Notes of lectures on Multivariable Calculus

G.L. Luke

October 14, 2007

1 Introduction

Let U be an open subset of $\mathbb{R}$, $a \in U$ and $f : U \to \mathbb{R}$. When the limit in the definition exists, we define the *derivative of* f *at* a by

$$\frac{\mathrm{d}f}{\mathrm{d}x}(a) = \lim_{\Delta \to 0} \frac{f(a+\Delta) - f(a)}{\Delta}.$$

This can be generalized in various ways.

- Again let U be an open subset of $\mathbb{R}$, $a \in U$ and now $f : U \to \mathbb{R}^n$. For each $i = 1, 2, \cdots, n$, the projections (co-ordinates) $p_i : \mathbb{R}^n \to \mathbb{R}$,

  $$p_i \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = a_i,$$

  are linear and continuous. If $f_i = p_i \circ f$ then

  $$f(u) = \begin{pmatrix} f_1(u) \\ \vdots \\ f_n(u) \end{pmatrix}.$$

  We define the derivative of f at a by a similar formula

  $$\frac{\mathrm{d}f}{\mathrm{d}x}(a) = \lim_{\Delta \to 0} \frac{1}{\Delta}(f(a+\Delta) - f(a))$$

  which yields

  $$\frac{\mathrm{d}f}{\mathrm{d}x}(a) = \begin{pmatrix} \frac{\mathrm{d}f_1}{\mathrm{d}x}(a) \\ \vdots \\ \frac{\mathrm{d}f_n}{\mathrm{d}x}(a) \end{pmatrix}$$

  in terms of the components of f, because the p_i are both continuous and linear.
  [1]

  In this case the derivative is vector-valued. The map gives a *path* and the derivative at a is the *tangent vector* to the path at the point $f(a)$.

[1] To see this: $p_i\left(\lim_{\Delta \to 0} \frac{1}{\Delta}(f(a+\Delta) - f(a))\right) = \lim_{\Delta \to 0} p_i\left(\frac{1}{\Delta}(f(a+\Delta) - f(a))\right) = \lim_{\Delta \to 0} \frac{1}{\Delta}\left(p_i \circ f(a+\Delta) - p_i \circ f(a)\right)$ the first equality follows because p_i is continuous and the second because p_i is linear.

- Suppose now that U is an open subset of $\mathbb{R}^k$ and $f : U \to \mathbb{R}$. Here we can define *directional derivatives.* Let $a \in U$ and $v \in \mathbb{R}^k$. The *directional derivative* of f at a in direction v is given by

$$\lim_{t \to 0} \frac{f(a+tv) - f(a)}{t}.$$

This is the derivative of the function $t \mapsto f(a+tv)$ at $t = 0$.

- Special cases are the *partial derivatives.* Let

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ o \end{pmatrix} = (0 \cdots 1 \cdots 0)^T,$$

where 1 is in the i^{th} position. We write

$$\frac{\partial f}{\partial x_i}(a) = \lim_{t \to 0} \frac{f(a+te_i) - f(a)}{t},$$

the directional derivative in the coordinate direction e_i. You will probably already have seen this expressed in the coordinate notation:

$$\frac{\partial f}{\partial x_i}(a) = \lim_{t \to 0} \frac{1}{t}(f \begin{pmatrix} a_1 \\ \vdots \\ a_i + t \\ \vdots \\ a_k \end{pmatrix} - f \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_k \end{pmatrix}).$$

There is a more general notion relating these various derivatives with a more economical notation in which to calculate. In order to understand this new idea we will rework some of the definitions just given above.

Let $U \subset \mathbb{R}$ be an open set, $a \in U$ and $f : U \to \mathbb{R}$. We let $l = \frac{\mathrm{d}f}{\mathrm{d}x}(a)$ and then

$$\lim_{\Delta \to 0} \frac{f(a+\Delta) - f(a) - l.\Delta}{\Delta} = 0,$$

that is,

$$f(a+\Delta) - f(a) - l\Delta = o(\Delta),$$

or

$$f(a+\Delta) = f(a) + l\Delta + o(\Delta).$$

[2] This short Taylor expansion suggests a way of making sense of a derivative in the more general situation where U is an open subset of $\mathbb{R}^k$, $a \in U$ and $f : U \to \mathbb{R}^n$, multiple dimensions in both the domain and

[2]Recall that we say $g(\Delta) = o(\|\Delta\|)$ if $\lim_{\Delta \to 0} \frac{\|g(\Delta)\|}{\|\Delta\|} = 0$.
Another way of putting this is to write $h(\Delta) = \frac{g(\Delta)}{\|\Delta\|}$ and then $g(\Delta) = \|\Delta\| h(\Delta)$ with $\lim_{\Delta \to 0} h(\Delta) = 0$.

range of the function f. The idea is that the *total derivative* is the best linear approximation to the map $\Delta \mapsto f(a+\Delta) - f(a)$. It is a linear map $L : \mathbb{R}^k \to \mathbb{R}^n$ satisfying

$$\lim_{\Delta \to 0} \frac{f(a+\Delta) - f(a) - L(\Delta)}{\|\Delta\|} = 0,$$

or writing this as a short Taylor expansion,

$$f(a+\Delta) = f(a) + L(\Delta) + o(\|\Delta\|).$$

To understand this properly we need to make sense of the symbol $\|\Delta\|$ which has been used to mean the 'size' of Δ.

Definition 1 *Let V be a real vector space. A function $\| \ \| : V \to [0, \infty)$ is a* norm *if:*

(i) $\|v\| = 0$ *if and only if* $v = 0$*;*

(ii) $\|cv\| = |c|\|v\|$ *for any* $c \in \mathbb{R}$*, and* $v \in V$*;*

(iii) $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$*, for any* $v_1,\ v_2 \in V$ *(sub-additivity property).*

Examples:

1. On $\mathbb{R}$ or $\mathbb{C}$, the *absolute value* is a norm.

2. On $\mathbb{R}^n$, we define the Euclidean norm, by

$$\|v\|_2 = \| \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \|_2 = \sqrt{v_1^2 + \cdots + v_n^2}.$$

The proof of the sub-additivity property follows from the Cauchy-Schwarz inequality: $|v.w| \leq \|v\|\|w\|$.

$$\begin{array}{rcl} \|v_1 + v_2\|_2^2 & = & (v_1 + v_2).(v_1 + v_2) \\ & = & v_1.v_1 + v_1.v_2 + v_2.v_1 + v_2.v_2 \\ & \leq & \|v_1\|_2^2 + \|v_1\|_2\|v_2\|_2 + \|v_2\|_2\|v_1\|_2 + \|v_2\|_2^2 \\ & = & (\|v_1\|_2 + \|v_2\|_2)^2, \end{array},$$

that is, $\|v_1 + v_2\|_2 \leq \|v_1\|_2 + \|v_2\|_2$.

3. If V is an abstract n-dimensional vector space with basis $e_1, e_2, \ldots, e_n$, we can define:

$$\|v\| = \|\sum_{i=1}^{n} v_i e_i\| = \sqrt{v_1^2 + \cdots + v_n^2}.$$

This norm depends on the basis.

4. On $\mathbb{R}^n$ there are a number of natural norms, for example, the 'sup' norm, usually denoted $\| \, \|_\infty$, is defined by

$$\| \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \|_\infty = \sup_{1 \leq i \leq n} |v_i|.$$

5. On $M_{n\times k}(\mathbb{R})$, the vector space of real $n \times k$ matrices, there are a number of norms available but for these lectures we will use the Euclidean norm: for $A = (a_{ij})$ we take

$$\|A\| = \left(\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}^2\right)^{\frac{1}{2}}.$$

For this norm and $A \in M_{n\times k}(\mathbb{R})$, $B \in M_{k\times l}(\mathbb{R})$, we have

$$\|AB\| \le \|A\|\|B\|$$

(see Problem Sheet 1).

Definition 2 *Let V be a real vector space, then norms $\| \ \|$ and $| \ |$ are* equivalent *if there exist constants $K \ge C > 0$ such that*

$$C\| \ \| \le | \ | \le K\| \ \|,$$

equivalently,

$$\frac{1}{K}| \ | \le \| \ \| \le \frac{1}{C}| \ |.$$

It is easy to check that this condition gives an equivalence relation on the set of norms on V.

Theorem 1 *All norms on a finite dimensional vector space are equivalent.*

Proof:(Not examinable) Suppose that $\| \ \|_a$ and $\| \ \|_b$ are norms on a finite dimensional space V. Let $e_1, e_2, \ldots, e_n$ be a basis for V and define norm $\| \ \|_2$ by

$$\|v\|_2 = \|\sum_{n=1}^{n} v_i e_i\|_2 = \sqrt{v_1^2 + \cdots v_n^2}$$

Let $S = \{v \in V : \|v\|_2 = 1\}$. Then S is a closed, bounded subset of V where bounded, closed and (later) continuous are all defined in terms of the norm $\| \ \|_2$. Then

$$\begin{aligned}
\|v\|_a &= \|v_1e_1 + \cdots + v_ne_n\|_a \\
&\le |v_1|\|e_1\|_a + \cdots + |v_n|\|e_n\|_a \\
&\le (|v_1| + \cdots + |v_n|)\sup_{1\le i\le n}\|e_i\|_a \\
&\le \sqrt{n}\sqrt{v_1^2 + \cdots + v_n^2}\sup_{1\le i\le n}\|e_i\|_a
\end{aligned}$$

by Cauchy-Schwarz inequality. Put $K = \sqrt{n}\sup_{1\le i\le n}\|e_i\|_a$ and then $\|v\|_a \le K\|v\|_2$.

Note $|\|v\|_a - \|w\|_a| \le \|v - w\|_a \le K\|v - w\|_2$ so that the function $\| \ \|_a$ is continuous on V. Restrict $\| \ \|_a$ to S. Then $\| \ \|_a$ achieves its infimum at some $s_0 \in S$, that is, $\|s_0\|_a \le \|s\|_a$ if $\|s\|_2 = 1$. If $v \in V$, $v \ne 0$, then $\dfrac{v}{\|v\|_2} \in S$ and $\|s_0\|_a \le \|\dfrac{v}{\|v\|_2}\|_a$. Hence $\|s_0\|_a\|v\|_2 \le \|v\|_a$. Putting $C = \|s_0\|_a$ gives

$$C\|v\|_2 \le \|v\|_a \le K\|v\|_2.$$

Thus $\| \ \|_a$ is equivalent to $\| \ \|_2$. Similarly, $\| \ \|_b$ is equivalent to $\| \ \|_2$ and hence $\| \ \|_a$ and $\| \ \|_b$ are equivalent norms.

