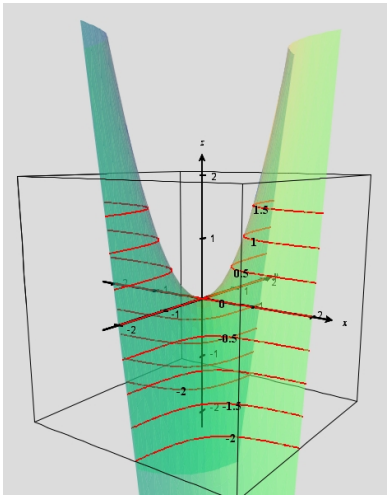
I just solved $u - v = 0$ to give $v = u$ choose $u = 1$ then normalize to get the vector above. Let $P = [\vec{u}_1|\vec{u}_2]$ and introduce new coordinates $\vec{y} = [\bar{x}, \bar{y}]^T$ defined by $\vec{y} = P^T\vec{x}$. Note these can be inverted by multiplication by P to give $\vec{x} = P\vec{y}$. Observe that

$$P = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} x &= \tfrac{1}{2}(\bar{x} + \bar{y}) \\ y &= \tfrac{1}{2}(-\bar{x} + \bar{y}) \end{aligned} \quad or \quad \begin{aligned} \bar{x} &= \tfrac{1}{2}(x - y) \\ \bar{y} &= \tfrac{1}{2}(x + y) \end{aligned}$$

The proposition preceding this example shows that substitution of the formulas above into Q yield:

$$\tilde{Q}(\bar{x}, \bar{y}) = -2\bar{x}^2 + 2\bar{y}^2$$

It is clear that in the barred coordinate system the level curve $Q(x,y) = k$ is a hyperbola. If we draw the barred coordinate system superposed over the xy-coordinate system then you'll see that the graph of $Q(x,y) = 4xy = k$ is a hyperbola rotated by 45 degrees. The graph $z = 4xy$ is thus a hyperbolic paraboloid:



The fascinating thing about the mathematics here is that if you don't want to graph $z = Q(x,y)$, but you do want to know the general shape then you can determine which type of quadraic surface you're dealing with by simply calculating the eigenvalues of the form.

Remark 5.2.20.

I made the preceding triple of examples all involved the same rotation. This is purely for my lecturing convenience. In practice the rotation could be by all sorts of angles. In addition, you might notice that a different ordering of the e-values would result in a redefinition of the barred coordinates. [6]

We ought to do at least one 3-dimensional example.

Example 5.2.21. *Consider the quadric form Q defined below:*

$$Q(x,y,z) = [x,y,z]\begin{bmatrix} 6 & -2 & 0 \\ -2 & 6 & 0 \\ 0 & 0 & 5 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Denote the matrix of the form by A and calculate the e-values/vectors:

$$\begin{aligned} det(A-\lambda I) &= det\begin{bmatrix} 6-\lambda & -2 & 0 \\ -2 & 6-\lambda & 0 \\ 0 & 0 & 5-\lambda \end{bmatrix} \\ &= [(\lambda-6)^2-4](5-\lambda) \\ &= (5-\lambda)[\lambda^2-12\lambda+32](5-\lambda) \\ &= (\lambda-4)(\lambda-8)(5-\lambda) \end{aligned}$$

Therefore, the e-values are $\lambda_1 = 4,\ \lambda_2 = 8$ and $\lambda_3 = 5$. After some calculation we find the following orthonormal e-vectors for A:

$$\vec{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad \vec{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \qquad \vec{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

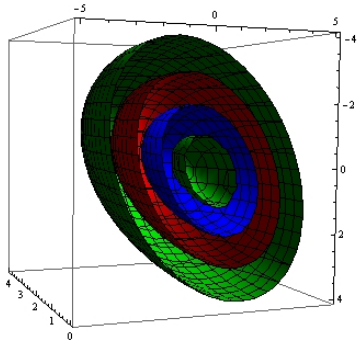
Let $P = [\vec{u}_1|\vec{u}_2|\vec{u}_3]$ and introduce new coordinates $\vec{y} = [\bar{x},\bar{y},\bar{z}]^T$ defined by $\vec{y} = P^T\vec{x}$. Note these can be inverted by multiplication by P to give $\vec{x} = P\vec{y}$. Observe that

$$P = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} x &= \tfrac{1}{2}(\bar{x}+\bar{y}) \\ y &= \tfrac{1}{2}(-\bar{x}+\bar{y}) \\ z &= \bar{z} \end{aligned} \quad or \quad \begin{aligned} \bar{x} &= \tfrac{1}{2}(x-y) \\ \bar{y} &= \tfrac{1}{2}(x+y) \\ \bar{z} &= z \end{aligned}$$

The proposition preceding this example shows that substitution of the formulas above into Q yield:

$$\tilde{Q}(\bar{x},\bar{y},\bar{z}) = 4\bar{x}^2 + 8\bar{y}^2 + 5\bar{z}^2$$

It is clear that in the barred coordinate system the level surface $Q(x,y,z) = k$ is an ellipsoid. If we draw the barred coordinate system superposed over the xyz-coordinate system then you'll see that the graph of $Q(x,y,z) = k$ is an ellipsoid rotated by 45 degrees around the $z-axis$. Plotted below are a few representative ellipsoids:



In summary, the behaviour of a quadratic form $Q(x) = x^T Ax$ is governed by it's set of eigenvalues[7] $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. Moreover, the form can be written as $Q(y) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_k y_k^2$ by choosing the coordinate system which is built from the orthonormal eigenbasis of $col(A)$. In this coordinate system the shape of the level-sets of Q becomes manifest from the signs of the e-values.)

Remark 5.2.22.

If you would like to read more about conic sections or quadric surfaces and their connection to e-values/vectors I reccommend sections 9.6 and 9.7 of Anton's linear algebra text. I have yet to add examples on how to include translations in the analysis. It's not much more trouble but I decided it would just be an unecessary complication this semester. Also, section 7.1,7.2 and 7.3 in Lay's linear algebra text show a bit more about how to use this math to solve concrete applied problems. You might also take a look in Gilbert Strang's linear algebra text, his discussion of tests for positive-definite matrices is much more complete than I will give here.

5.3 second derivative test in many-variables

There is a connection between the shape of level curves $Q(x_1, x_2, \ldots, x_n) = k$ and the graph $x_{n+1} = f(x_1, x_2, \ldots, x_n)$ of f. I'll discuss $n = 2$ but these comments equally well apply to $w = f(x, y, z)$ or higher dimensional examples. Consider a critical point (a, b) for $f(x, y)$ then the Taylor expansion about (a, b) has the form

$$f(a + h, b + k) = f(a, b) + Q(h, k)$$

where $Q(h, k) = \frac{1}{2}h^2 f_{xx}(a, b) + hk f_{xy}(a, b) + \frac{1}{2}h^2 f_{yy}(a, b) = [h, k][Q](h, k)$. Since $[Q]^T = [Q]$ we can find orthonormal e-vectors $\vec{u}_1, \vec{u}_2$ for $[Q]$ with e-values λ_1 and λ_2 respective. Using $U = [\vec{u}_1|\vec{u}_2]$ we can introduce rotated coordinates $(\bar{h}, \bar{k}) = U(h, k)$. These will give

$$Q(\bar{h}, \bar{k}) = \lambda_1 \bar{h}^2 + \lambda_2 \bar{k}^2$$

Clearly if $\lambda_1 > 0$ and $\lambda_2 > 0$ then $f(a, b)$ yields the local minimum whereas if $\lambda_1 < 0$ and $\lambda_2 < 0$ then $f(a, b)$ yields the local maximum. Edwards discusses these matters on pgs. 148-153. In short, supposing $f \approx f(p) + Q$, if all the e-values of Q are positive then f has a local minimum of $f(p)$ at p whereas if all the e-values of Q are negative then f reaches a local maximum of $f(p)$ at p. Otherwise Q has both positive and negative e-values and we say Q is non-definite and the function has a saddle point. If all the e-values of Q are positive then Q is said to be **positive-definite** whereas if all the e-values of Q are negative then Q is said to be **negative-definite**. Edwards gives a few nice tests for ascertaining if a matrix is positive definite without explicit computation of e-values. Finally, if one of the e-values is zero then the graph will be like a trough.

