Analysis 3

Based on lectures by
Notes taken by yehelip

Winter 2025

These notes are not endorsed by the lecturers. I have revised them outside lectures to incorporate supplementary explanations, clarifications, and material for fun. While I have strived for accuracy, any errors or misinterpretations are most likely mine.

Contents

1 Introduction to Topology in Euclidean Spaces

1.1 Norms and Metrics

We will start by giving basic topological definitions in the Euclidean space $\mathbb{R}^d$. First we set

$$\mathbb{R}^d := \left\{(x_1, x_2, \ldots, x_d) \mid {}^{1\leq i\leq d}_{x_i\in\mathbb{R}}\right\}$$

And now we can continue to define some more topological terms:

Definition 1.1 (Euclidean norm)**.** The Eulidean norm is defined as:

$$\|x\| = \|x\|_2 := \sqrt{\sum_{i=1}^{d} x_i^2}$$

We can similarly define the L_p norm as:

$$\|x\| = \|x\|_p := \sqrt[p]{\sum_{i=1}^{d} x_i^p}$$

which satisfies all properties of the norm.

Definition 1.2 (Euclidean metric)**.** The Euclidean metric is defined as:

$$d_2(P_1, P_2) := \|P_1 - P_2\|_2 = \sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}$$

Notice that it is induced by the Euclidean norm and similarly we can induce L_p metric using L_p norms.

Remark 1.1. The Euclidean metric is induced by the Euclidean norm, and the Euclidean norm is induced by the standard inner product defined on $\mathbb{R}^n$ given by

$$\langle x, y\rangle = \sum_{i=i}^{n} x_i y_i, \quad \forall x, y \in \mathbb{R}^n.$$

1.2 Sequences

Up until now we didn't have many problems using subscript for indexes of sequences, but now since we have coordinates we must denote a sequence in another similar way called $^{\text{superscript}}$ as such $(x^n)_{n=1}^{\infty}$. We define convergence of such sequences in the following way:

$$\lim_{n\to\infty} x^n = x \iff \forall i \left(\lim_{n\to\infty} x_i^n = x_i\right)$$

Definition 1.3 (Cauchy sequence)**.** A sequence $(x^n)_{n=1}^{\infty}$ is called a Cauchy sequence if and only if:

$$\lim_{n,m\to\infty} \|x^n - x^m\| = 0$$

1.3 Definitions

Definition 1.4 (Complete metric space)**.** A complete metric space is a metric space M such that every Cauchy sequence in M converges to some limit in M.

Definition 1.5 (Open set)**.** An open set in a Euclidean space is a subset U such that for any $x \in U$ exists $\varepsilon > 0$ such that any y such that any $y \in B_\varepsilon(x)$ satisfies $y \in U$

Definition 1.6 (Neighbourhood)**.** In a topological space X a space a neighbourhood of a point $x \in X$ is a subset such that exists an open set U such that $p \in U \subset V$.

Definition 1.7 (Closed set)**.** A closed set E in a Euclidean space X is a subset of X such that:

$$(x^n)_{n=1}^{\infty} \subseteq E \quad x^n \xrightarrow{n\to\infty} x \implies x \in E$$

Definition 1.8 (Compact set)**.** A set $X \subset \mathbb{R}^d$ is called compact if every open cover of X has a finite subcover.

Definition 1.9 (Sequencially compact set)**.** A set $X \subset \mathbb{R}^d$ is called sequentially compact if every sequence $(x^n)_{n=1}^{\infty} \subset X$ has a subsequence $(x^{n_k})_{k=1}^{\infty}$ that converges to a point x in X.

Remark 1.2. In Euclidean spaces, being sequentially compact is equivalent to being compact.

In the following definitions, let E be a subset of a Euclidean space.

Definition 1.10 (Closure)**.** The closure of a set $E \subseteq \mathbb{R}^n$ is can be

$$\mathrm{Cl}(E) = \left\{ x \mid \exists (x^n) \text{ s.t. } x^n \xrightarrow{n\to\infty} x \right\}$$

Proposition 1.1. *The closure of $E \subseteq \mathbb{R}^n$ is the smallest closed subset of $\mathbb{R}^n$ containing E.*

Definition 1.11 (Interior)**.** The interior of E is defined as:

$$\mathrm{Int}(E) = \left\{ x \mid \exists r > 0 \text{ s.t. } B_r(x) \subseteq X \right\}$$

Definition 1.12 (Boundary)**.** The boundary of E is defined as:

$$\partial E = \left\{ x \in \mathbb{R}^d \mid \forall r > 0 \ \exists x \in X \wedge \exists y \in X^c \text{ s.t. } y, z \in B_r(x) \right\}$$

1.4 Continuous functions

Definition 1.13 (Continuous function)**.** A function $f\colon \underset{\subseteq \mathbb{R}^n}{A} \to \underset{\subseteq \mathbb{R}^m}{A}$ is said to be continuous at $x \in A$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\|x - y\| < \delta$

$$\big\| f(y) - f(x) \big\| < \varepsilon$$

And we say that a function is continuous on X if it is continuous for every $x \in X$.

An important, equivalent, more general definition for continuity is that if f is a function from A to B then if U is an open set in B implies $f^{-1}(U)$ is an open set we say that f is continuous on A.

Definition 1.14 (Uniformly continuous function)**.** A function $f\colon \underset{\subseteq \mathbb{R}^n}{A} \to \underset{\subseteq \mathbb{R}^m}{A}$ is said uniformly continuous on $X \subseteq A$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\|x - y\| < \delta$

$$\big\| f(x) - f(y) \big\| < \varepsilon.$$

Definition 1.15 (Connectedness)**.** A set $A \subset \mathbb{R}^n$ is said to be connected if it cannot be written as a union of two or more disjoint nonempty open subsets. A set that is not connected is said to be disconnected.

Example 1.1. Any interval $I \subset \mathbb{R}$ is connected.

Example 1.2. The set $A = (0, 1) \cup (2, 3) \subset \mathbb{R}$ is disconnected because it can be written as the union of $(0, 1)$ and $(2, 3)$.

Definition 1.16 (Path)**.** A path from $x \in X$ to $y \in X$ is a continuous function $f\colon [0,1] \to X$ such that $f(0) = 0$ and $f(1) = y$.

Definition 1.17 (Path connectedness)**.** A set $A \subset \mathbb{R}^n$ is said to be path connected if there exists a path between any two points $x, y \in A$.

Definition 1.18 (Convex set)**.** A set $C \subset \mathbb{R}^n$ is called convex if and only if

$$\forall x, y \in C \quad \forall t \in [0,1] \quad tx + (1-t)y = y + t(x-y) \in C$$

Definition 1.19 (Convex hull)**.** Let $A \subset \mathbb{R}^n$ be a set. The convex hull of A is defined as

$$\mathrm{Conv}(A) = \bigcap_{\substack{A \subset S \\ S \text{ convex}}} S.$$

Remark 1.3. This is the smallest convex set containing A.

2 Differentiability

2.1 Definitions

Let $A \in \mathbb{R}^{m\times n}$. We define the linear map $T_A \colon \mathbb{R}^n \to \mathbb{R}^m$ as such:

$$T(x) := Ax.$$

Definition 2.1 (Operator norm)**.** Let $T\colon V \to W$ be a linear transformation between inner product spaces, we define the operator norm to be:

$$\|T\|_{\text{op}} = \|T\| = \sup_{v\neq 0} \frac{\|Tv\|_W}{\|v\|_V}$$

and T is said to be bounded if $\|T\| < \infty$.

Remark 2.1. An important result to prove is that if T is a bounded linear transformation then:

$$\|T\|_{\text{op}} \leq \left(\sum_{i,j} a_{ij}^2\right)^{\frac{1}{2}} < \infty$$

and since this gives that:

$$\big\|T(x) - T(y)\big\| \leq \|T\|_{\text{op}}\|x-y\|$$

we get that T is continuous and even Lipschitz continuous on its domain.

Definition 2.2 (Affine function)**.** An affine function is a function fo the form:

$$T_{A,b}(x) = Ax + b$$

such that $A \in \mathbb{R}^{m\times n}$ and $b \in x \in \mathbb{R}^m$.

Definition 2.3 (Differentiable function)**.** Let $U \subseteq \mathbb{R}^n$ be an open set, and let $f\colon U \to \mathbb{R}^m$. Given $a \in U$ we say that f is differentiable at a if there exists a linear transformation $T\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

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

We say that f is differentiable if f is differentiable at every $a \in U$.

Remark 2.2. Note that since U is open there exists a δ small enough such that $B(a,\delta) \subseteq U$.