Corollary: **In finite dimensional real vector spaces, open sets, bounded sets and continuity are independent of the norm used to define them.**

2 The Total Derivative

Definition 3 *Let V and W be finite dimensional real vector spaces. Let U be an open subset of V, $a \in U$ and $f : U \to W$. The* total derivative *of f at a (when it exists) is a linear map $L : V \to W$ such that*

$$\lim_{\Delta \to 0} \frac{f(a + \Delta) - f(a) - L(\Delta)}{\|\Delta\|} = 0,$$

that is,

$$f(a + \Delta) - f(a) - L(\Delta) = o(\|\Delta\|).$$

[Remember that on a finite dimensional space the property of a function being $o(\|\Delta\|)$ does not depend on the norm chosen.]

The total derivative of f at a, when it exists, will be denoted by Df_a. Thus

$$f(a + \Delta) = f(a) + Df_a(\Delta) + o(\|\Delta\|).$$

[In some texts this is called the *differential* of f at a and written as df_a or $df(a)$.]

What is the relationship between the total derivative and the directional derivatives?

If $v \in V$, $v \neq 0$, then putting $\Delta = t.v$ we have from the definitions

(i) the directional derivative of f at a in the direction v is given by

$$\lim_{t \to 0} \frac{f(a + tv) - f(a)}{t};$$

(ii) the total derivative satisfies

$$\lim_{tv \to 0} \frac{f(a + tv) - f(a) - Df_a(tv)}{\|tv\|} = 0.$$

We rewrite this expression for the total derivative as

$$\lim_{t \to 0} \frac{f(a + tv) - f(a) - tDf_a(v)}{t} \frac{t}{\|tv\|} = 0,$$

and note $\left| \frac{t}{\|tv\|} \right| = \left| \frac{t}{|t|\|v\|} \right| = \frac{1}{\|v\|}$. Hence

$$\lim_{t \to 0} \frac{f(a + tv) - f(a)}{t} - Df_a(v) = 0,$$

because $\|v\| \neq 0$ and so $Df_a(v)$ is equal to the directional derivative at a in the direction v.

Beware: There are examples where the total derivative fails to exist at a point even though all directional derivatives exist.

Next we find a matrix representation for Df_a. Let $e_1, \ldots, e_k$ be a basis for V and $\tilde{e}_1, \ldots, \tilde{e}_n$ be a basis for W. Then, by our calculation above, $Df_a(e_j) = \frac{\partial f}{x_j}(a)$. Further, the function f has n components:

$$f(v) = \sum_{i=1}^{n} f_i(v)\tilde{e}_i,$$

where each $f_i : U \to \mathbb{R}$,

$$f \sim \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

Hence

$$Df_a(e_j) = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_j}(a)\tilde{e}_i.$$

Definition 4 *In terms of the given bases, the matrix corresponding to the total derivative Df_a is*

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_k(a)} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \cdots & \frac{\partial f_n}{\partial x_k}(a) \end{pmatrix},$$

more briefly,

$$\left(\frac{\partial f_i}{\partial x_j}\right),$$

known as the Jacobian matrix.

The Jacobian matrix corresponding to a real valued function $f : U \to \mathbb{R}$ is a row vector. It is known as the *gradient of f at a* and denoted by $\operatorname{grad} f(a)$:

$$\operatorname{grad} f(a) = (\frac{\partial f}{\partial x_1}(a), \cdots, \frac{\partial f}{x_k}(a)),$$

and

$$Df_a(v) = (\frac{\partial f}{\partial x_1}(a), \cdots, \frac{\partial f}{x_k}(a)) \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix} = \sum_{i=1}^{k} v_i \frac{\partial f}{\partial x_i}(a).$$

Examples

(i) Let $f : \mathbb{R}^k \to \mathbb{R}^n$ be a *linear* map. Then

$$f(a + \Delta) = f(a) + f(\Delta) + 0.$$

Hence $Df_a = f$ as we would expect. (We will often use the fact that the total derivative of an identity map on a vector space is the same identity map.)

(ii) Let $s : \mathbb{R}^k \to \mathbb{R}$ be the length squared of a vector, defined using the dot product: $s(v) = v.v = v^T v$ for $v \in \mathbb{R}^k$. Then $s(a + \Delta) = (a + \Delta).(a + \Delta) = (a + \Delta)^T(a + \Delta) = a^T a + \Delta^T a + a^T\Delta + \Delta^T\Delta$. But note $\Delta^T\Delta = \|\Delta\|^2 = o\|\Delta\|$. Hence

$$Ds_a(\Delta) = \Delta^T a + a^T\Delta = 2\Delta^T a = 2\Delta.a.$$

Alternatively, $s(\vec{x}) = \sum_{i=1}^{k} x_i^2$. So $\operatorname{grad} s(\vec{x}) = (2x_1, \ldots, 2x_k)$. Then $Ds_a(v) = \operatorname{grad} s(a).v = \sum_{i=1}^{k} 2a_i v_i = 2a.x$. Occasionally, it is difficult to identify the linear term even though you know it exists. There is a trick: simply calculate the directional derivatives by substituting tv for Δ and read off the coefficient of the first power of t. Thus $s(a + tv) = (a + tv)^T(a + tv) = a^T a + t(a^T v + v^T a) + t^2 v^T v$. Thus $Ds_a(v) = a^T v + v^T a = 2a^t v$.

(iii) Let $p : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be given by $p(v, w) = v.w = v^T w$, so the function p is the dot product. Here $\Delta \in \mathbb{R}^n \times \mathbb{R}^n$ has two components: $\Delta = (\Delta_1, \Delta_2)$ and $(v, w) + (\Delta_1, \Delta_2) = (v + \Delta_1, w + \Delta_2)$. Also $\|\Delta\|^2 = \|(\Delta_1, \Delta_2)\|^2 = \|\Delta_1\|^2 + \|\Delta_2\|^2$.

Then $p(v + \Delta_1, w + \Delta_2) = (v + \Delta_1).(w + \Delta_2) = v.w + \Delta_1.w + v.\Delta_2 + \Delta_1.\Delta_2$.

Now $\Delta_1.\Delta_2 = o(\|\Delta\|)$ because $|\Delta_1.\Delta_2| \leq \|\Delta_1\|\|\Delta_2\| \leq \|\Delta\|^2$. Thus

$$Dp_{(v,w)}(\Delta_1, \Delta_2) = \Delta_1.w + v.\Delta_2.$$

(iv) Let $f : M_{n\times n}(\mathbb{R}) \to M_{n\times n}(\mathbb{R})$ given by $f(X) = X^2$.
Then $(A + \Delta)^2 = A^2 + A\Delta + \Delta A + \Delta^2$ and $\|\Delta^2\| \leq \|\Delta\|^2 = o(\|\Delta\|)$. Hence $Df_A(\Delta) = A\Delta + \Delta A$.

NOTE: The Jacobian matrix for this example is a $n^2 \times n^2$ matrix.

We try the directional derivative trick on this example:

$$f(A + t\Delta) = f(A) + Df_A(t\Delta) + o(\|t\Delta\|).$$

If we fix Δ then a function of order $o(\|t\Delta\|)$ is of order $o(|t|)$ so

$$f(A + t\Delta) = f(a) + tDf_A(\Delta) + o(|t|).$$

Using the linearity of Df_A we identify the derivative as the coefficient of t. So $f(A+t\Delta) = (A+t\Delta)^2 = A^2 + t(A\Delta + \Delta A) + t^2\Delta^2$ and we deduce that $Df_A(\Delta) = A\Delta + \Delta A$.

(v) Let $f : U \to \mathbb{C}$ be a holomorphic function defined on a domain U in the complex plane. Then the definition for the complex derivative gives

$$\lim_{\Delta\to 0} \frac{f(a + \Delta) - f(a)}{\Delta} - \frac{\mathrm{d}f}{\mathrm{d}z}(a) = 0.$$

Here $\Delta \in \mathbb{C}$. We can write this expression as

$$\lim_{\Delta\to 0} \frac{f(a + \Delta) - f(a) - \frac{\mathrm{d}f}{\mathrm{d}z}(a)\Delta}{\Delta} = 0.$$

Thus $Df_a(\Delta) = \frac{\mathrm{d}f}{\mathrm{d}z}(a)\Delta$, the product being multiplication of complex numbers.

If we identify $\mathbb{C}$ with $\mathbb{R}^2$ by $z = x + \mathrm{i}y \leftrightarrow \begin{pmatrix} x \\ y \end{pmatrix}$, then $f(x + \mathrm{i}y) = u(x, y) + \mathrm{i}v(x, y)$, $f(x + \mathrm{i}y) \leftrightarrow \begin{pmatrix} u(x, y) \\ v(x, y) \end{pmatrix}$ and

$$Df_a \sim \begin{pmatrix} \frac{\partial u}{\partial x}(a) & \frac{\partial u}{\partial y}(a) \\ \frac{\partial v}{\partial x}(a) & \frac{\partial v}{\partial y}(a) \end{pmatrix}.$$

The Cauchy-Riemann equations tell us that this matrix has the form $\begin{pmatrix} c & d \\ -d & c \end{pmatrix}$.

(vi) The total derivative of the determinant $\det : M_{n\times n}(\mathbb{R}) \to \mathbb{R}$ is an interesting and important example (explored further on Sheet 1).

The vector space $M_{n\times n}(\mathbb{R})$ has a canonical basis: E_{ij}, $1 \leq i, j \leq n$, where E_{ij} is the $n \times n$ matrix with 1 in the ij^{th} entry and 0 in all other entries. We calculate the gradient of det with respect to this basis.

The matrix $X = (x_{ij}) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} E_{ij}$.

Using Lagrange's formula for calculating the determinant and expanding along the i^{th} row we get

$$\det((x_{ij})) = \sum_{j=1}^{n} (-1)^{i+j} x_{ij} X_{ij},$$

where X_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix formed from X by deleting the i^{th} row and the j^{th} column. The coordinate functions x_{ij} are independent variables (in fact, linearly independent functionals on $M_{n \times n}(\mathbb{R})$) so

$$\frac{\partial}{\partial x_{kl}} \det((x_{ij})) = (-1)^{k+l} X_{kl},$$

found by expanding along the k^{th} row. The matrix with $(-1)^{i+j} X_{ij}$ in the ij^{th} entry is called the matrix of cofactors which we denote by $\operatorname{cofac}(X)$. According to our conventions, the gradient should be written as a row vector, but if we arrange the components into a matrix the gradient becomes the matrix, $\operatorname{cofac}(X)$.

Recall the formula for inverting an invertible matrix X:

$$X^{-1} = \frac{1}{\det X} (\operatorname{cofac} X)^T .$$

On the problem sheet you are led through a calculation of $D \det_X(A)$ when X is invertible. You should keep in mind that this is an approximation for $\det(X + A) - \det(X)$.