[7] this set is called the spectrum of the matrix

Example 5.3.1. *Suppose* $f(x,y) = exp(-x^2 - y^2 + 2y - 1)$ *expand* f *about the point* $(0,1)$*:*

$$f(x,y) = exp(-x^2)exp(-y^2 + 2y - 1) = exp(-x^2)exp(-(y-1)^2)$$

expanding,

$$f(x,y) = (1 - x^2 + \cdots)(1 - (y-1)^2 + \cdots) = 1 - x^2 - (y-1)^2 + \cdots$$

Recenter about the point $(0,1)$ *by setting* $x = h$ *and* $y = 1 + k$ *so*

$$f(h, 1+k) = 1 - h^2 - k^2 + \cdots$$

If (h,k) *is near* $(0,0)$ *then the dominant terms are simply those we've written above hence the graph is like that of a quadraic surface with a pair of negative e-values. It follows that* $f(0,1)$ *is a local maximum. In fact, it happens to be a global maximum for this function.*

Example 5.3.2. *Suppose* $f(x,y) = 4 - (x-1)^2 + (y-2)^2 + Aexp(-(x-1)^2 - (y-2)^2) + 2B(x-1)(y-2)$ *for some constants* A, B*. Analyze what values for* A, B *will make* $(1,2)$ *a local maximum, minimum or neither. Expanding about* $(1,2)$ *we set* $x = 1 + h$ *and* $y = 2 + k$ *in order to see clearly the local behaviour of* f *at* $(1,2)$*,*

$$\begin{aligned} f(1+h, 2+k) &= 4 - h^2 - k^2 + Aexp(-h^2 - k^2) + 2Bhk \\ &= 4 - h^2 - k^2 + A(1 - h^2 - k^2) + 2Bhk \cdots \\ &= 4 + A - (A+1)h^2 + 2Bhk - (A+1)k^2 + \cdots \end{aligned}$$

There is no nonzero linear term in the expansion at $(1,2)$ *which indicates that* $f(1,2) = 4 + A$ *may be a local extremum. In this case the quadratic terms are nontrivial which means the graph of this function is well-approximated by a quadraic surface near* $(1,2)$*. The quadratic form* $Q(h,k) = -(A+1)h^2 + 2Bhk - (A+1)k^2$ *has matrix*

$$[Q] = \begin{bmatrix} -(A+1) & B \\ B & -(A+1)^2 \end{bmatrix}.$$

The characteristic equation for Q *is*

$$det([Q] - \lambda I) = det \begin{bmatrix} -(A+1) - \lambda & B \\ B & -(A+1)^2 - \lambda \end{bmatrix} = (\lambda + A + 1)^2 - B^2 = 0$$

We find solutions $\lambda_1 = -A - 1 + B$ *and* $\lambda_2 = -A - 1 - B$*. The possibilities break down as follows:*

1. *if* $\lambda_1, \lambda_2 > 0$ *then* $f(1,2)$ *is local minimum.*
2. *if* $\lambda_1, \lambda_2 < 0$ *then* $f(1,2)$ *is local maximum.*
3. *if just one of* λ_1, λ_2 *is zero then* f *is constant along one direction and min/max along another so technically it is a local extremum.*
4. *if* $\lambda_1 \lambda_2 < 0$ *then* $f(1,2)$ *is not a local etremum, however it is a saddle point.*

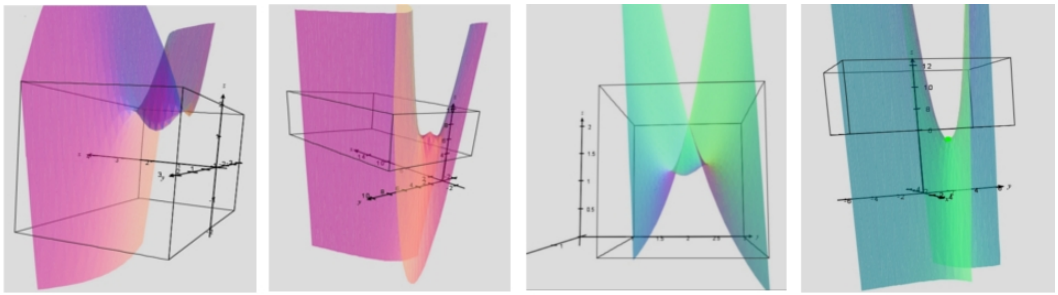
In particular, the following choices for A, B *will match the choices above*

1. *Let* $A = -3$ *and* $B = 1$ *so* $\lambda_1 = 3$ *and* $\lambda_2 = 1$*;*
2. *Let* $A = 3$ *and* $B = 1$ *so* $\lambda_1 = -3$ *and* $\lambda_2 = -5$

3. *Let $A = -3$ and $B = -2$ so $\lambda_1 = 0$ and $\lambda_2 = 4$*

4. *Let $A = 1$ and $B = 3$ so $\lambda_1 = 1$ and $\lambda_2 = -5$*

Here are the graphs of the cases above, note the analysis for case 3 is more subtle for Taylor approximations as opposed to simple quadraic surfaces. In this example, case 3 was also a local minimum. In contrast, in Example 5.2.18 the graph was like a trough. The behaviour of f away from the critical point includes higher order terms whose influence turns the trough into a local minimum.



Example 5.3.3. *Suppose $f(x,y) = \sin(x)\cos(y)$ to find the Taylor series centered at $(0,0)$ we can simply multiply the one-dimensional result $\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \cdots$ and $\cos(y) = 1 - \frac{1}{2!}y^2 + \frac{1}{4!}y^4 + \cdots$ as follows:*

$$\begin{aligned} f(x,y) &= (x - \tfrac{1}{3!}x^3 + \tfrac{1}{5!}x^5 + \cdots)(1 - \tfrac{1}{2!}y^2 + \tfrac{1}{4!}y^4 + \cdots) \\ &= x - \tfrac{1}{2}xy^2 + \tfrac{1}{24}xy^4 - \tfrac{1}{6}x^3 - \tfrac{1}{12}x^3y^2 + \cdots \\ &= x + \cdots \end{aligned}$$

The origin $(0,0)$ is a critical point since $f_x(0,0) = 0$ and $f_y(0,0) = 0$, however, this particular critical point escapes the analysis via the quadratic form term since $Q = 0$ in the Taylor series for this function at $(0,0)$. This is analogous to the inconclusive case of the 2nd derivative test in calculus III.