Remark 2.3. We can also say that f is differentiable at a if there exists an "error" function $\epsilon(h)$ defined around 0 such that $f(a+h) = f(a) + Th + \epsilon(h)$ and

$$\lim_{h\to 0} \frac{\epsilon(h)}{\|h\|} = 0.$$

This property is equivalent to requiring $\epsilon(h) = o\big(\|h\|\big)$.

Remark 2.4. We denote the derivative of f at a either as $f'(a)$ or $\mathrm{D}f(a)$.

Many basic properties of derivatives also satisfied by this definition of differentiablity. For example, the chain rule.

Proposition 2.1 (The chain rule)**.** *Let f be differentiable at a, and suppose that g is differentiable at $b := f(a)$. Then $g \circ f$ is differentiable at a and*

$$(g\circ f)'(a) = g'(f(a))f'(a) \text{ or } \mathrm{D}(g\circ f)(a) = \mathrm{D}g(b) \circ \mathrm{D}f(a).$$

We can even define directional derivatives.

Definition 2.4 (Directional derivative)**.** Let $f\colon U \to \mathbb{R}$ with $U \subset \mathbb{R}^n$, $a \in U$, $0 \neq v \in \mathbb{R}^n$. Then, if the following limit exists:

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

we say that f is differentiable in the direction of v, and $D_v f(a)$ is called the directional derivative of f at a.

Remark 2.5. Note that if $\mathcal{E} = \{e_1, \ldots, e_n\}$ is the standard basis for $\mathbb{R}^n$ then we may denote the directional derivative of f at a in the direction of e_i in one of the following ways:

$$D_{e_i} f(a) = D_i f(a) = \frac{\partial f}{\partial e_i}(a) = f_{x_i}(a).$$

These are also called the partial derivatives of f.

Proposition 2.2. *If* $f\colon \underset{\subset \mathbb{R}^n}{U} \to \mathbb{R}$ *for is differentiable at* a *then:*

$$D_v f(a) = Df(a)v$$

for every $v \in \mathbb{R}^n$*.*

The proof is using the fact that:

$$f(a+h) = f(a) + Df(a)h + o(h).$$

An important corollary of this proposition is

$$D_v f(a) = Df(a)v = \nabla f(a) \cdot v = \sum_i^n \frac{\partial f}{\partial x_i}(a) v_i.$$

Remark 2.6. Notice that we get the biggest value for $\|v\| = 1$ when

$$v = \frac{\nabla f(a)}{\|\nabla f(a)\|}$$

which implies that the direction of the gradient is the direction of the steepest ascent. Conversely, the direction orthogonal to it, is the direction of least change.

Proposition 2.3. *Let* $f\colon U \to \mathbb{R}$ *for* $U \subset \mathbb{R}^n$ *be differetiable in* U*. Then if* f *has a local minimum or maximum at* $a \in U$ *then* $\nabla f(a) = 0$*.*

Remark 2.7. This is simply a reiteration of Fermat's theorem for the n-dimensional case.

Proposition 2.4. *A linear function* $f\colon \mathbb{R}^n \to \mathbb{R}^m$ *is differentiable at* $a \in U$ *if and only if* $f_i = f \circ \pi_i$ *is differentibale for every* $1 \leq i \leq n$ *where* π_i *is the projection function on the ith coordinate.*

Proof. Suppose that f_i is differentiable for every $1 \leq i \leq n$ then we can define a new function:

$$Tv := \begin{pmatrix} Df_1(a)v \\ Df_1(a)v \\ \vdots \\ Df_m(a)v \end{pmatrix}$$

And see that indeed:

$$\|f(a+h)-f(a)-Th\| \leq \sum_{i=1}^{m} |f(a+h)-f(a)-Df_i(a)h| = \sum_{i=1}^{m} |\epsilon_i(h)| = o(h)$$

As wanted. On the other hand, if f is differentiable we can use the chain rule to show that each f_i is differentiable with the desired derivative. $\square$

Definition 2.5 (Jacobian)**.** Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ such that each of its first-order partial derivatives exists. We define the Jacobian of f as such:

$$J_f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

2.2 Properties of the operator norm

For convenience we identify any matrix $A \in \mathbb{R}^{n\times n}$ as the operator it defines $T_A\colon \mathbb{R}^n \to \mathbb{R}^n$.

Recall that we defined the operator norm as

$$\|A\| = \|A\|_{\mathrm{op}} = \sup\left\{\|Ax\| : \|x\| = 1\right\} = \sup\left\{\frac{\|Ax\|}{\|x\|} : x \neq 0\right\}.$$

We can easily prove that $A \mapsto \|A\|$ is a norm, and that $\|A\|$ is the smallest constant that satisfies the ineqaility $\|Ax\| \leq \|A\|\|x\|$ for all $x \in \mathbb{R}^n$.

The operator norm also satifies

$$\|AB\| \leq \|A\|\|B\|$$

because we have

$$\left\|(AB)x\right\| \leq A\|Bx\| \leq AB\|x\|.$$

Proposition 2.5. *A matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if there exists a constant $c > 0$ such that $\|Ax\| \geq c\|x\|$ for all $x \in \mathbb{R}^n$. Moreover, the constant $c = \left\|A^{-1}\right\|^{-1}$ is the smallest such constant, and the is also the biggest constant that satisfies $\|Ax\| \geq c\|c\|$ for all $x \in \mathbb{R}^n$.*

Proof. Suppose A is invertible, then we have

$$\|x\| = \left\|A^{-1}Ax\right\| \leq \left\|A^{-1}\right\| \cdot \|Ax\|,$$

and because $\left\|A^{-1}\right\| \neq 0$ we get that

$$c\|x\| \leq \|Ax\|$$

for $c = \left\|A^{-1}\right\|^{-1}$ for all $x \in \mathbb{R}^n$. It is the biggest constant the satisfies this inequality because

$$r\|x\| \leq \|Ax\| \iff \left\|A^{-1}Ax\right\| \leq r^{-1} \cdot \|Ax\| \iff r \leq \left\|A^{-1}\right\|^{-1} = c$$

Next suppose there exists $c > 0$ such that $\|Ax\| \geq c\|x\|$ for all $x \in \mathbb{R}^n$. From this we get that

$$Ax = 0 \iff \|Ax\| = 0 \implies c\|x\| = 0 \implies x = 0,$$

which shows that A is invertible. $\square$

Proposition 2.6. *Let $A, B \in \mathbb{R}^{n\times n}$. If A is invertible, and $\|A-B\| < \frac{1}{2}\left\|A^{-1}\right\|^{-1}$ then B is also invertible, and $\left\|B^{-1}\right\| \le 2\left\|A^{-1}\right\|$.*

Proof. We want to somehow get to $B-A$ so we may notice that for all $x \in \mathbb{R}^n$ we have that

$$Bx = Ax + (B-A)x.$$

Thus,

$$\|Bx\| \ge \|Ax\| - \left\|(B-A)x\right\| \ge \left\|A^{-1}\right\|^{-1}\|x\| - \frac{1}{2}\left\|A^{-1}\right\|^{-1}\|x\| \ge \frac{1}{2}\left\|A^{-1}\right\|^{-1}\|x\|.$$

From the previous proposition, we get that B is invertible and that $\left\|B^{-1}\right\| \le 2\left\|A^{-1}\right\|$ which completes the proof. ☐

Definition 2.6 (General linear group)**.** We denote $\mathrm{GL}(\mathbb{R}^n)$ the collection of invertible matrices in $\mathbb{R}^{n\times n}$.

Remark 2.8. It is clear that the set $\mathrm{GL}(\mathbb{R}^n)$ coupled with matrix multiplication forms a group.

Proposition 2.7. *The group $\mathrm{GL}(\mathbb{R}^n)$ is an open set in $\mathbb{R}^{n\times n}$, and the operations of multiplication and inverse are continuous.*

Proof. Let $A \in \mathrm{GL}(\mathbb{R}^n)$. According to Proposition 2.6 we have that for $r = \frac{1}{2}\left\|A^{-1}\right\|^{-1}$ that $B(A,r) \subseteq \mathrm{GL}(\mathbb{R}^n)$, so $\mathrm{GL}(\mathbb{R}^n)$ is indeed open in $\mathbb{R}^n$.

To prove that the operator $m\colon \mathrm{GL}(\mathbb{R}^n)\times\mathrm{GL}(\mathbb{R}^n) \to \mathrm{GL}(\mathbb{R}^n)$ given by $m(A,B) = AB$ is continuous, let $A_k \xrightarrow{k\to\infty} A$ and $B_k \xrightarrow{k\to\infty} B$. Then

$$\|A_kB_k - AB\| = \|A_kB_k - A_kB + A_kB - AB\| \le \|A_k\|\|B_k - B\| + \|A_k - A\|\|B\| \xrightarrow{k\to\infty} 0.$$

Notice that we used the fact that a convergent sequence is bounded.