Aside: A point to remember when considering matrices is that if we forget that $A = (a_{ij})$ and $B = (b_{ij})$ are matrices, regard them as vectors taking their dot product, then

$$A.B = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} b_{ij} = A.B = \operatorname{trace} A^T B.$$

Then $D \det_X(A) = \text{grad } \det(X).A = \text{cofac}X.A = \text{trace}(\text{cofac}X)^T A$.

3 Properties of Total Derivatives

1. By definition, the total derivative of a map f at a point a is a linear map:
   $Df_a(v_1 + v_2) = Df_a(v_1) + Df_a(v_2)$ and $Df_a(cv) = cDf_a(v)$, $c \in \mathbb{R}$ and $v, v_1, v_2 \in \mathbb{R}$.
2. The correspondence, $f \mapsto Df$, of a map to its total derivative is linear:
   $D(cf)_a = cD(f)_a$ and $D(f + g)_a = D(f)_a + D(g)_a$
3. **Chain Rule** $D(g \circ f)_a = D(g)_{f(a)} \circ D(f)_a$.

   Theorem 2 *Let V, W and Z be a finite dimensional vector spaces, $U_1 \subseteq V$ and $U_2 \subseteq W$ open subsets, $a \in U$, $f : U_1 \to W$ and $g : U_2 \to Z$ such that $f(U_1) \subseteq U_2$:*

   $$U_1 \xrightarrow{f} U_2 \xrightarrow{g} Z.$$

   If Df_a and $Dg_{f(a)}$ exist, then so does $D(g \circ f)_a$ with $D(g \circ f)_a = D(g)_{f(a)} \circ Df_a$.

   Proof:(Not examinable)
   We start with
   $$f(a + \Delta) = f(a) + Df_a(\Delta) + \|\Delta\| h_f(\Delta)$$
   where $h_f(\Delta) \to 0$ as $\Delta \to 0$ and
   $$g(f(a) + \tilde{\Delta}) = g(f(a)) + Dg_{f(a)}(\tilde{\Delta}) + \|\tilde{\Delta}\| h_g(\tilde{\Delta}).$$
   where $h_g(\tilde{\Delta}) \to 0$ as $\tilde{\Delta} \to 0$. But
   $$g(f(a + \Delta)) = g(f(a) + Df_a(\Delta) + \|\Delta\| h_f(\Delta)).$$
   Take $\tilde{\Delta} = Df_a(\Delta) + \|\Delta\| h_f(\Delta)$ and note
   $$\begin{aligned} \|\tilde{\Delta}\| &\leq \|Df_a\|\|\Delta\| + \|\Delta\|\|h_f(\Delta)\| \\ &= \|\Delta\| \left(\|Df_a\| + \|h_f(\Delta)\|\right) \end{aligned}$$
   and so $\tilde{\Delta} \to 0$ as $\Delta \to 0$. Next
   $$\begin{aligned} g(f(a + \Delta)) &= g(f(a) + \tilde{\Delta}) \\ &= g(f(a)) + Dg_{f(a)}(\tilde{\Delta}) + \|\tilde{\Delta}\| h_g(\tilde{\Delta}), \end{aligned}$$
   and
   $$\begin{aligned} Dg_{f(a)}(\tilde{\Delta}) &= Dg_{f(a)}(Df_a(\Delta) + \|\Delta\| h_f(\Delta)) \\ &= Dg_{f(a)}(Df_a(\Delta)) + Dg_{f(a)}(\|\Delta\| h_f(\Delta)). \end{aligned}$$
   We write this as
   $$g(f(a + \Delta)) = g(f(a)) + Dg_{f(a)} \circ Df_a(\Delta) + I_1 + I_2.$$
   where $I_1 = Dg_{f(a)}(\|\Delta\| h_f(\Delta))$ and $I_2 = \|\tilde{\Delta}\| h_g(\tilde{\Delta})$.
   We need to show that $I_1 + I_2 = o(\|\Delta\|)$:
   (a) $I_1 = \|\Delta\| Dg_{f(a)}(h_f(\Delta))$ so $\|I_1\| \leq \|\Delta\|\|Dg_{f(a)}\|\|h_f(\Delta)\| = o(\|\Delta\|)$;
   (b) $\|I_2\| \leq \|\Delta\| \left(\|Df_a\| + \|h_f(\Delta)\|\right) \|h_g(\tilde{\Delta})\| = o(\|\Delta\|)$ because $h_g(\tilde{\Delta}) \to 0$ as $\tilde{\Delta} \to 0$ and $\tilde{\Delta} \to 0$ as $\Delta \to 0$.

This completes proof that $D(g \circ f)_a = Dg_{f(a)} \circ Df_a$.

Notice that when multiplying Jacobian matrices, the $(i,j)^{th}$ component comes from the dot product of a row and column:

$$\begin{aligned} \frac{\partial(g_i \circ f)}{\partial x_j} &= \left(\frac{\partial g_i}{\partial f_1}, \cdots, \frac{\partial g_i}{\partial f_m}\right) \begin{pmatrix} \frac{\partial f_1}{\partial x_j} \\ \vdots \\ \frac{\partial f_m}{\partial x_j} \end{pmatrix} \\ &= \sum_{k=1}^{m} \frac{\partial g_i}{\partial f_k} \frac{\partial f_k}{\partial x_j}, \end{aligned}$$

the form of the chain rule with which you are already familiar.

Corollary: $D(f^{-1})_{f(a)} = (Df_a)^{-1}$.

Proof: $f^{-1} \circ f = I$. Thus $Df^{-1}_{f(a)} \circ Df_a = I$ because the derivative of the (linear) identity map is the same (linear)identity map.

Example Let $F : (0,\infty) \times (0, 2\pi) \to \mathbb{R}^2 \setminus (0,\infty)$ be given by

$$F(r,\theta) = \begin{pmatrix} x(r,\theta) \\ y(r,\theta) \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}.$$

It is easy to calculate DF and then use it to calculate $D(F^{-1})$ because $D(F^{-1})_{F(r,\theta)} = \left(DF_{(r,\theta)^T}\right)^{-1}$. The Jacobian matrix for $DF_{(r,\theta)^T}$ is $\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$. Thus the Jacobian matrix of $D(F^{-1})_{F(r,\theta)}$ is $\frac{1}{r}\begin{pmatrix} r\cos\theta & r\sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$.

4. We see from the chain rule formula that if $p : (c,d) \to U \subseteq \mathbb{R}^k$ is a differentiable path and $f : U \to \mathbb{R}^n$ has total derivative Df then $p'(t_0) = \begin{pmatrix} p_1'(t_0) \\ \vdots \\ p_k'(t_0) \end{pmatrix}$ is the tangent to the path at $p(t_0)$ and $(f \circ p)'(t_0) = Df_{p(t_0)}(p'(t_0))$. So $Df_{p(t_0)}$ maps the tangent vector to the path p at $p(t_0)$ to the tangent vector to the path $f \circ p$ at $f(p(t_0))$.

5. **Existence of total derivatives**.

Theorem 3 *Let $U \subseteq \mathbb{R}^k$ be open and $f : U \to \mathbb{R}^n$ be such that all partial derivatives $\frac{\partial f}{\partial x_1} \ldots, \frac{\partial f}{\partial x_k}$ are continuous on U. Then Df_a exists for all $a \in U$ and $a \mapsto Df_a : U \to \mathrm{Lin}(\mathbb{R}^k, \mathbb{R}^n) \cong M_{n\times k}(\mathbb{R})$ is continuous.*

Proof: We will restrict ourselves to proving the case when $n = 1$, but the general case follows from this.

Let $e_1, \ldots, e_k$ be the standard basis for $\mathbb{R}^k$ and then for $a \in U$,

$$\begin{aligned} f(a+\Delta) - f(a) &= f(a + \textstyle\sum_{i=1}^{k} \Delta_i e_i) \\ &= f(a + \textstyle\sum_{i=1}^{k} \Delta_i e_i) - f(a + \sum_{i=1}^{k-1} \Delta_i e_i) \\ &+ f(a + \textstyle\sum_{i=1}^{k-1} \Delta_i e_i) - f(a + \sum_{i=1}^{k-2} \Delta_i e_i) \\ &\vdots \\ &+ f(a + \Delta_1 e_1) - f(a) \\ &= \frac{\partial f}{\partial x_n}(c_n)\Delta_n + \frac{\partial f}{\partial x_{n-1}}(c_{n-1})\Delta_{n-1} + \cdots + \frac{\partial f}{\partial x_1}(c_1)\Delta_1 \end{aligned}$$

where each c_r lies on the line joining $a + \sum_{i=1}^{r} \Delta_i e_i$ and $a + \sum_{i=1}^{r-1} \Delta_i e_i$.

Thus

$$\begin{aligned} f(a+\Delta) - f(a) &= \tfrac{\partial f}{\partial x_n}(a)\Delta_n + \tfrac{\partial f}{\partial x_{n-1}}(a)\Delta_{n-1} + \cdots + \tfrac{\partial f}{\partial x_1}(a)\Delta_1 \\ &\quad +(\left(\tfrac{\partial f}{\partial x_n}(c_n) - \tfrac{\partial f}{\partial x_n}(a)\right)\Delta_n + \cdots + \left(\tfrac{\partial f}{\partial x_1}(c_1) - \tfrac{\partial f}{\partial x_1}(a)\right)\Delta_1 \\ &= \operatorname{grad} f(a)\Delta + o(\|\Delta\|). \end{aligned}$$

Hence Df_a exists and equals $\operatorname{grad} f(a)$.

Example: The total derivative does not exists in the following example.

Let $f(x,y) = \dfrac{x^2 y}{x^4 + y^2}$ when $(x,y) \neq (0,0)$ and take the value 0 at $(0,0)$. Let $u = (h,k)$ and calculate the directional derivative in the direction u,

$$\frac{f(0+tu) - f(0)}{t} = \frac{1}{t}\frac{t^3 h^2 k}{t^4 h^4 + t^2 k^2} = \frac{h^2 k}{t^2 h^4 + k^2} \to \frac{h^2}{k},$$

if $k \neq 0$, and the fraction vanishes if $k = 0$. But if $Df_{(0,0)}(u)$ were to exist then it would depend linearly on (h,k) and hence must be of the form $ah + bk$ for some $a, b \in \mathbb{R}$. It follows that although all directional derivatives exist at $(0,0)$, the total derivative does not.

6. **Second Derivatives**

Now suppose that V and W are a finite dimensional vector spaces, $U \subseteq V$ is open and $f : U \to W$ is such that Df_a exists for all $a \in U$. Then $Df : U \to L(V,W)$, which is also a finite dimensional space. We consider the second derivative $D(DF)_a$.

If $u, v \in V$ then $D(Df)_a(u) \in L(V,W)$ and $(D(Df)_a(u)(v) \in W$.