Example 5.3.4. *Suppose $f(x,y,z) = xyz$. Calculate the multivariate Taylor expansion about the point $(1,2,3)$. I'll actually calculate this one via differentiation, I have used tricks and/or calculus II results to shortcut any differentiation in the previous examples. Calculate first derivatives*

$$f_x = yz \qquad f_y = xz \qquad f_z = xy,$$

and second derivatives,

$$\begin{aligned} f_{xx} &= 0 & f_{xy} &= z & f_{xz} &= y \\ f_{yx} &= z & f_{yy} &= 0 & f_{yz} &= x \\ f_{zx} &= y & f_{zy} &= x & f_{zz} &= 0, \end{aligned}$$

and the nonzero third derivatives,

$$f_{xyz} = f_{yzx} = f_{zxy} = f_{zyx} = f_{yxz} = f_{xzy} = 1.$$

It follows,

$$\begin{aligned} &f(a+h,b+k,c+l) = \\ &\quad = f(a,b,c) \ + \ f_x(a,b,c)h \ + \ f_y(a,b,c)k \ + \ f_z(a,b,c)l \ + \\ &\qquad \tfrac{1}{2}(\ f_{xx}hh + f_{xy}hk + f_{xz}hl + f_{yx}kh + f_{yy}kk + f_{yz}kl + f_{zx}lh + f_{zy}lk + f_{zz}ll\ ) + \cdots \end{aligned}$$

Of course certain terms can be combined since $f_{xy} = f_{yx}$ etc... for smooth functions (we assume smooth in this section, moreover the given function here is clearly smooth). In total,

$$f(1+h, 2+k, 3+l) = 6 + 6h + 3k + 2l + \frac{1}{2}\big(3hk + 2hl + 3kh + kl + 2lh + lk\big) + \frac{1}{3!}(6)hkl$$

Of course, we could also obtain this from simple algebra:

$$f(1+h, 2+k, 3+l) = (1+h)(2+k)(3+l) = 6 + 6h + 3k + l + 3hk + 2hl + kl + hkl.$$

Chapter 6

introduction to variational calculus

6.1 history

The problem of variational calculus is almost as old as modern calculus. Variational calculus seeks to answer questions such as:

Remark 6.1.1.

1. what is the shortest path between two points on a surface ?
2. what is the path of least time for a mass sliding without friction down some path between two given points ?
3. what is the path which minimizes the energy for some physical system ?
4. given two points on the x-axis and a particular area what curve has the longest perimeter and bounds that area between those points and the x-axis?

You'll notice these all involve a variable which is not a real variable or even a vector-valued-variable. Instead, the answers to the questions posed above will be **paths** or **curves** depending on how you wish to frame the problem. In variational calculus the variable is a function and we wish to find extreme values for a **functional**. In short, a functional is an abstract function of functions. A functional takes as an input a function and gives as an output a number. The space from which these functions are taken varies from problem to problem. Often we put additional **contraints** or **conditions** on the **space of admissable solutions**. To read about the full generality of the problem you should look in a text such as Hans Sagan's. Our treatment is introductory in this chapter, my aim is to show you why it is plausible and then to show you how we use variational calculus.

We will see that the problem of finding an extreme value for a functional is equivalent to solving the Euler-Lagrange equations or Euler equations for the functional. Euler predates Lagrange in his discovery of the equations bearing their names. Eulers's initial attack of the problem was to chop the hypothetical solution curve up into a polygonal path. The unknowns in that approach were the coordinates of the vertices in the polygonal path. Then through some ingenious calculations he arrived at the Euler-Lagrange equations. Apparently there were logical flaws in Euler's original treatment. Lagrange later derived the same equations using the viewpoint that the variable was a function and the **variation** was one of shifting by an arbitrary function. The treatment of

variational calculus in Edwards is neither Euler nor Lagrange's approach, it is a refined version which takes in the contributions of generations of mathematicians working on the subject and then merges it with careful functional analysis. I'm no expert of the full history, I just give you a rough sketch of what I've gathered from reading a few variational calculus texts.

Physics played a large role in the development of variational calculus. Lagrange was a physicist as well as a mathematician. At the present time, every physicist takes course(s) in *Lagrangian Mechanics*. Moreover, the use of variational calculus is fundamental since Hamilton's principle says that all physics can be derived from the principle of least action. In short this means that nature is lazy. The solutions realized in the physical world are those which minimize the action. The action

$$S[y] = \int L(y, y', t)\, dt$$

is constructed from the Lagrangian $L = T - U$ where T is the kinetic energy and U is the potential energy. In the case of classical mechanics the Euler Lagrange equations are precisely Newton's equations. The Hamiltonian $H = T + U$ is similar to the Lagrangian except that the fundamental variables are taken to be momentum and position in contrast to velocity and position in Lagrangian mechanics.

Hamiltonians and Lagrangians are used to set-up new physical theories. Euler-Lagrange equations are said to give the so-called *classical limit* of modern field theories. The concept of a force is not so useful to quantum theories, instead the concept of energy plays the central role. Moreover, the problem of quantizing and then renormalizing field theory brings in very sophisiticated mathematics. In fact, the math of modern physics is not understood. In this chapter I'll just show you a few famous classical mechanics problems which are beatifully solved by Lagrange's approach. We'll also see how expressing the Lagrangian in non-Cartesian coordinates can give us an easy way to derive forces that arise from geometric contraints.

I am following the typical physics approach to variational calculus. Edwards' last chapter is more natural mathematically but I think the math is a bit much for your first exposure to the subject. The treatment given here is close to that of Arfken and Weber's Mathematical Physics text, however I suspect you can find these calculations in dozens of classical mechanics texts. More or less our approach is that of Lagrange.

6.2 the variational problem

Our goal in what follows here is to maximize or minimize a particular function of functions. Suppose $\mathcal{F}_o$ is a set of functions with some particular property. For now, we may could assume that all the functions in $\mathcal{F}_o$ have graphs that include (x_1, y_1) and (x_2, y_2). Consider a functional $J : \mathcal{F}_o \to \mathcal{F}_o$ which is defined by an integral of some function f which we call the **Lagrangian**,

$$J[y] = \int_{x_1}^{x_2} f(y, y', x)\, dx.$$

We suppose that f is given but y is a variable. Consider that if we are given a function $y^* \in \mathcal{F}_o$ and another function η such that $\eta(x_1) = \eta(x_2) = 0$ then we can reach a whole family of functions indexed by a real variable α as follows (relabel $y^*(x)$ by $y(x, 0)$ so it matches the rest of the family of functions):

$$y(x, \alpha) = y(x, 0) + \alpha\eta(x)$$

Note that $x \mapsto y(x,\alpha)$ gives a function in $\mathcal{F}_o$. We define the **variation** of y to be

$$\boxed{\delta y = \alpha\eta(x)}$$

This means $y(x,\alpha) = y(x,0) + \delta y$. We may write J as a function of α given the variation we just described:

$$J(\alpha) = \int_{x_1}^{x_2} f(y(x,\alpha), y(x,\alpha)', x)\, dx.$$

It is intuitively obvious that if the function $y^*(x) = y(x,0)$ is an extremum of the functional then we ought to expect

$$\left[\frac{\partial J(\alpha)}{\partial \alpha}\right]_{\alpha=0} = 0$$

Notice that we can calculate the derivative above using multivariate calculus. Remember that $y(x,\alpha) = y(x,0) + \alpha\eta(x)$ hence $y(x,\alpha)' = y(x,0)' + \alpha\eta(x)'$ thus $\frac{\partial y}{\partial \alpha} = \eta$ and $\frac{\partial y'}{\partial \alpha} = \eta' = \frac{d\eta}{dx}$. Consider that:

$$\begin{aligned}
\frac{\partial J(\alpha)}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\left[\int_{x_1}^{x_2} f(y(x,\alpha), y(x,\alpha)', x)\, dx \right] \\
&= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\frac{\partial y}{\partial \alpha} + \frac{\partial f}{\partial y'}\frac{\partial y'}{\partial \alpha} + \frac{\partial f}{\partial x}\frac{\partial x}{\partial \alpha} \right) dx \\
&= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\frac{d\eta}{dx} \right) dx \qquad (6.1)
\end{aligned}$$

Observe that

$$\frac{d}{dx}\left[\frac{\partial f}{\partial y'}\eta\right] = \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right]\eta + \frac{\partial f}{\partial y'}\frac{d\eta}{dx}$$

Hence continuing Equation 6.1 in view of the product rule above,

$$\begin{aligned}
\frac{\partial J(\alpha)}{\partial \alpha} &= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\eta + \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\eta\right] - \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right]\eta \right) dx \\
&= \frac{\partial f}{\partial y'}\eta\Big|_{x_1}^{x_2} + \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\eta - \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right]\eta \right) dx \qquad (6.2) \\
&= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} - \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right] \right)\eta\, dx
\end{aligned}$$