To prove that the operator $A \mapsto A^{-1}$ is continuous, set $A \in \mathrm{GL}(\mathbb{R}^n)$. Let $\epsilon > 0$. Choose $\delta = \min\left\{\frac{1}{2}\left\|A^{-1}\right\|^{-1}, \frac{1}{2}\left\|A^{-2}\right\|^{-1}\right\}\cdot\epsilon$. Now for any B such that $B \in B(A,\delta)$ then B is invertible and from Proposition 2.6 we have $\left\|B^{-1}\right\| \le 2\left\|A^{-1}\right\|$ so

$$\left\|A^{-1} - B^{-1}\right\| = \left\|A^{-1}(B-A)B^{-1}\right\| \le 2\left\|A^{-1}\right\|^2\delta < \epsilon$$

which shows that taking the inverse of A is a continuous operation and completes the proof. ☐

2.3 Continuous and differentiable partial derivatives

Recall that a map $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is called differentiable if it has an affine approximation. In the case of a single variable functions

$$f(x) = f(a) + f'(a)(x-a) + o(x-a).$$

In the case of multivariate functions, we have

$$f'(a) = J_f(a) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

If all the partial derivatives are continuous, then f is differetiable at a.

Proposition 2.8. *Let $f\colon \underset{\subseteq\mathbb{R}^n}{U} \to \mathbb{R}^m$ be a function such that $f = (f_1, \dots, f_m)$. Suppose that for all i, j the partial derivative $\dfrac{\partial f_i}{\partial x_j}$ is continuous in a neighbourhood of a. Then f is differentiable at a.*

Proof. We know that it suffices to prove that f_i is differentiable for all i, so suppose $f\colon U \to \mathbb{R}$ and let $h \in \mathbb{R}^n$ then

$$h = (h_1, \dots, h_n) = \sum_{j=1}^{n} h_j e_j$$

where e_j are the unit vectors. We then have that

$$f(a+h) - f(a) = \sum_{j=1}^{n} \left(f\left(a + \sum_{i=1}^{j} h_i e_i\right) - f\left(a + \sum_{i=1}^{j-1} h_i e_i\right)\right)$$

because this is a telescopic sum, and since the elements in this sum are the partial derivatives $\dfrac{\partial f}{\partial x_j}$ at the point $a + \sum_{i=1}^{j-1} h_i e_i$ we get that this is equal to

$$\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}\left(a + \sum_{i=1}^{j-1} h_i e_i\right) h_j + o(h_j) = \nabla f(a)\cdot h + \sum_{j=1}^{n}\left(\frac{\partial f}{\partial x_j}\left(a + \sum_{i=1}^{j-1} h_i e_i\right) - \frac{\partial f}{\partial x_j}(a)\right) h_j + o(h)$$

and now we only need to show that

$$\sum_{j=1}^{n}\left(\frac{\partial f}{\partial x_j}\left(a + \sum_{i=1}^{j-1} h_i e_i\right) - \frac{\partial f}{\partial x_j}(a)\right) h_j = o(h)$$

but this follows from the continuity of the partial derivatives in a neighbourhood of a. Thus for $h \to 0$ we get that f is differentiable at a which completes the proof. □

Remark 2.9. Let $f\colon \underset{\subset\mathbb{R}^n}{U} \to \mathbb{R}^m$ be a function. If $\dfrac{\partial f_i}{\partial x_j}$ is continuous for all i, j then we say $f \in C^1(U, \mathbb{R}^m)$. This is the collection of the continuously differentiable functions from U to $\mathbb{R}^m$.

Remark 2.10. A function $f\colon \underset{\subset\mathbb{R}^n}{U} \to \mathbb{R}^m$ is continuously differentiable if and only if the map $x \mapsto \mathrm{D}f(x)$ is continuous from U to the space of linear maps $L(\mathbb{R}^n, \mathbb{R}^m)$ with respect to the operator norm.

Example 2.1. We will show that the function $f\colon \mathrm{GL}_n(\mathbb{R}) \to \mathrm{GL}_(\mathbb{R})$ defined by $f(A) = A^{-1}$ is differentiable. We see that for any matrices $A, H \in \mathrm{GL}_n(\mathbb{R})$ we have

$$f(A+H) - f(A) = (A+H)^{-1} - A^{-1} = (A+H)^{-1}(A - (A+H))A^{-1} = -(A+H)^{-1}HA^{-1}.$$

We need to represent this expression as a linear function in H and an error function that is $o(H)$ so

$$-(A+H)^{-1}HA^{-1} = -A^{-1}HA^{-1} + (A^{-1} - (A+H)^{-1})HA^{-1}.$$

The function $H \mapsto -A^{-1}HA^{-1}$ is linear in H and since we know that $A \mapsto A^{-1}$ is continuous we get

$$\begin{aligned}\left\|\frac{(A^{-1} - (A+H)^{-1})HA^{-1}}{\|H\|}\right\| &\le \frac{\|f(A) - f(A+H)\|\cdot\|H\|\cdot\|A^{-1}\|}{\|H\|}\\ &= \|f(A) - f(A+H)\|\cdot\left\|A^{-1}\right\| \xrightarrow{H\to 0} 0\end{aligned}$$

which shows that $(A^{-1} - (A+H)^{-1})HA^{-1} = o(H)$. This means that f is differentiable with

$$\mathrm{D}f(A)H = -A^{-1}HA^{-1}.$$

From this formula we see that $\mathrm{D}f$ is also continuous so f is even continuously differentiable.

3 Least Squares and Gradient Descent

Let $f\colon U \subset \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function. Suppose we are intrested in finding a minimum (maybe a local minimum) of f, then there are two methods to solve find it using the tools we have aquired so far.

3.1 The standard method

We can use the fact that if f has a local minimum at a point $a \in U$ then $\nabla f(a) = 0$. We get n equations in n variables given by

$$\begin{aligned}\frac{\partial f}{\partial x_1}(a_1, \dots, a_n) &= 0\\ &\vdots\\ \frac{\partial f}{\partial x_n}(a_1, \dots, a_n) &= 0\end{aligned}$$

This method is good because if we solve these equations, and only get a single point where the gradient is 0 and know that there must be a minimum in U then we can find it quite easily.

3.2 Least squares

Consider the function $f(x) = \|Ax - b\|^2$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. When the rank of A is m we can solve the system of linear equations

$$(*)\; Ax = b$$

and the minimum will be 0. This is also the case when the rank of A is lower than m. However the problem gets way more intresting when $m > n$ because this means that it's possible that there is no solution to $(*)$ so we need to find the solution with the smallest error. This problem is called the "least squares" problem because we are trying to find the least sum of the following squares

$$\|Ax - b\|^2 = \sum_i \left(\sum_j a_{ij}x_j - b_i\right)^2 .$$

This problem is also sometimes called "linear regression and it is very important in all fields of science and engineering.

Since this is the simplest case of such a problem it has a closed solution. We notice that

$$f(x) = \langle Ax - b, Ax - b\rangle = \langle A^T Ax, Ax, -\rangle 2\langle Ax, b\rangle + \langle b, b\rangle := g(x) + h(x)$$

for

$$\begin{aligned}g(x) &= \langle A^T Ax, Ax, =\rangle x^T A^T Ax\\ h(x) &= -2\langle Ax, b\rangle + \langle b, b\rangle\end{aligned}$$

We therefore have that $\nabla f = \nabla g + \nabla h$. Since h is an affine function, its derivative is equal to its linear part:

$$\mathrm{D}h(x)\colon h \mapsto -2\langle Ah, b, =\rangle - 2b^T Ah \implies \nabla h = -2b^T A.$$

We also have that

$$g(x + h) - g(x) = \langle A^T A(x + h), x + h\rangle - \langle A^T Ax, x\rangle = 2\langle A^T Ax), h\rangle + \|Ah\|^2$$

and we see that $2\langle A^T Ax), h\rangle$ is linear in h, and that $\|Ah\|^2 \leq \|A\|^2\|h\|^2 = o\left(\|h\|\right)$ so it follows that

$$\mathrm{D}g(x)\colon h \mapsto 2\langle A^T Ax), h\rangle = 2x^T A^T Ah \implies \nabla g(a) = 2x^T A^T A.$$

It now follows that

$$\nabla f(x) = 2x^T A^T A - 2b^T A.$$

To find the minimum of f, we now need to solve the equation

$$\nabla f(x) = 2x^T A^T A - 2b^T A = 0 \quad \text{or} \quad A^T Ax = A^T b.$$

Whenever $m > n$ we have that $\operatorname{rank} A = n$ and thus the matrix $A^T A$ is also of rank n so it is invertible. In this case the solution to the equation is given by

$$x = (A^T A)^{-1} A^T b.$$

Remark 3.1. In practice we don't calculate the inverse of $A^T A$ since this is a rather complex numerical task, and instead we find the solution in other ways.