If $V = \mathbb{R}^k$ and $W = \mathbb{R}^n$ then $D(Df)_a(e_i) = \frac{\partial (Df)}{\partial x_i}(a)$ and

$$D(Df)_a(e_i)(e_j) = \frac{\partial (Df)_a}{\partial x_i}(a)(e_j)$$

This this equals $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and we have the following result.

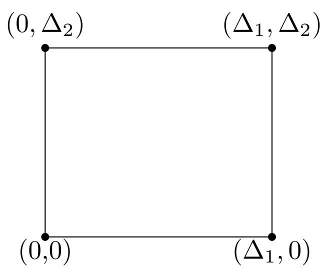
Theorem 4 *If* $\frac{\partial^2 f}{\partial x_i \partial x_j}$ *and* $\frac{\partial^2 f}{\partial x_j \partial x_i}$ *are continuous on* U *then*

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

The map $(u_1, u_2) \mapsto D(Df)_a(u_1)(u_2)$ is a symmetric bilinear form (known as the *hession* of the map f at a).

Proof: For simplicity we suppose that $U \subseteq \mathbb{R}^2$ and $f : U \to \mathbb{R}$. (The case $f : U \to \mathbb{R}^n$ is deduced from this case.)

Consider a rectangle:



Let $I(\Delta_1, \Delta_2) = f(\Delta_1, \Delta_2) - f(\Delta_1, 0) - f(0, \Delta_2) + f(0, 0)$.

(a) Put $g(\Delta_1) = f(\Delta_1, \Delta_2) - f(\Delta_1, 0)$. Then $g(\Delta_1) - g(0) = g'(c_1)\Delta_1$ where c_1 lies between Δ_1 and 0. Further,

$$g'(c_1)\Delta_1 = \left(\frac{\partial f}{\partial x_1}(c_1, \Delta_2) - \frac{\partial f}{\partial x_1}(c_1, 0)\right)\Delta_1 = \left(\frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_1}(c_1, c_2)\Delta_2\right)\Delta_1,$$

for some c_2 lying between Δ_2 and 0.
Note that

$$g(\Delta_1) - g(0) = f(\Delta_1, \Delta_2) - f(\Delta_1, 0) - f(0, \Delta_2) + f(0, 0) = I(\Delta_1, \Delta_2).$$

(b) Now put $h(\Delta_2) = f(\Delta_1, \Delta_2) - f(0, \Delta_2)$. Then

$$h(\Delta_2) - h(0) = h'(d_2)\Delta_2$$

for some d_2 between 0 and Δ_2, and

$$h'(d_2) = \left(\frac{\partial f}{\partial x_2}(\Delta_1, d_2) - \frac{\partial f}{\partial x_2}(0, d_2)\right)\Delta_2 = \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_2}(d_1, d_2)\Delta_2\Delta_1,$$

for some d_1 between Δ_1 and 0.
Note

$$h(\Delta_2) - h(0) = f(\Delta_1, \Delta_2) - f(0, \Delta_2) - f(\Delta_1, 0) + f(0, 0) = I(\Delta_1, \Delta_2).$$

Thus $\frac{I(\Delta_1,\Delta_2)}{\Delta_1\Delta_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}(c_1, c_2) = \frac{\partial^1 f}{\partial x_2 \partial x_1}(d_1, d_2)$.
As $\Delta_1 \to 0$ and $\Delta_2 \to 0$, $c_1 \to 0$ and $c_2 \to 0$, $d_1 \to 0$ and $d_2 \to 0$. So $\frac{\partial^2 f}{\partial x_2 \partial x_1}(c_1, c_2) \to \frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0)$, $\frac{\partial^1 f}{\partial x_2 \partial x_1}(d_1, d_2) \to \frac{\partial^1 f}{\partial x_2 \partial x_1}(0, 0)$. Therefore,

$$\frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0) = \frac{\partial^1 f}{\partial x_2 \partial x_1}(0, 0),$$

as required.

Example: Let

$$f(x_1, x_2) = x_1 x_2 \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2},$$

with $f(0, 0) = 0$. Then $\frac{\partial^2 f}{\partial x_1 \partial x_2}(0, 0) = 1$ and $\frac{\partial^2 f}{\partial x_2 \partial x_1}(0, 0) = -1$.
HINT: Show first that $\frac{\partial f}{\partial x_2}(\Delta_1, 0) = \Delta_1$ and then use the skew symmetry of f.

7. **Mean Value Theorem** When the target or image space has dimension greater than 1 we do not have a Mean Value Theorem as seen in the Moderations syllabus:

Theorem 5 (Mean Value Theorem) *: Let $f : [a, b] \to \mathbb{R}$ be continuous and such that $f'(x)$ exists for all $x \in (a, b)$. Then there exists $c \in (a, b)$ such that $f(b) - f(a) = f'(c)(b - a)$.*

It follows from this result that if the image space is of dimension greater than one :

for $f : [a,b] \to \mathbb{R}^n$ and $f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$, $f(b)-f(a) = \begin{pmatrix} f_1'(c_1)(b-a) \\ f_1'(c_2)(b-a) \\ \vdots \\ f_k'(c_k)(b-a) \end{pmatrix}$ for various points $c_1, c_2, \ldots, c_k \in (a,b)$. From this we conclude $\|f(b) - f(a)\| \leq \sqrt{n} \sup_{c\in[a,b]} \|Df_c(b-a)\|$.

However, we can get a better estimate if we use integration. We use,

$$f(b) - f(a) = g(1) - g(0) = \int_0^1 g'(t)\,\mathrm{d}t,$$

where $g(t) = f(a + t(b-a))$.

Hence,

$$\|f(b) - f(a)\| \leq \int_0^1 \|g'(x)\|\,\mathrm{d}t \leq \sup_{x\in[a,b]} \|Df_x\||b-a|.$$

This argument works for any norm and we fill in details for the Euclidean norm the end of the chapter. The proof is not examinable but the fact

$$\|f(b) - f(a)\| \leq \sup_{c\in[a,b]} \|Df_c\||b-a|$$

is very useful.

For the case where $f : U \subseteq \mathbb{R}^k \to \mathbb{R}$ is differentiable and $a, b \in U$ are such that $a + tb \in U$ for all $t \in [0,1]$ then putting $g(t) = f(a + t(b-a))$ we have $g(1) - g(0) = g'(t_0)(1-0)$ for some $t_0 \in (0,1)$, that is,

$$f(b) - f(a) = Df_{a+t_0 b}(b-a) = \operatorname{grad} f(a + t_0(b-a))(b-a),$$

(matrix multiplication of a row vector and a column vector. This can be written as

$$f(b) - f(a) = \operatorname{grad}(f)(c)(b-a) = \sum_{i=1}^{k} \frac{\partial f}{\partial x_i}(c)(b_i - a_i),$$

where c lies on the line segment joining a and b.

These weaker results are sufficient for our purposes and we will refer to them weak versions of the Mean Value Theorem.

Theorem 6 *Let $U \subseteq \mathbb{R}^n$, $V \subseteq \mathbb{R}^m$ and let $f : U \to V$ be continuously differentiable. Let $a, b \in U$ be such that $a + t(b-a) \in U$ for all $t \in [0,1]$. Then*

$$\|f(b) - f(a)\| \leq \sup_{t\in[0,1]} \|Df_{a+t(b-a)}\|\|b-a\|.$$

Proof:(not examinable) Let $g(t) = f(a + t(b-a))$. Then $g'(t) = Dg_t = Df_{a+t(b-a)}(b-a)$. The function g has m components $g = (g_1, g_2, \ldots, g_m)^T$ and $\int_0^1 g'(t)\,\mathrm{d}t$ means the vector

$$\left(\int_0^1 g_1'(t)\,\mathrm{d}t, \int_0^1 g_2'(t)\,\mathrm{d}t, \ldots, \int_0^1 g_m'(t)\,\mathrm{d}t\right)^T \in \mathbb{R}^m.$$

Then

$$f(b) - f(a) = g(1) - g(0) = \int_0^1 g'(t)\,\mathrm{d}t.$$

Now

$$\begin{aligned}\|\textstyle\int_0^1 g'(t)\,\mathrm{d}t\|^2 &= \textstyle\sum_{i=1}^m(\int_0^1 g_i'(t)\,\mathrm{d}t)^2\\ &\le \textstyle\sum_{i=1}^m \int_0^1 |g_i'(t)|^2\,\mathrm{d}t \text{ by the Cauchy-Schwartz inequality}\\ &= \textstyle\int_0^1 \|g'(t)\|^2\,\mathrm{d}t\\ &\le \sup_{t\in[0,1]}\|g'(t)\|^2\\ &= (\sup_{t\in[0,1]}\|g'(t)\|)^2.\end{aligned}$$

Thus

$$\|\int_0^1 g'(t)\,\mathrm{d}t\| \le \sup_{t\in[0,1]}\|g'(t)\| = \sup_{t\in[0,1]}\|Df_{a+t(b-a)}(b-a)\| \le \sup_{t\in[0,1]}\|Df_{a+t(b-a)}\|\,\|b-a\|.$$

A more intuitive proof of

$$\|\int f\| \le \int \|f\|$$

which applies for any norm goes as follows.

A continuous function f can be approximated uniformly closely by (vector-valued) step functions. So it is enough to prove

$$\|\int_a^b k(t)\mathrm{d}t\| \le \int_a^b \|k(t)\|\mathrm{d}t$$

for $k=\sum_1^m c_i\chi_{I_i}$ where the I_i are disjoint intervals. Then $\int_a^b k(t)\mathrm{d}t = \sum c_i\, length(I_i)$ and $\|\int_a^b k(t)\mathrm{d}t\| \le \sum \|c_i\|\, length(I_i) = \int_a^b \|k(t)\|\mathrm{d}t$ by the subadditivity of the norm.

4 Introduction to the Inverse Function Theorem and the Implicit Function Theorem

Consider

$$S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}.$$

At a point (x_0, y_0) such that $x_0^2 + y_0^2 = 1$ and $y_0 \neq 0$, there is a function $y(x)$ defined near x_0 for which $y(x_0) = y_0$ and $x^2 + y(x)^2 = 1$, namely $y(x) = \sqrt{1 - x^2}$ if $y_0 > 0$ or $y(x) = -\sqrt{1 - x^2}$ if $y_0 < 0$. However, if $y_0 = 0$ so that $x_0 = \pm 1$, we cannot define such a function on any open subset of $\mathbb{R}$ containing x_0. We can only say that locally when $\|x\| < 1$ the function $y(x)$ is determined implicitly as a function of x.

The Implicit Function Theorem describes conditions under which certain variables can be written as functions of the others.