Note we used the conditions $\eta(x_1) = \eta(x_2)$ to see that $\frac{\partial f}{\partial y'}\eta\big|_{x_1}^{x_2} = \frac{\partial f}{\partial y'}\eta(x_2) - \frac{\partial f}{\partial y'}\eta(x_1) = 0$. Our goal is to find the extreme values for the functional J. Let me take a few sentences to again restate our set-up. Generally, we take a function y then J maps to a new function $J[y]$. The family of functions indexed by α gives a whole ensemble of functions in $\mathcal{F}_o$ which are near y^* according to the formula,

$$y(x,\alpha) = y^*(x) + \alpha\eta(x)$$

Let's call this set of functions W_η. If we took another function like η, say ζ such that $\zeta(x_1) = \zeta(x_2) = 0$ then we could look at another family of functions:

$$y(x,\alpha) = y^*(x) + \alpha\zeta(x)$$

and we could denote the set of all such functions generated from ζ to be W_ζ. The total variation of y based at y^* should include all possible families of functions in $\mathcal{F}_o$. You could think of W_η and W_ζ be two different subspaces in $\mathcal{F}_o$. If $\eta \neq \zeta$ then these subspaces of $\mathcal{F}_o$ are likely disjoint except

for the proposed extremal solution y^*. It is perhaps a bit unsettling to realize there are infinitely many such subspaces because there are infinitely many choices for the function η or ζ. In any event, each possible variation of y^* must satisfy the condition $\left[\frac{\partial J(\alpha)}{\partial \alpha}\right]_{\alpha=0} = 0$ since we **assume** that y^* is an extreme value of the functional J. It follows that the Equation 6.2 holds for all possible η. Therefore, we ought to expect that any extreme value of the functional $J[y] = \int_{x_1}^{x_2} f(y, y', x)\, dx$ must solve the **Euler Lagrange Equations:**

$$\boxed{\frac{\partial f}{\partial y} - \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right] = 0 \quad \text{Euler-Lagrange Equations for} \quad J[y] = \int_{x_1}^{x_2} f(y, y', x)\, dx}$$

6.3 variational derivative

The role that η played in the discussion in the preceding section is somewhat similar to the role that the "h" plays in the definition $f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$. You might hope we could replace arguments in η with a more direct approach. Physicists have a heuristic way of making such arguments in terms of the variation δ. They would cast the arguments in the last page by just "taking the variation of J". Let me give you their formal argument,

$$\begin{aligned}
\delta J &= \delta\left[\int_{x_1}^{x_2} f(y, y', x)\, dx\right] \\
&= \left[\int_{x_1}^{x_2} \delta f(y, y', x)\, dx\right] \\
&= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial y'}\delta\left(\tfrac{dy}{dx}\right) + \frac{\partial f}{\partial x}\delta x\right) dx \\
&= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial y'}\frac{d}{dx}(\delta y)\right) dx \qquad (6.3)\\
&= \frac{\partial f}{\partial y'}\delta y\Big|_{x_1}^{x_2} + \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} - \frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right]\right)\delta y\, dx
\end{aligned}$$

Therefore, since $\delta y = 0$ at the endpoints of integration, the Euler-Lagrange equations follow from $\delta J = 0$. Now, if you're like me, the argument above is less than satisfying since we never actually defined what it means to "take δ" of something. Also, why could I commute the variational δ and $\frac{d}{dx}$)? That said, the formal method is not without use since it allows the focus to be on the Euler Lagrange equations rather than the technical details of the variation.

Remark 6.3.1.

The more adept reader at this point should realize the hypocrisy of me calling the above calculation formal since even my presentation here was formal. I also used an analogy, I assumed that the theory of extreme values for multivariate calculus extends to function space. But, $\mathcal{F}_o$ is not $\mathbb{R}^n$, it's much bigger. Edwards builds the correct formalism for a rigourous calculation of the variational derivative. To be careful we'd need to develop the norm on function space and prove a number of results about infinite dimensional linear algebra. Take a look at the last chapter in Edwards' text if you're interested. I don't believe I'll have time to go over that material this semester.

6.4 Euler-Lagrange examples

I present a few standard examples in this section. We make use of the calculation in the last section. Also, we will use a result from your homework which states an equivalent form of the Euler-Lagrange equation is

$$\frac{\partial f}{\partial x} - \frac{d}{dx}\left[f - y'\frac{\partial f}{\partial y'}\right] = 0.$$

This form of the Euler Lagrange equation yields better differential equations for certain examples.

6.4.1 shortest distance between two points in plane

If s denotes the arclength in the xy-plane then the pythagorean theorem gives $ds^2 = dx^2 + dy^2$ infinitesimally. Thus, $ds = \sqrt{1 + \frac{dy}{dx}^2}\, dx$ and we may add up all the little distances ds to find the total length between two given points (x_1, y_1) and (x_2, y_2):

$$J[y] = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\, dx$$

Identify that we have $f(y, y', x) = \sqrt{1 + (y')^2}$. Calculate then,

$$\frac{\partial f}{\partial y} = 0 \qquad \text{and} \qquad \frac{\partial f}{\partial y'} = \frac{y'}{\sqrt{1 + (y')^2}}.$$

Euler Lagrange equations yield,

$$\frac{d}{dx}\left[\frac{\partial f}{\partial y'}\right] = \frac{\partial f}{\partial y} \qquad \Rightarrow \qquad \frac{d}{dx}\left[\frac{y'}{\sqrt{1 + (y')^2}}\right] = 0 \qquad \Rightarrow \qquad \frac{y'}{\sqrt{1 + (y')^2}} = k$$

where $k \in \mathbb{R}$ is constant with respect to x. Moreover, square both sides to reveal

$$\frac{(y')^2}{1 + (y')^2} = k^2 \qquad \Rightarrow \qquad (y')^2 = \frac{k^2}{1 - k^2} \qquad \Rightarrow \qquad \frac{dy}{dx} = \pm\sqrt{\frac{k^2}{1 - k^2}} = m$$

where I have defined m is defined in the obvious way. We find solutions $y = mx + b$. Finally, we can find m, b to fit the given pair of points (x_1, y_1) and (x_2, y_2) as follows:

$$y_1 = mx_1 + b \qquad \text{and} \qquad y_2 = mx_2 + b \qquad \Rightarrow \qquad y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

provided $x_1 \neq x_2$. If $x_1 \neq x_2$ and $y_1 \neq y_2$ then we could perform the same calculation as above with the roles of x and y interchanged,

$$J[x] = \int_{y_1}^{y_2} \sqrt{1 + (x')^2}\, dy$$

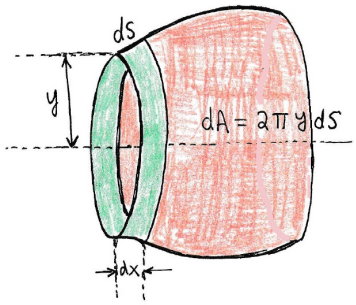
where $x' = dx/dy$ and the Euler Lagrange equations would yield the solution

$$x = x_1 + \frac{x_2 - x_1}{y_2 - y_1}(y - y_1).$$

Finally, if both coordinates are equal then $(x_1, y_1) = (x_2, y_2)$ and the shortest path between these points is the trivial path, the armchair solution. Silly comments aside, we have shown that a straight line provides the curve with the shortest arclength between any two points in the plane.