3.3 Gradient descent

In the example above, finding a solution to the system of linear equations $\nabla f(x) = 0$ was easy, however, in practice not all equations are linear, and we can't always find a closed solution.

The gradient descent method uses the fact that the direction of the gradient at a point is the direction of the steepest slope in order to approximate the minimum.

In general, given a function $f\colon U \subset \mathbb{R}^n \to \mathbb{R}$ the algorithm goes like this

(1) Guess a solution x to the minimazation problem and calculate $\nabla f(x)$ at x.

(2) If $\left\|\nabla f(x)\right\|$ is 0 or "very close" to 0 then x is the minimum we are looking for and we stop the algorithm.

(3) Otherwise, we set $x \mapsto x - c\nabla f(x)$. This moves x down the steepest slope, and since we usually set $0 < c < 1$ it doesn't go too far in that direction as well.

(4) Return to step (2).

The steps above only describe the basic idea, and the implementation is usually more complex. For example, we need to define what "very close" means. We also set a maximal number of iterations so the computation will stop if it doesn't find a minimum.

Here is a code example of the gradient descent method:

4 Taylor's Theorem

4.1 Higher Order Partial Derivatives

Theorem 4.1. (Schwart'z Theorem). *A function $f\colon \Omega \to \mathbb{R}$ defined on a set $\Omega \subset \mathbb{R}^n$, then if $p \in \Omega$ is a point with some neighbourhood contained in Ω, and f has continuous second partial derivatives in that neighbourhood then for all i, j in $\{1, 2, \ldots, n\}$,*

$$\frac{\partial^2}{\partial x_i\, \partial x_j} f(p) = \frac{\partial^2}{\partial x_j\, \partial x_i} f(p).$$

4.2 Multi-index Notation

An n-dimensional multi-index is an n-tuple

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$$

of non-negative integers. Define the following operations on some multi-indices $\alpha, \beta \in \mathbb{N}_0^n$ and a vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$:

Addition and Subtraction

$$\alpha \pm \beta = (\alpha_1 \pm \beta_1,\, \alpha_2 \pm \beta_2, \ldots,\, \alpha_n \pm \beta_n)$$

Partial Order

$$\alpha \leq \beta \quad \Leftrightarrow \quad \alpha_i \leq \beta_i \quad \forall\, i \in \{1, \ldots, n\}$$

Absolute Value

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$$

Factorial

$$\alpha! = \alpha_1! \cdot \alpha_2! \cdots \alpha_n!$$

Binomial Coefficient

$$\binom{\alpha}{\beta} = \binom{\alpha_1}{\beta_1}\binom{\alpha_2}{\beta_2} \cdots \binom{\alpha_n}{\beta_n} = \frac{\alpha!}{\beta!(\alpha - \beta)!}$$

Multinomial Coefficient

$$\binom{k}{\alpha} = \frac{k!}{\alpha_1! \alpha_2! \cdots \alpha_n!} = \frac{k!}{\alpha!}, \quad k := |\alpha| \in \mathbb{N}_0$$

Power

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$$

Higher Order Partial Derivatives

$$\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_n^{\alpha_n},$$

4.3 Taylor's Theorem

(Multivariate Version of Taylor's Theorem). Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a k-times continuously differentiable function at the point $\boldsymbol{a} \in \mathbb{R}^n$. Then there exist functions $h_\alpha\colon \mathbb{R}^n \to \mathbb{R}$, where $|\alpha| = k$, such that:

$$f(\boldsymbol{x}) = \sum_{|\alpha| \leq k} \frac{D^\alpha f(\boldsymbol{a})}{\alpha!} (\boldsymbol{x} - \boldsymbol{a})^\alpha + \underbrace{\sum_{|\alpha| = k} h_\alpha(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{a})^\alpha}_{o(\|x - \alpha\|^k)},$$

$$\text{and} \quad \lim_{\boldsymbol{x} \to \boldsymbol{a}} h_\alpha(\boldsymbol{x}) = 0.$$

5 Determining Critical Points

Suppose that $f\colon U \to \mathbb{R}$ such that $U \subset \mathbb{R}^n$ is open and $a \in U$.

Definition 5.1. If exists $r > 0$ such that for any point $x \in B_r(a)$ we have $f(a) \leq f(x)$ then the point is called a **local weak minimum**.

Remark 5.1. If the inequality is strict it is called a **local strong maximum**.

We define a local weak/strong maximum in the same manner.

Definition 5.2. If f is differentiable at a and $\nabla f(a) = 0$ then a is called a **critical point** of f.

Definition 5.3. If a is a critical point of f and for any $r > 0$ exist $x, y \in B_r(a)$ such that $f(x) < f(a) < f(y)$ then a is called a **saddle point** of f.

Definition 5.4. Let $f \in C^2(U)$. We define the **Hessian** of f at a to be the matrix:

$$H_f = H_f(a) = \left[\frac{\partial^2 f(a)}{\partial x_i \partial x_i}\right]_{i,j=1}^{n}$$

Notice that by Clairaut's theorem we have that the Hessian is symmetrical and thus it defines a quadratic form as follows:

$$\langle H_f(a)v, v\rangle = \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_i}(a) v_i v_j$$

Using Taylor's theorem and the multi-index notation explained earlier we can express the function as such:

$$f(x+h) = f(x) + \sum_{i=1}^{n} D_i f(x) h_i + \sum_{|\alpha|=2} \frac{D^\alpha f(x) h^\alpha}{\alpha!} + o(\|h\|^2)$$

We can notice that:

$$|\alpha| = 2 \Rightarrow \alpha_1 + \cdots + \alpha_n = 2$$

Which means that either way $\alpha! = 2$, so after some algebraic manipulation:

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle H_f(x)h, h\rangle + o(\|h\|^2)$$

Now if assume that x is a critical point of f, we can see that magic happens. We have $\nabla f(x) = 0$ and then:

$$f(x+h) - f(x) = \langle 0, h\rangle + \frac{1}{2}\langle H_f(x)h, h\rangle + o(\|h\|^2) \approx \frac{1}{2}\langle H_f(x)h, h\rangle$$

Since $H_f(x)$ is symmetrical it is similar to a diagonal matrix which we may denote $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and under the coordinates system corresponding to the eigenvectors of the Hessian we get that:

$$f(x+h) - f(x) \approx \frac{1}{2}\langle H_f(x)h, h\rangle = \sum_{i=1}^{n} \lambda_i h_i^2$$

And in case $n = 2$ we have more simply:

$$f(x + (a,b)) - f(x) \approx \lambda_1 a^2 + \lambda_2 b^2$$

And we can use this formula to determine the type of critical points.

Suppose that $a \in U$ is a critical point of f, then the following hold:

1. If $H_f(a)$ is **positive definite** then a is a **local strong minimum.**
2. If $H_f(a)$ is **negative definite** then a is a **local strong maximum.**
3. If $H_f(a)$ is **indefinite** then a is a **saddle point.**

In other cases we can't tell anything about the point and it might be of any type.

6 The Inverse Function Theorem

Recall that a set $C \subset \mathbb{R}^n$ is called convex if and only if

$$\forall x, y \in C \quad \forall t \in [0,1] \quad tx + (1-t)y = y + t(x-y) \in C$$

Intuitively, we say that C is convex if for every two points $a, b \in C$ the interval connecting the points is a subset of C.

Proposition 6.1. *Let $U \subset \mathbb{R}^n$ be convex and open, let $f\colon U \to \mathbb{R}^n$ be differentiable. If $f'(x)$ is bounded by M then,*

$$\forall x, y \in U \quad \|f(x) - f(y)\| \leq M\|x - y\|$$

Proof. If $f(x) = f(y)$ then we are trivially finished, otherwise we can denote $v = \frac{f(x)-f(y)}{\|f(x)-f(y)\|}$ and the following functions:

$$\begin{aligned} &g\colon \mathbb{R}^n \to \mathbb{R} && g(u) = \langle v, u\rangle \\ &h\colon [0,1] \to U && h(t) = y + t(x-y) \end{aligned}$$

Differentiating gives:

$$Dg = v \quad \text{and} \quad Dh = x - y.$$

We can now define the function:

$$\varphi = g \circ f \circ h$$

and apply Lagrange's theorem to get that exists $c \in (0,1)$ such that:

$$\varphi'(c) = \frac{\varphi(1) - \varphi(0)}{1 - 0} = \langle v, f(x)\rangle - \langle v, f(y)\rangle = \|f(x) - f(y)\|.$$

On the other hand, we have:

$$|\varphi'(c)| = \|g'(f \circ h)(f \circ h)'(h)'\| \leq \|v\| \cdot M \cdot \|x - y\|,$$

but we know that $\|v\| = 1$ so we get:

$$\|f(x) - f(y)\| \leq M\|x - y\|$$

as wanted. □

Try showing this proposition still stands for the $|\cdot|_\infty$ norm.