Definition 5 *Let $U \subseteq \mathbb{R}^k$, $V \subseteq \mathbb{R}^n$. A function $f : U \to V$ is a* homeomorphism *if*

(i) f is a bijection;

(ii) both $f : U \to V$ and $f^{-1} : V \to U$ are continuous.

Definition 6 *Let U and V be open subsets of $\mathbb{R}^n$. A function $f : U \to V$ is a* diffeomorphism *if*

(i) f is a homeomorphism;

(ii) both f and f^{-1} are continuously differentiable.

In these lectures we deduce the Implicit Function Theorem from the Inverse Function Theorem. The proof of the Inverse Function Theorem is not examinable (but we give a proof in an appendix at the end of these notes). The proof is quite technical but the statement of the tearem is clear and it is easy to gain some appreciation of the ideas involved in the proof.

Theorem 7 (Inverse Function Theorem) *Let U be an open subset of $\mathbb{R}^n$, $a \in U$ and $f : U \to \mathbb{R}^n$ be a continuously differentiable function such that $Df_a : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. Then there exists an open set $V \subseteq U$ with $a \in V$ such that $f(V)$ is an open subset of $\mathbb{R}^n$ and $f : V \to f(V)$ is a diffeomorphism.*

The proof of the Inverse Function Theorem is not examinable but is included in an appendix at the end of these notes.

To illustrate the *local* nature of this theorem, consider

$$f : \mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}$$

given by $f(z) = z^2$. Then $Df_a(z) = 2az$, complex multiplication of the two complex numbers $2a$ and z, and Df_a is invertible if $a \neq 0$. The mapping f is not one-to-one. In fact, it is a two-to-one mapping which is locally one-to-one: if $a \in \mathbb{C} \setminus \{0\}$ and $V = \{z \in \mathbb{C} \setminus \{0\} : |\arg z - \arg a| < \pi/2\}$, then $f|_V$ is one-to-one with a branch of $z \mapsto \sqrt{z}$ as inverse.

We now move to the Implicit Function Theorem. To gain some appreciation of the statement we first examine the case when the map is linear.

Suppose that linear $L : \mathbb{R}^{n+k} \to \mathbb{R}^n$ has rank n. Then $L^{-1}(0) = \ker L$ has dimension k. If $L(a) = b$ then $L^{-1}(b) = a + \ker L$ is a translation of a k-dimensional vector space (an affine space).

Suppose that

$$L \sim \left(\begin{array}{ccc|ccc} l_{11} & \cdots & l_{1n} & l_{1n+1} & \cdots & l_{1n+k} \\ \vdots & & \vdots & \vdots & & \vdots \\ l_{n1} & \cdots & l_{nn} & l_{nn+1} & \cdots & l_{nn+k} \end{array} \right) = (L'|L''),$$

where L' is an $n \times n$ matrix and L'' is an $n \times k$ matrix. **We assume that the first block L' is invertible.**

Let

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ x_{n+1} \\ \vdots \\ x_{n+k} \end{pmatrix} = \begin{pmatrix} x' \\ x'' \end{pmatrix} \in L^{-1}(b).$$

Then

$$(L'|L'') \begin{pmatrix} x' \\ x'' \end{pmatrix} = b,$$

that is, $L'x' + L''x'' = b$, $L'x' = b - L''x''$. Thus

$$x' = (L')^{-1} (b - L''x'').$$

In this formula where $L(x) = b$, the first n components $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ are determined by (or a function of) the last k components $\begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+k} \end{pmatrix}$. Thus

$$\underline{x} = \begin{pmatrix} x'(x'') \\ x'' \end{pmatrix}.$$

The Implicit Function Theorem extends this result to a non-linear setting.

Theorem 8 (Implicit Function Theorem) *Let U be an open subset of $\mathbb{R}^{n+k}$, $a \in U$, $f : U \to \mathbb{R}^n$ a continuously differentiable function, $f(a) = b$ and Df_a of rank n. Suppose that the Jacobian matrix has the form*

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) & \frac{\partial f_1}{\partial x_{n+1}}(a) & \cdots & \frac{\partial f_1}{\partial x_{n+k}}(a) \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \cdots & \frac{\partial f_n}{\partial x_n}(a) & \frac{\partial f_n}{\partial x_{n+1}}(a) & \cdots & \frac{\partial f_n}{\partial x_{n+k}}(a) \end{pmatrix}$$

where the first n columns are linearly independent. Then if

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ a_{n+1} \\ \vdots \\ a_{n+k} \end{pmatrix},$$

there exists an open set $V \subseteq \mathbb{R}^k$ containing $\begin{pmatrix} a_{n+1} \\ \vdots \\ a_{n+k} \end{pmatrix}$ and continuously differentiable function $g : V \to \mathbb{R}^n$ such that

$$g \begin{pmatrix} a_{n+1} \\ \vdots \\ a_{n+k} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

and

$$f \begin{pmatrix} g(v) \\ v \end{pmatrix} = b$$

for all $v \in V$.

Further the map $v \mapsto \begin{pmatrix} g(v) \\ v \end{pmatrix}$ is a differentiable homeomorphism from the open set V onto an open subset of $f^{-1}(b)$. Thus $v \mapsto \begin{pmatrix} g(v) \\ v \end{pmatrix}$ parametrises a neighbourhood in $f^{-1}(b)$ of the point a.

(Near a, $f^{-1}(b)$ 'looks like' the graph of a function.)

We have stated the theorem for the case when the first n columns are linearly independent. Of course, similar results hold when any n of the columns are linearly independent. Then **the variables corresponding to the chosen set of linearly independent columns can be written as functions of the complementary set of variables**. We illustrate this in the following example.

Example Show that near $x = 0$, $y = 0$, $z(0,0) = 1$ there exists a continuously differentiable function $(x, y) \mapsto z(x, y)$ such that $x^2 + y^2 + z^2 = 1$.

Consider the function $f \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x^2 + y^2 + z^2 - 1$. Then $Df_(x, y, z)^T = \text{grad}\,(x, y, z)^T = (2x, 2y, 2z)$. So $\text{grad}\,(0, 0, 1)^T = (0, 0, 2) \neq 0$. The third component does not vanish and so near $(x, y) = (0, 0)$, according to the Implicit Function Theorem, we can write z as a function of x and y. In fact, $z(x, y) = \sqrt{1 - x^2 - y^2}$.

Example Can the 'surface' given by the equation $xy - z \log y + e^{xz} = 1$ be represented in either of the forms $z = f(x, y)$ or $y = g(x, z)$ in a neighbourhood of $(0, 1, 1)$?

Let $h(x,y,z) = xy - z\log y + \mathrm{e}^{xz} - 1$. To answer the question we need to calculate the $\operatorname{grad} h(0,1,1)$ (writing vectors as row vectors for convenience because there is no danger of confusion). Now

$$\operatorname{grad} h(x,y,z) = (y + z\mathrm{e}^{xz}, x - \frac{z}{y}, -\log y + x\mathrm{e}^{xz})$$

which equals $(2,-1,0)$ at the point $(0,1,1)$. The Implicit Function Theorem tells us that near $(0,1,1)$ the 'surface' can be written in the form $y = g(x,z)$ because the second component of the gradient is -1 and not 0, but it does **not** tell us that it can be written in the form $z = f(x,y)$ because the third component is 0. In fact, if such a differentiable function f were to exist then $xy - f(x,y)\log y + \mathrm{e}^{xf(x,y)} - 1 \equiv 0$. Taking the partial derivative with respect to the x-variable we get $y - f_x(x,y)\log y + (f(x,y) + xf_x(x,y))\mathrm{e}^{xf(x,y)} \equiv 0$; evaluating at $(0,1,1)$ and noting that we must have $f(0,1) = 1$, we get a contradiction.

5 The Implicit Function Theorem and Applications

Example Let $f(x,y,z) = xy - y\log z + \sin xz$ and note that $f(0,2,1) = 0$. We consider the 'surface' given by $f(x,y,z) = 0$ and wish to show that it can be represented (locally, near $(0,2,1)$) as the graph of a function $z = z(x,y)$.

Note that $\text{grad}\, f(x,y,z) = (y + z\cos xz, x - \log z, -\frac{y}{z} + x\cos xz)$ and so $\text{grad}\, f(0,2,1) = (3,0,-2)$.

The third component is non-zero and hence, by the Implicit Function Theorem, there exists a differentiable real-valued function $z(x,y)$ defined on an open neighbourhood of $(0,2)$ such that $(x,y) \mapsto (x,y,z(x,y))$ which parametrizes the 'surface' near $(0,2,1)$). [We define surfaces properly next lecture.]

Example The point $(1,-1,1)$ lies on the 'surfaces' given by $x^3(y^3+z^3) = 0$ and $(x-y)^3 - z^2 = 7$. We will show that that, in a neighbourhood of this point, the curve of intersection of these surfaces can be described by a set of equations of the form $y = y(x)$ and $z = z(x)$.

In fact, this question can be done 'by hand'. On the surface and near $x = 1$, $y^3 + z^3 = 0$ and so $z = -y$. Hence $(x-y)^3 - y^2 = 7$. Thus $x^3 + (3x-1)y^2 - 3x^2y + x^3 - 7 = 0$ and then we apply the formula for the root of a cubic to find $y(x)$ and hence $z(x)$.

However, existence can be established using the Implicit Function Theorem:

$$f\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x^3(y^3+z^3) \\ (x-y)^3 - z^2 - 7 \end{pmatrix},$$

$$Df_{\begin{pmatrix} x \\ y \\ z \end{pmatrix}} = \begin{pmatrix} 3x^2(y^3+z^3) & 3x^3y^2 & 3x^3z^2 \\ 3(x-y)^2 & -3(x-y)^2 & -2z \end{pmatrix},$$

$$Df_{\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}} = \begin{pmatrix} 0 & 3 & 3 \\ 12 & -12 & -2 \end{pmatrix}.$$

But

$$\det\begin{vmatrix} 3 & 3 \\ -12 & 2 \end{vmatrix} = 42 \neq 0.$$

Hence, there exist continuously differentiable real-valued $y(x)$ and $z(x)$ defined in an open neighbourhood of 1 such that $y(1) = -1$, $z(1) = 1$ and $\begin{pmatrix} x \\ y(x) \\ z(x) \end{pmatrix}$ parametrizes part of the curve near $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.

In order to be more economical in the following statement and proof, we will use x' to indicate the first n components and x'' the last k components of a vector x with $n+k$ components.