6.4.2 surface of revolution with minimal area

Suppose we wish to revolve some curve which connects (x_1, y_1) and (x_2, y_2) around the x-axis. A surface constructed in this manner is called a **surface of revolution**. In calculus we learn how to calculate the surface area of such a shape. One can imagine deconstructing the surface into a sequence of ribbons. Each ribbon at position x will have a "radius" of y and a width of dx however, because the shape is tilted the area of the ribbon works out to $dA = 2\pi y ds$ where ds is the arclength. I made a ribbon green in the picture below. You can imagine many ribbons approximating the surface, although, I made no attempt to draw those here:



If we choose x as the parameter this yields $dA = 2\pi y\sqrt{1+(y')^2}\,dx$. To find the surface of minimal surface area we ought to consider the functional:

$$A[y] = \int_{x_1}^{x_2} 2\pi y\sqrt{1+(y')^2}\,dx$$

Identify that $f(y, y', x) = 2\pi y\sqrt{1+(y')^2}$ hence $f_y = 2\pi\sqrt{1+(y')^2}$ and $f_{y'} = 2\pi y y'/\sqrt{1+(y')^2}$. The usual Euler-Lagrange equations are not easy to solve for this problem, it's easier to work with the equations you derived in homework,

$$\frac{\partial f}{\partial x} - \frac{d}{dx}\left[f - y'\frac{\partial f}{\partial y'}\right] = 0.$$

Hence,

$$\frac{d}{dx}\left[2\pi y\sqrt{1+(y')^2} - \frac{2\pi y(y')^2}{\sqrt{1+(y')^2}}\right] = 0$$

Dividing by 2π and making a common denominator,

$$\frac{d}{dx}\left[\frac{y}{\sqrt{1+(y')^2}}\right] = 0 \qquad \Rightarrow \qquad \frac{y}{\sqrt{1+(y')^2}} = k$$

where k is a constant with respect to x. Squaring the equation above yields

$$\frac{y^2}{1+(\frac{dy}{dx})^2} = k^2 \qquad \Rightarrow \qquad y^2 - k^2 = k^2(\tfrac{dy}{dx})^2$$

Solve for dx, integrate, assuming the given points are in the first quadrant,

$$x = \int dx = \int \frac{k\,dy}{\sqrt{y^2-k^2}} = k\cosh^{-1}(\tfrac{y}{k}) + c$$

Hence,

$$\boxed{y = k\cosh\left(\frac{x-c}{k}\right)}$$

generates the surface of revolution of least area between two points. These shapes are called **Catenoids** they can be observed in the formation of soap bubble between rings. There is a vast literature on this subject and there are many cases to consider, I simply exhibit a simple solution. For a given pair of points it is not immediately obvious if there exists a solution to the Euler-Lagrange equations which fits the data. (see page 622 of Arfken).

6.4.3 Braichistochrone

Suppose a particle slides freely along some curve from (x_1, y_1) to $(x_2, y_2) = (0, 0)$ under the influence of gravity where we take y to be the vertical direction. **What is the curve of quickest descent?** Notice that if $x_1 = 0$ then the answer is easy to see, however, if $x_1 \neq 0$ then the question is not trivial. To solve this problem we must first offer a functional which accounts for the time of descent. Note that the speed $v = ds/dt$ so we'd clearly like to minimize $J = \int_{(0,0)}^{(x_1,y_1)} \frac{ds}{v}$. Since the object is assumed to fall freely we may assume that energy is conserved in the motion hence

$$\frac{1}{2}mv^2 = mg(y - y_1) \qquad \Rightarrow \qquad v = \sqrt{2g(y_1 - y)}$$

As we've discussed in previous examples, $ds = \sqrt{1 + (y')^2}dt$ so we find

$$J[y] = \int_0^{x_1} \underbrace{\sqrt{\frac{1 + (y')^2}{2g(y_1 - y)}}}_{f(y,y',x)} \, dx$$

Notice that the modified Euler-Lagrange equations $\frac{\partial f}{\partial x} - \frac{d}{dx}\left[f - y'\frac{\partial f}{\partial y'}\right] = 0$ are convenient since $f_x = 0$. We calculate that

$$\frac{\partial f}{\partial y'} = \frac{1}{2\sqrt{\frac{1+(y')^2}{2g(y_1-y)}}} \frac{2y'}{2g(y_1 - y)} = \frac{y'}{\sqrt{2g(y_1 - y)(1 + (y')^2)}}$$

Hence there should exist some constant $1/(k\sqrt{2g})$ such that

$$\sqrt{\frac{1 + (y')^2}{2g(y_1 - y)}} - \frac{(y')^2}{\sqrt{2g(y_1 - y)(1 + (y')^2)}} = \frac{1}{k\sqrt{2g}}$$

It follows that,

$$\frac{1}{\sqrt{(y_1 - y)(1 + (y')^2)}} = \frac{1}{k} \qquad \Rightarrow \qquad \Big(y_1 - y\Big)\left(1 + \left(\frac{dy}{dx}\right)^2\right) = k^2$$

We need to solve for dy/dx,

$$(y_1 - y)\left(\frac{dy}{dx}\right)^2 = k^2 - y_1 + y \qquad \Rightarrow \qquad \left(\frac{dy}{dx}\right)^2 = \frac{y + k^2 - y_1}{y_1 - y}$$

Or, relabeling constants $a = y_1$ and $b = k^2 - y_1$ and we must solve

$$\frac{dy}{dx} = \pm\sqrt{\frac{b + y}{a - y}} \qquad \Rightarrow \qquad x = \pm \int \sqrt{\frac{a - y}{b + y}} \, dy$$

The integral is not trivial. It turns out that the solution is a cycloid (Arfken p. 624):

$$\boxed{x = \frac{a+b}{2}\Big(\theta + \sin(\theta)\Big) - d \qquad y = \frac{a+b}{2}\Big(1 - \cos(\theta)\Big) - b}$$

This is the curve that is traced out by a point on a wheel as it travels. If you take this solution and calculate $J[y_{cycloid}]$ you can show the time of descent is simply

$$T = \frac{\pi}{2}\sqrt{\frac{y_1}{2g}}$$

if the mass begins to descend from (x_2, y_2). But, this point has no connection with (x_1, y_1) except that they both reside on the same cycloid. It follows that the period of a pendulum that follows a cycloidal path is indpendent of the starting point on the path. This is not true for a circular pendulum in general, we need the small angle approximation to derive simple harmonic motion.

It turns out that it is possible to make a pendulum follow a cycloidal path if you let the string be guided by a frame which is also cycloidal. The neat thing is that even as it loses energy it still follows a cycloidal path and hence has the same period. The "Brachistochrone" problem was posed by Johann Bernoulli in 1696 and it actually predates the variational calculus of Lagrange by some 50 or so years. This problem and ones like it are what eventually prompted Lagrange and Euler to systematically develop the subject. Apparently Galileo also studied this problem however lacked the mathematics to crack it.

See this Geogebra demonstration to compare and contrast lines, verses parabolas, verses the cycloid. A google search will show you dozens of these.

6.5 Euler-Lagrange equations for several dependent variables

We still consider problems with just one independent parameter underlying everything. For problems of classical mechanics this is almost always time t. In anticipation of that application we choose to use the usual physics notation in the section. We suppose that our functional depends on functions $y_1, y_2, \ldots, y_n$ of time t along with their time derivatives $\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_n$. We again suppose the functional of interest is an integral of a **Lagrangian** function f from time t_1 to time t_2,

$$J[(y_i)] = \int_{t_1}^{t_2} f(y_i, \dot{y}_i, t)\, dt$$

here we use (y_i) as shorthand for $(y_1, y_2, \ldots, y_n)$ and $(\dot{y}_i)$ as shorthand for $(\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_n)$. We suppose that n-conditions are given for each of the endpoints in this problem; $y_i(t_1) = y_{i1}$ and $y_i(t_2) = y_{i2}$. Moreover, we define $\mathcal{F}_o$ to be the set of paths from $\mathbb{R}$ to $\mathbb{R}^n$ subject to the conditions just stated. We now set out to find necessary conditions on a proposed solution to the extreme value problem for the functional J above. As before let's assume that an extremal solution $y* \in \mathcal{F}_o$ exists. Moreover, imagine varying the solution by some variational function $\eta = (\eta_i)$ which has $\eta(t_1) = (0, 0, \ldots, 0)$ and $\eta(t_2) = (0, 0, \ldots, 0)$. Consequently the family of paths defined below are all in $\mathcal{F}_o$,