Corollary 6.2. *If $f \in C^1$ then f is Lifschitz continuous on any compact subset of its domain.*

Theorem 6.3. (Inverse Function Theorem). *Let $U \subset \mathbb{R}^n$ be open, $f \in C^1(U, \mathbb{R}^n)$, and $a \in U$. If $f'(a)$ is invertible then exist open sets $a \in V \subset U$ and $f(a) \in W$ such that $f\colon V \to W$ is a bijection and f^{-1} is also continuously differetiable and:*

$$(f^{-1})'(f(x)) = [f'(x)]^{-1}$$

7 Newton–Raphson Method

8 The Open Mapping Theorem

Definition 8.1. Suppose $U \subset \mathbb{R}^n$. A function $f \in C^1(U, \mathbb{R}^m)$ is called **regular** in a point $a \in U$ if $\text{rank} J_f(a) = m$. The function is called regular in U if it is regular for any $a \in U$.

Definition 8.2. A mapping $f\colon U \to V$ is called **open** if it send any open set W to an open set. That is for any open set $W \subset U$ then $f(W) \subset V$ is also open.

Theorem 8.1. (The Open Mapping Theorem). *Let $U \subset \mathbb{R}^n$ be an open set, and let $f \in C^1(U, \mathbb{R}^m)$ be regular in U. Then f is an open mapping.*

Proof. In the case of $m = n$ the theorem follows directly from the inverse function theorem. Because U is open, and because we know that for any $a \in U$ that $J_f(a)$ is invertible, we get from the inverse function theorem that exist open sets $a \in V_a \subset U$ and $f(a) \in W_a = f(V_a)$. We then have clearly that:

$$f(U) = \bigcup_{a \in U} W_a$$

Since W_a are open for any $a \in U$ we get that $f(U)$ is open as a union of open sets. For any open set $V \subset U$ we can use this exact proof using $f|_V$ instead.
Otherwise we must have $m < n$ because if $n < m$ the rank of the Jacobian can't be m. Considering our case, the rank of the Jacaobian is m which means it has m linearly independent columns. Without lose of generality we assume they are the first m columns and denote:

$$\frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_m)} = \left[\frac{\partial f_i}{\partial x_j}\right]_{(i,j)=1}^{m}$$

We now define a new transformation as such:

$$\begin{aligned} &F\colon U \to \mathbb{R}^n \\ &F(x) = (f(x), x_{m+1}, \dots, x_n) \end{aligned}$$

We see that the Jacobian of F is:

$$J_F = \begin{pmatrix} \dfrac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_m)} & \star \\ 0 & I_{(n-m)\times(n-m)} \end{pmatrix}$$

We notice that F satisfies the conditions for the inverse function theorem and thus it is an open mapping. We also notice that for every open set $V \subset U$ that:

$$f(V) = \pi(F(V))$$

Where π is the projection $\pi\colon (x_1, \dots, x_n) \mapsto (x_1, \dots, x_m)$. Since it is clear that π is an open mapping we have that f is also an open mapping as a composition of two open mappings as wanted. $\square$

9 Constrained Optimization

Proposition 9.1. *Let $g \in C^1(U, \mathbb{R}^m)$ be regular, $M = \{x \in U \mid g(x) = 0\}$, and $f \in C(U)$. If $a \in M$ and $f(a) \leq f(x)$ for any $x \in M$ then:*

$$\nabla f(a) \in \operatorname{span}\{\nabla g_1(a), \dots, \nabla g_m(a)\}$$

To actually find the minimum we solve the system of equations:

$$\begin{cases} \nabla f(a) = \sum_{i=1}^m \lambda_i \nabla g_i(a) \\ g(a) = 0 \end{cases}$$

Notice that the first equation gives n equations in $n+m$ variables that are $a_1, \dots, a_n, \lambda_1, \dots, \lambda_m$ and the second equation gives m equations in n variables, thus we have $n+m$ equation in $n+m$ variables.

Remark 9.1. The variables λ_i for $1 \leq i \leq m$ are called **Lagrange multipliers**.

Here's a simple example to how we can use Lagrange's multiplies to find the distance of a plane from the origin. The function we want to minimize is:

$$f(x, y, z) = \|(x, y, z)\|_2 = \sqrt{x^2 + y^2 + z^2}$$

But this is actually equivalent to finiding the minimum of:

$$f(x, y, z) = x^2 + y^2 + z^2$$

Under the constraints of a plane:

$$g(x, y, z) = ax + by + cz - d = 0$$

We notice that g is indeed regular. To find the minimum we will solve the system of equations:

$$\begin{cases} \nabla f(x, y, z) = \sum_{i=1}^m \lambda_i \nabla g_i(x, y, z) \\ g(x, y, z) = 0 \end{cases}$$

Since g is a function to $\mathbb{R}$ we only have one Lagrange multiplier. We can calculate the gradients of the functions:

$$\nabla f(x, y, z) = (2x, 2y, 2z)$$
$$\nabla g(x, y, z) = (a, b, c)$$

And put them in the system to get:

$$(2x, 2y, 2z) = \lambda(a, b, c) \quad \text{thus} \quad (x, y, z) = \frac{\lambda}{2}(a, b, c)$$

from the first equation. Putting that in the second equation gives:

$$g\left(\frac{\lambda}{2}(a, b, c)\right) = \frac{\lambda}{2}(a^2 + b^2 + c^2) - d = 0$$

Finally we get that the Lagrange multiplier is:

$$\lambda = \frac{2d}{a^2 + b^2 + c^2}$$

And the minimum point is:

$$(x, y, z) = \frac{d}{a^2 + b^2 + c^2}(a, b, c)$$

And the minimal distance of the plane from the origin is:

$$\|(x, y, z)\| = \left\| \frac{d}{a^2 + b^2 + c^2}(a, b, c) \right\| = \frac{d}{\|(a, b, c)\|}$$

In fact what we have shown so far is not a complete proof because the theorem only implies that at a minimum point the equations hold but it doensn't necessarily mean that the point we found is a minimum point. To complete the proof we can choose an arbitrary point on the plane $P_0 = (x_0, y_0, z_0)$, and denote $R = 2\|P_0\|$, and consider the set:

$$S = \{x \in \mathbb{R}^3 \mid g(x) = 0\} \cap \overline{B_R(0)}$$

This set is clearly compact and since $P_0 \in S$ we know that it's not empty and thus we know that f has a minimum in that set, and moreover, we know a priori that this minimum must be the global minimum of the function. Since we only found one point satisfying the equations given by the theorem it must be that minimum point and now the proof is complete.

Proof. To prove the theorem we will prove the contrapositive. Suppose that:

$$\nabla f(a) \notin \operatorname{span}\{\nabla g_1(a), \ldots, \nabla g_m(a)\}$$

We need to show that a is not a minimum point. Define the function:

$$F \colon U \to \mathbb{R}^{m+1}$$
$$F(x) = (f(x), g(x))$$

The Jacobian will be:

$$J_F(a) = \begin{bmatrix} \nabla f(x) \\ \nabla_1(x) \\ \vdots \\ \nabla_m(x) \end{bmatrix}$$

Because we assumed $\nabla f(a)$ in linearly independent from the rest of the gradients at a the rank of the Jacobian at a must be $m+1$ and since the rank function is continuous we know that exists a neighborhood U of a where the Jacobian rank is $m+1$ and thus we can allpy the open mapping theorem and get that F is an open mapping in that neighborhood. We know that $F(a) = (f(a), 0)$ is in the open set $F(U)$ and also that for a small enough epsilon that $(f(a) - \varepsilon, 0) \in F(U)$ and thus exists $x \in U$ such that $F(x) = (f(a) - \varepsilon), 0) = (f(x), g(x))$ which means that:

1. $g(x) = 0$
2. $f(x) < f(a)$

It follows that $x \in M$ and $f(x) < f(a)$ and thus a is not a minimum point in M. $\square$

10 The Implicit Function Theorem

In mathematics we have so far talked about explicit function - functions that take the elements of one set and transform them into an element of the other. In contrast, we can consider a new way of looking at functions using implicit functions. First, we will consider implicit equations.