Theorem 9 (Implicit Function Theorem) *Let U be an open subset of $\mathbb{R}^{n+k}$, $a \in U$, and $f : U \to \mathbb{R}^n$ a continuously differentiable function with $f(a) = b$ and such that Df_a has rank n. Assume for convenience that the first n columns of the Jacobian matrix at a are linearly independent. Then there exists an open set $V \subseteq \mathbb{R}^k$ with $a'' \in V$ and a continuously differentiable function $g : V \to \mathbb{R}^n$ such that $g(a'') = a'$, $f\begin{pmatrix} g(x'') \\ x'' \end{pmatrix} = b$ and the map $x'' \mapsto \begin{pmatrix} g(x'') \\ x'' \end{pmatrix}$ maps V homeomorphically onto an open subset of $f^{-1}(b)$ containing a.*

Proof [Not examinable]

We define $F : U \to \mathbb{R}^{n+k}$ by $F(\begin{pmatrix} x' \\ x'' \end{pmatrix}) = \begin{pmatrix} f(x) \\ x'' \end{pmatrix}$, remembering that $f(x)$ has n components. The total derivative of F has Jacobian matrix:

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} & \frac{\partial f_1}{\partial x_{n+1}} & \cdots & \frac{\partial f_1}{\partial x_{n+k}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} & \frac{\partial f_n}{\partial x_{n+1}} & \cdots & \frac{\partial f_n}{\partial x_{n+k}} \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix}$$

where the bottom right $k \times k$ block is the identity matrix.

At $x = a$, DF_a is invertible because the top left $n \times n$ block is invertible. By the Inverse Function Theorem, there exists open $\tilde{U} \subseteq \mathbb{R}^{n+k}$ with $a \in \tilde{U} \subseteq U$ and

(i) $F(\tilde{U})$ an open subset of $\mathbb{R}^{n+k}$;

(ii) $F : \tilde{U} \to F(\tilde{U})$ a bijection;

(iii) F and F^{-1} are diffeomorphisms between $\tilde{U}$ and $F(\tilde{U})$.

Keep in mind that $F(x) = \begin{pmatrix} f(x) \\ x'' \end{pmatrix}$ and $x = F^{-1}\begin{pmatrix} f(x) \\ x'' \end{pmatrix}$. Let W be an open subset of $\mathbb{R}^n$ with $b \in W$, V be an open subset of $\mathbb{R}^k$ with $a'' \in V$, and such that $W \times V \subseteq F(\tilde{U})$. Then $b \in W$ and for $x'' \in V$ we can define the function g by

$$F^{-1}\begin{pmatrix} b \\ x'' \end{pmatrix} = \begin{pmatrix} g(x'') \\ x'' \end{pmatrix},$$

so g is made up of the first n components of F^{-1} on $\{b\} \times V$. Thus g is it is continuously differentiable as a restriction of the continuously differentiable function F^{-1} composed with a projection onto the first n components. Then

$$F\begin{pmatrix} g(x'') \\ x'' \end{pmatrix} = F\left(F^{-1}\begin{pmatrix} b \\ x'' \end{pmatrix}\right) = \begin{pmatrix} b \\ x'' \end{pmatrix}$$

with

$$f\begin{pmatrix} g(x'') \\ x'' \end{pmatrix} = b.$$

as required.

6 Submanifolds of $\mathbb{R}^N$

A k-dimensional submanifold of $\mathbb{R}^N$ is a (curved) generalization of a k-dimensional subspace of $\mathbb{R}^N$ formed by patching together with local diffeomeorphisms pieces of k-dimensional subspaces of $\mathbb{R}^N$. This has proved to be a very fruitful idea providing a large interesting class of spaces on which the techniques of differential calculus can be applied.

In Chapter 6, §3, Shifrin, Multivariable Calculus (Wiley) this is set out well.

Definition 7 *Let k and N be non-negative integers with $k < N$. A subset M of $\mathbb{R}^N$ is a k-dimensional submanifold if for each $m \in M$ there exists*

(i) an open subset U of $\mathbb{R}^N$ with $m \in U$;

(ii) a k-dimensional subspace X of $\mathbb{R}^N$ and an open set V of $\mathbb{R}^N$;

(iii) a diffeomorphism $\phi : V \to \phi(V) = U$ such that $\phi(V \cap X) = U \cap M$.

The map $\phi|V \cap X$ is a local parametrisation of $U \cap M$.

ASIDE: The map ϕ is like a restriction of the map F^{-1} in the proof of the Implicit Function Theorem and $x'' \mapsto F^{-1}\begin{pmatrix} b \\ x'' \end{pmatrix} = \begin{pmatrix} g(x'') \\ x'' \end{pmatrix}$ is like $\phi|_{V\cap X}$.

Example: Let $\mathbb{S}^2 = \{m \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$. Then $\mathbb{S}^2$ is a 2-dimensional submanifold of $\mathbb{R}^3$.

If $m_0 \in \mathbb{S}^2$, $m_0 = (x_0, y_0, z_0)$ with, for example, $z_0 > 0$, and V is a small open subset of $\mathbb{R}^3$ containing $(x_0, y_0, 0)$. We take $\{(x, y, 0) : x, y \in \mathbb{R}\}$ as the 2-dimensional subspace of $\mathbb{R}^3$ and define $\phi : V \to \mathbb{R}^3$ by $\phi(x, y, z) = (x, y, z + \sqrt{1 - x^2 - y^2})$. Then in V, if $x^2 + y^2 < 1$, this map is well defined and continuously differentiable; if in addition $z = 0$, $\phi(x, y, z) \in \mathbb{S}^2$.

Example More generally, suppose that $f : \mathbb{R}^3 \to \mathbb{R}$ is continuously differentiable and $b \in \mathbb{R}$ is such that $f^{-1}(b) \neq \emptyset$. Suppose that for all $a \in f^{-1}(b)$, $\operatorname{grad} f(a) \neq 0$. Then by the proof of the Implicit Function Theorem, $f^{-1}(b)$ is a 2-dimensional submanifold of $\mathbb{R}^3$.

Definition 8 *: A subset X of $\mathbb{R}^N$ is called a* surface *if it is a 2-dimensional submanifold of $\mathbb{R}^N$.*

(An abstract surface is a related but more general concept.)

Example: Let $\mathbb{T}^2 = \{(z_1, z_2) \in \mathbb{C} \times \mathbb{C} : |z_1| = |z_2| = 1\}$. Then $\mathbb{T}^2$ is a 2-dimensional submanifold of $\mathbb{C} \times \mathbb{C} \leftrightarrow \mathbb{R}^4$. In real terms, let

$$f(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}) = \begin{pmatrix} x_1^2 + x_2^2 \\ x_3^2 + x_4^2 \end{pmatrix}.$$

Then $\mathbb{T}^2 = f^{-1}(\begin{pmatrix} 1 \\ 1 \end{pmatrix})$ and the Jacobian of f is

$$\begin{pmatrix} 2x_1 & 2x_2 & 0 & 0 \\ 0 & 0 & 2x_3 & 2x_4 \end{pmatrix}.$$

Notice that $x_1^2 + x_2^2 = 1 \Rightarrow (2x_1, 2x_2) \neq (0,0)$, $x_3^2 + x_4^2 = 1 \Rightarrow (2x_3, 2x_4) \neq (0,0)$. Hence on $\mathbb{T}^2 = f^{-1}(\begin{pmatrix} 1 \\ 1 \end{pmatrix})$, the rank of the Jacobian is 2 and thus $\mathbb{T}^2$ is a 2-dimensional submanifold of $\mathbb{R}^4$.

Example (of a set which fails at just one point to be a submanifold)
Let $f : \mathbb{R}^3 \to \mathbb{R}$ be given by

$$f(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}) = x_1^2 + x_2^2 - x_3^2.$$

Then $f^{-1}(0)$ is a 'double cone'. We see, when we sketch $f^{-1}(0)$, that the origin in $\mathbb{R}^3$ is a 'singular point' at which the definition of a submanifold fails to hold. [In fact, removing the origin from any open subset of the double cone containing the origin yields a disconnected set. Thus no open neighbourhood of the origin in the double cone is homeomorphic to an open disc.]

In this example, $\text{grad} f(x_1, x_2, x_3)^T = (2x_1, 2x_2, -2x_3)$ and hence the grad $(0,0,0)^T = (0,0,0)$ which is not of rank 1.

An equivalent definition of a submanifold is as follows (the equivalence can be established by using the Injective Mapping Theorem, proved in an appendix at the end of these notes):

Definition 9 *A set $M \subseteq \mathbb{R}^N$ is a k-dimensional submanifold if for each $m \in M$ there exists U an open subset of $\mathbb{R}^N$ with $m \in U$, W an open subset of $\mathbb{R}^k$ and a continuously differentiable homeomorphism $\mathbf{r} : W \to r(W) = U \cap M$ such that $D\mathbf{r}_w$ has rank k for each $w \in W$.*

The map $\mathbf{r}$ is called a parametrisation *of $U \cap M$, its inverse is a* coordinate map *on $U \cap M$.*

The map $x'' \mapsto \begin{pmatrix} g(x'') \\ x'' \end{pmatrix}$ in the Implicit Function Theorem is an example of a parametrization of $f^{-1}(b) \cap U$.

Definition 10 *If M is a k-dimensional submanifold of $\mathbb{R}^N$, $m \in M$, $p : (a,b) \to M \subset \mathbb{R}^N$ a continuously differentiable map (path) with $p(c) = m$ for some $c \in (a,b)$, then $p'(c)$ is a* tangent vector *to M at m. The set of all tangent vectors at m is called the* tangent space *to M at m and denoted by $T_m(M)$.*

7 Tangent spaces, normal vectors and Lagrange multipliers.

Lemma 1 *Let M be a k-dimensional submanifold of $\mathbb{R}^N$ and let $m \in M$. Then $T_m(M)$, the set of all tangent vectors at m, is a k-dimensional subspace of $\mathbb{R}^N$.*

Proof: By definition, there exists an open subset U of $\mathbb{R}^N$ with $m \in U$, a k-dimensional subspace X of $\mathbb{R}^N$, an open subset W of $\mathbb{R}^N$ and a diffeomorphism $\phi : W \to \phi(W) = U$ such that $\phi(W \cap X) = U \cap M$. Say $\phi(d) = m$. We will prove that $T_m(M) = D\phi_d(X)$.

Suppose that $p : (a, b) \to U \cap M$ is a continuously differentiable path with $p(c) = m$. Let $q = \phi^{-1} \circ p$. Then q is the corresponding continuously differentiable path in $W \cap X$ such that $p = \phi \circ q$. All tangent vectors to q lie in X because the path q lies in the vector space X, and $p'(c) = D\phi_d(q'(c)) \subseteq D\phi_d(X)$.

Now suppose that $v \in X$. Then $t \mapsto d + tv$ is a path in X with tangent vector v at d (when $t = 0$). For t sufficiently small, $d + tv \in W \cap X$ defining $p(t) = \phi(d + tv)$ a continuously differentiable path in M with $p(0) = m$. Note that $p'(t) = D\phi_{d+tv}(v)$ and $D\phi_d$ is an isomorphism. Hence $T_m(M) = D\phi_d(X)$, a k-dimensional subspace of $\mathbb{R}^N$.