$$y(t, \alpha) = y^*(t) + \alpha\eta(t)$$

Thus $y(t, 0) = y^*$. In terms of component functions we have that

$$y_i(t, \alpha) = y_i^*(t) + \alpha\eta_i(t).$$

You can identify that $\delta y_i = y_i(t,\alpha) - y_i^*(t) = \alpha\eta_i(t)$. Since y^* is an extreme solution we should expect that $\left(\frac{\partial J}{\partial \alpha}\right)_{\alpha=0} = 0$. Differentiate the functional with respect to α and make use of the chain rule for f which is a function of some $2n+1$ variables,

$$\begin{aligned}
\frac{\partial J(\alpha)}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\left[\int_{t_1}^{t_2} f(y_i(t,\alpha), \dot{y}_i(t,\alpha), t)\, dt \right] \\
&= \int_{t_1}^{t_2} \sum_{j=1}^{n} \left(\frac{\partial f}{\partial y_j}\frac{\partial y_j}{\partial \alpha} + \frac{\partial f}{\partial \dot{y}_j}\frac{\partial \dot{y}_j}{\partial \alpha} \right) dt \\
&= \int_{t_1}^{t_2} \sum_{j=1}^{n} \left(\frac{\partial f}{\partial y_j}\eta_j + \frac{\partial f}{\partial \dot{y}_j}\frac{d\eta_j}{dt} \right) dt \qquad (6.4) \\
&= \sum_{j=1}^{n} \frac{\partial f}{\partial \dot{y}_j}\eta \Big|_{t_1}^{t_2} + \int_{t_1}^{t_2} \sum_{j=1}^{n} \left(\frac{\partial f}{\partial y_j} - \frac{d}{dt}\frac{\partial f}{\partial \dot{y}_j} \right)\eta_j\, dt
\end{aligned}$$

Since $\eta(t_1) = \eta(t_2) = 0$ the first term vanishes. Moreover, since we may repeat this calculation for all possible variations about the optimal solution y^* it follows that we obtain a set of Euler-Lagrange equations for each component function of the solution:

$$\boxed{\frac{\partial f}{\partial y_j} - \frac{d}{dt}\left[\frac{\partial f}{\partial \dot{y}_j}\right] = 0 \quad j = 1, 2, \ldots n \quad \text{Euler-Lagrange Eqns. for} \quad J[(y_i)] = \int_{t_1}^{t_2} f(y_i, \dot{y}_i, t)\, dt}$$

Often we simply use $y_1 = x$, $y_2 = y$ and $y_3 = z$ which denote the position of particle or perhaps just the component functions of a path which gives the geodesic on some surface. In either case we should have 3 sets of Euler-Lagrange equations, one for each coordinate. We will also use non-Cartesian coordinates to describe certain Lagrangians. We develop many useful results for set-up of Lagrangians in non-Cartesian coordinates in the next section.

6.5.1 free particle Lagrangian

For a particle of mass m the kinetic energy K is given in terms of the time derivatives of the coordinate functions x, y, z as follows:

$$K = \tfrac{m}{2}\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right)$$

Construct a functional by integrating the kinetic energy over time t,

$$S = \int_{t_1}^{t_2} \tfrac{m}{2}\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right)\, dt$$

The Euler-Lagrange equations for this functional are

$$\frac{\partial K}{\partial x} = \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{x}}\right] \qquad \frac{\partial K}{\partial y} = \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{y}}\right] \qquad \frac{\partial K}{\partial z} = \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{z}}\right]$$

Since $\frac{\partial K}{\partial \dot{x}} = m\dot{x}$, $\frac{\partial K}{\partial \dot{y}} = m\dot{y}$ and $\frac{\partial K}{\partial \dot{z}} = m\dot{z}$ it follows that

$$\boxed{0 = m\ddot{x} \qquad 0 = m\ddot{y} \qquad 0 = m\ddot{z}.}$$

You should recognize these as Newton's equation for a particle with no force applied. The solution is $(x(t), y(t), z(t)) = (x_o + tv_x, y_o + tv_y, z_o + tv_z)$ which is uniform rectilinear motion at constant velocity (v_x, v_y, v_z). The solution to Newton's equation minimizes the integral of the Kinetic energy. Generally the quantity S is called the **action** and Hamilton's Principle states that the laws of physics all arise from minimizing the action of the physical phenomena. We'll return to this discussion in a later section.

6.5.2 geodesics in $\mathbb{R}^3$

A **geodesic** is the path of minimal length between a pair of points on some manifold. Note we already proved that geodesics in the plane are just lines. In general, for $\mathbb{R}^3$, the square of the infinitesimal arclength element is $ds^2 = dx^2 + dy^2 + dz^2$. The arclength integral from $p = 0$ to $q = (q_x, q_y, q_z)$ in $\mathbb{R}^3$ is most naturally given from the parametric viewpoint:

$$S = \int_0^1 \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, dt$$

We assume $(x(0), y(0), z(0)) = (0, 0, 0)$ and $(x(1), y(1), z(1)) = q$ and it should be clear that the integral above calculates the arclength. The Euler-Lagrange equations for x, y, z are

$$\frac{d}{dt}\left[\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}\right] = 0, \qquad \frac{d}{dt}\left[\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}\right] = 0, \qquad \frac{d}{dt}\left[\frac{\dot{z}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}\right] = 0.$$

It follows that there exist constants, say a, b and c, such that

$$a = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}, \qquad b = \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}, \qquad c = \frac{\dot{z}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}.$$

These equations are said to be **coupled** since each involves derivatives of the others. We usually need a way to uncouple the equations if we are to be successful in solving the system. We can calculate, and equate each with the constant 1:

$$1 = \frac{\dot{x}}{a\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} = \frac{\dot{y}}{b\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} = \frac{\dot{z}}{c\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}.$$

But, multiplying by the denominator reveals an interesting identity

$$\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} = \frac{\dot{x}}{a} = \frac{\dot{y}}{b} = \frac{\dot{z}}{c}$$

The solution has the form, $x(t) = tq_x$, $y(t) = tq_y$ and $z(t) = tq_z$. Therefore,

$$(x(t), y(t), z(t)) = t(q_x, q_y, q_z) = tq.$$

for $0 \leq t \leq 1$. These are the parametric equations for the line segment from the origin to q.

6.6 the Euclidean metric

The square root in the functional of the last subsection certainly complicated the calculation. It is intuitively clear that if we add up squared line elements ds^2 to give a minimum then that ought to correspond to the minimum for the sum of the positive square roots ds of those elements. Let's check if my conjecture works for $\mathbb{R}^3$:

$$S = \int_0^1 (\underbrace{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}_{f(x,y,z,\dot{x},\dot{y},\dot{z})})\, dt$$

This gives us the Euler Lagrange equations below:

$$\ddot{x} = 0, \qquad \ddot{y} = 0, \qquad \ddot{z} = 0$$

The solution of these equations is clearly a line. In this formalism the equations were uncoupled from the outset.

Definition 6.6.1.