Definition 10.1. An **implicit equation** is a relation of the form $R(x_1, \ldots, x_n) = 0$ where R is a multivariable function. A vector $x \in \mathbb{R}^n$ that satisfies $R(x) = 0$ is called a **solution**.

We say that the implicit equation $F(x, y)$ where $(x, y) \in U \subset \mathbb{R}^n \times \mathbb{R}^m$ defines y as an implicit function of x if exist $V_1 \times V_2 \subset U$ such that for any $x \in V_1$ exists a unique $y \in V_2$ such that $F(x, y) = 0$. Intuitively the function maps each $x \in V_1$ to its corresponding unique y.

Theorem 10.1. (The Open Mapping Theorem). *Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$ be open, $f \in C^1(U, \mathbb{R}^m)$, (a, b) a solution of $f(x, y) = 0$ and also assume that $\det \frac{\partial(f_1,\ldots,f_m)}{(x_1,\ldots,x_m)}|_{(a,b)} \neq 0$. Then exists a neighborhood $(a, b) \in V$ and a function $g \in C^1$ defined around a such that:*

$$\forall (x, y) \in V \quad f(x, y) \iff y = g(x)$$

And $g(a) = b$ and $f(x, g(x)) = 0$ for every x near a.

Theorem 10.2. (Implicit Function Differentiation). *Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$ be an open set, $f \in C^1(U, \mathbb{R}^m)$, (a, b) a solution of $f(x, y) = 0$, and also assume that $\det \frac{\partial(f_1,\ldots,f_m)}{(x_1,\ldots,x_m)}|_{(a,b)} \neq 0$. Then the derivative of the function g from the implicit function theorem is:*

$$g'(a) = -[D_y f(a, b)]^{-1} D_x f(a, b)$$

Notice that this thoerem is very useful because it allows us to find $g'(a)$ without knowing what is $g(a)$ itself.

11 Finding Roots of Polynomials

So it turns out that given a polynomial $p(x) = \sum_{i=0}^{n} a_n x^n$ then its root are continuously dependent on the coefficients.

Definition 11.1. Let $p(x) = \sum_{i=0}^{n} a_n x^n$ be a polynomial. A root of the polynomial with multiplicity of 1 is called a **simple root**.

We want to show that if x_0 is a simple root of a polynomial $P_a(x) = \sum_{i=0}^{n} a_n x^n$ such that $a = (a_0, \dots, a_n)$ then exists a neighborhood V of a in $\mathbb{R}^{n+1}$ such that one of the roots of the polynomial:

$$P_b(x) = \sum_{i=0}^{n} b_n x^n$$

For any $b \in V$ are given as a function $g \in C^1(\mathbb{R}^{n+1})$.

Proof. Define the function:

$$\begin{aligned} &f : \mathbb{R}^{n+2} \to \mathbb{R} \\ &f(a_0, \dots, a_n, x) = P_a(x) \end{aligned}$$

We notice that the implicit equation $f(a, x) = 0$ gives all the roots of the polynomial $P_a(x)$. Denote x_0 a simple root of $P_a(x)$ where a is the coefficients vector of the polynomial. We can calculate that:

$$f(a, x_0) = 0$$

And since x_0 is a simple root we also know that:

$$\left.\frac{df}{dx}\right|_{(a,x_0)} = \left.\frac{dP_a}{dx}\right|_{x_0} \neq 0$$

Therefore we satisfied the conditions for the implicit function theorem and can conclude that exists a neighborhood V of (a, x_0) and a function $g \in C^1$ defined around a such that:

$$\forall (a, x) \in V \quad P_a(x) = 0 \iff x = g(a)$$

$\square$

This theorem of great importance because when using computers and numerical methods to find roots using approximate coefficients we want to know that the roots we find are a good approximation of the real roots we need to find and this theorem shows exactly that.

12 Manifolds

Definition 12.1. Let $k, n \in \mathbb{N}$ such that $k \leq n$. A subset $M \subset \mathbb{R}^n$ is called a C^1 **manifold of dimension** k if for every $a \in M$ exist open sets $U, V \subset \mathbb{R}^n$ such that $a \in U$ and $V \cap \mathbb{R}^k \times \{0_{n-k}\} \neq \emptyset$ and exists a regular function $f \in C^1(U, V)$ such that:

$$f(M \cap U) = (V \cap \mathbb{R}^k \times \{0_{n-k}\}) = \{x \in V \mid x_{k+1} = \cdots = x_n = 0\}$$

Recall that the definition is regular is that in any point in the domain the derivative at that point is a surjective linear transformation - or in other words - the rank of the Jacobian is equal to its row number. Because f is a transformation between two open sets of the same dimension we know that f is regular if and only if its derivative is invertible at any point.

Definition 12.2. A function $f \in C^1$ that is invertible and regular is called a **diffeomorphism**. From the inverse function theorem every diffeomorphism's inverse function is a diffeomorphism in itself.

Remark 12.1. A 1-dimensional manifold is sometimes called a **differential curve** and a 2-dimensional manifold is sometimes called a **differential surface**.

The title of this section is indeed "manifolds" but in fact we are only talking about **embedded manifolds**. That is to say, we are not talking about manifolds in the abstract sense at all, only about those who are specifically embedded in the Euclidean space.

Proposition 12.1. *The following conditions are equivalent:*

1. *M is a C^1 manifold of dimension k*
2. *For every $a \in M$ exists a neighborhood U_a and a regular function $g \in C^1(U, \mathbb{R}^{n-k})$ such that:*
$$M \cap U = \{x \in U \mid g(x) = 0\}$$
3. *Up to permutation of the variables, for every $a \in M$ exists a neighborhood $a \in V \times W$ such that $V \subset \mathbb{R}^k$ and $W \subset \mathbb{R}^{n-k}$ and exists $h \in C^1(V, \mathbb{R}^{n-k})$ such that:*
$$M \cap (V \times W) = \mathrm{graph}(h) = \{(x, h(x)) \mid x \in V\}$$
4. *For each $a \in M$ exists a neighborhood U, an open set $V \subset \mathbb{R}^k$ and an injective function $H \in C^1(V, \mathbb{R}^n)$ such that:*
   (a) $\mathrm{rank} DH = k$
   (b) $M \cap U = H(V)$
   (c) *The function $H^{-1} \colon H(V) \to V$ is continuous according to the topology on $H(V)$ induced by $\mathbb{R}^n$.*

Remark 12.2. The function H from 4 is called a parametrization of $M \cap U$.

Example 12.1. Let $S^n = \partial \mathbb{B}_{n+1} = \{x \mid \|x\| = 1\} \subset \mathbb{R}^{n+1}$ be the n-dimensional sphere. We will show that this sphere is a n-dimensional manifold.

According to 2
For any $a \in M$ we can choose the open set $U = \mathbb{R}^{n+1} \setminus \{0\}$ such that $x \in U$. Define the function:

$$g \colon U \to \mathbb{R}$$
$$g(x) = \left(\sum_{i=1}^{n+1} x_i^2 \right) - 1$$

We see that:

$$\nabla g = 2(x_1, \ldots, x_{n+1})$$

So g is regular and continuously differentiable in U. Finally we check that:

$$S^n \cap U = S^n = \{x \in U \mid g(x) = 1\}$$

So indeed this verifies that S^n is an n-dimensional manifold.
According to 3
Let $a \in M$, without loss of generality and allowing permutation of variables we can assume that $a_{n+1} < 0$. Now choose $V = \mathbb{B}_n$ and $W = (-\infty, 0)$. Define:

$$\begin{aligned} &h\colon V \to \mathbb{R} \\ &h(v_1, \ldots, v_n) = -\sqrt{1 - \sum_{i=1}^{n} v_i^2} \end{aligned}$$

We can verify that indeed:

$$S^n \cap (\mathbb{B}_n \times (-\infty, 0)) = \{(v, h(v)) \mid v \in V\}$$

According to 4
We will show only the case of $S^n = S^2$. Suppose we want to find a parametrization at the point $(1, 0, 0)$. We define the sets:

$$\begin{aligned} V &= (0, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \\ U &= \{(x, y, z) \mid x > 0\} \end{aligned}$$

And the function:

$$H(\theta, \phi) = (\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta))$$

We can see that H is indeed a continuously differentiable continuous injection. We see that:

$$DH = J_H = \begin{pmatrix} \cos(\theta)\cos(\phi) & -\sin(\theta)\sin(\phi) \\ \cos(\theta)\sin(\phi) & \sin(\theta)\cos(\phi) \\ -\sin(\theta) & 0 \end{pmatrix}$$

In our case $k = 2$, so seeing that the determinant of the upper is $\sin(\theta)\cos(\theta)$ which is 0 only when $\theta = \frac{\pi}{2}$ but then the matrix is:

$$\begin{pmatrix} 0 & \sin(\phi) \\ 0 & \cos(\phi) \\ -1 & 0 \end{pmatrix}$$

Which is a matrix of rank 2 so we conclude that $\text{rank} DH = 2$ for every value pair (θ, ϕ) in V. Conditions $\langle b \rangle$ and $\langle c \rangle$ also hold after some algebraic manipulation.