The existence of a tangent space is a fundamental property of a sub-manifold. Just as the derivative at a point is the best linear approximation to a map, locally, the tangent space is an approximation to the submanifold.

Definition 11 *Let M be a k-dimensional submanifold of $\mathbb{R}^N$ and let $m \in M$. A vector $v \in \mathbb{R}^N$ is* normal *to M at m if $v \in T_m(M)^\perp$, that is, v is orthogonal to all tangent vectors to M at m.*

Note that the set of normal vectors to M at m forms an $(N - k)$-dimensional subspace of $\mathbb{R}^N$.

We now explore these notions in the context of submanifolds defined implicitly by a set of constraints.

Let U be an open subset of $\mathbb{R}^{n+k}$, $f : U \to \mathbb{R}^n$ a continuously differentiable map, $b \in \mathbb{R}^n$ such that $f^{-1}(b) \neq \emptyset$ and at each point $m \in f^{-1}(b)$, the total derivative Df_m has rank n. Then, by the methods used in the proof of the Implicit Function Theorem, $M = f^{-1}(b)$ is a k-dimensional submanifold of $\mathbb{R}^{n+k}$. Suppose that $p : (r, s) \to f^{-1}(b)$ is a continuously differentiable path. Then $f(p(t)) = b$ for all $t \in (r, s)$. Hence $Df_{p(t)}(p'(t)) = 0$, that is, $p'(t) \in \ker Df_{p(t)}$. Thus $T_m(f^{-1}(b)) = \ker Df_m$, both k-dimensional vector spaces.

Expressing $T_m(f^{-1}(b)) = \ker Df_m$ in terms of the Jacobian matrix yields for $v \in T_m(f^{-1}(b))$, $v = \begin{pmatrix} v_1 \\ \vdots \\ v_{n+k} \end{pmatrix}$,

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(m) & \cdots & \frac{\partial f_1}{\partial x_{n+k}}(m) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(m) & \cdots & \frac{\partial f_n}{\partial x_{n+k}}(m) \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_{n+k} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Thus $v \perp \operatorname{grad} f_i(m)$ for $i = 1, \ldots, n$ and we have established that **the n linearly independent gradient vectors of the components of f at m, form a basis for the n-dimensional orthogonal space in $\mathbb{R}^{n+k}$ to the k dimensional tangent space $T_m(f^{-1}(b))$.**

Thus any vector, orthogonal to the tangent vectors at m, lies in the linear span of the gradient vectors of the components of f at m.

Theorem 10 (Lagrange Multipliers) *Let U be an open subset of $\mathbb{R}^{n+k}$, let $f : U \to \mathbb{R}^n$ be a continuously differentiable function and $b \in \mathbb{R}^n$ such that $f^{-1}(b) \neq \emptyset$. Suppose that for each $m \in f^{-1}(b)$, the total derivative Df_m has rank n and that $gU \to \mathbb{R}$ has a maximum or minimum on $f^{-1}(b)$ at $m_0 \in f^{-1}(b)$. Then* $\operatorname{grad} g(m_0) \perp T_m(f^{-1}(b))$, *equivalently,* $\operatorname{grad} g(m_0)$ *lies in the linear span of* $\operatorname{grad} f_1(m_0), \cdots, \operatorname{grad} f_n(m_0)$.

The components $f_1, \cdots, f_n$ are known as the constraints of the problem.

Proof: Let $p : (r, s) \to f^{-1}(b)$ be a continuously differentiable path such that $p(c) = m_0$. Then $g \circ p : (r, s) \to \mathbb{R}$ has a maximum or minimum at c. Thus $(g \circ p)'(c) = 0$, that is, $\operatorname{grad} g(m_0)p'(c) = 0$ (row vector multiplied by a column vector, or dot product of two vectors). Thus $\operatorname{grad} g(m_0)$ is orthogonal to $p'(c)$ which is an arbitrary tangent to $f^{-1}(b)$ at m_0, that is, $\operatorname{grad} g(m_0) \perp T_{m_0}(f^{-1}(b))$. Hence $\operatorname{grad} g(m_o)$ lies in the linear span of the vectors $\operatorname{grad} f_1(m_0), \cdots, \operatorname{grad} f_n(m_0)$.

It follows that there exist $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$ such that

$$\operatorname{grad} g(m_0) = \lambda_1 \operatorname{grad} f_1(m_0) + \cdots + \lambda_n \operatorname{grad} f_n(m_0)$$

These scalars $\lambda_1 \cdots, \lambda_k$ are known as the Lagrange multiplyers.

8 Applications of the technique of Lagrange Multipliers.

A number of exercises using Lagrange Multipliers are given on the second problem sheet. For our last lecture we will study several more advanced applications.

Example In this first example we will prove a well known result concerning the distance between submanifolds. (Another method of deducing this same result is set out in the second problem sheet. You are recommended to master both approaches.)

Let $f_1, f_2 : \mathbb{R}^3 \to \mathbb{R}$ be such that $M_1 = f^{-1}(b_1)$ and $M_2 = f^{-1}(b_2)$ are disjoint 2-dimensional submanifolds of $\mathbb{R}^3$, that is, surfaces. Show that if $a_1 \in M_1$, $a_2 \in M_2$ are such that

$$\|a_1 - a_2\| \leq \|m_1 - m_2\|$$

for all $m_1 \in M_1$, $m_2 \in M_2$, then the line joining a_1 and a_2 is normal to M_1 at a_1 and M_2 at a_2.

We set this up in such a way that we can apply the technique of Lagrange multipliers. Let $F : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R} \times \mathbb{R}$ be given by

$$F(\begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix}) = \begin{pmatrix} f_1(\underline{x}) \\ f_2(\underline{y}) \end{pmatrix}.$$

Then $M_1 \times M_2 = F^{-1}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ and we try to minimize

$$g\begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix} = \|\underline{x} - \underline{y}\|^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2$$

on $M_1 \times M_2$.

The Jacobian matrix of F is:

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \frac{\partial f_2}{\partial y_3} \end{pmatrix};$$

and the gradient vector of g has six components:

$$\operatorname{grad} g(\begin{matrix} \underline{x} \\ \underline{y} \end{matrix}) = 2(x_1 - y_1, x_2 - y_2, x_3 - y_3, -(x_1 - y_1), -(x_2 - y_2), -(x_3 - y_3)).$$

Thus if $\operatorname{grad} g$ is a linear combination of the rows of the Jacobian matrix at a point $(\underline{x}, \underline{y})$ then $\operatorname{grad} f_1(\underline{x})$, $\operatorname{grad} f_2(\underline{y})$ and $\underline{x} - \underline{y}$ are parallel. According to the technique of Lagrange multipliers, this occurs at the 'closest points', $\underline{x} = a_1$ and $\underline{y} = a_2$.

Example
Suppose that $A, B \in M_{n\times n}(\mathbb{R})$ are symmetric and such that $x^T Ax = \sum_{ij} x_i a_{ij} x_j > 0$ if $x \neq 0$. We say that A is a positive matrix. (Note that this implies that A is invertible.) Show that $\max x^T Bx$ subject to the constraint $x^T Ax = 1$ is an eigenvalue of $A^{-1}B$.

Solution: Let $f(x) = x^T Ax = \sum_{ij} x_i a_{ij} x_j$ and $g(x) = x^T Bx$. By the technique of Lagrange multipliers, if $g(a) \geq g(x)$ on $\{x \in \mathbb{R}^n : f(x) = 1\}$ then $\operatorname{grad} g(a) = \lambda \operatorname{grad} f(a)$ for some $\lambda \in \mathbb{R}$.
First we calculate $\operatorname{grad} f(a)$:

$$f(a + \Delta) = (a + \Delta)^T A(a + \Delta) = a^T Aa + \Delta^T Aa + a^T A\Delta + \Delta^T A\Delta.$$

Thus

$$Df_a(\Delta) = \Delta^T Aa + a^T A\Delta = 2a^T A\Delta$$

as A is symmetric, and we see that grad $f(a)$ is the row vector $2a^T A$.

Similarly, $2a^T B$ is the row vector grad $g(a)$. Therefore, $2a^T B = \lambda 2a^T A$, or equivalently, $Ba = \lambda Aa$ as the matrices A and B are symmetric. Thus $A^{-1}Ba = \lambda a$. Further, $g(a) = a^T Ba = \lambda a^T Aa = \lambda$ because $a^T Aa = 1$.

Thus λ is an eigenvalue for $A^{-1}B$ and the maximum is achieved at a corresponding eigenvector a.

If $(B - \lambda A)a = 0$, $a \neq 0$, then λ is called a *relative eigenvalue* and a a *relative eigenvector*.

We should note that there is a purely algebraic way of doing this problem: the matrix A is symmetric and positive and we can diagonalize A and B simultaneously.

Example

In this example we will show that the $n \times n$ orthogonal matrices form an $n(n-1)/2$-dimensional submanifold of the n^2-dimensional vector space $M_{n\times n}(\mathbb{R})$. So for example, $O(3)$ and $SO(3)$ are 3-dimensional submanifolds of $\mathbb{R}^9$.

Let $\mathrm{Sym}_{n\times n}(\mathbb{R})$ denote the $n \times n$ symmetric real matrices and let

$$f : M_{n\times n}(\mathbb{R}) \to \mathrm{Sym}_{n\times n}(\mathbb{R})$$

be given by $f(X) = X^T X$. Then $f^{-1}(I) = \{X \in M_{n\times n}(\mathbb{R}) : X^T = X^{-1}\}$ is the group of orthogonal matrices, denoted $O(n)$. The columns of the an orthogonal matrix form an orthonormal basis for $\mathbb{R}^n$ and multiplication by an orthogonal matrix preserves the length of a vector.

Note that $\det(X^T X) = 1$ when X is an orthogonal matrix so $(\det X)^2 = 1$, $\det X = \pm 1$. The subgroup of orthogonal matrices of determinant 1 is known as the the special orthogonal group and denoted by $SO(n)$. The group $SO(n)$ is both an open and a closed subset of $O(n)$.

[We show that the set $O(n)$, given implicitly, is a submanifold by showing that the Jacobian of the constraint map (in this case f) is of constant rank.]

First we calculate Df_I.

$$f(I + \Delta) = (I + \Delta)^T (I + \Delta) = I^2 + \Delta^T I + I\Delta + \Delta^T \Delta.$$

Hence, $Df_I(\Delta) = \Delta^T + \Delta$, a symmetric matrix. If $\tilde{\Delta}$ is an arbitrary symmetric matrix then $Df_I(\tilde{\Delta}/2) = \tilde{\Delta}$. The symmetric matrices, $\mathrm{Sym}_{n\times n}(\mathbb{R})$, form an $\frac{n(n+1)}{2}$-dimensional vector space. So $Df_I : M_{n\times n}(\mathbb{R}) \to \mathrm{Sym}_{n\times n}(\mathbb{R})$ is a surjective linear map of rank $\frac{n(n+1)}{2}$.