The Euclidean metric is $ds^2 = dx^2 + dy^2 + dz^2$. Generally, for orthogonal curvelinear coordinates u, v, w we calculate $ds^2 = \frac{1}{||\nabla u||^2}du^2 + \frac{1}{||\nabla v||^2}dv^2 + \frac{1}{||\nabla w||^2}dw^2$. We use this as a guide for constructing functionals which calculate arclength or speed

The beauty of the metric is that it allows us to calculate in other coordinates, consider

$$x = r\cos(\theta) \qquad y = r\sin(\theta)$$

For which we have implicit inverse coordinate transformations $r^2 = x^2 + y^2$ and $\theta = \tan^{-1}(y/x)$. From these inverse formulas we calculate:

$$\nabla r \ = < x/r, y/r > \qquad \nabla\theta \ = < -y/r^2, x/r^2 >$$

Thus, $||\nabla r|| = 1$ whereas $||\nabla\theta|| = 1/r$. We find that the metric in polar coordinates takes the form:

$$\boxed{ds^2 = dr^2 + r^2 d\theta^2}$$

Physicists and engineers tend to like to think of these as arising from calculating the length of infinitesimal displacements in the r or θ directions. Generically, for u, v, w coordinates

$$dl_u = \frac{1}{||\nabla u||}du \qquad dl_v = \frac{1}{||\nabla v||}dv \qquad dl_w = \frac{1}{||\nabla w||}dw$$

and $ds^2 = dl_u^2 + dl_v^2 + dl_w^2$. So in that notation we just found $dl_r = dr$ and $dl_\theta = rd\theta$. Notice then that cylindircal coordinates have the metric,

$$\boxed{ds^2 = dr^2 + r^2 d\theta^2 + dz^2.}$$

For spherical coordinates $x = r\cos(\phi)\sin(\theta)$, $y = r\sin(\phi)\sin(\theta)$ and $z = r\cos(\theta)$ (here $0 \leq \phi \leq 2\pi$ and $0 \leq \theta \leq \pi$, physics notation). Calculation of the metric follows from the line elements,

$$dl_r = dr \qquad dl_\phi = r\sin(\theta)d\phi \qquad dl_\theta = rd\theta$$

Thus,

$$\boxed{ds^2 = dr^2 + r^2\sin^2(\theta)d\phi^2 + r^2 d\theta^2.}$$

We now have all the tools we need for examples in spherical or cylindrical coordinates. What about other cases? In general, given some p-manifold in $\mathbb{R}^n$ how does one find the metric on that manifold? If we are to follow the approach of this section we'll need to find coordinates on $\mathbb{R}^n$ such that the manifold S is described by setting all but p of the coordinates to a constant. For example, in $\mathbb{R}^4$ we have generalized cylindircal coordinates (r, ϕ, z, t) defined implicitly by the equations below

$$x = r\cos(\phi), \qquad y = r\sin(\phi), \qquad z = z, \qquad t = t$$

On the hyper-cylinder $r = R$ we have the metric $ds^2 = R^2 d\theta^2 + dz^2 + dw^2$. There are mathematicians/physicists whose careers are founded upon the discovery of a metric for some manifold. This is generally a difficult task.

6.7 geodesics

A **geodesic** is a path of smallest distance on some manifold. In general relativity, it turns out that the solutions to Eistein's field equations are geodesics in 4-dimensional curved spacetime. Particles that fall freely are following geodesics, for example projectiles or planets in the absense of other frictional/non-gravitational forces. We don't follow a geodesic in our daily life because the earth pushes back up with a normal force. Also, do be honest, the idea of length in general relativity is a bit more abstract that the geometric length studied in this section. The metric of general relativity is non-Euclidean. General relativity is based on semi-Riemannian geometry whereas this section is all Riemannian geometry. The metric in Riemannian geometry is positive definite. The metric in semi-Riemannian geometry can be written as a quadratic form with both positive and negative eigenvalues. In any event, if you want to know more I know some books you might like.

6.7.1 geodesic on cylinder

The equation of a cylinder of radius R is most easily framed in cylindrical coordinates (r, θ, z); the equation is merely $r = R$ hence the metric reads

$$ds^2 = R^2 d\theta^2 + dz^2$$

Therefore, we ought to minimize the following functional in order to locate the parametric equations of a geodesic on the cylinder: note $ds^2 = \left(R^2 \frac{d\theta^2}{dt^2} + \frac{dz^2}{dt^2}\right) dt^2$ thus:

$$S = \int (R^2 \dot{\theta}^2 + \dot{z}^2)\ dt$$

Euler-Lagrange equations for the dependent variables θ and z are simply:

$$\ddot{\theta} = 0 \qquad \ddot{z} = 0.$$

We can integrate twice to find solutions

$$\boxed{\theta(t) = \theta_o + At \qquad z(t) = z_o + Bt}$$

Therefore, the geodesic on a cylinder is simply the line connecting two points in the plane which is curved to assemble the cylinder. Simple cases that are easy to understand:

1. Geodesic from $(R\cos(\theta_o), R\sin(\theta_o), z_1)$ to $(R\cos(\theta_o), R\sin(\theta_o), z_2)$ is parametrized by $\theta(t) = \theta_o$ and $z(t) = z_1 + t(z_2 - z_1)$ for $0 \leq t \leq 1$. Technically, there is some ambiguity here since I never declared over what range the t is to range. Could pick other intervals, we could use z at the parameter is we wished then $\theta(z) = \theta_o$ and $z = z$ for $z_1 \leq z \leq z_2$

2. Geodesic from $(R\cos(\theta_1), R\sin(\theta_1), z_o)$ to $(R\cos(\theta_2), R\sin(\theta_2), z_o)$ is parametrized by $\theta(t) = \theta_1 + t(\theta_2 - \theta_1)$ and $z(t) = z_o$ for $0 \leq t \leq 1$.

3. Geodesic from $(R\cos(\theta_1), R\sin(\theta_1), z_1)$ to $(R\cos(\theta_2), R\sin(\theta_2), z_2)$ is parametrized by

   $$\theta(t) = \theta_1 + t(\theta_2 - \theta_1) \qquad z(t) = z_1 + t(z_2 - z_1)$$

   You can eliminate t and find the equation $z = \frac{z_2 - z_1}{\theta_2 - \theta_1}(\theta - \theta_1)$ which again just goes to show you this is a line in the curved coordinates.

6.7.2 geodesic on sphere

The equation of a sphere of radius R is most easily framed in spherical coordinates (r, ϕ, θ); the equation is merely $r = R$ hence the metric reads

$$ds^2 = R^2 \sin^2(\theta) d\phi^2 + R^2 d\theta^2.$$

Therefore, we ought to minimize the following functional in order to locate the parametric equations of a geodesic on the sphere: note $ds^2 = \left(R^2 \sin^2(\theta) \frac{d\phi^2}{dt^2} + R^2 \frac{d\theta^2}{dt^2}\right) dt^2$ thus:

$$S = \int (\underbrace{R^2 \sin^2(\theta)\dot{\phi}^2 + R^2\dot{\theta}^2}_{f(\theta,\phi,\dot{\theta},\dot{\phi})}) \, dt$$

Euler-Lagrange equations for the dependent variables ϕ and θ are simply: $f_\theta = \frac{d}{dt}(f_{\dot{\theta}})$ and $f_\phi = \frac{d}{dt}(f_{\dot{\phi}})$ which yield:

$$2R^2 \sin(\theta)\cos(\theta)\dot{\phi}^2 = \tfrac{d}{dt}(2R^2\dot{\theta}) \qquad 0 = \frac{d}{dt}\Big(2R^2 \sin^2(\theta)\dot{\phi}\Big).$$

We find a **constant of motion** $L = 2R^2 \sin^2(\theta)\dot{\phi}$ inserting this in the equation for the azmuthial angle θ yields:

$$2R^2 \sin(\theta)\cos(\theta)\dot{\phi}^2 = \tfrac{d}{dt}(2R^2\dot{\theta}) \qquad 0 = \frac{d}{dt}\Big(2R^2 \sin^2(\theta)\dot{\phi}\Big).$$

If you can solve these and demonstrate through some reasonable argument that the solutions are great circles then I will give you points. I have some solutions but nothing looks too pretty.