Remark 12.3. Every open set $U \subset \mathbb{R}^n$ is a manifold of dimension n and every vector space $V \subset \mathbb{R}^n$ is a manifold of dimension $\dim(V)$.

Definition 12.3. Let $M \subset \mathbb{R}^n$ be a manifold of dimension k. A function $g\colon M \to \mathbb{R}^m$ is called **continuously differentiable** if for every $a \in M$ and every parametrization (H, V) in a we have:

$$g \circ H \in C^1(V, \mathbb{R}^m)$$

For practice, show that the definition is not dependent on the choice of the coordiante system. Also, show that this definition is equivalent to exists a neighborhood $a \in U$ and a function $G \in C^1(U, \mathbb{R}^m)$ such that $G|_{M\cap U} = g|_{M\cap U}$ where g is the C^1 function.

13 The Tangent Space

Definition 13.1. A continuous function $f\colon [a,b] \to \mathbb{R}^n$ is called a **continuous path**. If f is also differentiable we say it is a **continuously differentiable path**. The "1- dimensional" set $f([a,b])$ is called a **curve**.

Definition 13.2. For each point $p \in \mathbb{R}^n$ we define the **tangent space** to $\mathbb{R}^n$ in p as:

$$T_p(\mathbb{R}^n) = \{v_p = (p,v) \mid v \in \mathbb{R}^n\}$$

In manifolds we have the following definition

Definition 13.3. Let M be a C^1 manifold of dimension k, and let $p \in M$ be a point on the manifold. The **tangent space** to M in point p is defined as:

$$\left\{ p + \gamma'(t_0) \middle|_{t_0 \in (a,b) \quad \gamma(t_0)=p}^{C^1 \ni \gamma\colon (a,b) \to M} \right\}$$

And we also denote:

$$T_p(M) = \left\{ \gamma'(t_0) \middle|_{t_0 \in (a,b) \quad \gamma(t_0)=p}^{C^1 \ni \gamma\colon (a,b) \to M} \right\}$$

Remark 13.1. Let $H\colon V \subset \mathbb{R}^k \to \mathbb{R}^n$ be a parametrization of a manifold M around p, and $q \in V$ a point such that $H(q) = p$. Then $T_p(M) = [\mathrm{Im}(DH(q))]_p$.

Example 13.1. Consider S^1 and the parametrization:

$$H\colon \left(\frac{\pi}{2}, \frac{3\pi}{2}\right)$$
$$H(t) = (\cos(t), \sin(t))$$

Choose $p = (-1, 0) = H(\pi)$. We see that:

$$DH(\pi) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}_{t=\pi} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} : \mathbb{R} \to \mathbb{R}^2$$

We notice that indeed:

$$\mathrm{Im}(DH(\pi)) = \mathrm{Sp}\left\{ \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\}$$

$$\Rightarrow T_p(M) = \left\{ p + a \begin{pmatrix} 0 \\ -1 \end{pmatrix} \middle| a \in \mathbb{R} \right\}$$

Definition 13.4. Let $N \subset \mathbb{R}^3$ be a smooth surface. A vector $0 \neq v \in \mathbb{R}^3$ is said to be **normal** to M at a point $p \in M$ if it is orthogonal to $T_p(M)$.

Remark 13.2. If M is equal to the set of roots of a regular function g in a neighborhood of a then a vector normal to M must be co-linear with $\nabla g(p)$. That is because

14 Manifolds With a Boundary

Defining manifolds is very convenient, but many intresting sets we want to analyze are not manifolds like any closed interval or half a sphere. That is why in this section we introduce a new object that is almost like a manifold, but slightly different.

Definition 14.1. The **half k-dimensional space** is defined as:

$$\mathcal{H}_k = \{x \in \mathbb{R}^n \mid x_n \geq 0\}$$

Definition 14.2. A subset $M \subset \mathbb{R}^n$ is called a C^1 **manifold with boundary** of dimension k if for every $a \in M$ exists a neighborhood $a \in U$, an open set $V \subset \mathbb{R}^k$ and an injection $H \in C^1(V, \mathbb{R}^n)$ such that:

1. $\text{rank} DH = k$.
2. $M \cap U = H(V \cap \mathcal{H}_k)$.
3. $H^{-1} \colon M \cap U \to V \cap \mathcal{H}_k$ is continuous with respect to the topology induced by $\mathbb{R}^n$.

The function H is called a **parametrization** of M at a.

Definition 14.3. Let $M \subset \mathbb{R}^n$ be a manifold with boundary with dimension k. A point $a \in M$ is called a **boundary point** if $a = H(b)$ where $b \in \partial\mathcal{H}_k$. We denote the set of the boundary points of M as ∂M. The rest of the points are called **inner points** and we denote them as $\text{int}(M)$.

Remark 14.1. Watch out! the topological definition of a boundary is different from the definition of a boundary for a manifold!

For practice here are some exercises:

1. Prove that if $V \subset \mathbb{R}^k$ is an open set such that $\mathcal{H}_k \cap V \neq \emptyset$ and if $H \in C^1(V \cap \mathcal{H}_k), \mathbb{R}^n)$ then exists a function $\tilde{H} \in C^1(V, \mathbb{R}^n)$ such that $\tilde{H}|_{V \cap \mathcal{H}_k)} = H$
2. Find explicit formulas for a normal vector and a tangent space. Find examples of manifolds with and without boundaries.

15 Length of a Path

Definition 15.1. Let $f\colon [a,b] \to \mathbb{R}^n$ be a continuous path, and let $P = (t_0 = a < \cdots < t_k = b)$. Define:

$$l(f,P) = \sum_{i=1}^{k} \|f(t_i) - f(t_{i-1})\|$$

And also define:

$$l(f) = \sup_{p} l(f,p)$$

If $l(f) < \infty$ we say that f has a **length** of $l(f)$.

Proposition 15.1. *If $f \in C^1([a,b], \mathbb{R}^n)$ then the path f has length. Moreover, for every partition P we have $l(f,P) \leq \int_a^b \|f'(t)\|\, dt$, and $l(f) = \int_a^b \|f'(t)\|\, dt$.*

Definition 15.2. A **simple smooth curve** is a set $C \subset \mathbb{R}^n$ that is the image of a path $f \in C^1([a,b], \mathbb{R}^n)$. That satisfies:

1. The path f is an injection on (a,b).
2. For every $t \in [a,b]$ we have $f'(t) \neq 0$.

Show that if f is a injective path then the curve it defines is a C^1 manifold.

Definition 15.3. A curve is called **directed** if we differentiate between its beginning and end.

Definition 15.4. A curve is called **closed** or a **loop** if it satisfies $f(a) = f(b)$.

Definition 15.5. We define the **length** of a curve to be the length of the path it's defined by.

Prove that definition 14.5 is well defined.

Definition 15.6. A parametrization of a curve g such that $\|g'(s)\| = 1$ is called an **arc length paramtrization**.

16 Line Integrals

Definition 16.1. Let $C \subset \mathbb{R}^n$ be a smooth simple curve with parametrization $C^1 \ni f\colon [a,b] \to \mathbb{R}^n$. If ρ is a continuous function on C then we define the **scalar integral** of ρ on C or **type one line integral** as:

$$\int_C \rho \, ds = \int_a^b \rho(f(t)) \|f'(t)\| \, dt$$

Notice that when $p(x) = 1$ we get the length of the curve. We can use this type of integral when we want to sum values over that curve. For example, a wire can actually be parameterized using a function that gives us its concentration of mass at any given point and then the integral would give the total mass of the wire.

Definition 16.2. Let $S \subset \mathbb{R}^n$. A function $f\colon S \to \cup_{p \in S} T_p(\mathbb{R}^n)$ is called a **vector field** if for all $p \in S$ we have $f(p) \in T_p(\mathbb{R}^n)$.

A similar definition exists for smooth manifolds instead of $\mathbb{R}^n$ but there's no reason to get into it right now.