What is $\ker Df_I$?

A matrix $\Delta \in \ker Df_I$ iff $\Delta^T + \Delta = 0$, that is, $\Delta^T = -\Delta$. So $\ker Df_I = \mathrm{Skew}_{n\times n}(\mathbb{R})$ - the skew symmetric matrices.

Let $A \in f^{-1}(I)$. Then $Df_A(\Delta) = A^T\Delta + \Delta^T A$. We will show that $Df_A : M_{n\times n}(\mathbb{R}) \to \mathrm{Sym}_{n\times n}(\mathrm{R})$ is surjective. Let $\tilde{\Delta} \in \mathrm{Sym}_{n\times n}(\mathbb{R})$ and try to solve $Df_A(\Delta) = \tilde{\Delta}$, that is, $A^T\Delta + \Delta^T A = \tilde{\Delta}$. Note that $A^T = A^{-1}$ and try $\Delta = A\tilde{\Delta}/2$. Then

$$A^T A\tilde{\Delta}/2 + (A\tilde{\Delta}/2)^T A = A^{-1}A\tilde{\Delta}^T/2 + (\tilde{\Delta}/2)A^{-1}A = \tilde{\Delta}$$

because $\tilde{\Delta}$ is symmetric. Thus Df_A is surjective of rank $n(n+1)/2$.

We have shown that the map Df is of constant rank $n(n+1)/2$ on $f^{-1}(I) = O(n)$. Thus $O(n)$ and hence $SO(n)$ are $n(n-1)/2$-dimensional submanifolds of $M_{n\times n}(\mathbb{R})$.

What is $\ker Df_A$, when $A \in f^{-1}(I)$?

If $Df_A(\Delta) = \Delta^T A + A^T\Delta = 0$ then $A^T\Delta = A^{-1}\Delta \in \mathrm{Skew}_{n\times n}(\mathbb{R})$. Hence $\Delta \in A\,\mathrm{Skew}_{n\times n}(\mathbb{R})$. If $m_A : M_{n\times n}(\mathbb{R}) \to M_{n\times n}(\mathbb{R})$ is given by $m_A(X) = AX$ then the map m_A is linear, so $Dm_A = m_A$ and $Dm_A = m_A : T_I(O(n)) \to T_A(O(n)$.

This example is explored further on the problem sheet where you will find that $\exp : T_I(O(n)) \to O(n)$. So exp maps a tangent vector to the submanifold $O(n)$ at I to $O(n)$, that is, a skew symmetric matrix maps to an orthogonal matrix and the map exp gives a parametrisation of the submanifold $O(n)$ near $I \in O(n)$. This story is taken up later in your course in the study of Lie groups.

9 Appendices

9.1 Appendix A: Inverse Function Theorem

Theorem (Inverse Function Theorem - normalised version): *Let U, V be open subsets of $\mathbb{R}^n$, $0 \in \mathbb{R}^n$, $f : U \to V$ a continuously differentiable map with $f(0) = 0$ and $Df_0 = I$. Then there exist open sets $U' \subseteq U$, $V' \subseteq V$ with $0 \in U'$ such that $f|_{U'} : U' \to V'$ is a diffeomorphism, that is, $f|_{U'}$ is a homeomorphism from U' onto V' and $f^{-1} : V' \to U'$ is continuously differentiable.*

Proof:

- Let $g(x) = x - f(x)$. Then $Df_0 = 0$ and, by the continuity of $x \mapsto Df_x$, there exists $r > 0$ such that $\|x - 0\| < 2r$ when $\|Dg_x - 0\| < \frac{1}{2}$. Thus, by a waek version of the Mean Value Theorem, $\|g(x) - g(0)\| < \frac{1}{2}\|x - 0\|$ when $\|x\| < 2r$ and so $g\left(\overline{B_r(0)}\right) \subseteq \overline{B_{\frac{r}{2}}(0)}$.

- Next we show that $f\left(\overline{B_r(0)}\right) \supseteq \overline{B_{\frac{r}{2}}(0)}$.

  Let $y \in \overline{B_{\frac{r}{2}}(0)}$ so $\|y\| \leq \frac{r}{2}$. Define $g_y(x) = y + g(x) = y + x - f(x)$. If $\|y\| \leq \frac{r}{2}$ and $\|x\| \leq r$ then $\|g_y(x)\| \leq \frac{r}{2} + \frac{r}{2}$. Hence $g_y\left(\overline{B_r(0)}\right) \subseteq \overline{B_r(0)}$. Note that $D(g_y)_x = Dg_x$ so

  $$\|g_y(x_1) - g_y(x_2)\| = \|g(x_1) - g(x_2)\| \leq \frac{1}{2}\|x_1 - x_2\|$$

  when $x_1, x_2 \in \overline{B_r(0)}$. Hence, by the Contraction Mapping Theorem, there exists *unique* $x \in \overline{B_r(0)}$ such that $g_y(x) = x$, that is, $x = y + x - f(x)$ so $y = f(x)$. We have now shown that $f\left(\overline{B_r(0)}\right) \supseteq \overline{B_{\frac{r}{2}}(0)}$ and we will define $\phi : \overline{B_{\frac{r}{2}}(0)} \to \overline{B_r(0)}$ to be such that $f(\phi(y)) = y$. (We show that ϕ is well defined shortly)

- *Continuity of ϕ on $B_{\frac{r}{2}}(0)$.*

  Note that $x = g(x) + f(x)$. If $x_1, x_2 \in B_r(0)$,

  $$\|x_1 - x_2\| \leq \|g(x_1) - g(x_2)\| + \|f(x_1) - f(x_2)\| \leq \frac{1}{2}\|x_1 - x_2\| + \|f(x_1) - f(x_2)\|.$$

  Therefore, $\frac{1}{2}\|x_1 - x_2\| \leq \|f(x_1) - f(x_2)\|$, $\|x_1 - x_2\| \leq 2\|f(x_1) - f(x_2)\|$. This shows that f is one-to-one on $B_r(0)$and that its local inverse ϕ is well defined on $B_{\frac{r}{2}}(0)$. If if $y_1, y_1 \in B_{\frac{r}{2}}(0)$ then $\|\phi(y_1) - \phi(y_2)\| \leq 2\|y_1 - y_2\|$ and so ϕ is continuous on $B_{\frac{R}{2}}(0)$.

- *Differentiability of ϕ*

  Let $y_1, y_2 \in B_{\frac{r}{2}}(0)$ be such that $y_1 = f(x_1)$, $y_2 = f(x_2)$ where $x_1, x_2 \in \overline{B_r(0)}$. So $\phi(y_1) = x_1$ and $\phi(y_2) = x_2$. Then

  $$\|\phi(y_1) - \phi(y_2) - (Df_{x_2})^{-1}(y_1 - y_2)\| = \|x_1 - x_2 - (Df_{x_2})^{-1}(f(x_1) - f(x_2))\|.$$

  The continuity of $x \mapsto Df_x$ implies

  $$\begin{aligned} \|(x_1 - x_2) - (Df_{x_2})^{-1}(f(x_1) - f(x_2))\| &= \|(Df_{x_2}^{-1}(Df_{x_2}(x_1 - x_2) - f(x_1) - f(x_2))\| \\ &\leq \|Df_{x_2}^{-1}\|\|Df_{x_2}(x_1 - x_2) - (f(x_1) - f(x_2))\| \\ &= o(\|x_1 - x_2\|) \\ &= o(\|y_1 - y_2\|) \end{aligned}$$

  because $\frac{\|y_1 - y_1\|}{\|x_1 - x_2\|} \geq \frac{1}{2}$.

  Hence ϕ is differentiable and $D(\phi)_y = (Df_{\phi(y)})^{-1}$ for $y \in B_{\frac{r}{2}}(0)$.

  Finally, $x \mapsto Df_x$ is continuous which implies that $y \mapsto D(\phi)_y$ is continuous on $B_{\frac{r}{2}}(0)$.

  Let $U' = \phi(B_{\frac{r}{2}}(0))$ and $V' = B_{\frac{r}{2}}(0)$.

Theorem (Inverse Function Theorem): *Suppose that U, V are open subsets of $\mathbb{R}^n$ and $f : U \to V$ is continuously differentiable. Suppose that for some $a \in U$, Df_a is invertible. Then there exist open $U' \subseteq U$ with $a \in U'$, $V' \subseteq V$ such that $f(U') = V'$ and $f|U' : U' \to V'$ is a diffeomeophism.*

Proof: Define $\tilde{f}(x) = (Df_a)^{-1}(f(x+a) - f(a))$. The theorem above implies that $\tilde{f}$ is a diffeomorphism near 0. Hence so is $x \mapsto f(x+a) - f(a)$. Thus f is a diffeomphism near a.

9.2 Appendix B: Injective Mapping Theorem

Theorem 11 *Let U be an open subset of $\mathbb{R}^n$, $a \in U$ and $f : U \to \mathbb{R}^{n+k}$ be continuously differentiable function such that Df_a is injective, that is, a one-to-one linear map. Then there exist open $U' \subseteq U$ and V an open subset of $\mathbb{R}^k$ with $0 \in V$ and a continuously differentiable function $F : U' \times V \to \mathbb{R}^{n+k}$ such that $F|_{U' \times 0} = f|U'$, $F(U' \times V$ is an open subset of $\mathbb{R}^{n+k}$ and $F : U' \times V \to F(U' \times V)$ is a diffeomorphism.*

Note that it now follows that $f(U')$ is an n-dimensional submanifold of $\mathbb{R}^{n+k}$.

Proof

Let B be an $(n+k) \times n$ real matrix such that the columns of the Jacobian of Df_a and the columns of B together are linearly independent. Then the matrix (Df_a, B) is invertible. We will define a continuously differentiable function $F : U \times \mathbb{R}^n \to \mathbb{R}^{n+k}$ with the above matrix as the Jacobian at $\begin{pmatrix} a \\ 0 \end{pmatrix}$. Let

$$\underline{x}' = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \qquad \underline{x}'' = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_{n+k} \end{pmatrix}$$

and $\underline{x} = \begin{pmatrix} x' \\ x'' \end{pmatrix} \in \mathbb{R}^{n+k}$. Let

$$F(\underline{x}) = f(\underline{x}') + B\underline{x}''.$$

Then

$$DF_{\begin{pmatrix} a \\ 0 \end{pmatrix}} = (Df_a, B)$$

The required result now follows from the Inverse Function Theorem.

Corollary (Injective Mapping Theorem) The function $f|_{U'}$ is injective.