6.8 Lagrangian mechanics

6.8.1 basic equations of classical mechanics summarized

Classical mechanics is the study of massive particles at relatively low velocities. Let me refresh your memory about the basics equations of Newtonian mechanics. Our goal in this section will be to rephrase Newtonian mechanics in the variational langauge and then to solve problems with the Euler-Lagrange equations. Newton's equations tell us how a particle of mass m evolves through time according to the net-force impressed on m. In particular,

$$m\frac{d^2\vec{r}}{dt^2} = \vec{F}$$

If m is not constant then you may recall that it is better to use momentum $\vec{P} = m\vec{v} = m\frac{d\vec{r}}{dt}$ to set-up Newton's 2nd Law:

$$\frac{d\vec{P}}{dt} = \vec{F}$$

In terms of components we have a system of differential equations with indpendent variable time t. If we use position as the dependent variable then Newton's 2nd Law gives three second order ODEs,

$$m\ddot{x} = F_x \qquad m\ddot{y} = F_y \qquad m\ddot{z} = F_z$$

where $\vec{r} = (x, y, z)$ and the dots denote time-derivatives. Moreover, $\vec{F} =< F_x, F_y, F_z >$ is the sum of the forces that act on m. In contrast, if you work with momentum then you would want to solve six first order ODEs,

$$\dot{P}_x = F_x \qquad \dot{P}_y = F_y \qquad \dot{P}_z = F_z$$

and $P_x = m\dot{x}$, $P_y = m\dot{y}$ and $P_z = m\dot{z}$. These equations are easiest to solve when the force is not a function of velocity or time. In particular, if the force $\vec{F}$ is conservative then there exists a potential energy function $U : \mathbb{R}^3 \to \mathbb{R}$ such that $\vec{F} = -\nabla U$. We can prove that in the case the force is conservative the total energy is conserved.

6.8.2 kinetic and potential energy, formulating the Lagrangian

Recall the kinetic energy is $T = \frac{1}{2}m||\vec{v}||^2$, in Cartesian coordinates this gives us the formula:

$$T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2).$$

If $\vec{F}$ is a conservative force then it is independent of path so we may construct the potential energy function as follows:

$$U(\vec{r}) = -\int_{\mathcal{O}}^{\vec{r}} \vec{F} \cdot d\vec{r}$$

Here $\mathcal{O}$ is the origin for the potential and we can prove that the potential energy constructed in this manner has $\vec{F} = -\nabla U$. We can prove that the total (mechanical) energy $E = T + U$ for a conservative system is a constant; $dE/dt = 0$. Hopefully these comments are at least vaguely familiar from some physics course in your distant memory. If not relax, calculationally this chapter is self-contained, read onward.

We already calculated that if we use T as the Lagrangian then the Euler-Lagrange equations produce Newton's equations in the case that the force is zero (see 6.5.1). Suppose that we define the Lagrangian to be $L = T - U$ for a system governed by a conservative force with potential energy function U. We seek to prove the Euler-Lagrange equations are precisely Newton's equations for this conservative system[1] Generically we have a Lagrangian of the form

$$L(x, y, z, \dot{x}, \dot{y}, \dot{z}) = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - U(x, y, z).$$

We wish to find extrema for the functional $S = \int L(t)\, dt$. This yields three sets of Euler-Lagrange equations, one for each dependent variable x, y or z

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{x}}\right] = \frac{\partial L}{\partial x} \qquad \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{y}}\right] = \frac{\partial L}{\partial y} \qquad \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{z}}\right] = \frac{\partial L}{\partial z}.$$

Note that $\frac{\partial L}{\partial \dot{x}} = m\dot{x}$, $\frac{\partial L}{\partial \dot{y}} = m\dot{y}$ and $\frac{\partial L}{\partial \dot{z}} = m\dot{z}$. Also note that $\frac{\partial L}{\partial x} = -\frac{\partial U}{\partial x} = F_x$, $\frac{\partial L}{\partial y} = -\frac{\partial U}{\partial y} = F_y$ and $\frac{\partial L}{\partial z} = -\frac{\partial U}{\partial z} = F_z$. It follows that

$$\boxed{m\ddot{x} = F_x \qquad m\ddot{y} = F_y \qquad m\ddot{z} = F_z.}$$

Of course this is precisely $m\vec{a} = \vec{F}$ for a net-force $\vec{F} =< F_x, F_y, F_z >$. We have shown that **Hamilton's principle** reproduces Newton's Second Law for conservative forces. Let me take a moment to state it.

[1] don't mistake this example as an admission that Lagrangian mechanics is limited to conservative systems. Quite the contrary, Lagrangian mechanics is actually more general than the orginal framework of Newton!

Definition 6.8.1. Hamilton's Principle:

If a physical system has generalized coordinates q_j with velocities $\dot{q}_j$ and Lagrangian $L = T - U$ then the solutions of physics will minimize the action S defined below:

$$\boxed{S = \int_{t_1}^{t_2} L(q_j, \dot{q}_j, t)\, dt}$$

Mathematically, this means the variation $\delta S = 0$ for physical trajectories.

This is a necessary condition for solutions of the equations of physics. Sufficient conditions are known, you can look in any good variational calculus text. You'll find analogues to the second derivative test for variational differentiation. As far as I can tell physicists don't care about this logical gap, probably because the solutions to the Euler-Lagrange equations are the ones for which they are looking.

6.8.3 easy physics examples

Now, you might just see this whole exercise as some needless multiplication of notation and formalism. After all, I just told you we just get Newton's equations back from the Euler-Lagrange equations. To my taste the impressive thing about Lagrangian mechanics is that you get to start the problem with energy. Moreover, the Lagrangian formalism handles non-Cartesian coordinates with ease. If you search your memory from classical mechanics you'll notice that you either do constant acceleration, circular motion or motion along a line. What if you had a particle constrained to move in some frictionless ellipsoidal bowl. Or what if you had a pendulum hanging off another pendulum? How would you even write Newtons' equations for such systems? In contrast, the problem is at least easy to set-up in the Lagrangian approach. Of course, solutions may be less easy to obtain.

Example 6.8.2. Projectile motion: *take z as the vertical direction and suppose a bullet is fired with initial velocity $v_o =< v_{ox}, v_{oy}, v_{oz} >$. The potential energy due to gravity is simply $U = mgz$ and kinetic energy is given by $T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$. Thus,*

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - mgz$$

Euler-Lagrange equations are simply:

$$\frac{d}{dt}\left[m\dot{x}\right] = 0 \qquad \frac{d}{dt}\left[m\dot{y}\right] = 0 \qquad \frac{d}{dt}\left[m\dot{z}\right] = \frac{\partial}{\partial z}(-mgz) = -mg.$$

Integrating twice and applying initial conditions gives us the (possibly familiar) equations

$$x(t) = x_o + v_{ox}t, \qquad y(t) = y_o + v_{oy}t, \qquad z(t) = z_o + v_{oz}t - \tfrac{1}{2}gt^2.$$

Example 6.8.3. Simple Pendulum: *let θ denote angle measured off the vertical for a simple pendulum of mass m and length l. Trigonmetry tells us that*

$$x = l\sin(\theta) \qquad y = l\cos(\theta) \qquad \Rightarrow \qquad \dot{x} = l\cos(\theta)\dot{\theta} \qquad y = -l\sin(\theta)\dot{\theta}$$

Thus $T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) = \frac{1}{2}ml^2\dot{\theta}^2$. Also, the potential energy due to gravity is $U = -mgl\cos(\theta)$ which gives us

$$L = \frac{1}{2}ml^2\dot{\theta}^2 + mgl\cos(\theta)$$

Then, the Euler-Lagrange equation in θ is simply:

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\theta}}\right] = \frac{\partial L}{\partial \theta} \qquad \Rightarrow \qquad \frac{d}{dt}(ml^2\dot{\theta}) = -mgl\sin(\theta) \qquad \Rightarrow \qquad \ddot{\theta} + \frac{g}{l}\sin(\theta) = 0.$$

In the small angle approximation, $\sin(\theta) = \theta$ then we have the solution $\theta(t) = \theta_o\cos(\omega t + \phi_o)$ for angular frequency $\omega = \sqrt{g/l}$