Definition 16.3. Let $f\colon [a,b] \to \mathbb{R}^n$ be a continuously differentiable parametrization of a directed, smooth, simple curve C. Let $f\colon C \to \mathbb{R}^n$ be a continuous vector field. The line integral of F over C is called a **type two line integral** and is defined to be:

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C \sum_{i=1}^n F_i \, dx_i := \int_a^b F(f(t)) \cdot f'(t) \, dt$$

17 Integration

Definition 17.1. A **box** in $\mathbb{R}^n$ is a set $R \in \mathbb{R}^n$ such that:

$$R = \prod_{i=1}^{k} [a_i, b_i]$$

For real numbers $a_i \leq b_i$. If exists i such that $a_i = b_i$ then we say that the box is **degenerate**.

We define the volume of a box to be:

$$\mathrm{Vol}(A) = \prod_{i=1}^{k} (b_i - a_i)$$

A partition of an n-dimensional box is similar to the partitions of a 2-dimensional box and thus its definition is omitted. If P is a partition of a box R we denote $B \sim Q$ if B is a subbox of the partition. Since the rest of the basic definitions for a Darboux or Riemann integrals are very similar to the second dimensional case, we will only see the way to denote what's necessary. Let f be a bounded function on a box R. We denote the upper and lower Darboux sums with regards to a partition P of R as such:

$$U(f, P) = \sum_{B \sim R} M_Q \cdot \mathrm{Vol}(Q)$$

Where $M_Q = \sup\{f(x) \mid x \in Q\}$ as usual. We say that f is Darboux integrable if:

$$\inf U(f, P) = \sup L(f, P) = I$$

We denote the integral of f over the box as:

$$I = \int_R f \, dV = \int_R f(x) \, dx = \int_R f$$

Most theorems from analysis 2 also apply here.

Definition 17.2. Fubini's theorem - Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ and suppose all the functions:

$$\begin{aligned}
F_1(x_1, \ldots, x_{n-1}) &= \int_{a_n}^{b_n} f(x) \, dx_n \\
F_2(x_1, \ldots, x_{n-2}) &= \int_{a_{n-1}}^{b_{n-1}} f(x) \, dx_{n-1} \int_{a_n}^{b_n} f(x) \, dx_n \\
&\vdots \\
F_n \equiv I &= \int_{a_1}^{b_1} f(x) \, dx_1 \cdots \int_{a_{n-1}}^{b_{n-1}} f(x) \, dx_{n-1} \int_{a_n}^{b_n} f(x) \, dx_n
\end{aligned}$$

Are integrable then $I = \int_R f \, dV$.

To calculate the volume of a space Ω we can calculate the integral $\int_\Omega f \, dV$ for the function $f \equiv 1$. For a simple space in $\mathbb{R}^3$ we can use Fubini's theorem and get:

$$I = \iiint_\Omega f(x, y, z) \, dxdydz = \int_a^b dx \int_{y_1(x)}^{y_2(x)} dy \int_{z_1(x)}^{z_2(x)} f(x, y, z) \, dz$$

For the relevant functions.

18 Zero Volume

Remark 18.1. If $f\colon R \to \mathbb{R}^n$ is a continuous function on R then it is integrable.

Definition 18.1. A set $A \subset \mathbb{R}^n$ is said to have **volume zero** if for each $\varepsilon > 0$ exists a finite number of boxes $R_1, \ldots, R_k$ such that:

$$\sum_{i=1}^{k} \mathrm{Vol}(R_i) < \varepsilon \quad \text{and} \quad A \subset \bigcup_{i=1}^{k} \mathrm{int}(R_i)$$

We see that any point $x \in \mathbb{R}^n$ is of volume zero because we can always choose one box of arbitrarily small volume such that x is inside it. It is also trivial to prove that a finite union of sets of volume zero is of volumes zero.

Theorem 18.1. *Let $f\colon R \to \mathbb{R}^n$ be bounded and suppose:*

$$X = \{x \in R \mid f \text{ is not continuous in } x\}$$

be a set of volume 0. Then f is integrable over R.

Proof. We know that f is bounded so we can denote its bound M, and also we know that for any $\varepsilon > 0$ exist $R_1, \ldots, R_k$ such that:

$$\sum_{i=1}^{k} \mathrm{Vol}(R_i) < \frac{\varepsilon}{4M} \quad \text{and} \quad X \subset \bigcup_{i=1}^{k} \mathrm{int}(R_i)$$

Thus we can choose a partition P_X such that for every i we have $R_i \in P$ and then we have:

$$U(f,P) - L(f,P) = \sum_{i=1}^{k} (M_Q - m_Q)\mathrm{Vol}(R_i) < \frac{\varepsilon}{2} \quad \text{and} \quad X \subset \bigcup_{i=1}^{k} \mathrm{int}(R_i)$$

Now if we consider the set $Y = R \setminus \{R_i\}_i$ then it is clearly closed and bounded and thus it is integrable and we can choose a partition P_Y of Y such that:

$$U(f,P_Y) - L(f,P_Y) = \sum_{Q \sim P_Y} (M_Q - m_Q)\mathrm{Vol}(Q) < \frac{\varepsilon}{2}$$

And now it is clear that the partition $P = P_X \cup P_Y$ satisfies:

$$U(f,P) - L(f,P) < \varepsilon$$

□

Definition 18.2. We say that $A \subset \mathbb{R}^n$ is of **measure zero** if exist $\{R_i\}_{i \in I}$ such that $|I| < \aleph_0$ and:

$$\sum_{i \in I} \mathrm{Vol}(R_i) < \varepsilon \quad \text{and} \quad A \subset \bigcup_{i \in I} \mathrm{int}(R_i)$$

Remark 18.2. A set of volume zero is always of measure zero, but the contrary is not always true. For example the sets of the form $\mathbb{Q}^n \subset \mathbb{R}^n$ are all of measure zero but not of volume zero.

Theorem 18.2. *A compact set $D \subset \mathbb{R}^n$ is of volume zero if and only if it is of measure zero.*

The proof of this theorem is very direct and thus omitted.

Theorem 18.3. (Lebesgue's Theorem). *Let $R \subset \mathbb{R}^n$ be a box and let $f\colon R \to \mathbb{R}^n$ be bounded. Then f is integrable if and only if the set of discontinuity points of f is of meausre zero.*

Before proving the theorem we will define the **oscillation** of f over a set $B \subset \mathbb{R}^n$ as:

$$\omega(f, B) = \sup\{f(x) \mid x \in R \cap B\} - \inf\{f(x) \mid x \in R \cap B\}$$

And also define the **oscillation** of f at a point $x \in R$ as:

$$\omega(f, x) = \lim_{r \to 0^+} \omega(f, U_r(x))$$

Where $U_r(x) = (x_1 - r, x_1 + r) \times \cdots \times (x_n - r, x_n + r)$ which is the open ball with radius r around x with the infinity norm.

Lemma 18.4. *The function f is continuous at $x \in R$ if and only of $\omega(f, x) = 0$.*

Lemma 18.5. *For every $\varepsilon > 0$ the set:*

$$W_\varepsilon = \{x \in R \mid \omega(f, x) \geq \varepsilon\}$$

Is continuous.

We also define another set:

$$W = \{x \in R \mid \omega(f, x) \neq 0\}$$

Proof. □

Definition 18.3. A space $\Omega \subset \mathbb{R}^n$ is said to **have volume** if it is bounded and $\partial\Omega$ is of measure zero.

Definition 18.4. Let $\Omega \subset \mathbb{R}^n$ be a space with volume, let $f \colon R \to \mathbb{R}^n$ be a function, and assume that $\Omega \subset R$. We say that f is **integrable over** Ω if $f \cdot 1_\Omega$ is integrable over R and:

$$\int_\Omega f \, dV = \int_R f \cdot 1_\Omega \, dV$$

Theorem 18.6. *If $f \colon \Omega \to \mathbb{R}^n$ is continuous and Ω has volume then f is integrable over Ω.*

Definition 18.5. The volume of a set that has volume $A \subset \mathbb{R}^n$ is defined as:

$$\mathrm{Vol}(A) = \int_A 1 \, dv$$

We notice that this function is well defined because the constant function $f \equiv 1$ is always continuous on any domain that has volume.

Remark 18.3. A Riemann integral on a space with volume has all the basic properties as we defined them on intervals: linearity, additivity, the triangle inequlity, and more.

Theorem 18.7. *The graph of a Riemann integrable function is of measure zero, or volume zero if the function was defined on a compact space. A manifold has measure zero. The image of a manifold by a smooth transformation is of volume zero.*